DISCOVERING GROUP DIFFERENCES FROM QUALITATIVE AND QUANTITATIVE ATTRIBUTES USING CONTRAST SET MINING WITH DISCRETIZATION AND MEASURES OF INTERESTINGNESS

A Thesis

Submitted to the Faculty of Graduate Studies and Research

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Computer Science

University of Regina

By

Mondelle Simeon

Regina, Saskatchewan

December 2012

UNIVERSITY OF REGINA

FACULTY OF GRADUATE STUDIES AND RESEARCH

SUPERVISORY AND EXAMINING COMMITTEE

Mondelle Simeon, candidate for the degree of Doctor of Philosophy in Computer Science**,** has presented a thesis titled, ***Discovering Group Differences from Qualitative and Quantitative Attributes Using Contrast Set Mining with Discretization and Measures of Interestingness,*** in an oral examination held on September 24, 2012. The following committee members have found the thesis acceptable in form and content, and that the candidate demonstrated satisfactory knowledge of the subject material.

External Examiner:	Dr. Carson K. Leung, University of Manitoba
Supervisor:	Dr. Robert Hilderman, Department of Computer Science
Committee Member:	Dr. Michael Kozdron, Department of Mathematics and Statistics
Committee Member:	Dr. Malek Mouhoub, Department of Computer Science
Committee Member:	Dr. Cortney J. Butz, Department of Computer Science
Chair of Defense:	Dr. Dongyan Blachford, Faculty of Graduate Studies and Research

Abstract

Identifying differences between groups is a fundamental problem in many disciplines. Groups are defined by a selected property that distinguishes one group from the other. For example, gender (male and female students) or year of admission (students admitted from 2001 to 2010). The group differences sought are novel, indicating that they are not obvious or intuitive, potentially useful, suggesting that they can aid in decision-making, and understandable, implying that they are presented in a format easily understood by human beings. Contrast set mining has been developed as a data mining task which aims to efficiently identify differences between groups from observational multivariate data. Here we study two closely related steps in the contrast set mining process: the mining step, and the interpretation and evaluation step. In the mining step, the task to be performed is the discovery of valid contrast sets. A valid contrast set is a conjunction of attributes and values that differ meaningfully in their distribution across groups. We introduce a novel contrast set, called the λ-contrast set. A λ-contrast set has a ratio of maximum support to minimum support greater than a user-defined threshold. We introduce a novel discretization algorithm, called *Discretize*, for creating intervals for quantitative attributes in the dataset. We demonstrate how we build our search space of all possible contrast sets from the attributes, and attribute-values in the dataset. We then present the COSINE algorithm, for traversing the search space of possible contrast sets. We show how the COSINE algorithm generates all valid contrast sets, maximal valid contrast sets, and valid λ-contrast sets. We then show how removing the attributes and attribute-interval pairs that not highly correlated produces a smaller search space. We then present the GENCCS algorithm for traversing the search space of possible correlated contrast sets. We show how the GENCCS algorithm generates all correlated valid contrast sets, correlated maximal valid contrast sets, and correlated valid λ-contrast sets. In the interpretation and evaluation step, we investigate the problem of ranking the inter-

estingness of the contrast sets discovered using fourteen interestingness measures. Twelve measures have been previously utilized in various areas in the data mining community, such as association rule mining, emerging patterns, and subgroup discovery. Their use for ranking contrast sets is novel. The objective of this work is to gain insight into the behaviour that can be expected from the COSINE and GENCCS algorithms and the interestingness measures in practice. From our analysis, we found that COSINE is an effective technique for the efficient generation of all valid contrast sets, maximal valid contrast sets, and λ-contrast sets, and that it discovered more interesting contrast sets as compared to those obtained by two existing contrast set mining techniques, STUCCO and CIGAR. We also found that GENCCS is an effective technique for efficient generation of all valid correlated contrast sets, maximal correlated valid contrast sets, and correlated valid λ-contrast sets.

Acknowledgements

I would like to thank the many people who made this work possible. My deepest gratitude and appreciation goes to my supervisor, Dr. Robert J. Hilderman, who has given me great latitude, as well as remarkable guidance, to pursue research problems and their solutions. His great patience, willingness to help, and his constructive criticism have been invaluable in seeing this project through to a successful conclusion. Dr. Hilderman's efforts in helping me to become a more mature writer and researcher have been immeasurable.

Many thanks to my external examiner Dr. Carson Leung. It was an honour to have him participate in the process. Also, thanks to the members of my Ph.D committee: Dr. Malek Mouhoub, Dr. Cory Butz, and Dr. Michael Kozdron. Their comments, suggestions, and criticisms were greatly appreciated.

I would also like to thank Dr. Robert Hilderman, Dr. Howard Hamilton, the Department of Computer Science, and the Faculty of Graduate Studies and Research for financial support.

Also thanks to Dr. David Malloy, Phillip Sevigny, and Sean Targett for access to real-world datasets and for evaluating and providing feedback as an assessment of the usefulness of the results and general methodology from the perspective of an end-user.

Finally, I would like to thank my family, particularly my mother, Catherine Simeon, for her unconditional support and love and for giving me the courage to reach higher and dream bigger.

Dedication

I dedicate this thesis to my family. Their unconditional love and support have been the foundation of my life. To my mother especially, I love you and thanks for everything.

Table of Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Problem Statement and Motivation

Identifying the differences between groups is a fundamental problem in many disciplines. For example, a reproductive health study of demographic and socio-economic variables could be used by health care providers to characterize the differences in contraceptive methods used by married couples. A study of this kind could involve a dataset as the one shown in Table 1.1. In Table 1.1, the *Couple ID* column is a unique identifier assigned to each couple. The *Wife's Education* column describes the highest educational level attained by the wife. In this column, `H`, `U`, and `P` indicate that the wife has attained a high-school, university, and post-graduate level education, respectively. The *Wife's Occupation* describes the type of occupation that the wife currently has. In this column, `S`, `T`, and `P` indicate that the wife currently has a service, technical, and professional type of job, respectively. The *Family Income* column describes the annual income of the family in thousands. The *Contraceptive Method* column describes the type of contraceptive method being used by each couple, and defines the group to which each couple belongs. In this column, `N`, `S`, and `L` indicate membership in a group that is using no contraceptive method, a short-term contraceptive method, and a long-term contraceptive method, respectively.

Table 1.1: A Sample Dataset

Couple ID	*Wife's Education*	*Wife's Occupation*	*Family Income*	*Contraceptive Method*
1	U	T	44	S
2	H	S	56	S
3	H	S	55	N
4	U	T	37	N
5	P	P	43	L
6	H	S	56	N
7	P	P	57	S
8	P	P	56	L
9	U	T	44	S
10	P	P	57	N
11	H	S	45	L
12	H	S	58	N
13	P	P	45	N
14	U	T	57	N
15	U	T	34	N
16	P	P	42	L
17	P	P	43	S
18	H	S	57	L
19	U	T	36	S
20	P	P	36	L
21	P	P	56	L
22	P	P	44	N
23	H	S	57	L
24	U	T	38	N
25	H	S	57	N
26	U	T	36	N

From Table 1.1, the characteristics that differentiate the three contraceptive methods used by the couples can be represented as conditional probabilities, such as $P($`Contraceptive Method = N | Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]`$) = 0.363$, $P($`Contraceptive Method = S | Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$`Family Income = [47.9, 58]`$) = 0.125$, and $P($`Contraceptive Method = L | Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]`$) = 0.429$. These conditional probabilities are equivalent to rules of the form `Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]` $\Rightarrow$ `Contraceptive Method = N` (36.3%), `Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]` $\Rightarrow$

`Contraceptive Method = S` (12.5%), `Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]` $\Rightarrow$ `Contraceptive Method = L` (42.9%), known as association rules. Given a set of association rules, such as those shown above, where the consequent describes distinct groups (i.e., `Contraceptive Method = N`, `Contraceptive Method = S`, and `Contraceptive Method = L`), where the antecedent is shared by all groups (i.e., `Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]`), and where the percentage represents the number of instances in each group for which the association rule is true, `Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]` is called a contrast set if there is a significant difference across the percentages in each group.

Group differences are used in social sciences, medicine, psychiatry, and computer science. In the social sciences for instance, some of the research conducted involves comparing racial and ethnic groups [57] [59] [124], exploring gender differences [20] [104], and examining generational gaps [15] [107]. In medicine and psychiatry, there is a plethora of research examining patient differences for areas such as pain tolerance [26] [96], mortality rates [76], and drug response [33] [49] [52]. In computer science, most of the research in group differences is conducted in data mining.

There are several data mining tasks whose goal is to find differences between groups namely: emerging patterns, subgroup discovery, and contrast set mining. Emerging patterns was developed to capture useful contrasts between data classes or to capture emerging trends in time-stamped databases [27]. Emerging patterns have been used for classification [17] [30] [31] [117] [147], microarray analysis [74], customer analysis [116], and cancer detection [75] [139]. Subgroup discovery was developed to determine features that are characteristic for a selected class or group [61] [132]. Subgroup discovery has been applied to several real-life applications such as the analysis of brain ischaemia data [64], profiling sonogram examinations [9], mining of census data [63], and even string quartet movements [119]. Contrast set mining has been developed to detect all differences

between contrasting groups from observational multivariate data [11] [12]. Contrast set mining has been used to identify factors that distinguish between thrombolic and embolic stroke patients [65], identify the characteristics of news sites [138] and to contrast aircraft accident data with aircraft incident data in major safety databases then identify factors that are significantly associated with the accidents [88].

In this thesis, we develop and evaluate contrast set mining algorithms. Within this context, we also study techniques for discretizing continuous-valued attributes and measures for evaluating the interestingness of the contrast sets generated by our algorithms. Discretization, contrast set mining, and interestingness evaluation are just three possible steps in the much larger and comprehensive process known as knowledge discovery in databases. For an overview of all the steps involved in a typical knowledge discovery task, see [32] [34].

1.1.1 The Discretization Step

When the attributes in the dataset contain continuous values, discretization is the most frequently used method to create intervals for the attribute values. For example, in Table 1.1, *Family Income* is a quantitative attribute. We discretize (i.e., create intervals for) the quantitative attribute values that is based on their mean and standard deviation. The minimum, maximum, mean, and standard deviation for `Family Income` are: 34, 58, 47.9, and 8.5. The first cut-point in the discretization step us the mean, 47.9. The next cut-point is determined by subtracting the standard deviation, 8.5, from the mean, resulting in 39.4. This cut-point is more than half of the standard deviation away from the minimum value, so it is accepted. We subtract the standard deviation from the current cut-point, 39.4, resulting in 30.9 which is less than the minimum, so it is rejected. There are no more cut-points less than the mean. The next cut-point is determined by adding the standard deviation to the mean resulting in a value of 56.4. However, this cut-point is less than half of the standard deviation away from the maximum value, so it is re-

jected. Thus we have three intervals for `Family Income`, `Family Income = [34, 39.4)`, `Family Income = [39.4, 47.9)`, and `Family Income = [47.9, 58]`.

1.1.2 The Contrast Set Mining Step

Our contrast set mining algorithms seek to discover valid contrast sets, those contrast sets for which user-defined criteria regarding statistical significance, frequency, and specificity have been satisfied. The search space is represented as a tree structure. For example, part of the search space based on the dataset in Table 1.1 is shown in the tree in Figure 1.1, where each node represents a potential contrast set. For space considerations, and with no loss of clarity, we only show the expanded sub-tree for the potential contrast set {`Wife's Education = H`} at level one. In Figure 1.1, there are 16 possible contrast sets that could differentiate the couples where {`Wife's Education = H`}.

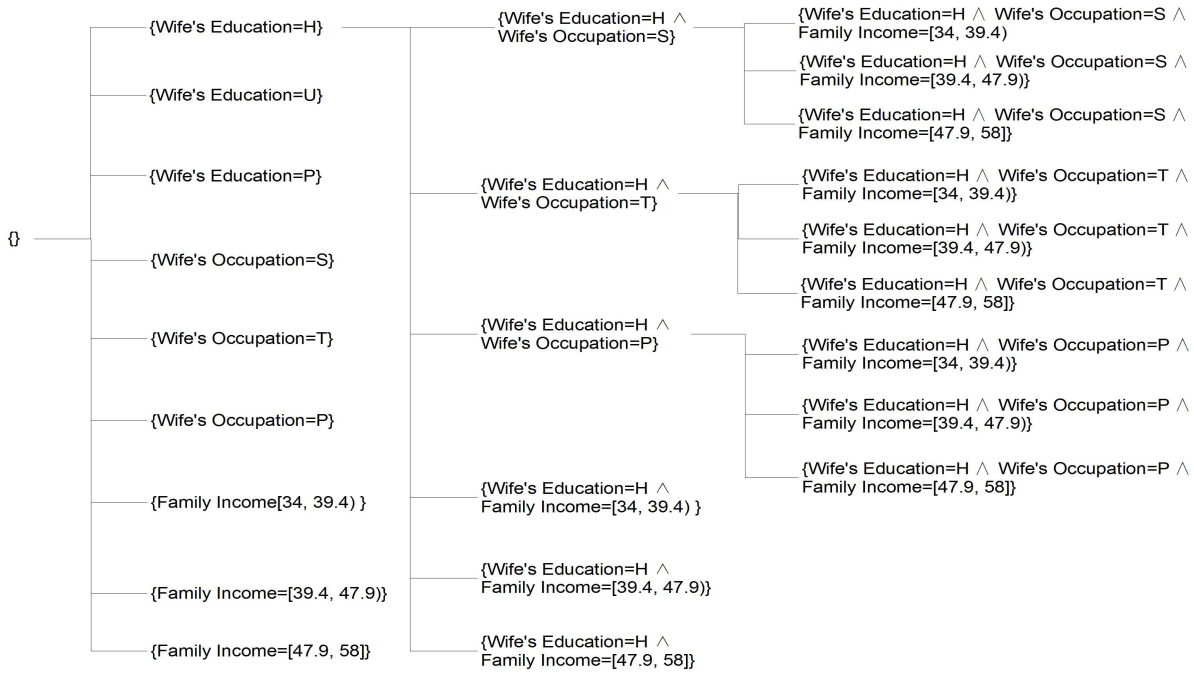


Figure 1.1: The partial search tree based upon Table 1.1

Levels one, two, and three contain all the potential contrast sets built from one, two,

and three attributes, respectively. For example, the one-attribute potential contrast sets at level one contain the eight attribute-value pairs `Wife's Education = H`, `Wife's Education = U`, `Wife's Education = P`, `Wife's Occupation = S`, `Wife's Occupation = T`, `Wife's Occupation = P`, `Family Income = [34, 39.4)`, `Family Income = [39.4, 47.9)`, and `Family Income = [47.9, 58]`. That is, one attribute-value pair for every combination of attribute and domain values. The potential two-attribute contrast sets at level two are derived by pairing each one attribute contrast set with all other contrast sets that follow it at level one, except those that share a common attribute. Consequently, `Wife's Education = H` is paired with `Wife's Occupation = S`, `Wife's Occupation = T`, `Wife's Occupation = P`, `Family Income = [34, 39.4)`, `Family Income = [39.4, 47.9)`, and `Family Income = [47.9, 58]`, at level two (i.e., follows `Wife's Education = H` in the tree), but is not paired with `Wife's Education = U` and `Wife's Education = P` because these attribute-value pairs share the attribute `Wife's Education`. Similarly, at level three, the potential three-attribute contrast sets at level two are derived by pairing each two attribute contrast set with all other contrast sets that follow it at level two, except those that share a common attribute. Consequently, `Wife's Education = H` $\wedge$ `Wife's Occupation = S` is paired with `Wife's Education = H` $\wedge$ `Family Income = [34, 39.4)`, `Wife's Education = H` $\wedge$ `Family Income = [39.4, 47.9)`, and `Wife's Education = H` $\wedge$ `Family Income = [47.9, 58]`, at level two (i.e., follows `Wife's Education = H` $\wedge$ `Wife's Occupation = S` in the tree), but is not paired with `Wife's Education = H` $\wedge$ `Wife's Occupation = T` and `Wife's Education = H` $\wedge$ `Wife's Occupation = P` because these attribute-value pairs share the attributes `Wife's Education` and `Wife's Occupation`.

In our search for valid contrast sets from the set of possible contrast sets, we are looking for contrast sets which are frequent (i.e., the percentage of instances in the dataset exceeds some threshold), large (i.e., the maximum difference between the percentage of instances in each group exceeds some threshold), specific (i.e., the contrast set is

built from at least two attributes), and significant (i.e., there is a statistically significant difference between the groups).

Beginning at the first level of the tree with the contrast set {`Wife's Education = H`}, we find the association rules `Wife's Education = H` $\Rightarrow$ `E = N` (36.4%), `Wife's Education = H` $\Rightarrow$ `E = S` (12.5%), and `Wife's Education = H` $\Rightarrow$ `E = L` (42.9%). Now, let's assume there is a significant difference between the groups at the 95% significance level. Also, since the contrast set occurs in 30.8% of the transactions, let's assume it is frequent, and the maximum difference between the percentage of transactions, which occurs between `Contraceptive Method = S` and `Contraceptive Method = L`, is 30.4%, so let's assume it's large. Thus, {`Wife's Education = H`} meets our criteria for a valid contrast set.

Continuing on to level two of the tree, we combine {`Wife's Education = H`} with {`Wife's Occupation = S`}, {`Wife's Occupation = T`}, {`Wife's Occupation = P`}, `Family Income = [34, 39.4)`, `Family Income = [39.4, 47.9)`, and `Family Income = [47.9, 58]` Consider the potential contrast set {`Wife's Education = H` $\wedge$ `Wife's Occupation = S`}. We find the association rules `Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\Rightarrow$ `E = N` (36.4%), `Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\Rightarrow$ `E = S` (12.5%), and `Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\Rightarrow$ `E = L` (42.9%).

Again, let's assume there is a significant difference between the groups and that the contrast set is frequent, specific, and large. Thus, {`Wife's Education = H` $\wedge$ `Wife's Occupation = S`} is valid. We then combine {`Wife's Education = H` $\wedge$ `Wife's Occupation = S`} at level two of the tree with {`Wife's Education = H` $\wedge$ `Family Income = [47.9, 58]`} and {`Wife's Education = H` $\wedge$ `Family Income = [47.9, 58]`}. Consider the potential combine set {`Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]`}. We find the association rules `Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]`

$\Rightarrow$ `E = N` (36.4%), `Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]` $\Rightarrow$ `E = S` (12.5%), and `Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]` $\Rightarrow$ `E = L` (28.6%).

Finally, let's assume there is a significant difference between the groups and that the contrast set is frequent, specific, and large. Thus, {`Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]`} is valid.

1.1.3 The Interestingness Evaluation Step

When the number of attributes in a dataset is large, many contrast sets may be discovered. A user may then be required to determine whether it is an interesting discovery. For example, consider the contrast sets shown in Table 1.2. Which contrast set should be considered the most interesting? Perhaps, given just eight contrast sets, this is not a daunting task. But given hundreds, or even thousands of contrast sets, it is simply not feasible for a user to manually identify interesting contrast sets.

Table 1.2: Which contrast sets are most interesting?

Contrast Set	*Coverage*
{`Wife's Education = U` $\wedge$ `Wife's Occupation = T`}	0.308
{`Wife's Education = P` $\wedge$ `Wife's Occupation = P` $\wedge$ `Family Income:[39.4, 47.9)` }	0.192
{`Family Income = [47.9, 58]`}	0.461
{`Wife's Education = U` $\wedge$ `Wife's Occupation = T` $\wedge$ `Family Income = [47.9, 58]`}	0.231
{`Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [34, 39.4)`}	0.231
{`Family Income = [39.4, 47.9)`}	0.269
{`Wife's Education = H` $\wedge$ `Wife's Occupation = S` $\wedge$ `Family Income = [47.9, 58]`}	0.269
{`Wife's Education = P` $\wedge$ `Wife's Occupation = P` $\wedge$ `Family Income = [39.4, 47.9)`}	0.192

What is needed is an objective measure of interestingness to assist in the interpretation and evaluation of the information discovered. Measures of interestingness have been widely studied and explored [39] [48]. Objective measures are based on the structure

of the discovered pattern, such as the frequency with which combinations of certain attribute-values appear in instances of the dataset [3]. Probability-based measures in particular, have been studied extensively as a means for evaluating the generality and reliability of association rules [39] [69] [120].

In this thesis, we study 14 objective measures and evaluate them for ranking search results which can then be used by a domain expert for further subjective assessment. For example, coverage, is one of the objective measures evaluated. Coverage measures the number of instances in the dataset where the contrast set occurs. The coverage of the contrast sets in Table 1.2, in order from top to bottom, are 0.398, 0.192, 0.461, 0.231, 0.231, 0.269, 0.269, and 0.192, respectively. If we consider the most interesting contrast set to simply be the one which has the highest coverage, then {`Family Income = [47.9, 58]`} is the most interesting.

1.1.4 Objective of Thesis

The objective of this thesis is to develop new algorithms and techniques for the efficient generation, ranking, and reporting of valid contrast sets. The following goals are necessary to meet the main objective:

- To introduce and evaluate the COSINE algorithm for efficiently generating contrast sets from both qualitative and quantitative data.
- To develop and evaluate techniques for reducing the size of the search space.
- To introduce and evaluate the GENCCS algorithm that traverses the reduced search space tree.
- To develop and evaluate objective measures as heuristic measures of interestingness for ranking contrast sets.

1.2 Contributions of Thesis

This thesis makes five original contributions:

- We introduced and evaluated the *Discretize* method for quantitative attributes. It is a novel, yet simple, discretization approach akin to equal-width binning, (i.e., it is not an expensive approach), but its main advantage over equal-width binning is that the width of the intervals do not need to be provided a priori, nor does it require that the number of intervals for each attribute be the same. *Discretize* uses the mean and standard deviation to determine both the width and number of intervals for continuous values. Consequently, the number of intervals for an attribute is merely dependent on the distribution of the values for that attribute.

- We introduced and evaluated the COSINE algorithm for contrast set mining [113]. COSINE can discover valid contrast sets on large datasets with categorical and quantitative attributes, where the association rules generated can have multiple discrete and continuous values in both the antecedent and consequent. No previous method can solve contrast set mining problems with these conditions. COSINE uses a vertical data format, based upon diffsets [143], which not only allows simultaneous generation and support counting of the contrast sets, but also an efficient method of storing the instances in the dataset which can then be easily propagated throughout the search space. It also uses a backtracking search paradigm in order to enumerate all possible configurations of contrast sets in the search space. Both techniques allow for efficient pruning of the search space.

- We introduced and evaluated the GENCCS algorithm for contrast set mining [112]. GENCCS builds on COSINE, in that, instead of enumerating the entire search space in order to identify valid contrast sets, it limits the size of the search space by reducing the number of attributes and attribute-values from which contrast sets

are derived. It utilizes mutual information and all confidence to select only the attributes and attribute-values that are highly correlated, resulting in a smaller search space.

- We introduced a novel contrast set, called the λ-contrast set (i.e., lambda contrast set) [115] and modified the COSINE and GENCCS algorithms to discover λ-contrast sets and correlated λ-contrast sets, respectively. A λ-contrast set is a contrast set where the ratio of maximum to minimum support across the groups is sufficiently large. A special case of the λ-contrast set, called the ∞-contrast set, (i.e., infinity contrast set) has a minimum support of zero in at least one group.

- We introduced and evaluated 14 measures for describing and ranking the interestingness of contrast sets. Although interestingness measures are used in many other data mining tasks, such as association rule mining, their use for ranking the interestingness of contrast sets is novel. We introduced the distribution difference measure for ranking contrast sets [114]. In the thesis, we also adapted the growth rate and unusualness measures from emerging patterns and subgroup discovery, respectively, for contrast set mining. Finally, we adapted 10 probability based measures, from association rule mining for contrast set mining, namely: coverage, lift, leverage, change of support, the Jaccard coefficient, ϕ coefficient, Yule's Q coefficient, conviction, and collective strength.

1.3 Outline of Thesis

The remainder of this thesis is organized as follows. In Chapter 2, we provide a general overview of classical association rule mining techniques and algorithms to highlight the important characteristics of each technique. We present a detailed survey of algorithms and techniques proposed for contrast set mining. Through this survey, we seek to impose a logical order on the important developments in contrast set mining, as well as to

highlight weaknesses and problems which we aim to address in this thesis. Finally, we present a general overview of interestingness measures from four data mining tasks which are closely related to contrast set mining.

In Chapter 3, we put forth our conceptual models for a contrast set, a correlated contrast set and our notions of a valid contrast set, a valid λ-contrast set, a valid correlated contrast set, and a valid λ-correlated contrast set. We describe our COSINE algorithm for mining valid and valid λ-contrast set, and our GENCCS algorithm for mining valid correlated valid and correlated valid λ-contrast sets.

In Chapter 4, we suggest 14 objective measures that evaluate the presence and absence of the contrast set in the various groups to generate a real-valued index which can be translated to an integer-valued rank that represents the interestingness of the contrast set relative to other contrast sets generated from the same dataset. We also present an additional objective measure that combines the ranks of a contrast set determined by multiple interestingness measures.

In Chapter 5, we describe experimental results obtained through an evaluation of both the COSINE and GENCCS algorithms. We also present experimental results describing the performance of GENCCS when mutual information and all-confidence are used to reduce the search space. Finally, we present experimental results evaluating COSINE and GENCCS for mining valid λ contrast sets and λ correlated contrast sets, respectively.

In Chapter 6, we present two case studies that demonstrate how COSINE and GENCCS are used in practice. In the first case study, a cross cultural study of health decision-making was conducted in order to better understand factors that influence ethical decision-making behaviour of physicians. In the second case study, a World of Warcraft user interface study was conducted in order to examine the effect that user created interfaces have had on the user community. The COSINE and GENCCS algorithms were evaluated on these two datasets and feedback was obtained and used as an informal assessment of the usefulness of the results and general methodology.

In Chapter 7, we provide a summary of our work and suggest areas for future work.

Chapter 2

Background and Related Work

Contrast sets have been shown to be a useful mechanism for describing differences between groups. These groups are defined by a selected property that distinguishes one from the other (e.g., customers who default on their mortgage versus those who don't). Previously proposed contrast set mining techniques have all been based on a horizontal mining approach [3] that is restricted to categorical attributes [11] [12] [47] or to a limited number of quantitative attributes [114]. They also enumerate the entire search space in order to identify significant differences amongst groups [11] [47]. On larger datasets, the search space becomes inordinately large, and the search prohibitively expensive, resulting in a large number of results which must eventually be analyzed and evaluated by a domain expert. Before we discuss our algorithms that address these weaknesses, we survey the existing algorithms and techniques proposed for contrast set mining.

This chapter is organized as follows. In Section 2.1, we present a brief overview of association rule mining and examine algorithms to mine association rules. In Section 2.2, we provide a detailed review of contrast set mining and survey the approaches to mining contrast sets. In Section 2.3, we describe methodologies and heuristics in areas closely related to contrast set mining.

2.1 Association Mining

The problem of mining contrast sets is based upon the problem of mining association rules. Let $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ be as a set of m attributes, called *items*. Let $\mathcal{D}$ be a set of database transactions, where each transaction T is described by a vector of m attribute-value pairs $a_i = v_1, a_2 = v_2, \ldots, a_m = v_m$, and each v_j is selected from a set of discrete values. The collection of items contained in transaction T is an *itemset*. Transaction T contains X, a set of items, if $X \subseteq T$. An *association rule* is as an implication of the form $X \Rightarrow Y$, where $X \subset A$, $Y \subset A$, and $X \cap Y = \emptyset$. X is referred to as the *antecedent* and Y, the *consequent* of the rule [3, 4].

Support and confidence are two metrics which are used to describe the quality and interestingness of an association rule. The *support* of an association rule is the fraction of the transactions in D that contain both X and Y. The *confidence* of an association rule is the fraction of transactions in D containing X that also contain Y. Association rules that satisfy both a minimum support threshold, and a minimum confidence threshold are called *strong*. A *frequent itemset* is an itemset whose support satisfies a predetermined minimum support threshold. *Association rule mining* is the process of finding all frequent itemsets and generating strong association rules from the frequent itemsets.

Most frequent itemset mining algorithms are variants of Apriori [3], the first algorithm for mining frequent itemsets. Apriori uses a bottom-up, breadth-first search technique to enumerate every single frequent itemset. The algorithm begins by computing the frequent itemset containing one attribute-value pair. This requires scanning all transactions in the dataset. These frequent items are combined to form candidate itemsets with two attribute-value pairs, called 2-itemsets. The database is then scanned again to determine the support for the candidate 2-itemsets. The frequent 2-itemsets are used then to form candidate 3-itemsets, and the process is repeated until all frequent itemsets have been enumerated. Apriori achieves good performance by pruning the search space using the

downward closure property, where subsets of a frequent itemset must themselves be frequent.

Apriori-inspired algorithms [18] [80] [92] [109] generally perform well with sparse datasets, where the frequent itemsets are very short. However, with dense datasets having many long frequent itemsets, these algorithms perform poorly. One common alternative is to mine only maximal itemsets that are frequent. A frequent itemset is maximal if it is not the subset of any other itemset. All-MFS [43], MaxMiner [13], Pincer-Search [77], Depth-Project[1], and Mafia [19] are some algorithms that discover maximal frequent itemsets. Another alternative is to mine only closed frequent itemsets. A frequent itemset is closed if it has no superset with the same support as the itemset. A-Close [93], Closet [94], Charm [145] are some algorithms that discover closed frequent itemsets.

Another alternative is to use vertical mining algorithms. Vertical mining algorithms [29] [46] [110] [143] [144] use a vertical format for data representation unlike the horizontal format used in traditional Apriori-like algorithms. In the vertical format, each item in the database is associated with the set of all transactions (tids) where it occurs, called the *tidset*. The frequency of the itemsets can be computed from the intersections of these tidsets, enabling itemset generation and frequency counting to occur in a single step. This is unlike Apriori which requires two steps. The intersection of these tidsets also facilitates the natural pruning of irrelevant transactions [143]. Mining algorithms using the vertical format have been shown to be very effective and outperform horizontal approaches [29] [108] [110] [144].

The main disadvantage of the vertical format is the high cardinality of the tidset for very frequent itemsets. This causes the intersecting time to be inordinately high and the memory required to store these tidsets and its intermediaries to be large. This causes the vertical approach to lose many of its advantages. An alternative vertical data representation called diffset, which only keeps track of differences in the tidsets of an itemset from

its generating frequent itemsets. Due to the significant reduction in size of diffsets over tidsets, intersection operations are performed extremely fast. Diffsets have been shown to improve by several orders of magnitude the running time of vertical algorithms like Eclat [142] which mines all frequent itemsets and GenMax [41] which mines maximal frequent itemsets. Diffset based vertical mining algorithms have been shown to outperform the classic frequent mining algorithm, Apriori, and traditional vertical mining algorithms VIPER [110], and FPGrowth [46].

2.2 Contrast Set Mining

In the problem of contrast set mining [11] [12], the database $\mathcal{D}$ is generalized to a set of multinomial *instances*, where each instance I is described by a vector of m attribute-value pairs $a_1 = v_1$, $a_2 = v_2, \ldots, a_m = v_m$, and each v_j is selected from a finite set of discrete domain values $\{v_{i1}, v_{i2}, ..., v_{im}\}$ associated with each A_i. A *contrast set* is a conjunction of attribute-value pairs defined on n mutually exclusive groups, $G_1, G_2, ..., G_n$ with no A_i occurring more than once and $G_i \cap G_j = \emptyset$, $\forall i \neq j$. These conjunctions produce rules of the form $X \Rightarrow G_1$, $X \Rightarrow G_2$, ..., $X \Rightarrow G_n$, where X is a conjunction of attribute-value pairs shared by all groups. The *support* of a contrast set with respect to a group G_i is the percentage of instances in G_i for which the contrast set is true. A contrast set, X, is called *significant* if $\exists ij \; P(X = True|G_i) \neq P(X = True|G_j)$ for groups G_i and G_j. A contrast set, X, is called *large* if the $\max_{ij} |support(X, G_i) - support(X, G_j)| \geq \delta$ for groups G_i and G_j, where δ is a user defined threshold. *Contrast set mining* is the process of finding all large and significant contrast sets.

2.2.1 Significant Difference

The statistical significance of a contrast set can be determined by testing the null hypothesis that contrast set support is equal across all groups, or, equivalently, if the contrast

set support is independent of group membership. The standard test for independence of variables in contingency tables is the χ^2 test where $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, with O_{ij} and E_{ij} equal to the observed frequency count and expected frequency count, respectively, for the cell in row i and column j of the contingency table, and r is the number of rows and c is the number of columns in the contingency table. E_{ij} is calculated as follows: $E_{ij} = \sum_j O_{ij} \sum_i \frac{O_{ij}}{N}$, where N is the total number of observations. The result is then compared to the χ^2 distribution when the null hypothesis is true. This comparison can be made because the χ^2 statistic has an approximate χ^2 distribution if the observed frequencies follow a multinomial distribution and the expected values are not too small.

In order to determine if the differences of support are significant, a test level, α, is chosen. The choice of α sets the maximum probability of rejecting the null hypothesis when it is true. For example, if we set $\alpha = 0.05$, a commonly used value, and we have a 2×2 contingency table where $\chi^2 = 5.76$, then we can calculate the degrees of freedom as $(2-1) \times (2-1) = 1$, and compare it 3.84, the value from the χ^2 distribution, at 0.05 with 1 degree of freedom. Since $5.76 > 3.84$, it can be inferred that the null hypothesis is likely false and that the contrast set support and group membership are not independent.

2.2.2 Controlling Search Error

In a hypothesis test, α sets the maximum probability of falsely rejecting the null hypothesis. Multiple hypothesis tests inflate this probability of rejection. For example, if the null hypothesis is always true, and 100000 tests of significance are made with $\alpha = 0.05$, we would accept, on average, 5000 differences that are not actually significant. Falsely rejecting the null hypothesis, i.e., concluding that there is a significant difference when none exists, is called a Type I error or a false positive. One way to control Type I errors is to use a more restrictive α value for the individual hypothesis tests, such as that used in the Bonferroni correction. The Bonferroni correction replaces α in the hypothesis tests with α/n, where n is the number of tests performed. The Bonferroni correction

can be applied using two approaches. The direct adjustment approach applies the Bonferroni correction directly to every hypothesis test. The hold-out approach [128] divides the dataset equally into exploratory and holdout sets. The Bonferroni correction is only applied to the hypothesis tests in the holdout set.

In general pattern mining, a variant of the Bonferroni correction is used which relies on multiple α values. For example, α_i, where i is size of the pattern, is adjusted by the set of all possible patterns with the same size. It has been demonstrated that using the set of all possible patterns with the same size is more effective in maintaining strict control over the risk of false discovery [128]. Specifically, $\alpha_i = \alpha/(i_{max} \times S_i)$, where i is the current pattern length, i_{max} is the maximum length of a pattern, and S_i is the number of patterns of length i. This ensures that $\alpha \geq \sum_{i=1}^{i_{max}} \alpha_i \times S_i$ [129].

2.2.3 Contrast Set Mining Algorithms

Several algorithms for contrast set mining have previously been proposed and evaluated. In this section, we survey these approaches.

2.2.3.1 STUCCO

The STUCCO (Search and Testing for Understandable Consistent COntrasts) algorithm [11] [12] which is based on the Max-Miner rule discovery algorithm [13], is the original technique for mining contrast sets. STUCCO finds significant contrast sets by testing the null hypothesis that 'contrast-set support is equal across all groups', or alternatively, 'contrast-set support is independent of group membership'. A potential contrast set X is discarded if it fails this test for independence, measured with the χ^2 test, with respect to the group variable, Y. The χ^2 statistic is measured on a $2 \times n$ contingency table, where n is the number of groups. In order to limit the possibly high probability of false rejections, STUCCO employs a variant of the Bonferroni correction for multiple hypothesis tests which uses more stringent critical values for the statistical tests as the

number of attribute-value pairs in the contrast set is increased. This variant has been found to be needlessly restrictive [128] and a modified alternative has been shown to be similarly effective [129].

STUCCO enumerates the space of all possible contrast sets using a set-enumeration tree [13, 106], which allows for each node in the tree to be visited at most once. The search space is traversed in a breadth-first manner beginning with the contrast sets which contain a single attribute-value pair, that is, the most general contrast sets. STUCCO then proceeds to more specific contrast sets which contain conjunctions of attribute-value pairs. At each level of the search space STUCCO scans the database and determines the support in each group for all the contrast sets at that level. Contrast sets that fail to meet the significant and large criteria are pruned from the search space.

STUCCO uses five pruning methods: effect size pruning, statistical significance pruning, interest based pruning, statistical surprise pruning, and maximum χ^2 pruning to disregard portions of the search space that fail to meet the search criteria. Effect size pruning requires that the maximum difference between the support of the groups in the contrast set, be greater than some predetermined threshold. Statistical significance pruning requires that there are enough data points to have a valid chi-square test, or that the maximum value the chi-squared statistic can take is large enough to be significant. Interest based pruning requires that the contrast sets selected are interesting enough, where interest is measured as representing new information from what is already known, with already selected contrast sets. Statistical surprise pruning requires that the observed frequencies are statistically different from the expected frequencies. Maximum χ^2 pruning requires that the maximum χ^2 value for all specializations at the next level of the search tree be greater than the χ^2 threshold for the current level.

2.2.3.2 MCR

The MCR (Mining Contrast Rules) algorithm [86] is a data mining technique that discovers a set of contrast rules by investigating three statistical measures: difference of confidence, difference of proportion, and correlation and χ^2. A contrast rule is defined as a 4-tuple consisting of base association rules, b, neighbourhood association rules to which the base rules are compared, $n(b)$, an ordered pair of methods, M, that measures the rules in the base, m_b and neighbourhood, $m_{n(b)}$, respectively, and a comparison function between the methods, $\triangle(m_b, m_{n(b)})$. A contrast rule is interesting, if and only if the comparison function is larger than a user-defined threshold. For the difference of confidence, $b = \{A \Rightarrow B\}$, $n(b) = \{A \Rightarrow \bar{B}\}$, $M = \langle$confidence, confidence$\rangle$, and $\triangle$=absolute difference. For the difference of confidence, $b = \{B \Rightarrow A\}$, $n(b) = \{B \Rightarrow \bar{A}\}$, $M = \langle$confidence, confidence$\rangle$, and $\triangle$=absolute difference. For the correlation and χ^2, $b = \{A \Rightarrow B, B \Rightarrow A, \bar{A} \Rightarrow \bar{B}, \bar{B} \Rightarrow \bar{A}\}$, $n(b) = \{A \Rightarrow \bar{B}\}$, $n(b) = \{A \Rightarrow \bar{B}, \bar{B} \Rightarrow A, \bar{A} \Rightarrow B, \bar{B} \Rightarrow \bar{A}\}$, $M = \langle$confidence, confidence$\rangle$, and $\triangle$=correlation. Since correlation= $\sqrt{\chi^2/N}$, both correlation and χ^2 are both measuring the same type of contrast. The definition of a contrast rule here is equivalent to the contrast set used in STUCCO when the comparison function is the chi-square measure.

MCR organizes the search space using a prefix tree which is similar to the set-enumeration tree used by STUCCO. It utilizes Borgelt's implementation [16] of the Apriori [3] algorithm and only finds contrast rules which are frequent and have no supersets with the same support. Additionally contrast rules which have very high support are removed in the preprocessing phase. However, MCR can be used with a very low minimum support threshold. This allows for the discovery of contrast rules which would normally have been overlooked. This results in a smaller set of more subtle contrast rules for analysis. These rules are ranked by a user-selected measure and the top k rules are chosen as the set of interesting rules, where k is also a user-defined measure.

2.2.3.3 Negative Contrast Sets

Contrast set mining is extended to discover negative contrast sets that can include negation of terms in the contrast set [131]. A contrast set, $Z = X \cup Y$, is a negative contrast set if support$(X) \geq$ minimum support, support$(Y) \geq$ minimum support, and either support$(X \cup \bar{Y}) \geq$ minimum support, or support$(\bar{X} \cup Y) \geq$ minimum support, or support$(\bar{X} \cup \bar{Y}) \geq$ minimum support, where $X = \emptyset, Y = \emptyset$, $X \cap Y = \emptyset$, and minimum support is a user-defined threshold. Additionally, either the interest$(X \cup \bar{Y})$, interest$(\bar{X} \cup Y)$, or interest$(\bar{X} \cup \bar{Y})$ must be $\geq$ minimum interest, where the minimum interest is a user-defined threshold and the interest$(X \cup \bar{Y}) = |\text{support}(X \cup \bar{Y}) - \text{support}(X)\text{support}(\bar{Y})|$, interest$(\bar{X} \cup Y) = |\text{support}(\bar{X} \cup Y) - \text{support}(\bar{X})\ \text{support}(Y)|$, and interest$(\bar{X} \cup \bar{Y}) = |\text{support}(\bar{X} \cup \bar{Y}) - \text{support}(\bar{X})\text{support}(\bar{Y})|$. The search algorithm is a modification of STUCCO to find negative contrast sets whose supports in the groups are significantly different. However, the Holm's sequential rejective method [50] is used for the independence test.

2.2.3.4 CIGAR

The CIGAR (Contrasting Grouped Association Rules) algorithm [47] is a contrast set mining technique that specifically identifies which pairs of groups are significantly different and whether the attributes in a contrast set are correlated. Correlation is measured with the Phi correlation co-efficient [25]. Like STUCCO, CIGAR seeks to identify contrast sets that are significant and large. In addition, CIGAR also aims to identify contrast sets that are frequent and strong. A contrast set is called *frequent*, if the support$(X, G_i) \geq \beta$ for group G_i, where δ is a user-defined threshold. A contrast set is called *strong*, if correlation$(X, G_i) \geq \gamma$ or group G_i, where γ is a user-defined threshold. A contrast set is called a *deviation* if it is significant, large, frequent, strong and the $|\text{correlation}(X, G_i, G_j) - \text{corelation}(\text{child}(X, G_i, G_j))| \geq \phi$ for groups, G_i and G_j, where ϕ is a user-defined threshold. Contrast sets that are not deviations, are called *spurious*

and are pruned from the search space.

CIGAR utilizes the same general approach as STUCCO in generating contrast sets and traversing the search space. For each contrast set it encounters in the search space, it determines if it is a deviation otherwise it is pruned from the search space. Unlike STUCCO, CIGAR uses a series of 2×2 contingency tables in determining whether a difference exists between the groups in the contrast set. For more than two groups, this is achieved by using a round-robin approach, whereby the differences are examined over all pairs of groups. This is unlike the one versus all that examines the differences between one group and all the other groups. Both approaches were experimented on in a contrast set mining task and it was found that the round robin approach was not appropriate when looking for differences between two similar groups [65]. In contrast to STUCCO, CIGAR focusses on controlling Type II error through increasing the significance level for the significance tests, and by not correcting for multiple corrections. Like STUCCO, CIGAR is only applicable to discrete-valued data.

CIGAR utilizes five pruning methods: Look-ahead χ^2 pruning, statistical significance pruning, minimum support pruning, minimum correlation pruning, and minimum correlation difference pruning. The look-ahead χ^2 pruning and the statistical significance pruning work similarly to the maximum χ^2 pruning and statistical significance pruning, respectively, employed by STUCCO. However, a Yates' correction is employed to adjust for very low expected frequencies in the statistical pruning done by CIGAR. The minimum support pruning removes all contrast sets from the search space that are not frequent. The minimum correlation pruning removes candidate sets from the search space that are not strong. The minimum correlation difference pruning removes contrast sets where the correlation difference between the parents and children in the search space are not greater than a predetermined threshold.

2.2.3.5 GEN_QCSets

Contrast set mining has also been applied to general real-valued data. The GEN_QCSets (Generate Quantitative Contrast Sets) algorithm [114] uses a modified equal-width binning interval approach to discretize continuous-values attributes where the approximate width of the intervals is provided as a parameter to the model. These intervals are then used in generating contrast sets where the consequent in the association rules contain up to two continuous-valued attributes. GEN_QCSets uses a depth-first search technique whereby each interval constitutes a potential contrast set, and this contrast set is extended if it meets the significant and large criteria, similarly to STUCCO. The Student's T-test and Z-test are used to determine significance. Both tests require at least two items for each group and this also served as the minimum support requirement. An objective measure for identifying and ranking potentially interesting contrast sets was also proposed.

2.2.4 Alternative Approaches to Contrast Set Mining

Several alternative approaches to contrast set mining have been undertaken. In this section, we survey these approaches.

2.2.4.1 Magnus Opus and C4.5

The C4.5 classification method [100] and the Magnus Opus algorithm [126] designed for mining association rules were utilized for the task of contrast set mining [130]. Magnus Opus unlike C4.5 is found to be suitable for contrast set mining. Magnus Opus searches for association rules of the form antecedent $\Rightarrow$ consequent where both the antecedent and consequent are conjunctions of attribute-value pairs. This is unlike other association rule discovery algorithms such as Max-Miner [13] where the consequent must be a single attribute-value pair. In order to use Magnus Opus for contrast set discovery, the conse-

quent is restricted to a single attribute-value pair. Magnus Opus searches the space of conjunctions of attribute-value pairs where the consequent is an attribute-value pair identifying a group and returns the maximum number of rules that maximizes the leverage. Magnus Opus prunes all association rules for which their confidence is not significantly different from the confidence of generalizations of the rules. Unlike STUCCO, significance is measured by the binomial sign test, and there are no corrections for multiple comparisons.

It has been argued that Magnus Opus only finds a subset of the contrast sets found by STUCCO and that it does a with-in groups instead of a between-group comparison [47]. Additionally, it most likely includes spurious contrast sets [131]. It has also been argued that the core contrast-set discovery task is equal to a special case of a more general association rule discovery task. However, the traditional association-rule measures of support, confidence, and lift are not directly appropriate for contrast sets, which left open the need for such measures, as well as methods to describe contrast sets to end-users [130].

2.2.4.2 Group-SAX

Contrast set mining has also been applied to time series data. The aim of which is to identify the key patterns that differentiate two sets of data [78]. This is unlike contrast set mining on discrete data which aims to identify rules that differentiate two sets of data. A time series $T = t_1, \ldots, t_m$ is an ordered set of m real-valued variables. A subsequence C in T is such that $C = t_p, \ldots, t_{p+n-1}, 1 \leq p \leq mn + 1$. A time series contrast set is the set $\{C, D\}$ such that $C = \text{TS} - \text{Diff}(T, S, n)$ and $D = \text{TS} - \text{Diff}(S, T, n)$, where S and T are time series, n is a user-defined parameter, and $\text{TS} - \text{Diff}(T_1, T_2, n)$ is a subsequence in T_1 of length n which has the largest euclidean distance to its closest match in T_2.

In order to find the contrast sets between two time series T and S, an algorithm Group-Sax [78] was developed which uses Symbolic Aggregation approXimation [79] to discretize

the real-valued time series data. In its outer loop Group-Sax places the subsequences in T in descending order of the distances to their closest match in S. In its inner loop Group-SAX places the subsequences in S in ascending order of the distances to the current subsequence in T. This ensures that the first $\mathrm{TS}-\mathrm{Diff}$ examined is the desired one, the maximum distance is set after the first iteration of the outer loop, and all subsequent invocations of the inner loop runs once.

2.2.4.3 Empirical Likelihood

An alternative approach to contrast set mining takes into account the structure of the contrasting groups, whether they are parametric, semi-parametric, or non-parametric. It also examines the imputation method for missing values in the contrasting groups. Finally, it examines how reliable the difference is between the groups [53] [99]. A semi-parametric model [97] of the groups is used, whereby one group is well known in terms of its distribution, and the other less so. A hot deck random imputation method [101] is used to replace the missing values for the groups. In order to measure the reliability of the difference between the groups with missing data, an empirical likelihood method [98] is used to construct confidence intervals for the mean and distribution differences between the groups.

2.3 Related Approaches

The goal of finding differences amongst groups is not limited to contrast set mining. There are several other areas which have examined this issue, albeit with different objectives. The focus here is on emerging patterns, subgroup discovery, correlation mining, and quantitative association rule mining as they are the most related to our work in contrast set mining. In Sections 2.3.1 and 2.3.2 the algorithms and heuristics of emerging patterns and subgroup discovery, respectively are described. In Section 2.3.3 we briefly overview

supervised descriptive rule discovery which aims to unify contrast set mining, emerging patterns, and subgroup discovery. In Section 2.3.4 an overview of correlation mining is presented. In Section 2.3.5 an overview of quantitative association rule mining is presented.

2.3.1 Emerging Pattern Mining

Emerging patterns were first defined as itemsets whose support increases significantly from one dataset to another [27]. Jumping Emerging Patterns (JEP) are emerging patterns with zero support in one dataset and greater than zero support in the other dataset [27]. The original purpose of emerging pattern mining was to "capture trends in time-stamped databases, or capture useful contrasts between classes of data". The first such algorithm, like STUCCO, was based on the Max-Miner Rule discovery algorithm [13]. Subsequent research has focused mainly on the use of discovered patterns for classification [72] [73] [30] [31]. Other areas of research include, selecting interesting emerging patterns [30], discovering emerging patterns in microarray data [17], and mining strong emerging patterns [117],.

The quality of an emerging pattern can be measured in terms of its growth rate. The growth rate of an itemset, X, from D_1 to D_2, denoted as $GrowthRate(X)$, is defined as $\frac{supp_{D_2}(X)}{supp_{D_1}(X)}$, where $supp_{D_i}(X)$ denotes the support of an itemset X in a dataset D_i. An emerging pattern is interesting if its minimum support is greater than a user-defined threshold, its minimum growth rate is also greater than a user-defined threshold, its growth rate is larger than its subset and it has a high correlation according to the chi-squared value [30].

Semantically, emerging patterns are association rules with an itemset in the rule antecedent, and a fixed consequent: ItemSet $\Rightarrow D_1$, for given data set D_1 being compared to another data set D_2 [89].

2.3.2 Subgroup Discovery

Subgroup discovery (SD) was first defined as follows: "Given a population of individuals and a property of those individuals that we are interested in, find population subgroups that are statistically 'most interesting', e.g., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest" [61] [132]. In subgroup discovery, rules have the form $X \Rightarrow Y$, where the property of interest for subgroup discovery is the class value Y that appears in the rule consequent, and the rule antecedent X is a conjunction of features (attribute-value pairs) selected from the features describing the items in a dataset.

The first system for subgroup discovery was Explora [61] which utilized both exhaustive and heuristic discovery algorithms. Subsequent research has focused on various areas including exhaustive subgroup discovery [7] [54], classification [36] [70], relational subgroup discovery [63] [132] [133], heuristics for subgroup discovery [54] [61] [63], and eliminating redundant subgroups [54] [63] [70].

The quality of the induced rules from subgroup discovery can be measured by the weighted relative accuracy, significance, coverage, and support [69] [121]. Weighted relative accuracy measures the distributive unusualness of the subgroups. Significance indicates how significant is a rule, if measured by this statistical criterion. It is computed in terms of the likelihood ratio of a rule, normalized with the likelihood ratio of the significance threshold. Coverage measures the percentage of instances covered on by one rule. Support measures the percentage of target examples (positives) covered by the rules, computed as the true positive rate for the union of subgroups.

In subgroup discovery, there is only one property of interest (class) for which subgroup descriptions (itemsets) are being built, while for contrast set mining, each group can be seen as a property of interest [65]. The two-subgroup discovery task has been shown to be equivalent to the two-group contrast mining task [64]. Along this vein, there has

been an approach to use subgroup discovery as a means to mine contrast sets [66]. The methodology proposed transformed a two-group contrast set mining task into a two-group subgroup discovery task, however the solution was targeted to a specific dataset which was quite small and where the groups were very similar. It leaves unanswered how effective it will be with a larger number of groups most importantly, as well as more dissimilar groups.

2.3.3 Supervised Descriptive Rule Discovery

The similarity of the three data mining tasks, contrast set mining, emerging pattern mining, and subgroup discovery has led to a proposal for their unification under a framework named supervised descriptive rule discovery [89]. This framework provides theoretical proof of the equality of some of the terms, heuristics and means of selecting subsets of induced patterns in each of the three discovery tasks. What is unclear is which of the algorithms developed independently in each of the three data mining tasks would be sufficient for supervised descriptive rule discovery, or whether new algorithms have to be developed. While we do not dispute the framework proposed, as it is incomplete we do not consider it further.

2.3.4 Correlation Mining

Correlation mining [18] [60] [71] [84] [146] is a data mining task that identifies underlying dependency in data. It is an alternative to traditional association rule mining that does not rely on the support measure to quantify the interestingness of the patterns; thus, correlated patterns are not restricted to frequently co-occurring attributes. As a result, those infrequent but significant patterns that are too expensive to be obtained by association rule mining can also be discovered.

There have been a number of measures for assessing correlation relationships. Interest and a χ^2 correlation measure were used in [18]. All-confidence and bond were introduced

in [90]. Both have the downward closure property [3]. H-confidence which is mathematically equivalent to all-confidence is introduced in [134] and [135]. All-confidence is shown to be a better measure for evaluating correlation relationships than χ^2 and interest [71] [60]. Mutual Information [24] [68] [136] [137] has also been used as a measure of dependency. Specifically normalized mutual information has been utilized in measuring dependency in mining for correlated patterns [55] [56].

2.3.5 Quantitative Association Rule Mining

The need for adapting traditional association rule mining algorithms for mining association rules with quantitative attributes was identified [102] [118] [148]. Discretization is the most frequently used method in dealing with quantitative attributes and quantitative rules mining. There have been numerous discretization algorithms developed. For a survey of discretization methods, see [28] [83]. Discretization is used as a tool for reduction of the theoretically infinite time complexity of quantitative association rule mining algorithms, as the number of quantitative values can theoretically be infinite. The first algorithm for mining quantitative association rules mining used equidepth [83] discretization followed by a decomposition of one quantitative attribute into several binary attributes [118]. In later work a distance-based approach is used for discretization [85]. Other approaches reduce the quantitative values to the boolean values [35], or completely omit the the discretization step with the argument that discretization causes the information loss [105].

Several measures to evaluate the quality of quantitative association rules have been proposed [42]. A new semantical understanding of quantitative association rules is given that makes current measures of quality for general association rules unsuitable. Statistical values such as the as mean value has also been proposed [10] [127]. None of the measures which were specially proposed for quantitative rules have become widely accepted and used. In most works and experiments presented in literature, the measures used are

based on classical measures such as support, confidence and lift.

Chapter 3

Mining Valid Contrast Sets

In this chapter, we introduce the data structures and algorithms utilized by our contrast set mining technique for discovering valid contrast sets.

This chapter is organized as follows. In Section 3.1, we describe the conceptual model for a contrast set and define various notions of a valid contrast set. In Section 3.2, we describe the COSINE algorithm for mining valid contrast sets. In Section 3.3, we provide a detailed walk-through of the COSINE algorithm. In Section 3.4, we describe the conceptual model for a correlated contrast set, define various notions of a valid correlated contrast set, and describe the GENCCS algorithm for mining valid correlated contrast sets. In Section 3.5, we provide a detailed walk-through of the GENCCS algorithm.

3.1 Definitions

In this section, we introduce the contrast set mining framework developed in this research. Let $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ be a set of n distinct attributes. Let $\mathcal{Q}$ and $\mathcal{C}$ denote the set of *quantitative* attributes and the set of *categorical* attributes, respectively. Let $\mathcal{V}(a_k)$ be the set of domain of values for a_k. Let $I^{(1)} = [v_{kl}, v_{kr}]$, $v_{kl}, v_{kr} \in \mathcal{V}(a_k)$ and $v_{kl} \leq v_{kr}$, be a closed interval. If $v_{kl} = v_{kr}$, then $I^{(1)}$ is simply the degenerate interval $I^{(1)} = [v_{kl}, v_{kr}] = \{v_{kl}\}$. Let $I^{(2)} = [v_{kl}, v_{kr})$, $v_{kl}, v_{kr} \in \mathcal{V}(a_k)$ and $v_{kl} < v_{kr}$, be a left-

closed, right-open interval. An *attribute-interval pair*, denoted as $a_k : I^{(x)}$, is an attribute a_k associated with an interval $I^{(x)}$, where $a_k \in \mathcal{A}$ and $x \in \{1, 2\}$. If $a_k \in \mathcal{Q}$, then $I^{(x)}$ is either a closed interval or a left-closed, right-open interval. If $a_k \in \mathcal{C}$, then $I^{(x)}$ is a degenerate interval. Let $\mathcal{T} = \{v_1, v_2, \ldots, v_p\}$, where $v_k \in \mathcal{V}(a_k)$, $1 \leq k \leq p$, be a *transaction*. Let $\mathcal{D}$ be a set of transactions, called the *database*. Let $\mathcal{U}(a_k)$ be a multi-set of $\mathcal{V}(a_k)$ containing the attribute values that actually occur in $a_k \in \mathcal{D}$. Let $\{i_1, i_2, \ldots, i_m\}$ be a set of m distinct values from the set $\{1, 2, \ldots, n\}$, $m < n$. Let $\mathcal{F} = \{a_{i_1}, a_{i_2}, \ldots, a_{i_m}\}$, $a_{i_k} \in \mathcal{A}$, be a set of m distinct class attributes.

Let $G = \{a_1 : I_1^{(x_1)}, \ldots, a_m : I_m^{(x_m)}\}$, $a_k \in \mathcal{F}, 1 \leq k \leq m$, $a_i \neq a_j$, $\forall i, j$ and $x_k \in \{1, 2\}$, be a set of m distinct class attribute-interval pairs, called a *group*. Let $X = \{a_1 : I_1^{(x_1)}, \ldots, a_q : I_q^{(x_q)}\}$, $a_i, a_j \in \mathcal{A} - \mathcal{F}$, $1 \leq k \leq q \leq n - m$, $a_i \neq a_j$, $\forall i, j$, and $x_k \in \{1, 2\}$, be a set of q distinct attribute-interval pairs, called a *quantitative contrast set*. Henceforth, we refer to a quantitative contrast set as simply a *contrast set*. A contrast set X is called k-specific if the number of attribute-interval pairs is equal to k. That is, if $|X| = k$. A contrast set X is a subset of a contrast set Y (or Y is a superset of X), denoted as $X \subseteq Y$ (or $Y \supseteq X$), if $\forall a_k : I^{(x)} \in X \Rightarrow a_k : I^{(x)} \in Y$. The *support of a contrast set X in a database $\mathcal{D}$*, denoted as $support(X)$, is the percentage of transactions in $\mathcal{D}$ containing X. The *support of a contrast set X in a group G*, denoted as $support(X, G)$, is the percentage of transactions in $\mathcal{D}$ containing $X \cup G$.

A contrast set X associated with b mutually exclusive groups $G_1, G_2, \ldots, G_b$ is called a *valid contrast set* if, and only if, the following four criteria are satisfied:

$$\exists i, j \; support(X, G_i) \neq support(X, G_j), \tag{3.1}$$

$$\exists i, j \; \max_{ij} |support(X, G_i) - support(X, G_j)| \geq \epsilon, \tag{3.2}$$

$$support(X) \geq \sigma, \tag{3.3}$$

and
$$\exists i \ \max_i \left\{ 1 - \frac{support(X, G_i)}{support(Y, G_i)} \right\} \geq \kappa, \tag{3.4}$$

where ϵ is the *minimum support difference threshold*, σ is the *minimum frequency threshold*, κ is the *minimum subset support ratio threshold*, and $X \subset Y$ such that $|X| = |Y| + 1$. Criteria 3.1 ensures that the contrast set represents a true difference between the groups. Contrast sets that meet this criteria are called *significant*. Criteria 3.2 ensures the effect size. Contrast sets that meet this criteria are called *large*. Criteria 3.3 ensures that the number of occurrences of the contrast set exceeds some number of transactions. Contrast sets that meet this criteria are called *frequent*. Criteria 3.4 ensures that the support of the contrast set in each group is different from that of its superset. Contrast sets that meet this criteria are called *specific*.

A valid contrast set is called a *maximal contrast set* if it is not a subset of any valid contrast set.

A valid contrast set is called a λ-*contrast set* (i.e., lambda contrast set) if, and only if,

$$\lambda(X) \geq \omega, \tag{3.5}$$

where ω is a user-defined *minimum support ratio threshold*, and the ratio of maximum to minimum support across the groups is sufficiently large. Formally, this ratio is defined as

$$\lambda(X) = \begin{cases} \infty, & \text{if } \min_{i=1}^{n} \{support(X, G_i)\} = 0, \\ \dfrac{\max_{i=1}^{n} \{support(X, G_i)\}}{\min_{j=1}^{n} \{support(X, G_j)\}}, & \text{otherwise.} \end{cases}$$

A large value for $\lambda(X)$ indicates that X occurs in significantly more transactions in one group G_i than in some other group G_j. Criteria 3.5 ensures that the ratio of maximum to minimum support across all groups is sufficiently large.

A λ-contrast set is called an ∞-*contrast set* (i.e., infinity contrast set) if, and only if,

$$\lambda(X) = \infty. \tag{3.6}$$

A value of ∞ indicates that X is present in at least one group G_i, and absent from at least one other group G_j.

3.1.1 A Representative Example

In this section, we provide a representative example describing the various components of the contrast set mining framework. Assume we are given a set of five distinct attributes $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}\}$ and the dataset $\mathcal{D}$ containing 10 unique transactions where for simplicity, and without loss of generality, each occurs 10 times as shown in Table 3.1.

Table 3.1: A Sample Transaction Dataset

TID	A	B	C	D	E	*Frequency*
$\mathtt{T}_1$	1	1	1	20	0	10
$\mathtt{T}_2$	0	0	1	22	1	10
$\mathtt{T}_3$	1	0	1	25	1	10
$\mathtt{T}_4$	1	1	1	21	0	10
$\mathtt{T}_5$	0	1	1	23	1	10
$\mathtt{T}_6$	0	0	1	20	0	10
$\mathtt{T}_7$	0	1	1	26	1	10
$\mathtt{T}_8$	1	1	1	26	0	10
$\mathtt{T}_9$	1	1	1	21	0	10
$\mathtt{T}_{10}$	0	0	1	25	1	10
Total						100

In Table 3.1, the *TID* column describes a unique transaction identifier, the $\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}$, and $\mathtt{E}$ columns correspond to the attributes in $\mathcal{A}$, and the *Frequency* column corresponds to the number of times the transaction occurs in $\mathcal{D}$. We assume that $\mathcal{Q} = \{\mathtt{D}\}$, where $\mathcal{U}(\mathtt{D}) = \{20, 21, 22, 23, 25, 26\}$, $\mathcal{C} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$, where $\mathcal{U}(\mathtt{A}) = \mathcal{U}(\mathtt{B}) = \mathcal{U}(\mathtt{C}) = \{\mathtt{0}, \mathtt{1}\}$, and $\mathcal{F} = \{\mathtt{E}\}$, where $\mathcal{U}(\mathtt{E}) = \{\mathtt{0}, \mathtt{1}\}$. That is, $\mathtt{D}$ is a quantitative attribute,

$\mathtt{A}, \mathtt{B}$, and $\mathtt{C}$ are categorical attributes, and $\mathtt{E}$ is the class attribute. Also, assume $X = \{\mathtt{A} : [0, 0], \mathtt{B} : [0, 0]\}$. That is, X is a quantitative contrast set containing two attribute-interval pairs. Finally, assume $\epsilon = 0.01$, $\sigma = 0.1$, $\kappa = 0.3$, and $\omega = 1.0$. Now $support(X) = \frac{30}{100} = 0.3 > \sigma = 0.1$, so X is frequent. Further, $G_1 = \{\mathtt{E} : [0, 0]\}$ and $G_2 = \{\mathtt{E} : [1, 1]\}$, so $support(X, G_1) = \frac{10}{50} = 0.2$, and $support(X, G_2) = \frac{20}{50} = 0.4$. Therefore, $|support(X, G_1) - support(X, G_2)| = |0.1 - 0.2| = 0.1 > \epsilon = 0.01$, so X is large. Since $X = \{\mathtt{A} : [0, 0], \mathtt{B} : [0, 0]\}$, its subset is $Y = \{\mathtt{A} : [0, 0]\}$,. Now, $support(Y, G_1) = \frac{10}{50} = 0.2$, and $support(Y, G_2) = \frac{40}{50} = 0.8$. Therefore, $\max\left\{\left(1 - \frac{support(X,G_1)}{support(Y,G_1)}\right), \left(1 - \frac{support(X,G_2)}{support(Y,G_2)}\right)\right\} = \left\{\left(1 - \frac{0.2}{0.2}\right), \left(1 - \frac{0.4}{0.8}\right)\right\} = 0.5 > \kappa = 0.3$, so X is specific. Since X is significant at $\alpha = 0.05$, X is a valid contrast set. Now $\lambda(X) = \frac{\max\{support(X,G_1), support(X,G_2)\}}{\min\{support(Y,G_1), support(X,G_2)\}} = \frac{\max\{0.2, 0.4\}}{\min\{0.2, 0.4\}} = \frac{0.4}{0.2} = 2$, so X is a λ-contrast set. Now, for demonstration purposes, assume $X = \{\mathtt{A} : [1, 1], \mathtt{B} : [1, 1]\}$. Then $\lambda(X) = \frac{\max\{support(X,G_1), support(X,G_2)\}}{\min\{support(Y,G_1), support(X,G_2)\}} = \frac{\max\{0, 0.4\}}{\min\{0, 0.4\}} = \frac{0.4}{0} = \infty$, so X is an ∞-contrast set.

3.1.2 A Notation for Describing a Valid Contrast Set

Here we describe a format for describing a valid contrast set. A contrast set, X, associated with b mutually exclusive groups $G_1, G_2, \ldots, G_b$ can be formally represented as

$$
\begin{aligned}
&\{X\ (support(X))\} \rightarrow \\
&\{G_1\ (support(X, G_1)) \gg \neg G_1\ (support(X, \neg G_1))\}, \\
&\{G_2\ (support(X, G_2)) \gg \neg G_2\ (support(X, \neg G_2))\}, \\
&\ldots, \\
&\{G_b\ (support(X, G_b)) \gg \neg G_b\ (support(X, \neg G_b))\}
\end{aligned}
$$

where X is the contrast set, $G_1, G_2, \ldots, G_b$ is the list of groups, $support(X)$ is the support of the contrast set, and $support(X, G_i)$ is the support of X in G_i. $G_i \gg \neg G_i$ means that the significant difference occurs between G_i and $\neg G_i$. It also means that the support of

X in G_i is significantly larger than the support of X in $\neg G_i$. This representation provides more information than STUCCO which only reports that a difference occurs but does not specify with specific group the significant difference lies.

Consider the following contrast set from the Adult dataset with Education as the group attribute with three values, Bachelors, Masters, and Doctorate:

$\{$`workingclass:[state-gov,state-gov]`, `income:[>50k, >50k]` $(0.031)\} \rightarrow$
$\{\neg$`education:[Bachelors,Bachelors]` $(0.213) \gg$ `education:[Bachelors, Bachelors]` $(0.017)\}, \{$`education:[Doctorate, Doctorate]` $(0.178) \gg \neg$`education:[Doctorate, Doctorate]` $(0.057)\}$

This contrast set is to be read as: *the contrast set "working for the state government and having an income of more than 50K" is significantly larger with persons not holding a Bachelors than those holding a Bachelors. A significant difference also occurs between those who hold a doctorate and those who do not, however it is significantly larger with persons holding a Doctorate than those not holding a Doctorate.* The group attribute `Masters` does not appear in the contrast set because there is no significant difference between those who do not hold a Masters and those who do. The following conclusions can be drawn from the contrast set:

- persons who work for the state government and have an income of more than 50K are less likely to hold a Bachelors by 92%.
- persons who work for the state government and have an income of more than 50K are more likely to hold a Doctorate by 68%.
- persons who work for the state government and have an income of more than 50K are equally likely to either hold a Masters or not.

3.2 Mining Valid Contrast Sets

Given a database $\mathcal{D}$, the COSINE (COntrast Set ExploratIon usiNg DiffsEts) algorithm, shown in Algorithm 1, generates valid contrast sets from $\mathcal{D}$.

Algorithm 1 COSINE$(\mathcal{D}, \epsilon, \sigma, \kappa, \omega, m, t)$

Input: Dataset $\mathcal{D}$, minimum support difference threshold ϵ, minimum frequency threshold σ, minimum subset support ratio threshold κ, minimum support ratio threshold ω, ranking method m, and mining type t
Output: Set of ranked valid contrast sets W

1: $W = \{\}$
2: L_0 =INITIALIZE($\mathcal{D}$)
3: L_1 =MERGE($L_0, W, \epsilon, \sigma, 0, \omega, t, 1$)
4: TRAVERSE($L_1, W, \epsilon, \sigma, \kappa, \omega, t, 1$)
5: RANK(W, m)
6: **return** W

3.2.1 General Overview

The initial call to COSINE is COSINE($\mathcal{D}, \epsilon, \sigma, \kappa, \omega, m, t$), where $\mathcal{D}$ is the dataset, ϵ is the minimum support difference threshold, σ is the minimum frequency threshold, κ is the minimum subset support ratio threshold, ω is the minimum support ratio threshold, m is the method used to rank the valid contrast sets, and t is the type of valid contrast sets to be mined.

The algorithm is described as follows. First, COSINE initializes the set of ranked valid contrast sets (line 1). Second, COSINE calls the INITIALIZE subroutine to create all 1-specific contrast sets, storing them in L_0 (line 2). Third, COSINE calls the MERGE subroutine to determine which of the 1-specific contrast sets in L_0 are valid, storing them in L_1 (line 3). Fourth, COSINE calls the TRAVERSE subroutine to find valid k-specific contrast sets using a modified generate-and-test approach, storing them in W (line 4). Fifth, COSINE calls the RANK subroutine to order the valid contrast sets in W according

to ranking measure m (line 5). Finally, COSINE returns the set of ranked valid contrast sets in W (line 6).

COSINE takes linearithmic time to create all the 1-specific contrast sets in L_0, quadratic time to determine which of them are valid, exponential time to traverse the search tree of the contrast sets in L_1, and linearithmic time to sort the valid contrast sets in W. Thus COSINE has an exponential time complexity.

3.2.2 A Primer on the COSINE Search Space

COSINE represents the search space as a tree and uses a modified generate-and-test, backtracking, depth-first approach for traversing the search space. A sample tree is shown in Figure 3.1. This tree can be used to generate all possible contrast sets from the sample dataset in Table 3.1. Each node of the tree consists of two parts: a prefix set and a combine set. For example, the topmost node at depth one contains the prefix set $\{\texttt{A} : [0, 0]\}$ and the combine set $\{\texttt{B} : [0, 0]$, $\texttt{B} : [1, 1]$, $\texttt{C} : [0, 0]$, $\texttt{C} : [1, 1]$, $\texttt{D} : [20, 22.9)$, $\texttt{D} : [22.9, 26]\}$. Similarly, the bottommost node at depth one contains the prefix $\{\texttt{D} : [22.9, 26]\}$ and the combine set $\{\}$.

The root node contains an empty prefix set and the combine set containing all possible attribute-interval pairs. For example, $\texttt{A} : [0, 0]$, $\texttt{A} : [1, 1]$, $\texttt{B} : [0, 0]$, $\texttt{B} : [1, 1]$, $\texttt{C} : [0, 0]$, and $\texttt{C} : [1, 1]$ correspond to the attribute-interval pairs associated with the categorical attributes A, B, and C, and $\texttt{D} : [20, 22.9)$ and $\texttt{D} : [22.9, 26]$ correspond to the attribute-interval pairs associated with the quantitative attribute D, where $\mathcal{U}(\texttt{D})$ has been discretized.

From the root node at depth zero, the nodes of the tree at depth one are generated as follows. The prefix set at each node is created by combining each of the attribute-interval pairs from the combine set at depth zero, in turn, with the empty prefix set. For example, the resulting prefix sets at depth one are $\{\texttt{A} : [0, 0]$, $\texttt{A} : [1, 1]$, $\texttt{B} : [0, 0]$, $\texttt{B} : [1, 1]$, $\texttt{C} : [0, 0]$, $\texttt{C} : [1, 1]$, $\texttt{D} : [20, 22.9)\}$, and $\{\texttt{D} : [22.9, 26]\}$. The combine set associated with each prefix

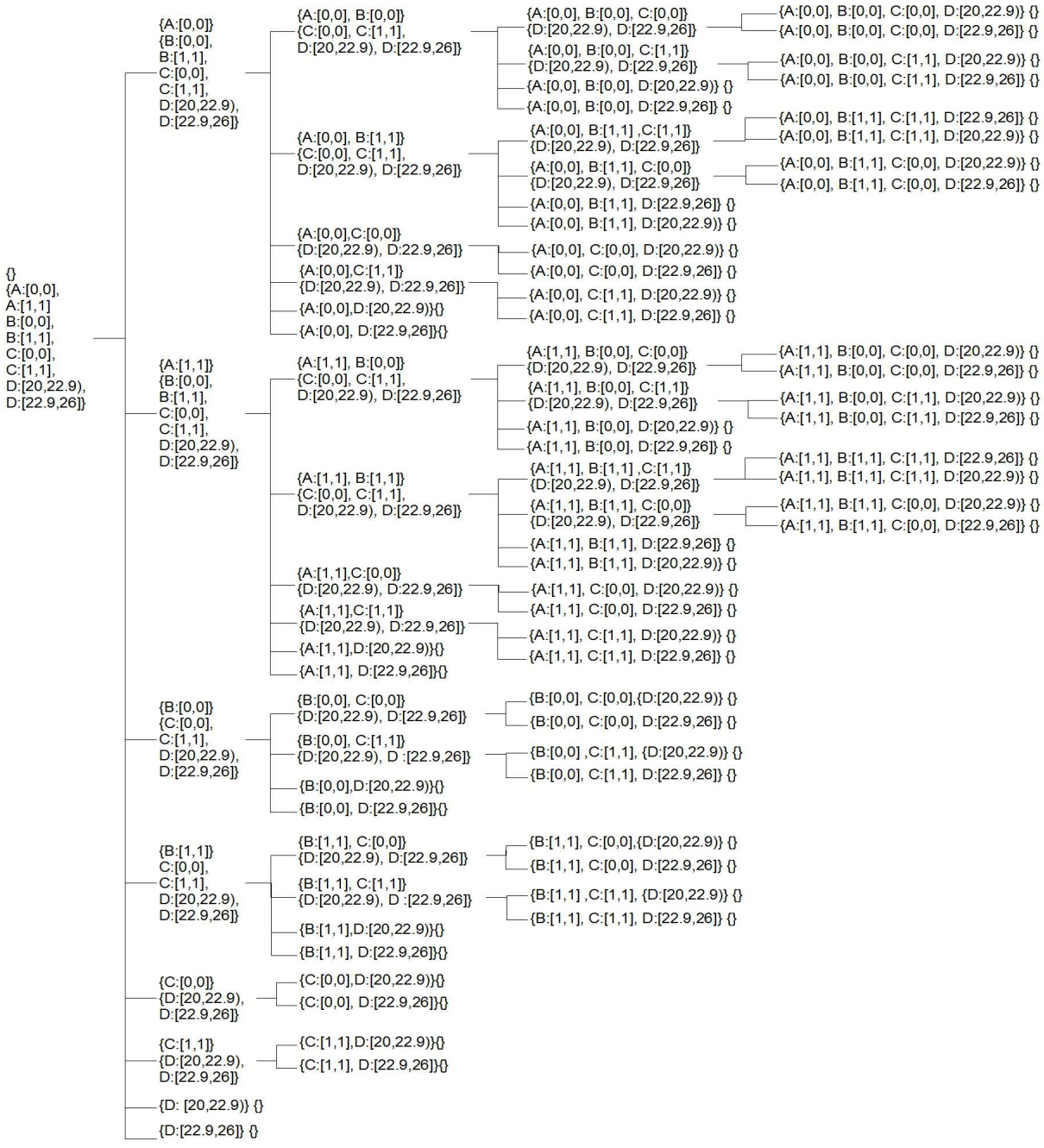


Figure 3.1: Tree representing the entire search space

set at each node at depth one is created by selecting all of the attribute-interval pairs that follow the prefix in the combine set at depth zero, except for those attribute-interval pairs that share the same attribute as the prefix. For example, the prefix set of the topmost node at depth one containing the prefix $\texttt{A} : [0, 0]$ is associated with the combine set $\{\texttt{B} : [0, 0], \texttt{B} : [1, 1], \texttt{C} : [0, 0], \texttt{C} : [1, 1], \texttt{D} : [20, 22.9), \texttt{D} : [22.9, 26]\}$. That is, all attribute-interval pairs following $\texttt{A} : [0, 0]$ in the combine set at depth zero, except $\texttt{A} : [1, 1]$ because $\texttt{A} : [1, 1]$ shares the same attribute as the prefix $\texttt{A} : [0, 0]$. Similarly, the prefix set at depth one containing $\texttt{B} : [1, 1]$ is associated with the combine set $\{\texttt{C} : [0, 0], \texttt{C} : [1, 1], \texttt{D} : [20, 22.9), \texttt{D} : [22.9, 26]\}$. That is, all attribute-interval pairs following $\texttt{B} : [1, 1]$ in the combine set at depth zero. Finally, the prefix sets at depth one containing $\texttt{D} : [20, 22.9)$ and $\texttt{D} : [22.9, 26]$ are associated with the empty combine set. The former because the only attribute-interval pair that follows it in the combine set at depth zero shares the same attribute, and the latter because there are no attribute-interval pairs following it in the combine set at depth zero.

The nodes of the tree at depth greater than one are generated in a similar manner. The prefix set for each node at depth k is created by combining each of the attribute-interval pairs from the combine set of the parent node at depth $k - 1$, in turn, with the prefix of the parent node at depth $k - 1$. The associated combine set for each node at depth k is created by selecting all of the attribute-interval pairs that follow the most recently selected attribute-interval pair from the combine set at depth $k-1$, except for those that share the same attribute. For example, consider the topmost node at depth two. The prefix sets at depth three are created from the topmost node at depth two by combining $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0]\}$ (i.e., the prefix set) with each of $\texttt{C} : [0, 0]$, $\texttt{C} : [1, 1]$, $\texttt{D} : [20, 22.9)$, and $\texttt{D} : [22.9, 26]$ (i.e, the attribute-interval pairs in the combine set). The resulting prefix sets at depth three are $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0], \texttt{C} : [0, 0]\}$, $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0], \texttt{C} : [1, 1]\}$, $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0], \texttt{D} : [20, 22.9)\}$, and $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0], \texttt{D} : [22.9, 26]\}$. These prefix sets are associated with combine sets $\{\texttt{D} : [20, 22.9), \texttt{D} : [22.9, 26]\}$, $\{\texttt{D} : [20, 22.9), \texttt{D} : [22.9, 26]\}$,

{}, and {}, respectively.

3.2.3 INITIALIZE

The INITIALIZE algorithm, shown in Algorithm 2, generates all 1-specific contrast sets. INITIALIZE iterates through n attributes, and has a worst-case time complexity of $O(n\ log\ n)$ when all the attributes are quantitative. It has a best-case time complexity of $O(1)$ when all the attributes are categorical.

Algorithm 2 INITIALIZE($\mathcal{D}$)

Input: Dataset $\mathcal{D}$
Output: Root node of the search space

```
1: Root.P = {}
2: Root.C = {}
3: for each a_i ∈ A, a_i ∉ F do
4:    if a_i ∈ Q then
5:       Root.C = Root.C ∪ DISCRETIZE(a_i, U(a_i))
6:    else
7:       Root.C = Root.C ∪ DEGENERATE(a_i, U(a_i))
8:    end if
9: end for
10: return Root
```

The algorithm is described as follows. INITIALIZE begins by initializing the prefix set Root.P and the combine set Root.C to the empty set (lines 1 and 2). It then generates the 1-specific contrast sets that are stored in the combine set of the Root (lines 3 to 9). For quantitative attributes, INITIALIZE calls the DISCRETIZE subroutine to generate 1-specific contrast sets containing closed intervals and/or left-closed, right-open intervals (line 5). For categorical attributes, INITIALIZE calls the DEGENERATE subroutine to generate 1-specific contrast sets containing degenerate intervals (line 7). Finally, INITIALIZE returns the root node of the search space (line 10).

3.2.4 DISCRETIZE

The DISCRETIZE algorithm, shown in Algorithm 3, uses the mean and standard deviation of the values in a quantitative attribute to determine the interval cut-points in the attribute-value pairs. DISCRETIZE finds n cut-points in constant time, but requires linearithmic time to sort them. Thus its time complexity is $O(n\ log\ n)$.

Algorithm 3 DISCRETIZE(a_i, V)

Input: Quantitative attribute a_i and the collection of values V
Output: Set of 1-specific contrast sets C

1: $H = \{\}$
2: Sort V
3: Calculate V_{min}, V_{max}, V_{mean}, V_{sd}
4: $H = H \cup \{V_{min}, V_{mean}, V_{max}\}$
5: $L_{cp} = V_{mean} - V_{sd}$
6: **while** $L_{cp} \geq V_{min} + (0.5 \times V_{sd})$ **do**
7: $\quad H = H \cup \{L_{cp}\}$
8: $\quad L_{cp} = L_{cp} - V_{sd}$
9: **end while**
10: $R_{cp} = V_{mean} + V_{sd}$
11: **while** $R_{cp} \leq V_{max} - (0.5 \times V_{sd})$ **do**
12: $\quad H = H \cup \{R_{cp}\}$
13: $\quad R_{cp} = R_{cp} + V_{sd}$
14: **end while**
15: Sort H
16: $C = \{\}$
17: **for** $k = 1$ to $|H| - 2$ **do**
18: $\quad C = C \cup \{a_i : [H(k), H(k+1)\}$
19: **end for**
20: $C = C \cup \{a_i : [H(k), H(k+1)]\}$
21: **return** C

DISCRETIZE begins by initializing the set of cut-points H (line 1), sorting V in ascending order (line 2), and calculating the minimum, maximum, mean and standard deviation of the values in V (V_{min}, V_{max}, V_{mean}, and, V_{sd}, respectively) (line 3), and adding V_{min}, V_{max}, and V_{mean} to H (line 4). DISCRETIZE then calculates additional cut-points, if they exist, for those cut-points less than the mean (lines 5 to 9), and those greater than

the mean (lines 10 to 14). For those cut-points less than the mean, the initial value for L_{cp} is determined by subtracting the standard deviation, V_{sd}, from the mean, V_{mean} (line 5). If L_{cp} is more than half the standard deviation away from the minimum value, V_{min} (line 6), it is added to H (line 7). L_{cp} is updated by subtracting the standard deviation, V_{sd} (line 8), and the process is repeated (line 6). Those cut-points greater than the mean are calculated similarly to those less than the mean (lines 10 to 14). DISCRETIZE then sorts the cut-points in H in ascending order (line 15). Attribute-interval pairs are created by in C by pairing each cut-point in H with its successor to form left closed, right-open intervals associated with attribute a_i (lines 17 to 19). The last interval in C is always a closed interval (line 20). Finally, DISCRETIZE returns the 1-specific contrast sets in C.

For example, assume that V_{min}, V_{max}, V_{mean}, and, V_{sd} for an attribute, A, are 19.4, 45.8, 28.5, and 3.5, respectively. Then, initially $L_{cp} = 28.5 - 3.5 = 25.0$, and $R_{cp} = 28.5 + 3.5 = 32.0$. Since both values are greater than a standard deviation away from the minimum and maximum values, they are added to H. The process is repeated, generating additional cut-points of 21.5, 35.5, 39, and 42.5. Thus, the 1-specific contrast sets generated are $A : [19.4, 21.5), A : [21.5, 35.5), A : [35.5, 39), A : [39, 42, 5)$, and $A : [42.5, 45.8]$. For the dataset shown in Table 3.1, we can determine the cut-points for the quantitative attribute, D. V_{min}, V_{max}, V_{mean}, and, V_{sd} for D are 20, 26, 22.9 and 2.4, respectively. L_{cp} and R_{cp} are 22.9-2.4=20.5, and 22.9+2.4=25.3, respectively. Since both values are less than half a standard deviation away from the minimum and maximum values, they are not added to H. Thus the only cut-point accepted is the mean, 22.9 and the 1-specific contrast sets generated in C are $D : [20, 22.9)$, and $D : [22.9, 26]$.

3.2.5 DEGENERATE

The DEGENERATE algorithm, shown in Algorithm 4, uses the values in a categorical attribute to determine the closed intervals in the attribute-interval pairs in C (lines 2 to 4). DEGENERATE then returns the 1-specific contrast sets in C (line 5). DEGENERATE

creates each attribute-interval pair in constant time, thus its time complexity is $O(1)$.

Algorithm 4 DEGENERATE(a_i, V)

Input: A categorical attribute a_i and the collection of values V
Output: Set of 1-specific contrast sets C

1: $C = \{\}$
2: **for** each $v \in V$ **do**
3: $\quad C = C \cup \{a_i : [v, v]\}$
4: **end for**
5: **return** C

3.2.6 MERGE

From a parent node at the current level in the search space, the MERGE algorithm, shown in Algorithm 5, generates the set of child nodes at the next level of the search space whose elements satisfy Criteria 3.1, 3.2, 3.3, 3.4, and 3.5.

3.2.6.1 General Overview

The initial call to MERGE is MERGE (Parent, $W, \epsilon, \sigma, \kappa, \omega, t, l$), where Parent is the node in the search space for which child nodes are currently being generated, W is the set of valid contrast sets found so far (initially the empty set), ϵ is the minimum support difference threshold, σ is the minimum frequency threshold, κ is the minimum subset support ratio threshold, ω is the minimum ratio threshold, t is the type of valid contrast sets to be mined, and l is the current level in the search space.

The algorithm is described as follows. MERGE begins by initializing the set of Children at the next level in the search space (line 1). It then combines prefix set Parent.P with each element c of the combine set Parent.C, in turn, to create new contrast sets and to determine whether they are valid (lines 2 to 24). Required values are stored in Child, a structured variable consisting of four components: a prefix set P, a diffset D, the number of transactions count, and a combine set C. Prefix set Parent.P

Algorithm 5 MERGE(Parent, $W, \epsilon, \sigma, \kappa, \omega, t, l$)

Input: Node from which child nodes are generated, Parent, the set of valid contrast sets, W, minimum support difference threshold ϵ, minimum frequency threshold σ, minimum subset support ratio threshold κ, minimum support ratio threshold ω, mining type t, and level l.

Output: Children, nodes at the next level of the search space, and a set of valid contrast sets W.

```
1: Children = {}
2: for each c ∈ Parent.C do
3:    Child.P = Parent.P ∪ {c}
4:    Child.D = DIFFSET(Child.P)
5:    if l == 1 then
6:       Child.count = COUNT(D) − |Child.D|
7:    else
8:       Child.count = Parent.count − |Child.D|
9:    end if
10:   Child.C = SELECT(c, Parent.C)
11:   if IS_VALID(Child.P, ε, σ, κ, l) then
12:      if t == LAMBDA then
13:         if λ(Child.P) == ∞ then
14:            if ∄Y ∈ W : Child.P ⊆ Y then
15:               W = W ∪ {Child.P}
16:            end if
17:         else if λ(Child.P) ≥ ω then
18:            Children = Children ∪ {Child.P}
19:         end if
20:      else
21:         Children = Children ∪ {Child.P}
22:      end if
23:   end if
24: end for
25: UPDATE_COMBINE_SET(Parent.P, Children)
26: Sort Children by |Child.C| in descending order, then by Child.count in ascending
    order
27: return Children
```

is combined with an element c from Parent.C and the resulting contrast set is stored in Child.P (line 3). The diffset, count, and combine set for Child.P are determined and stored in Child.D (line 4), Child.count (lines 6 and 8), and Child.C (line 10), respectively. MERGE calls the IS_VALID subroutine to determine whether Child.P satisfies Crite-

ria 3.1, 3.2, 3.3, and 3.4 (line 11). If λ-contrast sets are being mined (line 12), MERGE also checks whether Child.P satisfies Criteria 3.6 (line 13). If so, it is potentially maximal, and is added to W if it has no superset already in W (lines 14 and 15). Otherwise, if Child.P satisfies Equation 3.5 (line 17), MERGE adds Child to Children (line 18). If maximal or all contrast sets are being mined (line 20), MERGE adds Child to Children (line 21). After all combine sets in Parent.C have been considered, MERGE calls UPDATE_COMBINE_SET to update the combine sets of each Child in Children (line 25). MERGE then re-orders the contrast sets in descending order by the cardinality of the combine sets in Child.C, then in ascending order by Child.count (line 26). Ordering the combine set in increasing order of support has been used in several algorithms for mining frequent itemsets [13] [1] [19] [143]. Finally, MERGE returns Children, the child node at the next level of the search space (line 27).

MERGE iterates through n contrast sets, and takes linear time to determine which of the contrast sets are valid. It takes quadratic time to update the combine sets of each Child in Children. It takes linearithmic time to sort the valid contrast sets in descending order by the cardinality of the combine sets in Child.C. It also takes linearithmic time to then sort the valid contrast sets in ascending order by Child.count. Thus MERGE has a time complexity of $O(n^2)$.

3.2.6.2 A Primer on Diffsets

The *tidset* of a contrast set, X, is the set $t(X) \subseteq \mathcal{D}$, where $t(X)$ consists of all the transactions in $\mathcal{D}$ containing X. For example, in Table 3.1, $\{\mathtt{A} : [\mathtt{1}, \mathtt{1}]\}$ occurs in transactions $\mathtt{T_1}$, $\mathtt{T_3}$, $\mathtt{T_4}$, $\mathtt{T_8}$, and $\mathtt{T_9}$, so $t(X) = \{\mathtt{T_1}, \mathtt{T_3}, \mathtt{T_4}, \mathtt{T_8}, \mathtt{T_9}\}$. Here we use a *diffset* to describe the differences between the tidset of a contrast set and that of its parent. These differences are propagated from a parent to its children starting from the root. Formally, given a combine set, $\{Y, Z\}$, with a prefix set P, let $d(PY)$, which is $t(P) - t(Y)$, denote the diffset of PY. The count(PY) is equal to count$(P) - |d(PY)|$. This can be done similarly for

the count(PZ). This formulation is used for 1-specific contrast sets. However, at higher specificities, the diffsets for each element in the set is available. Thus, the diffset of PYZ, is given as $d(PZ) - d(PY)$, and its count, count(PYZ), is count(PY) $- |d(PYZ)|$.

Consider the dataset shown in Table 3.1. We can determine the diffsets for the 1-specific contrast sets, then we can use them to calculate the frequency of the 2-specific contrast sets, and so on. For example, $\{\mathtt{A} : [1, 1]\}$ is present in transactions $\mathtt{T_1}$,$\mathtt{T_3}$,$\mathtt{T_4}$,$\mathtt{T_8}$, and $\mathtt{T_9}$ so its diffset is $\{\mathtt{T_2}, \mathtt{T_5}, \mathtt{T_6}, \mathtt{T_7}, \mathtt{T_{10}}\}$, and its count is count($\mathcal{D}$) $- |d(\{\mathtt{A} : [1, 1]\})| = 10 - 5 = 5$. To calculate the count of $\{\mathtt{A} : [1, 1], \mathtt{B} : [1, 1]\}$, we need to first find its diffset. $d(\{\mathtt{A} : [1, 1], \mathtt{B} : [1, 1]\}) = d(\{\mathtt{B} : [1, 1]\}) - d(\{\mathtt{A} : [1, 1]\}) = \{\mathtt{T_2}, \mathtt{T_3}, \mathtt{T_6}, \mathtt{T_8}, \mathtt{T_{10}}\} - \{\mathtt{T_2}, \mathtt{T_5}, \mathtt{T_6}, \mathtt{T_7}, \mathtt{T_{10}}\} = \{\mathtt{T_3}, \mathtt{T_8}\}$. Thus count($\{\mathtt{A} : [1, 1], \mathtt{B} : [1, 1]\}$) = count($\{\mathtt{A} : [1, 1]\}$) $- |d(\{\mathtt{A} : [1, 1], \mathtt{B} : [1, 1]\})| = 5 - 2 = 3$. Checking the tidset of $\{\mathtt{A} : [1, 1], \mathtt{B} : [1, 1]\}$, we observe that this matches its size.

3.2.6.3 SELECT

Given an attribute-interval pair c_i and a combine set Parent.C, SELECT determines the combine set, Child.C, for the prefix, Child.P. For each element, c_j, in Parent.C, which is greater than (appears after) c_i, SELECT adds c_j to Child.C if c_j does not share an attribute with c_i. This ensures that all contrast sets contain only attribute-interval pairs from unique attributes.

3.2.6.4 UPDATE_COMBINE_SET

Given a prefix Parent.P and the set of nodes, Children, UPDATE_COMBINE_SET updates the combine set, Child.C, for each Child in Children.

UPDATE_COMBINE_SET removes any c from Child.C, if {Parent.P$\cup c$} is not equivalent to any prefix set in Children (line 4). Recall from MERGE in Section 3.2.6 that each c in Child.C is selected from Parent.C and each Child.P in Children is created by combining Parent.P with each c_1 in Parent.C. Thus, if {Parent.P $\cup\, c$} is not equal to

Algorithm 6 UPDATE_COMBINE_SET(Parent.P, Children)

Input: Children, nodes at the next level of the search space, and Parent.P, prefix of the parent node from which all the nodes in Children were generated
Output: Children, updated with attribute-interval pairs that are valid for each Child

```
1: for each Child.C ∈ Children do
2:   for each c ∈ Child.C do
3:     if ∄ Child.P ∈ Children : {Parent.P ∪ c} = Child.P then
4:       Child.C = Child.C − c
5:     end if
6:   end for
7: end for
8: return
```

any prefix set in Children, then $\{\text{Parent.P} \cup c\}$ (i.e., some Child.P in MERGE) would have failed to satisfy our criteria for a valid contrast set. Therefore, c cannot be in the combine set of any Child.P in Children. This ensures that all the prefix sets at the next level of the search space are associated with combine sets that only contain attribute-interval pairs that when combined with the prefix sets will themselves be valid. UPDATE_COMBINE_SET then returns to the invocation of MERGE from which it was called (line 8).

UPDATE_COMBINE_SET iterates through n nodes for the next level of the search space, and examines each of their m elements to determine if they are themselves valid. The validity of each of the m elements can be determined in constant time, however updating each node requires linear time. Thus UPDATE_COMBINE_SET has a time complexity of $O(nm)$.

3.2.6.5 IS_VALID

Given a contrast set X, a minimum support difference threshold ϵ, a minimum frequency threshold σ, a minimum subset support ratio threshold κ, and a level in the search space l, the IS_VALID algorithm, shown in Algorithm 7 determines whether X satisfies Criteria 3.1, 3.2, 3.3, and 3.4. IS_VALID has a time complexity of $O(1)$.

Algorithm 7 IS_VALID$(X, \epsilon, \sigma, \kappa, l)$

Input: Contrast set X, minimum support difference threshold ϵ, minimum frequency threshold σ, minimum subset support ratio threshold κ, and a level in the search space l
Output: true or false

```
1: if LARGE(X, ε) & FREQUENT(X, σ) & SPECIFIC(X, κ) then
2:    α_l = SIGNIFICANCE_LEVEL(l)
3:    if SIGNIFICANT(X, α_l) then
4:       ONE_VERSUS_ALL(X, α_l)
5:       return true
6:    else
7:       return false
8:    end if
9: else
10:   return false
11: end if
```

The algorithm is described as follows. IS_VALID begins by determining if the contrast set, X, is large, frequent, and specific according to Criteria 3.2, 3.3, and 3.4, respectively (line 1). It then determines the significance level, α_l, at which X will be tested (line 2) in order to determine whether it is significant according to Criteria 3.1 (line 3). If so, IS_VALID determines with which pairs of groups, X is significant (line 4), then returns true (line 5). Otherwise it returns false (lines 7 and 10).

In order to test the hypothesis that "contrast set support is equal across all groups", IS_VALID constructs a $2 \times G$ contingency table where the row represents the truth of the contrast set, and the column indicates group membership. Table 3.2 shows a sample $2 \times G$ contingency table.

Table 3.2: Sample $2 \times G$ contingency table

	G_1	G_2	...	G_n	$\sum X$
X	$n(X, G_1)$	$n(X, G_2)$	...	$n(X, G_n)$	$n(X)$
$\neg X$	$n(\neg X, G_1)$	$n(\neg X, G_2)$	...	$n(\neg X, G_n)$	$n(\neg X)$
$\sum G_i$	$n(G_1)$	$n(G_2)$	...	$n(G_n)$	N

We use the χ^2 statistic, the standard test for independence of variables in contingency tables. χ^2 tests give inaccurate results when the expected values are small. To correct

for this, if any of the expected values is less than 5, we use Fisher's Exact test if the number of groups is two and the number of transactions in the dataset is less than 1000, and the Yates correction, otherwise.

A hypothesis test return a p-value that indicates the probability that the observation occurs due to chance. When many hypothesis tests are performed, there is a greater probability that some of the observations will be falsely accepted as not due to chance. In setting α values to determine statistical significance, the IS_VALID subroutine uses a direct-adjustment technique that restricts the sum of all critical values to no more than the desired upper bound on the risk of any false discovery, α. The critical value used at each level of the search space is

$$\alpha_l = \frac{\alpha}{l_{max} \times S_l},$$

where l is the level of the search space, l_{max} is the maximum number of attributes, S_l is the size of the search space at l, $S_l = \binom{m}{l}$, and m is the number of attributes in the dataset. This ensures that

$$\alpha \geq \sum_{l=1}^{l_{max}} \alpha_l \times S_l.$$

In determining significance, when we reject the null hypothesis, we can conclude that a significant difference exists across the groups. However, when there are more than two groups, we do not have enough information to determine exactly between which groups the differences lie. IS_VALID constructs a set of 2×2 contingency tables representing the absence and presence of each group and determines with which pairs there is a significant difference. This is referred to as the *one versus rest* approach. Figure 3.2 show how the 2×3 contingency table in Table 3.2 is represented in the one versus rest approach.

3.2.7 TRAVERSE

The TRAVERSE algorithm, shown in Algorithm 8, traverses the search space to find contrast sets that satisfy Criteria 3.1, 3.2, 3.3, 3.4, and 3.5.

Figure 3.2: One versus rest representation of Table 3.2

(a) G_1 and $\neg G_1$

	G_1	$\neg G_1$	
X	$n(X, G_1)$	$n(X, \neg G_1)$	$n(X)$
$\neg X$	$n(\neg X, G_1)$	$n(\neg X, \neg G_1)$	$n(\neg X)$
	$n(G_1)$	$n(\neg G_1)$	N

(b) G_2 and $\neg G_2$

	G_2	$\neg G_2$	
X	$n(X, G_2)$	$n(X, \neg G_2)$	$n(X)$
$\neg X$	$n(\neg X, G_2)$	$n(\neg X, \neg G_2)$	$n(\neg X)$
	$n(G_2)$	$n(\neg G_2)$	N

(c)

$\vdots$

(d) G_n and $\neg G_n$

	G_n	$\neg G_n$	
X	$n(X, G_n)$	$n(X, \neg G_n)$	$n(X)$
$\neg X$	$n(\neg X, G_n)$	$n(\neg X, \neg G_n)$	$n(\neg X)$
	$n(G_n)$	$n(\neg G_n)$	N

3.2.7.1 General Overview

The initial call to TRAVERSE is TRAVERSE $(L_l, W_l, \epsilon, \sigma, \kappa, \omega, t, l)$, where L_l is the set of Parent nodes for the current level of the search space, W_l is the set of valid contrast sets for the previous level l of the search space, ϵ is the minimum support difference threshold, σ is the minimum frequency threshold, κ is the minimum subset support ratio threshold, ω is the minimum ratio threshold, t is the type of valid contrast sets to be mined, and l is the current level in the search space.

The algorithm is described as follows. If maximal or λ-contrast sets are being mined (line 2), and W_l is not empty (line 3), and the prefix set Parent.P is a subset of a contrast set in W_l (line 4), TRAVERSE returns to the previous invocation of the subroutine (line 5). Any node in the search space whose prefix set has a superset in W_l need not be explored further. Otherwise, if W_l is empty or Parent.P has no superset in

Algorithm 8 TRAVERSE$(L_l, W_l, \epsilon, \sigma, \kappa, \omega, t, l)$

Input: Set of Parent nodes L_l, set of valid contrast sets, W_l, minimum support difference threshold ϵ, minimum frequency threshold σ, minimum subset support ratio threshold κ, minimum support ratio threshold ω, mining type t, and a level in the search space l

Output: W_l, updated with valid contrast sets

```
1: for each Parent ∈ L_l do
2:   if t == MAXIMAL || t == LAMBDA then
3:     if |W_l| > 0 then
4:       if ∃Y ∈ W_l : Parent.P ⊆ Y then
5:         return
6:       end if
7:     end if
8:     L_{l+1} = MERGE(Parent, W_l, ε, σ, κ, ω, t, l + 1)
9:     if |L_{l+1}| > 0 then
10:      W_{l+1} = {w ∈ W_l : Parent.P ∈ w}
11:      TRAVERSE(L_{l+1}, W_{l+1}, ε, σ, κ, ω, t, l + 1)
12:      W_l = W_l ∪ W_{l+1}
13:    else
14:      if ∄Y ∈ W_l : Parent.P ⊆ Y then
15:        W_l = W_l ∪ Parent.P
16:      end if
17:    end if
18:  else
19:    L_{l+1} = MERGE(Parent, W_l, ε, σ, κ, ω, t, l + 1)
20:    W_l = W_l ∪ Parent.P
21:    if |L_{l+1}| > 0 then
22:      TRAVERSE(L_{l+1}, W_l, ε, σ, κ, ω, t, l + 1)
23:    end if
24:  end if
25: end for
26: return
```

W_l, TRAVERSE calls the MERGE subroutine to generate the set of child nodes for the next level of the search space, L_{l+1} (line 8). If L_{l+1} is not empty (line 9), then TRAVERSE creates a subset of W_l, W_{l+1}, that contains only contrast sets from W_l which can potentially subsume Parent.P (line 10). This allows the number of contrast sets of interest to be narrowed down to the most relevant maximal contrast sets as recursive calls are made. This technique is called progressive focusing [41, 143]. TRAVERSE

then calls itself to explore the next level of the search space (line 11). After recursion completes, W_l, is updated with the elements from W_{l+1} (line 12). Otherwise, if L_{l+1} is empty (line 13) and the prefix set Parent.P is not a subset of any contrast set in W_l (line 14), TRAVERSE adds Parent.P to W_l (line 15). If all valid contrast sets are being mined (line 18), TRAVERSE also calls the MERGE subroutine to generate L_{l+1} (line 19). It then adds Parent.P to W_l (line 20). If L_{l+1} is not empty (line 21), TRAVERSE also calls itself to explore the next level of the search space (line 22). Finally, TRAVERSE returns to the invocation of the subroutine from which it was called (line 26).

A parent node, Parent, in the current level of the search space has a combine set of size b. TRAVERSE builds a search tree of space complexity $O(bm)$, where m is the number of unique attributes from which the contrast sets in the combine set were created, and takes $O(b^m)$ time to traverse the search tree. Thus, to examine n nodes, TRAVERSE has a time complexity of $O(nb^m) = O(b^m)$.

3.3 Detailed Walk-through - COSINE

We now present a detailed walk-through of the COSINE algorithm. In Sections 3.3.1, 3.3.2, and 3.3.3 we use the COSINE algorithm to mine all, maximal, and λ-contrast sets, respectively.

3.3.1 Mining All Valid Contrast Sets

Consider again the dataset shown in Table 3.1, and the search space shown in Figure 3.1. COSINE is initially called with parameters $\mathcal{D}$ = dataset in Table 3.1, $\mathcal{F}$ = attribute E, $\epsilon = 0$, $\sigma = 0.1$, $\kappa = 0.0001$, $\omega = 0$, m = Coverage, and t = ALL.

Invocation - COSINE

We create an empty set W (line 1). We use the INITIALIZE subroutine to create 8 1-

specific contrast sets, {A : [0, 0], A : [1, 1], B : [0, 0], B : [1, 1], C : [0, 0], C : [1, 1], D : [20, 22.9)}, and {D : [22.9, 26]} and store them in L_0 (line 2). We call the MERGE subroutine with L_0, W, 0, 0.1, 0.0001, 0, Coverage, ALL, and 1 (line 3). We combine the empty prefix with each element in L_0. We determine the diffset and count for each element in L_0. We call SELECT to determine the combine sets for the 1-specific contrast sets in L_0. We check which of the 1-specific contrast sets satisfy Criteria 3.1, 3.2, and 3.3. C : [1, 1] fails to satisfy Criteria 3.1, and C : [0, 0] fails to satisfy Criteria 3.3 thus L_1 = {A : [0, 0], A : [1, 1]. B : [0, 0], B : [1, 1], D : [20, 22.9), D : [22.9, 26]}. We call UPDATE_COMBINE_SET to update the combine sets of the elements in L_1. The frequencies of the contrast sets are {50, 50, 40, 60, 50, 50} and cardinalities of the combine sets are {4, 4, 2, 2, 0, 0}, respectively of each element in L_1. We first reorder the elements in decreasing cardinality of their combine sets, L_1 = {A : [0, 0], A : [1, 1]. B : [0, 0], B : [1, 1], D : [20, 22.9), D : [22.9, 26]}, then in increasing frequency, L_1 = {A : [0, 0], A : [1, 1]. B : [0, 0], B : [1, 1], D : [20, 22.9), D : [22.9, 26]}. We call TRAVERSE with L_1, W, 0, 0.1, 0.0001, 0, Coverage, ALL, and 1 (line 4).

Invocation 1 - Iteration 1 - TRAVERSE

We call the MERGE subroutine with {A : [0, 0]}, W, 0, 0.1, 0.0001, 0, Coverage, ALL, and 2 (line 19). We combine A : [0, 0] with each element in its combine set. We call SELECT to determine the combine sets for the 2-specific contrast sets. We check which of the resulting 2-specific contrast sets satisfy Criteria 3.1, 3.2, 3.3, and 3.4. {A : [0, 0], D : [20, 22.9)} does not satisfy Criteria 3.1, thus L_2 = {{A : [0, 0],B : [0, 0]}, {A : [0, 0], B : [1, 1]}, {A : [0, 0], D : [22.9, 26]}}. We call UPDATE_COMBINE_SET to update the combine sets of the elements in L_2. We determine the frequencies and cardinalities of the combine sets of the elements in L_2. The frequencies of the contrast sets are {30, 20, 30} and cardinalities of the combine sets are {2, 2, 0}, respectively of each element in L_2. We first reorder the elements in decreasing cardinality of their combine sets, L_2 =

{{A : [0, 0],B : [0, 0]}, {A : [0, 0], B : [1, 1]}, {A : [0, 0], D : [22.9, 26]}}, then in increasing frequency, L_2 = {{A : [0, 0], B : [1, 1]}, {A : [0, 0], B : [0, 0]}, {A : [0, 0], D : [22.9, 26])}}. We also add {A : [0, 0]} to W (line 20). L_2 is not empty, thus we call TRAVERSE with L_2, W, 0, 0.1, 0.0001, 0, Coverage, ALL, and 2 (line 22).

Invocation 2 - Iteration 1 - TRAVERSE

We call the MERGE subroutine with {A : [0, 0], B : [1, 1]}, W, 0, 0.1, 0.0001, 0, Coverage, ALL, and 3. We combine A : [0, 0], B : [1, 1] with each element in its combine set. We check which of the resulting 3-specific contrast sets satisfy Criteria 3.1, 3.2, 3.3, and 3.4. {A : [0, 0], B : [1, 1], D : [22.9, 26]} does not satisfy Criteria 3.4, thus L_3 = {}. We also add {A : [0, 0], B : [1, 1]} to W. L_3 is empty thus the first iteration of the second invocation of TRAVERSE is complete.

Invocation 2 - Iteration 2 - TRAVERSE

We call the MERGE subroutine with {A : [0, 0], B : [0, 0]}, W, 0, 0.1, 0.0001, 0, Coverage, ALL, and 3. We combine {A : [0, 0], B : [0, 0]} with each element in its combine set. We call SELECT to determine the combine sets for the 3-specific contrast sets. We check which of the resulting 3-specific contrast sets satisfy Criteria 3.1, 3.2, 3.3, and 3.4. L_3 = {{A : [0, 0], B : [0, 0], D : [22.9, 26]}}. We call UPDATE_COMBINE_SET to update the combine set of the element in L_3. We also add {A : [0, 0] ,B : [0, 0]} to W. L_3 is not empty, thus we call TRAVERSE with L_3, W, 0, 0.1, 0.0001, 0, Coverage, ALL, and 3.

Invocation 3 - Iteration 1 - TRAVERSE

We call the MERGE subroutine with {A : [0, 0], B : [0, 0], D : [22.9, 26]}, W, 0, 0.1, 0.0001, 0, Coverage, ALL, and 4, but L_4 is empty because the combine set of {A : [0, 0], B : [0, 0], D : [22.9, 26]} is empty. We add {A : [0, 0], B : [0, 0], D : [22.9, 26]} to W (line 20), and the third invocation of TRAVERSE is complete.

Invocation 3 (continued) - TRAVERSE

L_3 has only one element, so the end of the first iteration of the third invocation of TRAVERSE is also the end of the third invocation of TRAVERSE. The algorithm returns to the second invocation, third iteration of TRAVERSE.

Invocation 2 - Iteration 3 - TRAVERSE

We call the MERGE subroutine with $\{\texttt{A} : [0, 0], \texttt{D} : [22.9, 26]\}$, W, 0, 0.1, 0.0001, 0, Coverage, ALL, and 3, but L_3 is empty because the combine set of $\{\texttt{A} : [0, 0], \texttt{D} : [22.9, 26]\}$ is empty. We add $\{A : [0, 0], D : [22.9, 26]\}$ to W, and the third iteration of the second invocation of TRAVERSE is complete.

Invocation 2 (continued) - TRAVERSE

L_2 has only three elements, so the end of the third iteration of the second invocation of TRAVERSE is also the end of the second invocation of TRAVERSE. The algorithm returns to the first invocation, second iteration of TRAVERSE.

Table 3.3 lists the top 7 valid contrast sets ranked by their coverage. We use Coverage as the ranking measure because its value is equal to the support of the contrast set, which we

The important aspects of the COSINE algorithm for mining all valid contrast sets have been clearly demonstrated.

3.3.2 Mining Maximal Valid Contrast Sets

Consider again the dataset shown in Table 3.1, and the search space tree shown in Figure 3.1. COSINE is initially called with parameters $\mathcal{D}$ = dataset in Table 3.1, $\mathcal{F}$ = attribute E, $\epsilon = 0$, $\sigma = 0.1$, $\kappa = 0.0001$, $\omega = 0$, m = Coverage, and t = MAXIMAL.

Table 3.3: Summary of all valid contrast sets discovered

Rank	*Contrast Set*	*Coverage*
1	$\{\mathtt{B}:[1,1]\ (0.6)\} \rightarrow \{\mathtt{E}:[0,0]\ (0.8) \gg \neg\mathtt{E}:[0,0]\ (0.4)\}$	0.6
2	$\{\mathtt{A}:[1,1]\ (0.5)\} \rightarrow \{\mathtt{E}:[0,0]\ (0.8) \gg \neg\mathtt{E}:[0,0]\ (0.2)\}$	0.5
2	$\{\mathtt{A}:[0,0]\ (0.5)\} \rightarrow \{\mathtt{E}:[1,1]\ (0.8)\} \gg \neg\mathtt{E}:[1,1]\ (0.2)$	0.5
2	$\{\mathtt{D}:[20,22.9)\ (0.5)\} \rightarrow \{\mathtt{E}:[0,0]\ (0.8) \gg \neg\mathtt{E}:[0,0]\ (0.2)\}$	0.5
2	$\{\mathtt{D}:[22.9,26.0]\ (0.5)\} \rightarrow \{\mathtt{E}:[1,1]\ (0.8) \gg \neg\mathtt{E}:[1,1]\ (0.2)\}$	0.5
6	$\{\mathtt{A}:[1,1],\mathtt{B}:[1,1]\ (0.4)\} \rightarrow \{\mathtt{E}:[0,0]\ (0.8) \gg \neg\mathtt{E}:[0,0]\ (0)\}$	0.4
6	$\{\mathtt{B}:[0,0]\ (0.4)\} \rightarrow \{\mathtt{E}:[1,1]\ (0.6) \gg \neg\mathtt{E}:[1,1]\ (0.2)\}$	0.4

Invocation - COSINE

The invocation of COSINE for mining maximal valid contrast sets is the same as all mining valid contrast sets described above, except that m = MAXIMAL, and W is referred to as W_1. So we call TRAVERSE with L_1 W_1, 0, 0.1, 0.0001, 0, Coverage, MAXIMAL, and 1 (line 4).

Invocation 1 - Iteration 1 - TRAVERSE

W_1 is empty so we call the MERGE subroutine with $\{\mathtt{A}:[0,0]\}$, W_1, 0, 0.1, 0.0001, 0, Coverage, MAXIMAL, and 2 (line 8). We combine $\mathtt{A}:[0,0]$ with each element in the combine set of $\{\mathtt{A}:[0,0]\}$. We call SELECT to determine the combine sets for the 2-specific contrast sets. We check which of the resulting 2-specific contrast sets satisfy Criteria 3.1, 3.2, 3.3, and 3.4. $\{\mathtt{A}:[0,0],\mathtt{D}:[20,22.9)\}$ does not satisfy Criteria 3.1, thus $L_2 = \{\{\mathtt{A}:[0,0],\mathtt{B}:[0,0]\}, \{\mathtt{A}:[0,0], \mathtt{B}:[1,1]\}, \{\mathtt{A}:[0,0], \mathtt{D}:[22.9,26]\}\}$. We call UPDATE_COMBINE_SET to update the combine sets of the elements in L_2. We determine the frequencies and cardinalities of the combine sets of the elements in L_2. The frequencies of the contrast sets are $\{30, 20, 30\}$ and cardinalities of the combine sets are $\{2, 2, 0\}$, respectively of each element in L_2. We first reorder the elements in decreasing cardinality of their combine sets, $L_2 = \{\{\mathtt{A}:[0,0],\mathtt{B}:[0,0]\}, \{\mathtt{A}:[0,0], \mathtt{B}:[1,1]\}, \{\mathtt{A}:[0,0],$

D : [22.9, 26]}}, then in increasing frequency, $L_2 = \{\{$A : [0, 0], B : [1, 1]$\}$, $\{$A : [0, 0], B : [0, 0]$\}$, $\{$A : [0, 0], D : [22.9, 26]$)\}\}$. Since L_2 is not empty (line 9), we create a new set of maximal elements W_2 which is empty, since W_1 is empty (line 10). We then call TRAVERSE with L_2, W_2, 0, 0.1, 0.0001, 0, Coverage, MAXIMAL, and 2 (line 11).

Invocation 2 - Iteration 1 - TRAVERSE

W_2 is empty so we call the MERGE subroutine with $\{$A : [0, 0], B : [1, 1]$\}$, W_2, 0, 0.1, 0.0001, 0, Coverage, MAXIMAL, and 3. We combine $\{$A : [0, 0], B : [1, 1]$\}$ with each element in the combine set of $\{$A : [0, 0], B : [1, 1]$\}$. We call SELECT to determine the combine sets for the 3-specific contrast sets. We check which of the resulting 3-specific contrast sets satisfy Criteria 3.1, 3.2, 3.3, and 3.4. $\{$A : [0, 0], B : [1, 1], D : [22.9, 26]$\}$ does not satisfy Criteria 3.4, thus $L_3 = \{\}$. Both L_3 and W_2 are empty, so we add $\{$A : [0, 0], B : [1, 1]$\}$ to W_2 (line 15). The first iteration of the second invocation of TRAVERSE is complete.

Invocation 2 - Iteration 2 - TRAVERSE

W_2 is not empty so we check whether $\{$A : [0, 0], B : [0, 0]$\}$ has a superset in W_2 (lines 3 and 4). It does not, since W_2 only contains $\{$A : [0, 0], B : [1, 1]$\}$ which cannot be a superset of $\{$A : [0, 0], B : [0, 0]$\}$, so we call the MERGE subroutine with $\{$A : [0, 0], B : [0, 0]$\}$, W_2, 0, 0.1, 0.0001, 0, Coverage, MAXIMAL, and 3. We combine $\{$A : [0, 0], B : [0, 0]$\}$ with each element in the combine set of $\{$A : [0, 0], B : [0, 0]$\}$. We call SELECT to determine the combine sets for the 3-specific contrast sets. We check which of the resulting 3-specific contrast sets satisfy Criteria 3.1, 3.2, 3.3, and 3.4. $L_3 = \{\{$A : [0, 0], B : [0, 0], D : [22.9, 26]$\}\}$. We call UPDATE_COMBINE_SET to update the combine set of the element in L_3. L_3 is not empty so we create a new set of maximal elements, W_3 (lines 9 and 10). W_2 only contains $\{$A : [0, 0], B : [1, 1]$\}$ which cannot be a superset of $L_3 = \{\{$A : [0, 0], B : [0, 0], D : [22.9, 26]$\}\}$, so W_3 is empty. We then call TRAVERSE with L_3,

W_3, 0, 0.1, 0.0001, 0, Coverage, MAXIMAL, and 3.

Invocation 3 - Iteration 1 - TRAVERSE

W_3 is empty so we call the MERGE subroutine with $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0], \texttt{D} : [22.9, 26]\}$, W_3, 0, 0.1, 0.0001, 0, Coverage, MAXIMAL, and 4, but L_4 is empty because $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0], \texttt{D} : [22.9, 26]\}$ has an empty combine set. Since W_3 is empty, we add $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0], \texttt{D} : [22.9, 26]\}$ to W_3 (line 15), and the first iteration of the third invocation of TRAVERSE is complete.

Invocation 3 (continued) - TRAVERSE

L_3 has only one element, so the end of the first iteration of the third invocation of TRAVERSE is also the end of the third invocation of TRAVERSE. The algorithm returns to the second invocation, third iteration of TRAVERSE. It then updates W_2 with the element, $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0], \texttt{D} : [22.9, 26]\}$, from W_3 (line 12).

Invocation 2 - Iteration 3 - TRAVERSE

W_2 is not empty so we check whether $\{\texttt{A} : [0, 0], \texttt{D} : [22.9, 26]\}$ has a superset in W_2 (lines 3 and 4). It does, $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0], \texttt{D} : [22.9, 26]\}$, so the third iteration of the second invocation of TRAVERSE. is complete.

Invocation 2 (continued) - TRAVERSE

L_2 has only three elements, so the end of the third iteration of the second invocation of TRAVERSE is also the end of the second invocation of TRAVERSE. The algorithm returns to the first invocation, second iteration of TRAVERSE. It also updates W_1 with the elements $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0], \texttt{D} : [22.9, 26]\}$ and $\{\texttt{A} : [0, 0], \texttt{B} : [1, 1]\}$ from W_2.

Table 3.4 lists the valid maximal contrast sets ranked by their coverage.

Table 3.4: Summary of valid maximal contrast sets discovered

Rank	*Contrast Set*	*Coverage*
1	$\{\mathtt{B} : [1,1]\mathtt{D} : [22.9,26]\ (0.3)\} \rightarrow$ $\{\mathtt{E} : [1,1]\ (0.4) \gg \neg\mathtt{E} : [1,1]\ (0.2)\}$	0.3
1	$\{\mathtt{B} : [1,1], \mathtt{D} : [20,22.9)\ (0.3)\} \rightarrow$ $\{\mathtt{E} : [0,0]\ (0.6) \gg \neg\mathtt{E} : [0,0]\ (0)\}$	0.3
1	$\{\mathtt{A} : [1,1], \mathtt{D} : [20,22.9)\ (0.3)\} \rightarrow$ $\{\mathtt{E} : [0,0]\ (0.6) \gg \neg\mathtt{E} : [0,0]\ (0)\}$	0.2
4	$\{\mathtt{A} : [0,0], \mathtt{B} : [1,1]\ (0.2)\} \rightarrow$ $\{\mathtt{E} : [0,0]\ (0.4) \gg \neg\mathtt{E} : [0,0]\ (0)\}$	0.2
5	$\{\mathtt{A} : [0,0], \mathtt{B} : [0,0], \mathtt{D} : [22.9,26]\ (0.1)\} \rightarrow$ $\{\mathtt{E} : [1,1]\ (0.2) \gg \neg\mathtt{E} : [0,0]\ (0)\}$	0.1

The important aspects of the COSINE algorithm for mining valid maximal contrast sets have been clearly demonstrated.

3.3.3 Mining Valid Maximal λ-Contrast Sets

Consider again the dataset shown in Table 3.1, and the search space tree shown in Figure 3.1. COSINE is initially called with parameters $\mathcal{D}$ = dataset in Table 3.1, $\mathcal{F}$ = attribute E, $\epsilon = 0$, $\sigma = 0.1$, $\kappa = 0.0001$, $\omega = 0$, m = Coverage, and t = LAMBDA.

COSINE - Invocation

The invocation of COSINE for mining valid maximal contrast sets is the same as mining valid maximal contrast sets described above, except that m = LAMBDA. So we call TRAVERSE with L_1 W_1, 0, 0.1, 0.0001, 0, Coverage, LAMBDA, and 1 (line 4).

Invocation 1 - Iteration 1 - TRAVERSE

W_1 is empty so we call the MERGE subroutine with $\{\mathtt{A} : [0,0]\}$, W_1, 0, 0.1, 0.0001, 0, Coverage, LAMBDA, and 2 (line 8). We combine $\mathtt{A} : [0,0]$ with each element in the combine set of $\{\mathtt{A} : [0,0]\}$. We call SELECT to determine the combine sets for the 2-

specific contrast sets. We check which of the resulting 2-specific contrast sets satisfy Criteria 3.1, 3.2, 3.3, and 3.4. $\{\texttt{A} : [0, 0], \texttt{D} : [20, 22.9)\}$ does not satisfy Criteria 3.1. We also check if the remaining contrast sets satisfy Criteria 3.6 and 3.5. $\{\texttt{A} : [0, 0], \texttt{B} : [1, 1]\}$ and $\{\texttt{A} : [0, 0], \texttt{D} : [22.9, 26]\}$ satisfy Criteria 3.6. Thus they are maximal and since W_1 is empty, we add them to W_1. $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0]\}$ satisfies Criteria 3.5, so we add it to L_2. Thus $L_2 = \{\{\texttt{A} : [0, 0], \texttt{B} : [0, 0]\}\}$. We call UPDATE_COMBINE_SET to update the combine sets of the elements in L_2. Since L_2 is not empty (line 9), we create a new set of maximal elements W_2 (line 10). None of the current elements in W_1, $\{\texttt{A} : [0, 0], \texttt{B} : [1, 1]\}$ and $\{\texttt{A} : [0, 0], \texttt{D} : [22.9, 26]\}$, can potentially be a superset of $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0]\}$, so W_2 is empty. We then call TRAVERSE with L_2, W_2, 0, 0.1, 0.0001, 0, Coverage, LAMBDA, and 2 (line 11).

Invocation 2 - Iteration 1 - TRAVERSE

W_2 is empty so we call the MERGE subroutine with $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0]\}$, W_2, 0, 0.1, 0.0001, 0, Coverage, LAMBDA, and 3, but L_3 is empty because $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0]\}$ has an empty combine set. Since W_2 is empty, we add $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0]\}$ to W_2 (line 15), and the first iteration of the second invocation of TRAVERSE is complete.

Invocation 2 (continued) - TRAVERSE

L_2 has only one element, so the end of the first iteration of the second invocation of TRAVERSE is also the end of the second invocation of TRAVERSE. The algorithm returns to the first invocation, second iteration of TRAVERSE. It updates W_1 with the element, $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0]\}$, from W_2.

Table 3.5 lists the valid maximal λ-contrast sets ranked by their coverage.

The important aspects of the COSINE algorithm for mining valid maximal λ- contrast sets have been clearly demonstrated.

Table 3.5: Summary of valid maximal λ-contrast sets discovered

Rank	*Contrast Set*	*Coverage*
1	$\{\mathtt{B}:[1,1],\mathtt{D}:[22.9,26)\ (0.3)\} \rightarrow$ $\{\mathtt{E}:[1,1]\ (0.4) \gg \neg\mathtt{E}:[1,1]\ (0.2)\}$	0.3
1	$\{\mathtt{A}:[0,0],\mathtt{D}:[22.9,26)\ (0.3)\} \rightarrow$ $\{\mathtt{E}:[1,1]\ (0.6) \gg \neg\mathtt{E}:[1,1]\ (0)\}$	0.3
1	$\{\mathtt{B}:[1,1],\mathtt{D}:[20,22.9)\ (0.3)\} \rightarrow$ $\{\mathtt{E}:[0,0]\ (0.6) \gg \neg\mathtt{E}:[0,0]\ (0)\}$	0.3
1	$\{\mathtt{A}:[1,1],\mathtt{D}:[20,22.9)\ (0.3)\} \rightarrow$ $\{\mathtt{E}:[0,0]\ (0.6) \gg \neg\mathtt{E}:[0,0]\ (0)\}$	0.3
1	$\{\mathtt{A}:[0,0],\mathtt{B}:[0,0]\ (0.3)\} \rightarrow$ $\{\mathtt{E}:[1,1]\ (0.4) \gg \neg\mathtt{E}:[0,0]\ (0.2)\}$	0.3
6	$\{\mathtt{A}:[0,0],\mathtt{B}:[1,1]\ (0.2)\} \rightarrow$ $\{\mathtt{E}:[0,0]\ (0.4) \gg \neg\mathtt{E}:[0,0]\ (0)\}$	0.2
7	$\{\mathtt{B}:[0,0],\mathtt{D}:[22.9,26)\ (0.2)\} \rightarrow$ $\{\mathtt{E}:[1,1]\ (0.4) \gg \neg\mathtt{E}:[1,1]\ (0)\}$	0.2

3.4 Mining Correlated Valid Contrast Sets

The task of finding all valid contrast sets from the set of all contrast sets becomes prohibitively expensive because our longest combine set contains all the distinct 1-specific contrast sets in the dataset. However, if we could narrow down which 1-specific contrast sets have the strongest relationships with each other, our search space would be smaller and more refined, resulting in a more relevant subset of valid contrast sets. We propose the notion of a correlated valid contrast set, which mutual information and all-confidence to measure the correlation between 1-specific contrast sets. Mutual information and all-confidence have previously been used in selecting attributes for correlation mining [55] [56].

In this section, we first review the concepts of mutual information and all-confidence in Section 3.4.1. In Section 3.4.2, we define the concept of a correlated valid contrast set. In Section 3.4.3, we provide an illustrative example of the correlated valid contrast set mining framework. In Section 3.4.4, we present our algorithm for mining correlated

valid contrast sets.

3.4.1 Mutual Information and All Confidence

Mutual information is a non-negative symmetric measure of the amount of information that one random variable contains about another random variable [24]. If we have two 1-specific contrast sets $X = P : I_i^{(x_i)}$ and $Y = Q : I_j^{(x_j)}$, we can represent our knowledge of X and Y in the contingency table shown in Table 3.6.

Table 3.6: Contingency table representing the relationship between X and Y

	Y	$\neg Y$	
X	$n(X, Y)$	$n(X, \neg Y)$	$n(X)$
$\neg X$	$n(\neg X, Y)$	$n(\neg X, \neg Y)$	$n(\neg X)$
	$n(Y)$	$n(\neg Y)$	N

The mutual information of X and Y, is defined as:

$$M_I(X;Y) = \log_2 \left(\frac{n(X,Y) * N}{(n(X,Y) + n(X,\neg Y) \times (n(X,Y) + n(\neg X,Y)} \right).$$

Mutual information is a good measure of dependency between attributes, so if $X = P : I_i^{(x_i)}$ and $Y = Q : I_j^{(x_j)}$ we measure $M_I(X : Y)$, $\forall i \in \mathcal{V}(P)$ and $\forall j \in \mathcal{V}(Q)$. Thus, the mutual information of P and Q is defined as:

$$M_I(P;Q) \quad = \quad \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{n(X,Y)}{N} \times M_I(X;Y)$$

where $n = |\mathcal{V}(P)|$ and $m = |\mathcal{V}(Q)|$.

Mutual information is symmetric, meaning $M_I(P;Q) = M_I(Q;P)$. Also $M_I(P;Q) \geq 0$. If $M_I(P;Q) = 0$, P and Q are independent [24]. In order to measure the relationship between pairs of 1-specific contrast sets we use all-confidence [45].

For the two 1-specific contrast sets X and Y, we can define the all-confidence as

follows:

$$A_C(X,Y) = \frac{n(X,Y)}{\max\left\{\left(n(X,Y)+n(X,\neg Y)\right),\left(n(X,Y)+n(\neg X,Y)\right)\right\}}.$$

3.4.2 Definitions

A valid contrast set is called *correlated* if every attribute-interval pair that it contains is correlated with every other attribute-interval pair in the contrast set, where correlation is determined by mutual information and all-confidence. Formally, a valid contrast set, $X = \{a_1 : I_1^{(x_1)}, a_2 : I_2^{(x_2)}, \ldots, a_n : I_n^{(x_n)}\}$, is a correlated contrast set if, and only if, the following two conditions are satisfied:

$$\forall a_i, a_j \in X, M_I(a_i; a_j) \geq \psi \tag{3.7}$$

$$\forall a_i, a_j \in X,\ \forall i \in \mathcal{V}(a_i),\ \forall j \in \mathcal{V}(a_j),\ A_C(a_i : I_i^{(x_i)}, a_j : I_j^{(x_j)}) \geq \xi \tag{3.8}$$

where ψ is the minimum mutual information threshold, and ξ is the minimum all-confidence threshold. Criteria 3.7 ensures that each attribute in a correlated contrast set is dependent on every other attribute. Criteria 3.8 ensures that each attribute-interval pair is highly correlated with every other attribute-interval pair.

3.4.3 Illustrative Example

In this section, we provide a representative example describing the various components of the correlated valid contrast set mining framework. We use the dataset from Table 3.1. Assume $X = \{\mathtt{A} : [0,0],\ \mathtt{B} : [0,0]\}$ and $\psi = 0.01$, and $\xi = 0.25$. Then, we can calculate

$M_I(\texttt{A}:[0,0];\texttt{B}:[0,0])$ as follows:

$$M_I(\texttt{A}:[0,0];\texttt{B}:[0,0]) \;=\; \log_2\left(\frac{30\times 100}{50\times 40}\right) = 0.585.$$

There are two possible values for each of A and B. Thus we can calculate $M_I(\texttt{A};\texttt{B})$ as follows:

$$\begin{aligned}
M_I(\texttt{A};\texttt{B}) &= M_I(\texttt{A}:[0,0];\texttt{B}:[0,0]) + M_I(\texttt{A}:[0,0];\texttt{B}:[1,1]) \\
&\quad + M_I(\texttt{A}:[1,1];\texttt{B}:[0,0]) + M_I(\texttt{A}:[1,1];\texttt{B}:[1,1]) \\
&= \frac{30}{100}\times\log_2\left(\frac{30\times 100}{50\times 40}\right) \\
&\quad + \frac{20}{100}\times\log_2\left(\frac{20\times 100}{50\times 60}\right) \\
&\quad + \frac{10}{100}\times\log_2\left(\frac{10\times 100}{50\times 40}\right) \\
&\quad + \frac{40}{100}\times\log_2\left(\frac{40\times 100}{50\times 60}\right) \\
&= 0.175 + (-0.117) + (-0.1) + 0.166 = 0.124.
\end{aligned}$$

Since $0.124 > 0.01$, then A and B, are dependent. We then find the all-confidence of X as follows:

$$\begin{aligned}
A_C(\texttt{A}:[0,0],\texttt{B}:[0,0]) &= \frac{30}{\max\{(50),(40)\}} \\
&= \frac{30}{50} = 0.6.
\end{aligned}$$

Since $0.6 > 0.25$, X is correlated. Having previously established in Section 3.1.1 that X is valid, we can now further conclude that X is a correlated contrast set.

Now, for demonstration purposes only, assume $X = \{\texttt{A}:[0,0],\ \texttt{C}:[0,0]\}$. Then $M_I(\texttt{A};\texttt{C}) = M_I(\texttt{A}:[0,0];\texttt{C}:[0,0]) + M_I(\texttt{A}:[0,0];\texttt{C}:[1,1]) + M_I(\texttt{A}:[1,1];\texttt{C}:[0,0]) + M_I(\texttt{A}:[1,1];\texttt{C}:[1,1]) = 0+0+0+0 = 0$. Since $M_I(\texttt{A};\texttt{C}) = 0$, then A and C are independent.

3.4.4 GENCCS

Given a database $\mathcal{D}$, a minimum support difference threshold ϵ, a minimum frequency threshold σ, a minimum subset support ratio threshold κ, and a minimum support ratio threshold ω, a minimum mutual information threshold, ψ, a minimum all-confidence threshold, ξ, and a ranking measure, m, the GENCCS (Generate Correlated Contrast Sets) algorithm, shown in Algorithm 9, traverses the search space to find valid contrast sets that satisfy Criteria 3.1, 3.2, 3.3, 3.4, 3.5, 3.7, and 3.8.

Algorithm 9 GENCCS$(\mathcal{D}, \epsilon, \sigma, \kappa, \omega, \psi, \xi, m, t)$

Input: Dataset $\mathcal{D}$, minimum support difference threshold ϵ, minimum frequency threshold σ, minimum subset support ratio threshold κ, minimum support ratio threshold ω, minimum mutual information threshold ψ, minimum all-confidence threshold ξ, ranking method m, and mining type t
Output: Set of ranked correlated valid contrast sets W

1: $W = \{\}$
2: L_0 =INITIALIZE($\mathcal{D}$)
3: L_1 =MERGE_G($L_0, W, \epsilon, \sigma, 0, \omega, \psi, \xi, t, 1$)
4: TRAVERSE_G($L_1, W, \epsilon, \sigma, \kappa, \omega, \psi, \xi, t, 1$)
5: RANK(W, m)
6: **return** W

3.4.5 General Overview

The initial call to GENCCS is GENCCS $(\mathcal{D}, \epsilon, \sigma, \kappa, \omega, \psi, \xi, m, t)$, where $\mathcal{D}$ is the dataset, ϵ is the minimum support difference threshold, σ is the minimum frequency threshold, κ is the minimum subset support ratio threshold, ω is the minimum support ratio threshold, ψ is the minimum mutual information threshold, ξ is the minimum all-confidence threshold, m is the method used to rank the valid contrast sets, and t is the type of valid contrast sets to be mined.

The algorithm is described as follows. First, GENCCS initializes the set of ranked correlated valid contrast sets (line 1). It then calls the INITIALIZE subroutine to create

all 1-specific contrast sets and their combine sets, storing them in L_0 (line 2). Third, GENCCS calls the MERGE_G subroutine to determine which of the 1-specific contrast sets in L_0 are valid, storing them in L_1 (line 3). Fourth, GENCCS calls the TRAVERSE_G subroutine to find correlated valid k-specific contrast sets using a modified generate-and-test approach, storing them in W (line 4). Fifth, GENCCS calls the RANK subroutine to order the correlated valid contrast sets in W according to a ranking measure m (line 5). Finally, GENCCS returns the set of ranked correlated valid contrast sets in W (line 6).

GENCCS like COSINE has an exponential time complexity. However, the use of mutual information and all-confidence reduces the average branching factor, b, of the nodes in the search space tree of maximum depth, m. Thus GENCCS builds a search tree of space complexity $O(bm)$ and has a time complexity of $O(b^m)$, such that $b < b_c$ and $m < m_c$, where b_c and m_c are the average branching factor and maximum depth, respectively, of the search tree built by COSINE.

3.4.6 A Primer on the GENCCS Search Space

GENCCS, like COSINE, represents the search space as a tree and uses a modified generate-and-test, backtracking, depth-first search approach for traversing the search space. A sample tree is shown in Figure 3.3. This tree can be used to generate all possible correlated contrast sets from the sample dataset in Table 3.1 with $\psi = 0.001$ and $\xi = 0.001$. Each node of the tree consists of two parts: a prefix set and the associated combine set.

The 1-specific contrast sets in the combine set of the root node are equivalent to the 1-specific contrast sets in the combine set of the root node in Figure 3.1.

From the root node at depth zero, the nodes of the tree at depth one are generated as follows. The prefix set at each node is created by combining each of the attribute-interval pairs from the combine set at depth zero, in turn, with the empty prefix set. For

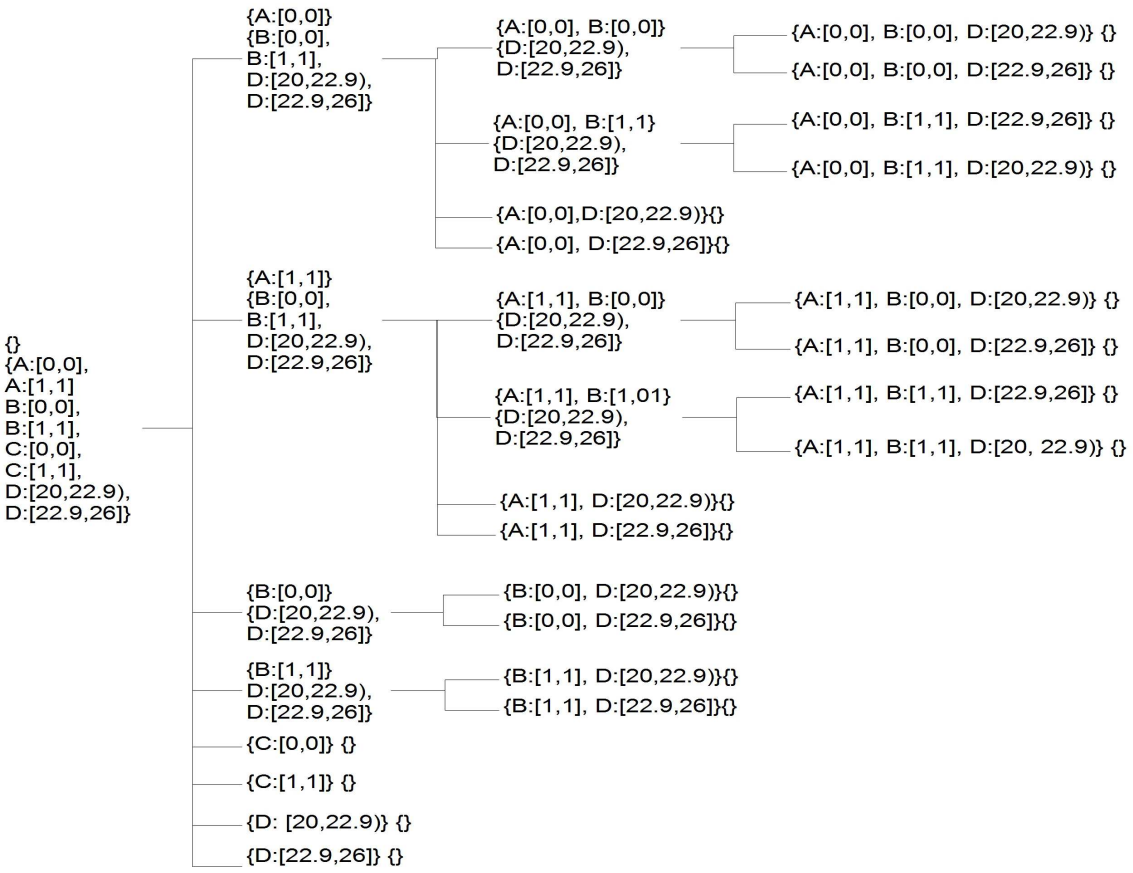


Figure 3.3: Tree representing the search space for correlated contrast sets

example, the resulting prefix sets at depth one are $\{\texttt{A} : [0, 0], \texttt{A} : [1, 1], \texttt{B} : [0, 0], \texttt{B} : [1, 1], \texttt{C} : [0, 0], \texttt{C} : [1, 1], \texttt{D} : [20, 22.9)\}$, and $\{\texttt{D} : [22.9, 26]\}$. The combine set associated with each prefix set at each node at depth one is created by selecting all of the attribute-interval pairs that follow the prefix in the combine set at depth zero, except for those attribute-interval pairs that share the same attribute as the prefix, and those attribute-interval pairs that when combined with the prefix fail to satisfy Criteria 3.7 and 3.8. For example, the prefix set of the top-most node at depth one containing the prefix $\texttt{A} : [0, 0]$ is associated with the combine set $\{\texttt{B} : [0, 0], \texttt{B} : [1, 1], \texttt{D} : [20, 22.9), \texttt{D} : [22.9, 26]\}$. That is, all attribute-interval pairs following $\texttt{A} : [0, 0]$ in the combine set at depth zero, except $\texttt{A} : [1, 1]$, $\texttt{C} : [1, 1]$ and $\texttt{C} : [1, 1]$. The first because $\texttt{A} : [1, 1]$ shares the same attribute as

the prefix, A : [0, 0], and the second and third because A : [0, 0] and C : [0, 0], A : [0, 0] and C : [1, 1] fail to satisfy Criteria 3.7.

The nodes of the tree at depth greater than one are generated in the same manner as with COSINE described in Section 3.2.2. For example, the prefix set at depth two containing {B : [1, 1], D : [20, 22.9)} is associated with the empty combine set, {}, because there are no attribute-interval pairs that follow D : [20, 22.9) in the combine set at depth one.

3.4.7 MERGE_G

From a parent node at the current level in the search space, the MERGE_G algorithm, shown in Algorithm 10, generates the set of child nodes at the next level of the search space whose elements satisfy Criteria 3.1, 3.2, 3.3, 3.4, 3.5, 3.7 and 3.8 .

3.4.7.1 General Overview

The initial call to MERGE_G is MERGE_G (Parent, $W, \epsilon, \sigma, \kappa, \omega, \psi, \xi, t, l$), where Parent is the node in the search space for which child nodes are currently being generated, W is the set of valid contrast sets found so far (initially the empty set), ϵ is the minimum support difference threshold, σ is the minimum frequency threshold, κ is the minimum subset support ratio threshold, ω is the minimum ratio threshold, ψ is the minimum mutual information threshold, ξ is the minimum all-confidence threshold, t is the type of valid contrast sets to be mined, and l is the current level in the search space.

MERGE_G is similar to MERGE, described in Section 3.2.6, except that on the first level of the search space it calls SELECT_G instead of SELECT to determine the combine sets of the prefix sets (lines 10 to 14). SELECT_G uses the parameters ψ and ξ in determining which attribute-interval pairs are correlated.

Algorithm 10 MERGE_G(Parent, $W, \epsilon, \sigma, \kappa, \omega, \psi, \xi, t, l$)

Input: Node from which child nodes are generated, Parent, the set of valid contrast sets, W, minimum support difference threshold ϵ, minimum frequency threshold σ, minimum subset support ratio threshold κ, minimum support ratio threshold ω, minimum mutual information threshold ψ, minimum all-confidence threshold ξ, mining type t, and level l.
Output: Children, nodes at the next level of the search space, and a set of valid contrast sets W.

```
Children = {}
for each c ∈ Parent.C do
    Child.P = Parent.P ∪ {c}
    Child.D = DIFFSET(Child.P)
    if l == 1 then
        Child.count = COUNT(D) − |Child.D|
    else
        Child.count = Parent.count − |Child.D|
    end if
    if l == 1 then
        Child.C = SELECT_G(c, Parent.C, ψ, ξ)
    else
        Child.C = SELECT(c, Parent.C)
    end if
    if IS_VALID(Child.P, ε, σ, κ, l) then
        if t == LAMBDA then
            if λ(Child.P) == ∞ then
                if ∄Y ∈ W : Child.P ⊆ Y then
                    W = W ∪ {Child.P}
                end if
            else if λ(Child.P) ≥ ω then
                Children = Children ∪ {Child.P}
            end if
        else
            Children = Children ∪ {Child.P}
        end if
    end if
end for
UPDATE_COMBINE_SET(Parent.P, Children)
Sort Children by |Child.C| in descending order, then by Child.count in ascending order
return Children
```

3.4.7.2 SELECT_G

Given an attribute-interval pair c, a combine set Parent.C, a minimum mutual information threshold ψ, and a minimum all-confidence threshold ξ, SELECT_G determines the combine set, Child.C, for the prefix, Child.P. For each element, c_1, in Parent.C, which is greater than (appears after) c, SELECT_G adds c_1 to Child.C if c_1 does not share an attribute with c and if the combination of c and c_1 satisfies Criteria 3.7 and 3.8. This ensures that the prefix Child.P is only subsequently combined with attribute-interval pairs with which it is correlated.

3.4.8 TRAVERSE_G

The TRAVERSE_G algorithm, shown in Algorithm 11, traverses the search space to find contrast sets that satisfy Criteria 3.1, 3.2, 3.3, 3.4, 3.5, 3.7, and 3.8.

3.4.8.1 General Overview

The initial call to TRAVERSE_G is TRAVERSE_G $(L_l, W_l, \epsilon, \sigma, \kappa, \omega, \psi, \xi, t, l)$, where L_l is the set of Parent nodes for the current level of the search space, W_l is the set of valid contrast sets for the previous level l of the search space, ϵ is the minimum support difference threshold, σ is the minimum frequency threshold, κ is the minimum subset support ratio threshold, ω is the minimum ratio threshold, ψ is the minimum mutual information threshold, ξ is the minimum all-confidence threshold, t is the type of valid contrast sets to be mined, and l is the current level in the search space.

TRAVERSE_G is similar to TRAVERSE, described in Section 3.2.7, except that it calls MERGE_G instead of MERGE to determine the validity of the correlated contrast sets (lines 8 and 19).

Algorithm 11 TRAVERSE_G$(L_l, W_l, \epsilon, \sigma, \kappa, \omega, \psi, \xi, t, l)$

Input: Set of Parent nodes L_l, set of valid contrast sets, W_l, minimum support difference threshold ϵ, minimum frequency threshold σ, minimum subset support ratio threshold κ, minimum support ratio threshold ω, minimum mutual information threshold ψ, minimum all-confidence threshold ξ, mining type t, and a level in the search space l
Output: W_l, updated with valid correlated contrast sets

1: **for** each Parent $\in L_l$ **do**
2: **if** $t ==$ MAXIMAL $||$ $t ==$ LAMBDA **then**
3: **if** $|W_l| > 0$ **then**
4: **if** $\exists Y \in W_l$: Parent.P $\subseteq Y$ **then**
5: **return**
6: **end if**
7: **end if**
8: L_{l+1} = MERGE_G(Parent, $W_l, \epsilon, \sigma, \kappa, \omega, \psi, \xi, t, l+1$)
9: **if** $|L_{l+1}| > 0$ **then**
10: $W_{l+1} = \{w \in W_l$: Parent.P $\in w\}$
11: TRAVERSE_G($L_{l+1}, W_{l+1}, \epsilon, \sigma, \kappa, \omega, \psi, \xi, t, l+1$)
12: $W_l = W_l \cup W_{l+1}$
13: **else**
14: **if** $\nexists Y \in W_l$: Parent.P $\subseteq Y$ **then**
15: $W_l = W_l \cup$ Parent.P
16: **end if**
17: **end if**
18: **else**
19: L_{l+1} = MERGE_G(Parent, $W_l, \epsilon, \sigma, \kappa, \omega, \psi, \xi, t, l+1$)
20: $W_l = W_l \cup$ Parent.P
21: **if** $|L_{l+1}| > 0$ **then**
22: TRAVERSE_G($L_{l+1}, W_l, \epsilon, \sigma, \kappa, \omega, \psi, \xi, t, l+1$)
23: **end if**
24: **end if**
25: **end for**
26: **return**

3.5 Detailed Walk-through - GENCCS

We now present a detailed walk-through of the GENCCS algorithm. Consider again the dataset shown in Table 3.1, and the search space tree shown in Figure 3.1. GENCCS is initially called with parameters $\mathcal{D}$ = dataset in Table 3.1, $\mathcal{F}$ = attribute E, $\epsilon = 0$, $\sigma = 0.1$, $\kappa = 0.0001$, $\omega = 0$, $\psi = 0.01$, $\xi = 0.25$, m = Coverage, and t = ALL.

Invocation - GENCCS

We create an empty set W (line 1). We use the INITIALIZE subroutine to create 8 1-specific contrast sets, $\{\texttt{A} : [0, 0], \texttt{A} : [1, 1], \texttt{B} : [0, 0], \texttt{B} : [1, 1], \texttt{C} : [0, 0], \texttt{C} : [1, 1], \texttt{D} : [20, 22.9)\}$, and $\{\texttt{D} : [22.9, 26]\}$ and store them in L_0 (line 2). We call the MERGE_G subroutine with L_0, W, 0, 0.1, 0.0001, 0, Coverage, ALL, and 1 (line 3). We combine the empty prefix with each element in L_0. We determine the diffset and count for each element in L_0. We call SELECT_G to determine the combine sets for the 1-specific contrast sets in L_0. We check which pairs of 1-specific contrast sets satisfy Criteria 3.7 and 3.8. We find that only $\{\texttt{A} : [0, 0], \texttt{B} : [0, 0]\}$, $\{\texttt{A} : [0, 0], \texttt{B} : [1, 1]\}$, and $\{\texttt{A} : [1, 1], \texttt{B} : [1, 1]\}$ satisfy both criteria. Thus $L_0 = \{\texttt{A} : [0, 0], \texttt{A} : [1, 1], \texttt{B} : [0, 0], \texttt{B} : [1, 1], \texttt{C} : [0, 0], \texttt{C} : [1, 1], \texttt{D} : [20, 22.9), \texttt{D} : [22.9, 26]\}$, where $\texttt{B} : [0, 0]$, $\texttt{B} : [1, 1]$, $\texttt{C} : [0, 0]$, $\texttt{C} : [1, 1]$, $\texttt{D} : [20, 22.9)$, and $\texttt{D} : [22.9, 26]$ have empty combine sets and $\texttt{A} : [0, 0]$ has a combine set of $\{\texttt{B} : [0, 0], \texttt{B} : [1, 1]\}$ and $\texttt{A} : [1, 1]$ has a combine set of $\{\texttt{B} : [1, 1]\}$. We check which of the 1-specific contrast sets satisfy Criteria 3.1, 3.2, and 3.3. $\texttt{C} : [1, 1]$ fails to satisfy Criteria 3.1, and $\{\texttt{C} : [0, 0]\}$ fails to satisfy Criteria 3.3 thus $L_1 = \{\texttt{A} : [0, 0], \texttt{A} : [1, 1], \texttt{B} : [0, 0], \texttt{B} : [1, 1], \texttt{D} : [20, 22.9), \texttt{D} : [22.9, 26]\}$. We call UPDATE_COMBINE_SET to update the combine sets of the elements in L_1. The frequencies of the contrast sets are $\{50, 50, 40, 60, 50, 50\}$ and cardinalities of the combine sets are $\{2, 1, 0, 0, 0, 0\}$, respectively of each element in L_1. We first reorder the elements in decreasing cardinality of their combine sets, $L_1 = \{\texttt{A} : [0, 0], \texttt{A} : [1, 1], \texttt{B} : [0, 0], \texttt{B} : [1, 1], \texttt{D} : [20, 22.9), \texttt{D} : [22.9, 26]\}$, then in increasing frequency, $L_1 = \{\texttt{A} : [0, 0], \texttt{A} : [1, 1], \texttt{B} : [0, 0], \texttt{B} : [1, 1], \texttt{D} : [20, 22.9), \texttt{D} : [22.9, 26]\}$. We call TRAVERSE_G with L_1, W, 0, 0.1, 0.0001, 0, 0.01, 0.25, Coverage, ALL, and 1 (line 4).

The invocation of TRAVERSE_G proceeds as previously described for COSINE.

Chapter 4

Ranking Methods

A contrast set mining process typically returns many valid contrast sets. Consequently, measures are needed to rank the relative interestingness of the valid contrast sets prior to presenting them to the end-user. This issue has previously been identified as an area that needs development in order for contrast set mining to mature as a data mining task [130]. Ideally, a measure would be used to rank the contrast sets, as well as describe them, akin to the support, confidence, leverage and lift measures used in association rule mining. In this chapter, we consider a variety of measures and demonstrate their use within the context of ranking contrast sets.

This chapter is organized as follows: In Section 4.1, we define the notation used in the remainder of the chapter. In Section 4.2, we describe a new interestingness measure that we have developed for ranking contrast sets. In Section 4.3, we describe two interestingness measures from emerging pattern mining and subgroup discovery that we have adapted for ranking contrast sets. In Section 4.4, we examine ten interestingness measures that have previously been developed for ranking patterns such as association rules, classification rules, and summaries and demonstrate their use for ranking contrast sets. In Section 4.5, we discuss an approach that combines the ranks generated from various measures to produce an *interestingness factor*, a value representative of the overall

interestingness of a contrast set.

4.1 Notation

A contrast set, X, is represented by a set of association rules, $X \rightarrow G_1, X \rightarrow G_2, \ldots, X \rightarrow G_n$, where $G_1, G_2, \ldots, G_n$ correspond to groups. Let $n(X, G_i)$ be the number of instances of the contrast set X in group G_i, $n(X, \neg G_i)$ be the number of instances of the contrast set X in groups other than G_i (that is, the number of times X occurs in groups $G_1, \ldots, G_{i-1}, G_{i+1}, G_n$), $n(\neg X, G_i)$ be the number of instances of contrast sets other than X in group G_i, and $n(\neg X, \neg G_i)$ be the number of instances of contrast sets other than X in groups other than G_i. Let N be the total number of instances.

The values $n(X, G_i), n(X, \neg G_i), n(\neg X, G_i)$, and $n(\neg X, \neg G_i)$ actually correspond to the observed frequencies at the intersection of the rows and columns in a 2×2 contingency table, such as the one shown in Table 4.1 for the association rule $X \rightarrow G_i$, where the rows represent the truth of the contrast set and the columns represent the groups.

Table 4.1: Contingency Table for $X \rightarrow G_i$

	G_i	$\neg G_i$	Σ *Row*
X	$n(X, G_i)$	$N(X, \neg G_i)$	$n(X)$
$\neg X$	$n(\neg X, G_i)$	$n(\neg X, \neg G_i)$	$n(\neg X)$
Σ *Column*	$n(G_i)$	$n(\neg G_i)$	N

In Table 4.2, the columns describe the number of instances in each of the groups A, B, C, and D, respectively. Each row corresponds to the number of instances in each contrast set V, W, X, Y, and Z broken down by the number of instances in each group.

For example, the number of instances of contrast set V in groups, A, B, C, and D corresponding to the rules $V \rightarrow A$, $V \rightarrow B$, $V \rightarrow C$, and $V \rightarrow D$ is 312, 198, 36, and 30, respectively. A contingency table for the rule $V \rightarrow A$ is shown in Table 4.3.

Table 4.2: Results Summary

Contrast Sets	Groups			
	A	B	C	D
V	312	198	36	30
W	62	10	23	13
X	104	66	12	10
Y	12	34	3	15
Z	36	108	18	30
Total	1210	384	69	65

Table 4.3: Contingency Table for $V \rightarrow A$

	A	$\neg A$	Σ *Row*
V	312	264	576
$\neg V$	898	254	1152
Σ *Column*	1210	518	1728

The values in Table 4.2 are used throughout the remainder of this chapter to illustrate calculations for the various interestingness measures.

4.2 Distribution Difference

The *Distribution Difference* of a contrast set measures how the group support in a contrast set differs from the group support in the entire dataset [114]. Formally, the Distribution Difference of a contrast set, X, in a group G_i is given as

$$\text{Distribution Difference}(X \rightarrow G_i) \quad = \quad \left| \frac{n(X, G_i)}{n(X)} \times \frac{N}{n(G_i)} - 1 \right|.$$

Background: Consider a dataset where 40% of the instances concern males and 60% concern females. Now consider a contrast set where 65% of the instances concern males and 35% concern females. Clearly, the difference between males and females is greater in the contrast set than it is in the dataset (i.e., if we consider the general population). If we now consider a contrast set where 42% of the instances concern males and 58% concern

females, the contrast set with the 65%/35% male/female distribution is more interesting because it deviates more from the distribution in the entire dataset. The Distribution Difference allows us to capture this deviation.

The values for Distribution Difference values occur in the interval $[0, \infty]$. A value of 0 indicates that there is no deviation in the distributions of the groups in the contrast set from the distribution in the entire dataset. Values further from 0 indicate strong deviation in the distributions.

The Distribution Difference of a contrast set X is the sum of the Distribution Difference values over all the groups. Formally,

$$\text{Distribution Difference}(X) = \sum_i \text{Distribution Difference}(X \rightarrow G_i).$$

For example, the Distribution Difference of contrast set V in group A is

$$\text{Distribution Difference}(V \rightarrow A) = \left| \frac{312}{576} \times \frac{1728}{1210} - 1 \right| = 0.226.$$

The Distribution Difference of the contrast set V in the four groups A, B, C, and D is 0.226, 0.547, 0.565, and 0.384, respectively. Therefore, the Distribution Difference of V is 1.723. The Distribution Difference of the contrast sets W, X, Y, and Z in groups A, B, C, and D is calculated similarly to that of V. Thus, the Distribution Difference of the five contrast sets V, W, X, Y, and Z is 1.723, 7.30, 1.723, 7.53, and 6.77, respectively. Therefore, using the Distribution Difference as the measure of interestingness, the order of the contrast sets, ranked from most interesting to least interesting is Y, W, Z, and V/Y (i.e., V and Y are tied).

4.3 Group Differences

Emerging pattern mining and subgroup discovery, are two data mining tasks which are similar to contrast set mining. Sections 4.3.1 and 4.3.2 demonstrate how the heuristics developed in the areas of emerging patterns and subgroup discovery can be used for ranking contrast sets.

4.3.1 Growth Rate

An emerging pattern is an itemset whose support in one dataset increases significantly in another related dataset [27]. The ratio of the two support values is called the growth rate of the emerging pattern. Formally, let $support_{D_i}(X)$ denote the support of an itemset X in a dataset D_i. The growth rate of an itemset, X, from D_1 to D_2, is given by

$$\text{Growth Rate}(X) = \begin{cases} 0, & \text{if } support_{D_1}(X) = 0 \text{ and } support_{D_2}(X) = 0, \\ \infty, & \text{if } support_{D_1}(X) = 0 \text{ and } support_{D_2}(X) \neq 0, \\ \frac{support_{D_2}(X)}{support_{D_1}(X)}, & \text{otherwise.} \end{cases}$$

Semantically, an emerging pattern is a set of association rules with an itemset in the antecedent and some class attribute in the consequent. For example, an itemset discovered in datasets D_1 and D_2 can be represented by the following pair of association rules:

$$\text{ItemSet} \rightarrow D_1 \ (x\%)$$

$$\text{ItemSet} \rightarrow D_2 \ (y\%)$$

Since a contrast set is also an itemset, and since all the instances belonging to some group G_i can be considered a dataset, the growth rate is also applicable to measuring

changes in support from one group to another. That is, given the association rules

$$\text{Contrast Set} \rightarrow G_1 \ (x\%)$$

$$\text{Contrast Set} \rightarrow G_2 \ (y\%)$$

and, if the order of the groups is irrelevant, then the *Growth Rate* of a contrast set, X, from G_1 to G_2 or G_2 to G_1 is given by

$$\text{Growth Rate}(X) = \begin{cases} 0, & \text{if } support_{G_1}(X) = 0 \\ & \text{and } support_{G_2}(X) = 0 \\ \infty, & \text{if } support_{G_1}(X) \neq 0 \\ & \text{or } support_{G_2}(X) \neq 0 \\ \max\left\{\dfrac{support_{G_2}(X)}{support_{G_1}(X)}, \dfrac{support_{G_1}(X)}{support_{G_2}(X)}\right\} & \\ = \dfrac{\max\{support_{G_1}(X), support_{G_2}(X)\}}{\min\{support_{G_1}(X), support_{G_2}(X)\}} & \text{otherwise,} \end{cases} \tag{4.1}$$

where the $support_{G_i}(X) = \frac{n(X,G_i)}{n(G_i)}$.

Equation 4.1 applies only to a two-group contrast set mining task. We can generalize Equation 4.1 so that it also applies to any number of groups. The Growth Rate of a contrast set of a contrast set X is given by

$$\text{Growth Rate}(X) = \begin{cases} \infty, & \text{if } \min\limits_i \{support_{G_i}(X)\} = 0, \\ \dfrac{\max\limits_i \{support_{G_i}(X)\}}{\min\limits_i \{support_{G_i}(X)\}}, & \text{otherwise.} \end{cases}$$

The values for the Growth Rate occur in the range $[0, \infty]$. A value of 0 indicates that the contrast set is not true in any of the groups. A value of ∞ indicates that there are no instances belonging to at least one of the groups where the contrast set is true. Values near $infty$ indicate large differences between the groups with respect to the maximum

and minimum support in the contrast set.

For example, the Growth Rate of contrast set V in group A is

$$\begin{aligned} \text{Growth Rate} &= \frac{support_C(V)}{support_A(V)} \\ &= \frac{\frac{36}{69}}{\frac{312}{1210}} = 2.02. \end{aligned}$$

The Growth Rate of the contrast sets W, X, Y, and Z is calculated similarly to that of V. The Growth Rate of the five contrast sets V, W, X, Y, and Z is 2.02, 12.80, 2.02, 23.27, and 15.51, respectively. Therefore, using the Growth Rate as the measure of interestingness, the order of the contrast sets, ranked from most interesting to least interesting is Y, Z, W, and V/X.

4.3.2 Unusualness

Unusualness if a measure of interestingness used in subgroup discovery [54]. Given a set of instances possessing some property of interest, a subgroup is a subset of instances in which the statistical characteristics of the property of interest are "unusual". Instances in the subgroup can be described by an association rule $X \rightarrow Y$, where the property of interest is represented by the consequent Y, and the antecedent X, is an itemset. Weighted relative accuracy (WRAcc) is used to evaluate the quality (i.e., unusualness) of the induced association rules.

Contrast set mining has been shown to be an extension of subgroup discovery, where each group represents a different property of interest [65]. Thus, we can use the weighted relative accuracy to measure the *Unusualness* (UN) of X in G_i. Formally, Unusualness is given by

$$\begin{aligned} UN(X, G_i) &= p(X) \times (p(G_i|X) - p(G_i)) \\ &\approx \frac{n(X)}{N} \times \left(\frac{n(X, G_i)}{n(X)} - \frac{n(G_i)}{N} \right). \end{aligned}$$

The values for Unusualness occur in the interval [0,1].

The Unusualness of a contrast set, X, is determined by the group, G_i, for which the Unusualness is the largest. Thus the Unusualness of X is given by

$$\text{Unusualness}(X) = \max_i \{\text{Unusualness}(X, G_i)\}.$$

For example, the Unusualness of contrast set V in group A is

$$\text{Unusualness}(V \to A) = \frac{576}{1728} \times \left[\frac{312}{576} - \frac{1210}{1728}\right] = -0.053.$$

The Unusualness of the contrast set V in the four groups A, B, C, and D is -0.053, 0.041, 0.008, and 0.005, respectively. Therefore, the maximum Unusualness is 0.041. The Unusualness of the contrast sets W, X, Y, and Z in groups A, B, C, and D is calculated similarly to that of V. The Unusualness of the five contrast sets V, W, X, Y, and Z is 0.041, 0.0108, 0.014, 0.011, and 0.038, respectively. Therefore, using the Unusualness as a measure of interestingness, the order of the contrast sets, ranked from most interesting to least interesting is V, Z, X and W/Y.

4.4 Probability-Based Interestingness Measures

Probability-based objective measures have been studied extensively as a means for evaluating the generality and reliability of association rules [39] [69] [120]. We examine 10 of these measures to show how they can be utilized in ranking contrast sets.

4.4.1 Coverage

The Coverage of an association rule, $X \rightarrow G_i$ is the proportion of instances in the dataset where X is true in G_i [4], and is given by

$$\text{Coverage}(X \rightarrow G_i) = p(X) = \frac{n(X, G_i)}{N}.$$

The values for Coverage range occur in the interval [0,1]. A value of 1 indicates that X occurs in all instances of the dataset. A value of 0 indicates that X does not occur in any of the instances of the dataset.

The Coverage of a contrast set for all groups is the sum of the individual Coverage values for the contrast set in each group. The Coverage of a contrast set X is given by

$$\text{Coverage}(X) = \sum_i \text{Coverage}(X \rightarrow G_i).$$

For example, the Coverage of contrast set V in group A is

$$\text{Coverage}(V \rightarrow A) = \frac{312}{1728} = 0.181.$$

For example, the Coverage of V in A, B, C, and D is 0.181, 0.115, 0.021, and 0.017, respectively. Therefore, the coverage for V is 0.334. The Coverage of W, X, Y, and Z in A, B, C, and D is calculated similarly to that of V. The Coverage of V, W, X, Y, and Z is 0.334, 0.063, 0.111, 0.040, and 0.111, respectively. Therefore, using Coverage as the measure of interestingness, the order of the contrast sets ranked from most interesting to least interesting is $V, W/Z, X$ and Y.

4.4.2 Lift

The Lift of an association rule $X \rightarrow G_i$ measures how many times X and G_i actually occur together compared to the number of times X and G_i would be expected to occur together if they where statistically independent [18] and is given by

$$\text{Lift}(X \rightarrow G_i) \quad = \quad \frac{p(X, G_i)}{p(X)p(G_i)} = \frac{N \times n(X, G_i)}{n(X) \times n(G_i)}.$$

The values for the Lift occur in the interval $[0, \infty]$. Values close to 1 indicate that X and G_i are independent. Values further from 1 indicate that X and G_i are strongly related.

The Lift for a contrast set across all groups is the sum of the Lift values for the contrast set in each group. Formally,

$$\text{Lift}(X) = \sum_i \text{Lift}(X \rightarrow G_i).$$

For example, the Lift of contrast set V in group A is

$$\text{Lift}(V \rightarrow A) = \frac{1728 \times 312}{576 \times 1210} = 0.774.$$

The Lift of the contrast set V in the four groups A, B, C, and D is 0.774, 1.547, 1.565, and 1.385, respectively. Therefore, the Lift of V is 5.27. The Lift of the contrast sets W, X, Y, and Z in groups A, B, C, and D is calculated similarly to that of V. The Lift of the five contrast sets V, W, X, Y, and Z is 5.27, 9.77, 5.27, 10.06, 9.30, respectively. Therefore, using Lift as a measure of interestingness, the order of the contrast sets, ranked from most interesting to least interesting is Y, W, X and V/X.

4.4.3 Leverage

The Leverage of an association rule $X \rightarrow G_i$ is the difference between the actual proportion of instances in which both X and G_i occur and the proportion that would be expected if X and G_i were statistically independent of each other [95] and is given by

$$\begin{aligned} \text{Leverage}(X \rightarrow G_i) &= p(G_i|X) - p(X)p(G_i) \\ &= \frac{n(X, G_i)}{n(X)} - \frac{n(X)}{N} \times \frac{n(G_i)}{N}. \end{aligned}$$

The values for leverage occur on the interval [-1, 1]. Values near one indicate strong independence between X and G_i.

The Leverage for a contrast set, X, is determined by the group G_i, for which the leverage is largest. Thus, the leverage of X is given by

$$\text{Leverage}(X) = \max_i \{\text{Leverage}(X \rightarrow G_i)\}.$$

For example, the Leverage of contrast set V in group A is

$$\text{Leverage}(V \rightarrow A) = \frac{312}{576} - \frac{576}{1728} \times \frac{1210}{1728} = 0.308.$$

The Leverage of the contrast set V in the four groups A, B, C, and D is 0.308, 0.110, -0.171, and -0.181, respectively. Therefore, the Leverage of V is 0.308. The Leverage of the contrast sets W, X, Y, and Z in groups A, B, C, and D is calculated similarly to that of V. The Leverage of the five contrast sets V, W, X, Y, and Z is 0.308, 0.530, 0.463, 0.523, and 0.538, respectively. Therefore, using the Leverage as a measure of interestingness, the order of the contrast sets, ranked from most interesting to least interesting is Z, W, Y, X and V.

4.4.4 Change of Support

The Change of Support of of an association rule $X \rightarrow G_i$ measures the correlation between X and G_i [120] and is given by

$$\begin{aligned} \text{Change of Support}(X \rightarrow G_i) &= p(G_i|X) - p(G_i) \\ &= \frac{n(X, G_i)}{n(X)} - \frac{n(G_i)}{N}. \end{aligned}$$

The values for Change of Support occur in the interval [-0.5, 1]. Values near 1 indicate strong correlation between X and G_i.

The Change of Support for a contrast set, X, is determined by the group, G_i, for which Change of Support is largest, and is given by

$$\text{Change of Support}(X) = \max_i \{\text{Change of Support}(X \rightarrow G_i)\}.$$

For example, the Change of Support of contrast set V in group A is

$$\text{Change of Support}(V \rightarrow A) = \frac{312}{576} - \frac{1210}{1728} = -0.159.$$

The Change of Support of the contrast set V in the four groups A, B, C, and D is -0.159, 0.122, 0.022, and 0.014, respectively. Therefore, the Change of Support for V is 0.122. The Change of Support of the contrast sets W, X, Y, and Z in groups A, B, C, and D is calculated similarly to that of V. The Change of Support of the five contrast sets V, W, X, Y, and Z is 0.122, 0.173, 0.122, 0.309 and 0.340, respectively. Therefore, using the Change of Support as the measure of interestingness, the order of the contrast sets, ranked from most interesting to least interesting is Z, Y, W and V/X.

4.4.5 Jaccard Coefficient

The *Jaccard Coefficient* of an association rule $X \rightarrow G_i$ measures the similarity between X and G_i [123] and is given by

$$\begin{aligned} \text{Jaccard}(X \rightarrow G_i) &= \frac{p(X, G_i)}{p(X) + p(G_i) - p(X, G_i)} \\ &= \frac{n(X, G_i)}{n(X) + n(G_i) - n(X, G_i)}. \end{aligned}$$

The values for the Jaccard Coefficient occur in the interval [-1, 1]. Values near 1 indicate strong similarity between X and G_i. Values near -1 indicate strong dissimilarity between X and G_i. 0 indicates that there is no relationship between X and G_i.

The Jaccard Coefficient for a contrast set, X, is determined by the group, G_i, for which Jaccard Coefficient is largest, and is given by

$$\text{Jaccard}(X) = \max_i \{\text{Jaccard}(X \rightarrow G_i)\}.$$

For example, the Jaccard Coefficient of contrast set V in group A is

$$\text{Jaccard}(V \rightarrow A) = \frac{312}{576 + 1210 - 312} = 0.212.$$

The Jaccard Coefficient of the contrast set V in the four groups A, B, C, and D is 0.212, 0.260, and 0.059, and 0.049, respectively. Therefore, the maximum Jaccard Coefficient is 0.260. The Jaccard Coefficient of the contrast sets W, X, Y, and Z in groups A, B, C, and D is calculated similarly to that of V. The Jaccard Coefficient of the five contrast sets V, W, X, Y, and Z is 0.260, 0.149, 0.129, 0.132 and 0.231, respectively. Therefore, using the Jaccard Coefficient as the measure of interestingness, the order of the contrast sets, ranked from most interesting to least interesting is V, Z, W, Y and X.

4.4.6 Yule's Q Coefficient

The *Yule's Q Coefficient* of an association rule $X \rightarrow G_i$ measures the degree to which X and G_i are associated with each other [140] and is given by

$$\begin{aligned} \text{Yule's Q}(X \rightarrow G_i) &= \frac{p(X,G_i)p(\neg X \neg G_i) - p(X,\neg G_i)p(\neg X, G_i)}{p(X,G_i)p(\neg X \neg G_i) + p(X,\neg G_i)p(\neg X, G_i)} \\ &= \frac{n(X,G_i)n(\neg X \neg G_i) - n(X,\neg G_i)n(\neg X, G_i)}{n(X,G_i)n(\neg X \neg G_i) + n(X,\neg G_i)p(\neg X, G_i)}. \end{aligned}$$

The values for Yule's Coefficient occur on the interval [-1,1]. Values near 1 indicate strong association between X and G_i. Values near -1 indicate complete dissociation between X and G_i. 0 indicates complete independence between X and G_i.

The Yule's Q coefficient for a contrast set, X, is determined by the group, G_i, for which Yule's Q coefficient is largest. Thus, the Yule's Q coefficient of X is given by

$$\text{Yule's Q}(X) = \max_i \left\{\text{Yule's Q}(X \rightarrow G_i)\right\}.$$

For example, the Yule's Q of contrast set V in group A is

$$\text{Yule's Q } (V \rightarrow A) = \frac{312 \times 254 - 264 \times 898}{312 \times 254 + 264 \times 898} = -0.498.$$

The Yule's Q Coefficient of the contrast set V in the four groups A, B, C, and D is -0.498, 0.462, 0.387, and 0.274, respectively. Therefore, the maximum Yule's Q Coefficient is 0.462. The Yule's Q Coefficient of the contrast sets W, X, Y, and Z in groups A, B, C, and D is calculated similarly to that of V. The Yule's Q Coefficient of the five contrast sets V, W, X, Y, and Z is 0.462, 0.805, 0.334, 0.816, and 0.776, respectively. Therefore, using the Yule's Q Coefficient as a measure of interestingness, the order of the contrast sets, ranked from most interesting to least interesting is Y, W, Z, V and X.

4.4.7 ϕ-coefficient

The *ϕ-coefficient* of an association rule $X \rightarrow G_i$ measures the degree to which X and G_i are associated with each other [5] and is given by

$$\begin{aligned}
\phi(X \rightarrow G_i) &= \frac{p(X, G_i) - p(X)p(G_i)}{\sqrt{p(X)p(G_i)p(\neg X)p(\neg G_i)}} \\
&= \frac{\frac{N \times n(X,G_i) - n(X) \times n(G_i)}{N^2}}{\frac{\sqrt{n(X) \times n(G_i) \times n(\neg X) \times n(\neg G_i)}}{N^2}} \\
&= \frac{N \times n(X, G_i) - n(X) \times n(G_i)}{\sqrt{n(X) \times n(G_i) \times n(\neg X) \times n(\neg G_i)}}.
\end{aligned}$$

The values for the ϕ-coefficient occur on the range [-1, 1]. Values near 1 indicate strong association between X and G_i. Values near -1 indicate strong disassociation between X and G_i. 0 indicates no relationship between X and G_i.

The ϕ-coefficient for a contrast set, X, is determined by the group, G_i, for which ϕ-coefficient is largest. Thus, the ϕ-coefficient of X is given by

$$\phi(X) = \max_i \{\phi(X \rightarrow G_i)\}.$$

For example, the *phi*-coefficient of contrast set V in group A is

$$\phi(V \rightarrow A) = \frac{1728 \times 312 - 576 \times 1210}{\sqrt{576 \times 1210 \times 1152 \times 518}} = -0.244$$

The $\phi-$coefficient of the contrast set V in the four groups A, B, C, and D is -0.244, 0.207, 0.082, and 0.054, respectively. Therefore, the maximum ϕ-coefficient is 0.207. The ϕ-coefficient of the contrast sets W, X, Y, and Z in groups A, B, C, and D is calculated similarly to that of V. The maximum ϕ-coefficient of the five contrast sets V, W, X, Y, and Z is 0.207, 0.228, 0.103, 0.203, and 0.290, respectively. Therefore, using ϕ-coefficient as the measure of interestingness, the order of the contrast sets, ranked from most interesting

to least interesting is Z, W, V, Y and X.

4.4.8 κ-coefficient

The κ-*coefficient* of an association rule $X \rightarrow G_i$ measures the degree to which X and G_i agree [23] and is given by

$$\begin{aligned}\kappa(X \rightarrow G_i) &= \frac{p(X, G_i) + p(\neg X, \neg G_i) - p(X)p(G_i) - p(\neg X)p(\neg G_i)}{1 - p(X)p(G_i) - p(\neg X)p(\neg G_i)} \\ &= \frac{N \times n(X, G_i) + N \times n(\neg X, \neg G_i) - n(X) \times n(G_i) - n(\neg X) \times n(\neg G_i)}{N^2 - n(X) \times n(G_i) - n(\neg X) \times n(\neg G_i)}.\end{aligned}$$

The values for the κ-coefficient occur on the interval [-1, 1]. Values close to 1 indicate strong agreement between X and G_i. Values close to -1 indicate strong disagreement between X and G_i. 0 indicates no agreement between X and G_i.

The κ-coefficient for a contrast set, X, is determined by the group, G_i, for which κ-coefficient is largest. Thus, the κ-coefficient of X is given by

$$\kappa(X) = \max_i \{\kappa(X \rightarrow G_i)\}.$$

For example, the κ-coefficient of contrast set V in group A is

$$\kappa(V \rightarrow A) = \frac{1728 \times 312 + 1728 \times 254 - 576 \times 1210 - 1152 \times 518}{1728^2 - 576 \times 1210 - 1152 \times 518} = -0.187$$

The κ-coefficient of the contrast set V in the four groups A, B, C, and D is -0.187, 0.199, 0.043, and 0.028, respectively. Therefore, the maximum κ-coefficient is 0.199. The κ-coefficient of the contrast sets W, X, Y, and Z in groups A, B, C, and D is calculated similarly to that of V. The maximum κ-coefficient of the five contrast sets V, W, X, Y, and Z is 0.199, 0.222, 0.095, 0.203, and 0.266, respectively. Therefore, using the κ-coefficient as a measure of interestingness, the order of the contrast sets, ranked from most interesting to least interesting is Z, W, Y, V and X.

4.4.9 Conviction

The *Conviction* of an association rule $X \rightarrow G_i$ measures true implication between X and G_i [18] and is given by

$$\text{Conviction}(X \rightarrow G_i) \quad = \quad \frac{p(X)p(\neg G_i)}{p(X, \neg G_i)} = \frac{n(X) \times n(\neg G_i)}{N \times n(X, \neg G_i)}$$

The values for Conviction occur in the interval $[1, \infty]$. A value equal to 1 indicates X and G_i are completely unrelated. Values further from 1 indicate that X and G_i are strongly related.

The Conviction of a contrast set for all groups is the sum of the individual Conviction values for the contrast set in each group. Thus the Conviction of a contrast set X is given by

$$\text{Conviction}(X) = \sum_i \text{Conviction}(X \rightarrow G_i).$$

For example, the Conviction of contrast set V in group A is

$$\text{Conviction}(V \rightarrow A) = \frac{576 \times 518}{1728 \times 264} = 0.654.$$

The Conviction of the contrast set V in the four groups A, B, C, and D is 0.654, 1.185, 1.024, and 1.015, respectively. Therefore, the Conviction of V is 3.88. The Conviction of the contrast sets W, X, Y, and Z in groups A, B, C, and D is calculated similarly to that of V. The Conviction of the five contrast sets V, W, X, Y, and Z is 3.88, 3.87, 3.88, 4.29, and 4.35, respectively. Therefore, using Conviction as the measure of interestingness, the order of the contrast sets ranked from most interesting to least interesting is $Z, Y, X/V$, and W.

4.4.10 Collective Strength

The *Collective Strength* of an association rule $X \rightarrow G_i$ measures the correlation between X and G_i [2] and is given by

$$\begin{aligned}\text{Collective Strength}(X \rightarrow G_i) &= \frac{p(X,G_i)+p(\neg G_i|\neg X)}{p(X)p(G_i)+p(\neg X)\times p(\neg G_i)} \times \\ &\quad \frac{1-p(X)p(G_i)-p(\neg X)\times p(\neg G_i)}{1-p(X,G_i)-p(\neg G_i|\neg X)} \\ &= \frac{\frac{n(X,G_i)}{N}+\frac{n(\neg X,\neg G_i)}{n(\neg X)}}{\frac{n(X)}{N}\frac{n(G_i)}{N}+\frac{n(\neg X)}{N}\times\frac{n(\neg G_i)}{N}} \times \\ &\quad \frac{1-\frac{n(X)}{N}\times\frac{n(G_i)}{N}-\frac{n(\neg X)}{N}\times\frac{n(\neg G_i)}{N}}{1-\frac{n(X,G_i)}{N}-\frac{n(\neg X,\neg G_i)}{n(\neg X)}}.\end{aligned}$$

The values for Collective Strength values occur in the interval [0, ∞]. A value of 0 indicates perfect negative correlation while a value of ∞ indicates perfect positive correlation between X and G_i. A value of 1 indicates independence between X and G_i.

The Collective Strength of a contrast set for all groups is the sum of the individual Collective Strength values for the contrast set in each group. The Collective Strength of a contrast set X is given by

$$\text{Collective Strength}(X) = \sum \text{Collective Strength}(X \rightarrow G_i).$$

For example, the Collective Strength of contrast set V in group A is

$$\begin{aligned}\text{Collective Strength}(X \rightarrow G_i) &= \frac{\frac{312}{1728}+\frac{254}{1152}}{\frac{576}{1728}\times\frac{1210}{1728}+\frac{1152}{1728}\times\frac{518}{1728}} \times \\ &\quad \frac{1-\frac{576}{1728}\times\frac{1210}{1728}-\frac{1152}{1728}\times\frac{518}{1728}}{1-\frac{312}{1728}-\frac{254}{516}} \\ &= 0.876.\end{aligned}$$

The Collective Strength of the contrast set V in the four groups A, B, C, and D is 0.876, 13.979, 67.380, and 40.079, respectively. Therefore, the Collective Strength of V is 3.88. The Collective Strength of the contrast sets W, X, Y, and Z in groups A, B, C, and D is calculated similarly to that of V. The Collective Strength of the five contrast sets V, W, X, Y, and Z is 122.32, 13.44, 13.54, 7.81, and 40.69, respectively. Therefore, using the Collective Strength as the measure of interestingness, the order of the contrast sets, ranked from most interesting to least interesting is V, Z, X, W and Y.

4.5 Interestingness Factor

The fourteen interestingness measures described above can be used individually to rank discovered contrast sets. However, they can can also be used in combination to determine the most interesting contrast sets based on multiple measures. The *Interestingness Factor* of a contrast set is the average of it's rank over a set of the selected interestingness measures and is given by

$$\text{Interestingness Factor}(X) = \frac{\sum_{i=1}^{n} r_i \times w_i}{\sum_{i=1}^{n} w_i \times m_i},$$

where n is the number of interestingness measures, r_i is the rank of the contrast set by a measure i, w_i is the weight of the ranking measure i (i.e., a user-defined parameter indicating the relative importance of measure i), and m_i is the maximum rank of measure i.

The values for the Interestingness Factor occur in the interval [0,1]. Values near to 0 indicate contrast sets which are more interesting, while values near to 1 indicate contrast sets which are less interesting. An Interestingness Factor of 1 indicates that the contrast set was ranked the lowest by each method used in calculating the Interestingness Factor.

For example, the calculation of the interestingness factor is illustrated below. Here we use the ranks of all five contrast sets for Coverage, Unusualness, and Lift as shown in Table 4.4.

Table 4.4: Ranking Values

Methods	Coverage	Unusualness	Lift
Weights	0.4	0.3	0.3
Maximum Rank	5	5	4
V	1	1	4
W	2	5	2
X	4	3	4
Y	5	4	1
Z	2	2	3

The calculation for contrast set V is representative for all other contrast sets.

$$\text{Interestingness Factor}(V) = \frac{1 \times 0.4 + 1 \times 0.3 + 4 \times 0.3}{5 \times 0.4 + 5 \times 0.3 + 4 \times 0.3} = 0.40.$$

The Interestingness Factor of the five contrast sets V, W, X, Y, and Z is 0.40, 0.62, 0.76, 0.74, and 0.49, respectively. Therefore, using the Interestingness Factor as the measure of interestingness, the order of the contrast sets ranked from most interesting to least interesting is V, Z, W, Y and X.

Chapter 5

Experimental Results

To evaluate the performance of the COSINE algorithm for generating maximal valid contrast sets and valid λ-contrast sets, and the GENCCS algorithm for generating correlated maximal valid contrast sets and valid λ-contrast sets, COSINE and GENCCS were implemented in Java and run under Windows 7 (64 bit) on an Intel dual core 2.40GHz processor with 4GB of memory. In this chapter, we present the results of our evaluation and experience.

This chapter is organized as follows. In Section 5.1, we describe the datasets used to evaluate COSINE and GENCCS. In Section 5.2, we evaluate COSINE on a variety of discovery tasks and compare the results obtained against those obtained using STUCCO and CIGAR. In Section 5.3, we evaluate the minimum subset support ratio threshold and the minimum frequency threshold as constraints on the COSINE algorithm. In Section 5.4, we evaluate GENCCS on a variety of discovery tasks and compare the results obtained against those obtained using COSINE. In Section 5.5, we evaluate the minimum support ratio threshold as a constraint on the COSINE and GENCCS algorithms. In Section 5.6, we evaluate the Discretize algorithm for the task of creating intervals for quantitative attributes.

5.1 Dataset Description

Discovery tasks were performed on four datasets, described in Table 5.1, obtained from the UCI Machine Learning Repository [6]. In Table 5.1, the *Dataset* column describes the name of each dataset, the *No. of Transactions*, *No. of Attributes*, and *No. of Groups* columns describe the number of tuples, the number of attributes, and the number of groups to which a tuple could be assigned for each dataset. The *Cardinality of the Longest Combine Set* Column describes the maximum number of attribute-value pairs that can be combined together to make contrast sets. These datasets were chosen because of the variation in the size of the longest combine set, the ability to mine correlated contrast sets with high specificity, and their use in evaluating other contrast set mining techniques (e.g., the Mushroom dataset and a subset of the Census dataset were previously utilized in the evaluation of STUCCO and CIGAR [11] and [47].)

Table 5.1: Dataset Description

Dataset	*No. of Transactions*	*No. of Attributes*	*No. of Groups*	*Cardinality of Longest Combine Set*
Census	48842	15	5	145
Mushroom	8124	23	2	198
Waveform	5000	41	3	390
Spambase	4601	58	2	560

The Census dataset describes demographic data extracted from the 1994 U.S. Census Bureau database. Each individual has one of five educational levels. The dataset consists of 48842 tuples describing a total of 15 attributes of which 9 are categorical and 6 are quantitative. The Mushroom dataset describes species of gilled mushrooms. Each species is identified as edible or poisonous. The dataset consists of 8124 tuples describing a total of 23 attributes of which all 23 are categorical. The Spambase dataset describes properties of emails. Emails are categorized as spam or non-spam. The dataset consists of 4601 tuples describing a total of 58 attributes of which 57 are quantitative and 1 is categorical.

The Waveform dataset describes properties of waves. There are three categories of waves. The dataset consists of 5000 tuples describing a total of 41 attributes of which 40 are quantitative and 1 is categorical.

5.2 COSINE vs. STUCCO and CIGAR

We ran a series of discovery tasks to compare the results obtained by COSINE with those obtained by STUCCO and CIGAR when mining for maximal valid contrast sets. For each dataset, we use a minimum frequency threshold (MFT) of 0 and a minimum subset support ratio threshold (MSSRT) of 0 for COSINE. CIGAR requires that four parameters, the minimum support difference threshold (δ), the minimum support threshold (β), minimum correlation threshold (λ), and minimum correlation difference threshold (υ) to be set, while STUCCO requires two values, the minimum support difference threshold (δ) and the specialization difference threshold (δ_s). The minimum support difference threshold (MSDT) is used by all three algorithms so we can compare the results obtained by COSINE with those obtained by STUCCO and CIGAR when mining for maximal valid contrast sets as the minimum support difference threshold varies. The MSDT serves as a constraint on the search space because no subtree of any contrast set that fails to meet Criteria 3.2 needs to be explored further. The values used by CIGAR for β, λ, and υ and by STUCCO for δ_s are shown in Table 5.2.

Table 5.2: Parameters used for CIGAR and STUCCO

	CIGAR			*STUCCO*
Dataset	β	λ	υ	δ_s
Census	0.01	0	0	0.01
Mushroom	0.02	0	0	0.01
Spambase	0.01	0	0	0.01
Waveform	0.01	0	0	0.01

For example, the results for the Mushroom dataset were generated using $\beta = 0.02$, $\lambda = 0$, and $\upsilon = 0$ for CIGAR and $\delta_s = 0.01$ for STUCCO. For each algorithm, the

results are averaged over 10 consecutive runs and a significance level of 0.95 is used in the statistical tests.

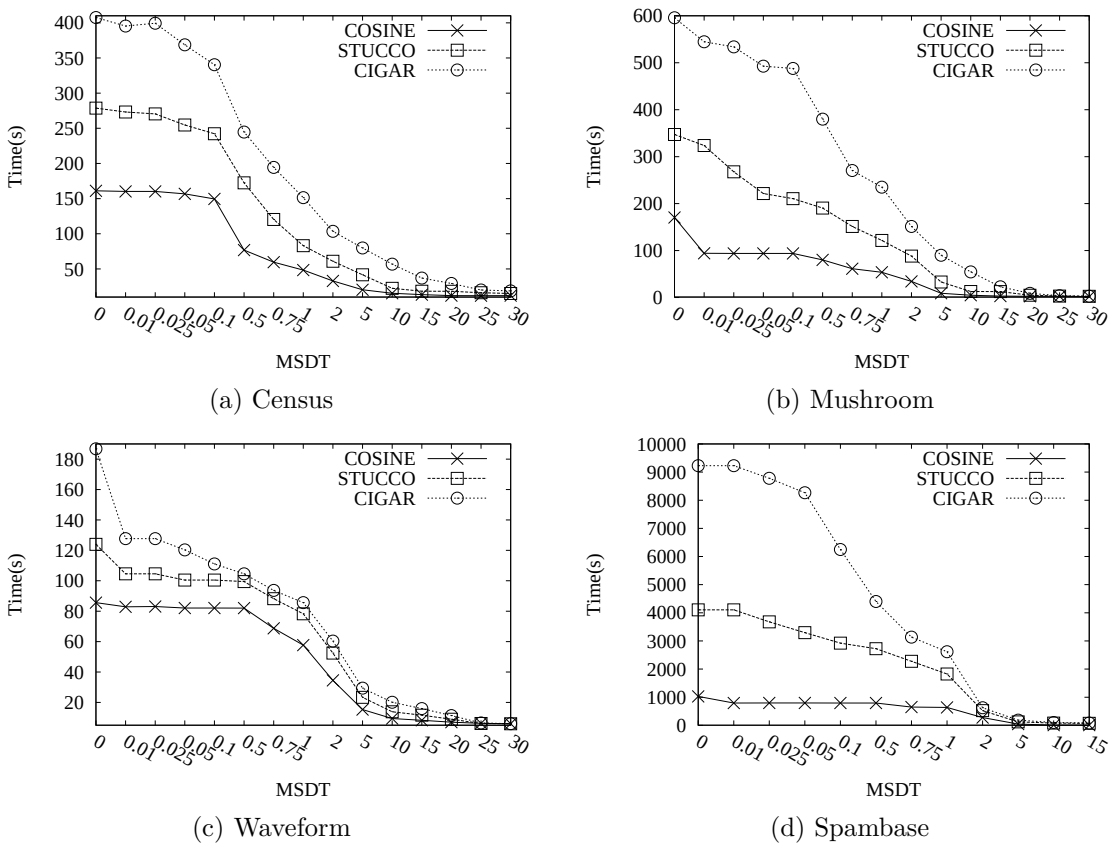
In Section 5.2.1, we compare the run time of the three algorithms as the MSDT is varied. In Section 5.2.2, we compare the interestingness of the contrast sets discovered. In Section 5.2.3, we compare the maximum and average specificities of the contrast sets discovered.

From our experiments, we draw two main conclusions. First, at low MSDT values, the run time of COSINE when mining for maximal valid contrast sets is significantly lower than that of STUCCO and CIGAR. Second, the average interestingness, average specificity, and maximum specificity of the maximal valid contrast sets discovered by COSINE are also significantly higher than that of those discovered by STUCCO, and CIGAR.

5.2.1 Runtime

We compare the runtime performance of COSINE to STUCCO and CIGAR when mining maximal valid contrast sets as the MSDT is varied. Figures 5.1(a) to 5.1(c) show the average runtime for the three algorithms on the Census, Mushroom, and Waveform datasets, respectively, as the MSDT is varied from 0 to 30. Figure 5.1(d) shows the average run time for the Spambase dataset as the MSDT is varied from 0 to 15, as values greater than 15 did not produce any contrast sets for all three methods.

For the Census, Mushroom, and Waveform datasets, COSINE has a significantly lower runtime than STUCCO and CIGAR for values of the MSDT less than 5. For the Spambase dataset, COSINE also has a significantly lower runtime than STUCCO and CIGAR for values of the MSDT less than 2. When the MSDT is 0, on the Spambase dataset, the run times of COSINE, STUCCO, and CIGAR are 1032, 4105, and 9228 seconds, respectively. That is COSINE takes 1/9th the time of CIGAR and 1/4 the time of STUCCO, when the MSDT is 0. The difference in run times is significant because the



(a) Census (b) Mushroom

(c) Waveform (d) Spambase

Figure 5.1: Run Time vs. MSDT

MSDT serves as a mechanism to limit the search space, thus as its value approaches 0, the improvements in run time for COSINE over STUCCO and CIGAR, are due to the fact that COSINE scans the dataset once to generate 1-specific contrast sets and their support then uses diffsets to determine the support of contrast sets with higher specificities as the search space is descended while STUCCO and CIGAR require a separate scan of the dataset to determine the support of contrast sets at each level of the search space. For the Census, Mushroom, and Waveform datasets, when the MSDT is greater than 5, the differences in runtime between COSINE, STUCCO, and CIGAR are negligible. For the Spambase dataset, the differences are negligible when the MSDT is greater than 2. As the MSDT increases, fewer contrast sets are large enough to meet the support difference threshold, so both the run times and the differences in run times between the algorithms decrease. For all four datasets, CIGAR consistently has the longest runtime, followed by

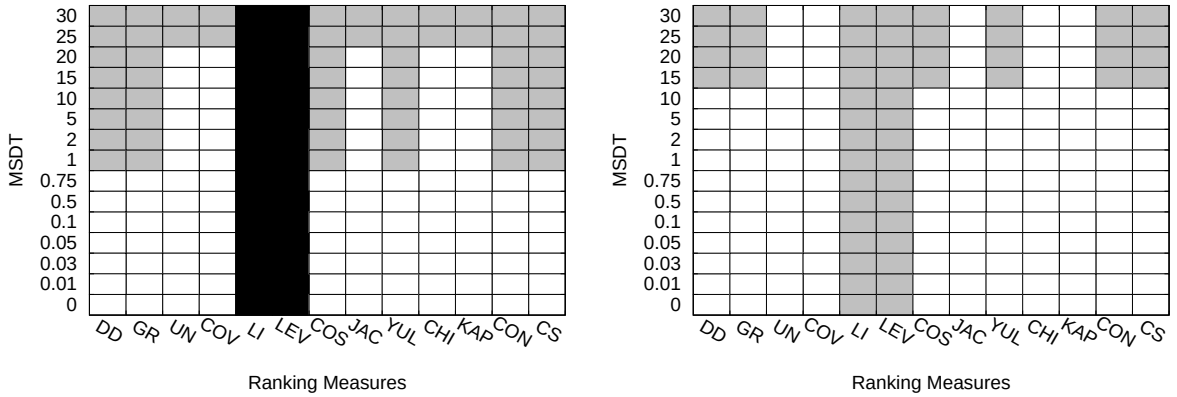
STUCCO, and then COSINE.

5.2.2 Interestingness

We compare the interestingness of the maximal valid contrast sets discovered by COSINE to STUCCO and CIGAR on the Census, Mushroom, Waveform, and Spambase datasets using thirteen ranking measures: distribution difference (DD), growth rate (GR), unusualness (UN), coverage (COV), lift (LI), leverage (LEV), change of support (COS), Jaccard coefficient (JAC), Yule's Q coefficient (YUL), ϕ-coefficient (CHI), κ-coefficient (KAP), conviction (CON), and collective strength (CS). Once again we vary the MSDT from 0 to 30 for the Census, Mushroom, and Waveform datasets, and from 0 to 15 for the Spambase dataset. We use a standard T-test at a confidence level of 95% to determine whether there is a statistically significant difference between the average interestingness values obtained by COSINE and STUCCO and those obtained by COSINE and CIGAR, respectively.

These results are shown in Figures 5.2 to 5.5, where statistical significance of the interestingness values obtained by COSINE versus STUCCO and CIGAR are plotted against the MSDT for each of the four datasets. In the graphs shown, the statistical significance is described by a three-step scale: white, black, and grey indicate that the interestingness values obtained by COSINE are significantly higher, significantly lower, and not significantly different, respectively, compared to those obtained by STUCCO and CIGAR. For example, in Figure 5.5(a), when the MSDT is 0, the average interestingness of the results obtained by COSINE are significantly higher than those obtained by STUCCO for DD, GR, UN, COV, JAC, CHI, KAP, CON, and CS, significantly lower for LI, LEV, and COS, and not significantly different than STUCCO for YUL.

Here we discuss the average interestingness results obtained for the Census dataset, as shown in Figure 5.2. In Figure 5.2(a), the average interestingness of the results obtained by COSINE are significantly higher than those obtained by STUCCO for DD, GR, COS,



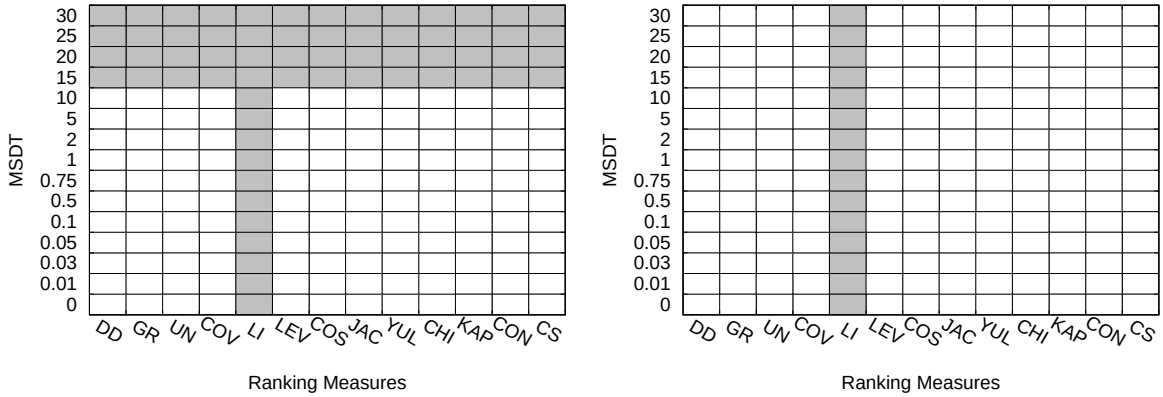
(a) COSINE vs. STUCCO (b) COSINE vs. CIGAR

Figure 5.2: Interestingness Comparison - Census

YUL, CON, and CS for MSDT values from 0 to 0.75 and for UN, COV, JAC, CHI, and KAP for MSDT values from 0 to 20. The results obtained by COSINE are significantly lower for LI and LEV for all MSDT values.

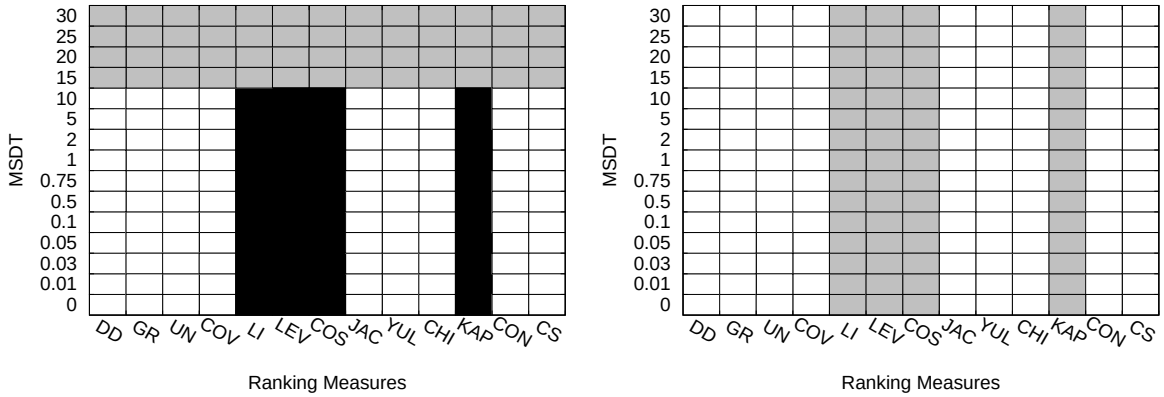
In Figure 5.2(b), the average interestingness of the results obtained by COSINE are generally significantly higher than those obtained by CIGAR for all measures except LI and LEV, where there is no significant difference.

The average interestingness of the results obtained by COSINE versus STUCCO and CIGAR for the Mushroom, Waveform, and Spambase datasets can be interpreted in a similar manner, and is left as an exercise to the reader.



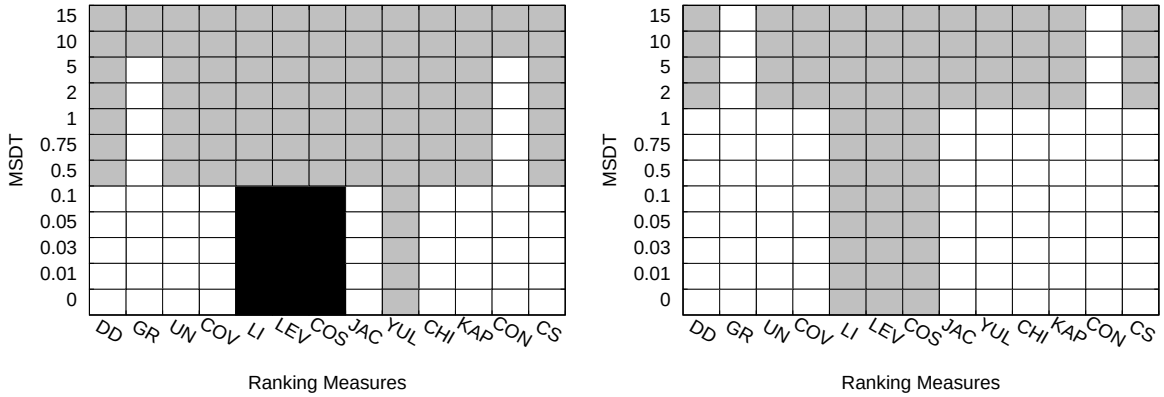
(a) COSINE vs. STUCCO (b) COSINE vs. CIGAR

Figure 5.3: Interestingness Comparison - Mushroom



(a) STUCCO (b) CIGAR

Figure 5.4: Interestingness Comparison - Waveform



(a) STUCCO (b) CIGAR

Figure 5.5: Interestingness Comparison - Spambase

Both STUCCO and CIGAR find a larger number of maximal contrast sets than COSINE (see Appendix A for the raw data). The former because it uses a higher α value than COSINE for significance tests particularly at higher levels of the search space, thus it finds a large number of less specific contrast sets. The latter because it uses a lower α value than COSINE at all levels of the search space, thus it finds a larger number of contrast sets at all specificities. Our results demonstrate that on average he contrast sets found by COSINE are more interesting than the additional contrast sets found by STUCCO and CIGAR.

5.2.3 Specificity

We compare the specificity of the maximal valid contrast sets discovered by COSINE to STUCCO and CIGAR within the context of average and maximum specificities as the MSDT is varied as described in the previous two sections. Once again we vary the MSDT from 0 to 30 for the Census, Mushroom, and Waveform datasets, and from 0 to 15 for the Spambase dataset. For the average specificity, we use a standard T-test at a confidence level of 95% to determine whether there is a significant difference between the average specificity values obtained by COSINE versus STUCCO and CIGAR, respectively.

The results are shown in Figure 5.6, where statistical significance of the average specificity values obtained by COSINE versus STUCCO and CIGAR are plotted against the MSDT for each of the four datasets. For the Spambase dataset, the MSDT is only varied from 0 to 15, so the area of the graph for MSDT values from 20 to 30 does not apply for the Spambase dataset. The raw data from which this figure was generated can be found in Appendix A. In the graph, the statistical significance is described by a three-step scale: white, black, and grey indicate that the average specificity values obtained by COSINE are significantly higher, significantly lower, and not significantly different, respectively, compared to those obtained by STUCCO and CIGAR. For example, when the MSDT is 0, the average specificity of the results obtained by COSINE are significantly higher than those obtained by STUCCO and CIGAR for all datasets.

For all four datasets, the average specificity of the maximal valid contrast sets discovered by COSINE is generally significantly higher than that of STUCCO and CIGAR for MSDT values from 0 to 5. For MSDT values from 10 to 30, the average specificity obtained by COSINE is generally not significantly different from that obtained by STUCCO and CIGAR, except for a few high MSDT values obtained by CIGAR on the Waveform dataset where COSINE is significantly lower.

Figures 5.7(a) to 5.7(d) show the maximum specificity of the maximal valid contrast

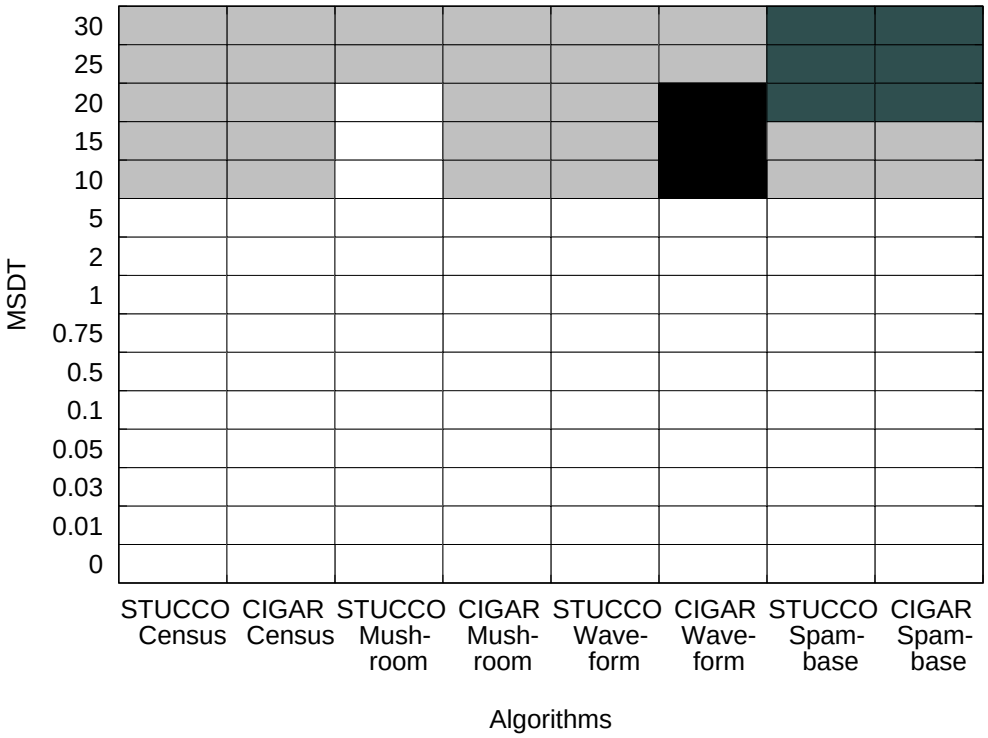
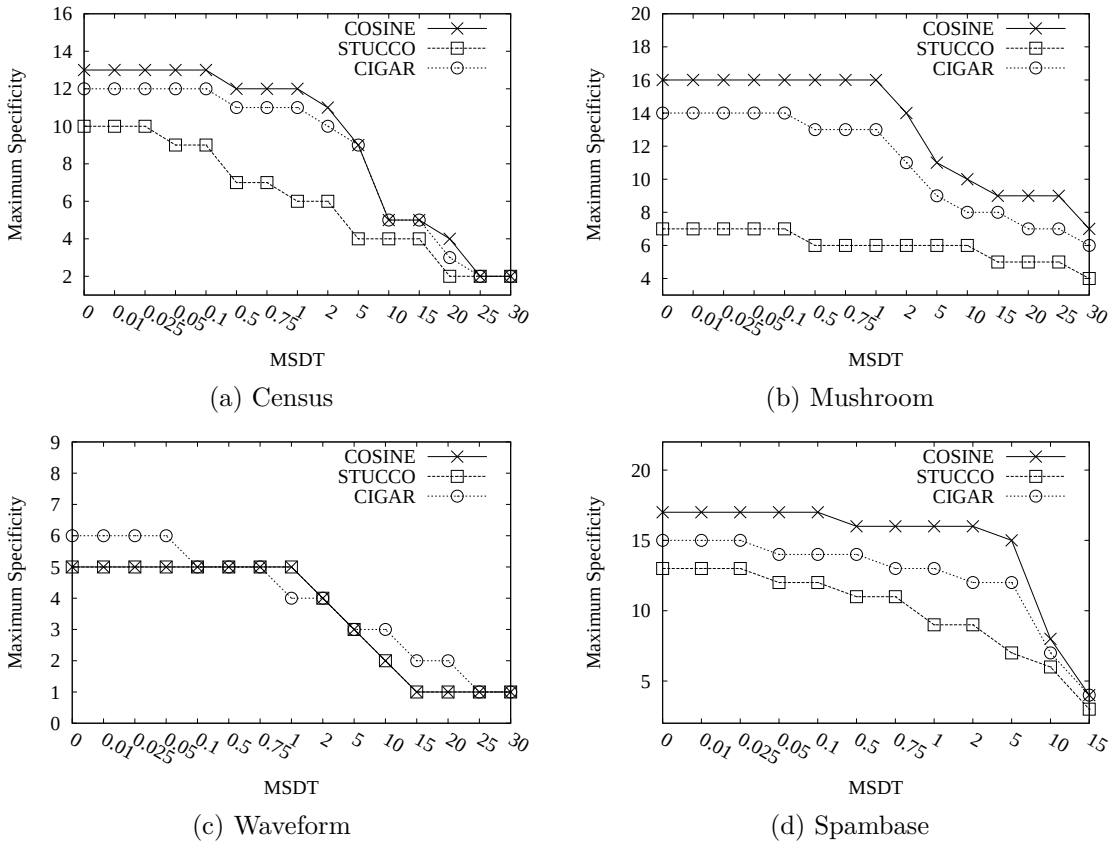


Figure 5.6: Comparison of Average Specificity

sets discovered by COSINE, STUCCO, and CIGAR on the Census, Mushroom, Waveform, and Spambase datasets, respectively, as the MSDT is varied.

For the Census, and Spambase datasets, the maximum specificity of the maximal contrast sets discovered by COSINE is higher than that of STUCCO and CIGAR for MSDT values from 0 to 2. For the Mushroom dataset, the maximum specificity of the maximal contrast sets discovered by COSINE is significantly higher than that of STUCCO for all MSDT values. For the Waveform dataset, there is no difference between the maximum specificities of the maximal contrast sets discovered by COSINE and STUCCO, while COSINE is only lower than that of CIGAR for MSDT values from 0 to 0.05 and 10 to 20. Above MSDT values of 2 on the Census and Spambase datasets, the maximum specificity of the maximal contrast sets discovered by COSINE is at least as high as than that of STUCCO and CIGAR.



(a) Census

(b) Mushroom

(c) Waveform

(d) Spambase

Figure 5.7: Maximum Specificity vs. MSDT

5.3 Evaluation of the COSINE Algorithm

COSINE utilizes two constraints, the minimum subset support ratio threshold (MSSRT) and the minimum frequency threshold (MFT), in the search for valid contrast sets. In this section we examine the effect of both these constraints on the run time, interestingness, and specificity of the maximal valid contrast sets discovered by COSINE. In Section 5.3.1, we examine the MSSRT and in Section 5.3.2, we examine the MFT.

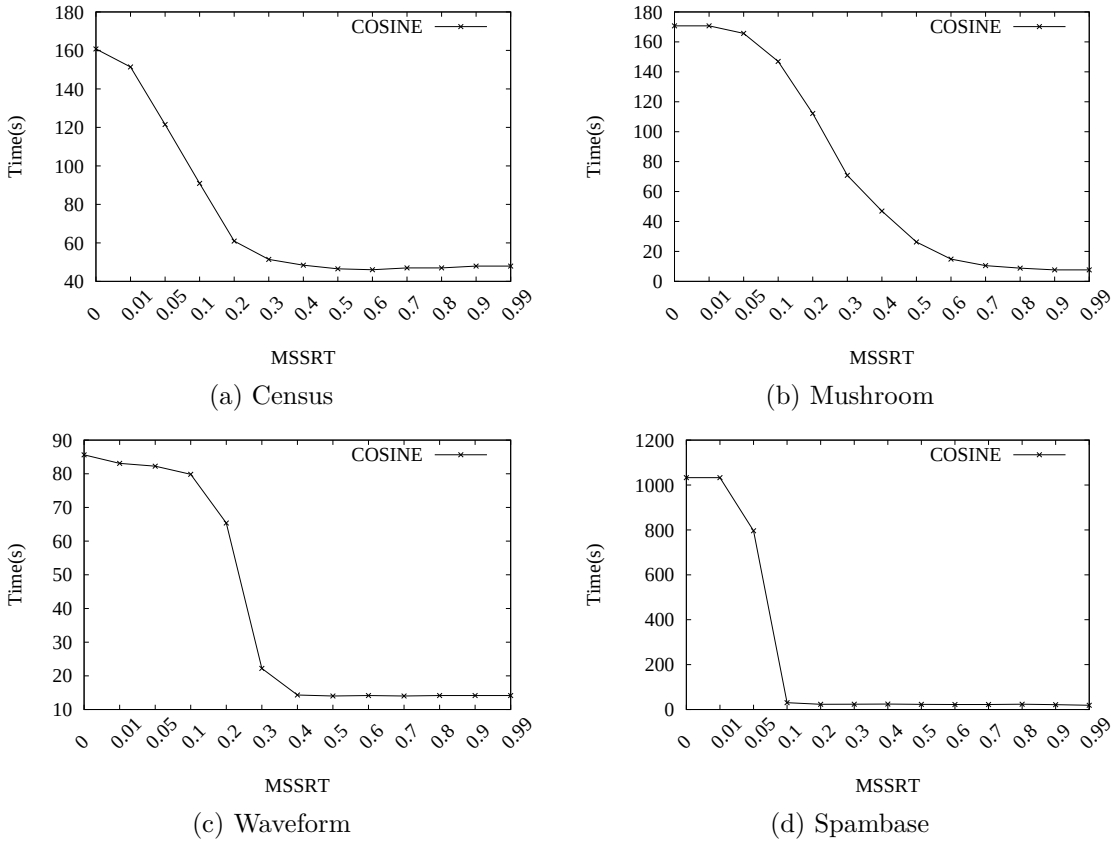
5.3.1 Minimum Subset Support Ratio Threshold (MSSRT)

We ran a series of discovery tasks on the Census, Mushroom, Waveform, and Spambase datasets to evaluate the effect of the MSSRT as a constraint on the COSINE algorithm for mining maximal contrast sets. For each dataset, we use an MFT of 0 and an MSDT of

0 for COSINE, we average the results over 10 consecutive runs, and we use a significance level of 0.95. In section 5.3.1.1, we examine the run time of COSINE as the MSSRT is varied. In Section 5.3.1.2, we examine the interestingness of the contrast sets discovered by COSINE as the MSSRT is varied. In Section 5.3.1.3, we examine the maximum and average specificity of the contrast sets discovered by COSINE as the MSSRT is varied.

5.3.1.1 Runtime

We evaluate the runtime performance of COSINE when mining maximal valid contrast sets as the MSSRT is varied. Figures 5.8(a) to 5.8(d) show the average runtime for COSINE on the Census, Mushroom, Waveform, and Spambase datasets, respectively, as the MSSRT is varied from 0 to 0.99.



(a) Census (b) Mushroom

(c) Waveform (d) Spambase

Figure 5.8: Run time vs. MSSRT

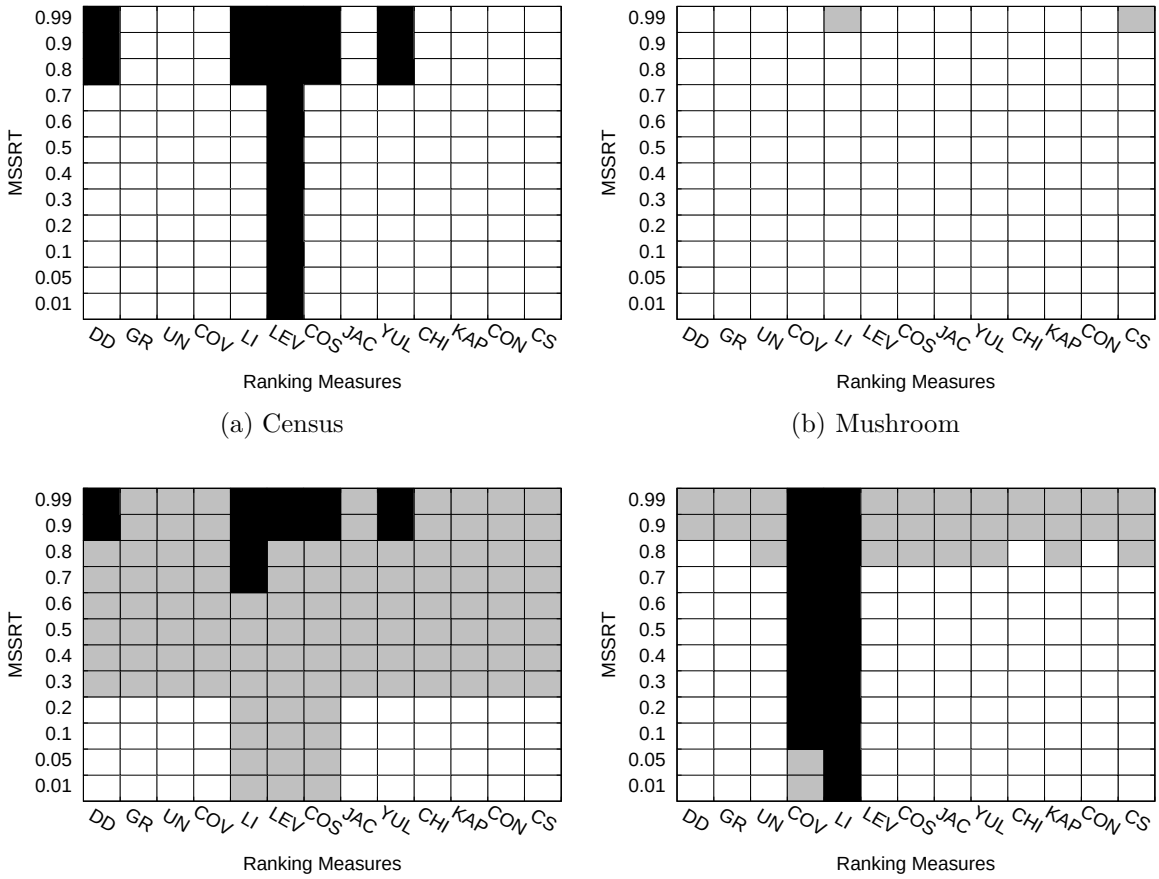
For all four datasets, the run time of COSINE decreases as the MSSRT is increased,

however, the run times begin to plateau when the MSSRT is greater than 0.2 for the Census dataset, 0.5 on the Mushroom dataset, 0.3 for the Waveform dataset and 0.05 for the Spambase dataset. Additionally, the run time of COSINE sharply decreases as the MSSRT increases from 0.05 to 0.1 on the Spambase dataset, and from 0.1 to 0.3 on the Waveform dataset. A decrease in runtime as the MSSRT increases occurs because the MSSRT serves as a constraint on the depth of the search space and as its value increases, less contrast sets will meet this threshold and therefore COSINE will take less time to complete the search.

5.3.1.2 Interestingness

We evaluate the interestingness of the maximal valid contrast sets discovered by COSINE on the Census, Mushroom, Waveform, and Spambase datasets using thirteen ranking measures: distribution difference (DD), growth rate (GR), unusualness (UN), coverage (COV), lift (LI), leverage (LEV), change of support (COS), Jaccard coefficient (JAC), Yule's Q coefficient (YUL), ϕ-coefficient (CHI), κ-coefficient (KAP), conviction (CON), and collective strength (CS). Once again, we vary the MSSRT from 0 to 0.99 for the Census, Mushroom, Waveform, and Spambase datasets. We use a standard T-test at a confidence level of 95% to determine whether there is a statistically significant difference between the average interestingness values obtained by COSINE when the MSSRT is 0, and those obtained by COSINE when the MSSRT is greater than 0.

The results are shown in Figures 5.9(a) to 5.9(d), where statistical significance of the interestingness values obtained by COSINE when the MSSRT is 0 versus the interestingness values obtained by COSINE when the MSSRT is greater than 0 are plotted against the MSSRT for each of the four datasets. In the graphs shown, the statistical significance is described by a three-step scale: white, black, and grey indicate that the interestingness values obtained by COSINE when the MSSRT is greater than 0 are significantly higher, significantly lower, and not significantly different, respectively, compared



(a) Census

(b) Mushroom

(c) Waveform

(d) Spambase

Figure 5.9: Significance Values vs. MSSRT

to those obtained by COSINE when the MSSRT is 0. For example, in Figure 5.9(a), the average interestingness of the results obtained by COSINE when the MSSRT is 0.01 is significantly higher than those obtained by COSINE when the MSSRT is 0 for DD, GR, UN, COV, LI, COS, JAC, YUL, CHI, KAP, CON, and CS, and significantly lower for LEV.

In Figure 5.9(a), the average interestingness of the results obtained by COSINE for MSSRT values from 0.01 to 0.7 is significantly higher than the results obtained by COSINE when the MSSRT is 0 for all measures except for LEV, where it is significantly lower. The average interestingness of the results obtained by COSINE for MSSRT values from 0.8 to 0.99 is significantly lower than the results obtained by COSINE when the

MSSRT is 0 for DD, LI, COS, and YUL, and significantly higher for all other measures.

The average interestingness of the results obtained by COSINE when the MSSRT is greater than 0 versus the interestingness values obtained by COSINE when the MSSRT is 0 for the Mushroom, Waveform, and Spambase datasets can be interpreted in a similar manner.

For three of the four datasets, the average interestingness of the results obtained by COSINE for MSSRT values from 0.01 to 0.7 is generally significantly higher than the interestingness values obtained by COSINE when the MSSRT is 0. These results demonstrate that contrast sets, where the group support deviates least from the group support of their immediate superset, have lower interestingness values, thus removing them produces a subset of contrast sets that is on average more interesting.

5.3.1.3 Specificity

We evaluate the specificity of the maximal valid contrast sets discovered by COSINE within the context of average and maximum specificities as the MSSRT is varied as described in the previous two sections. Tables 5.3 to 5.6 show the number of maximal valid contrast sets discovered by COSINE as the MSSRT is varied from 0 to 0.99.

Table 5.3: Summary of Contrast Sets discovered - Census

	MSSRT												
k	*0.00*	*0.01*	*0.05*	*0.10*	*0.20*	*0.30*	*0.40*	*0.50*	*0.60*	*0.70*	*0.80*	*0.90*	*0.99*
1	0	8	14	14	14	14	14	14	15	15	15	15	28
2	1	0	2	3	5	8	9	9	11	12	12	13	38
3	12	7	15	12	5	3	3	3	1	1	3	6	14
4	75	64	40	27	9	4	4	3	3	6	9	24	2
5	277	254	159	88	25	9	9	10	9	13	11	19	3
6	690	640	426	204	60	28	16	12	13	13	10	10	0
7	1052	1024	717	396	91	43	28	28	23	10	13	3	1
8	1290	1280	931	486	136	68	23	14	10	10	9	0	0
9	1084	1082	911	467	95	41	16	8	16	8	3	0	0
10	766	761	646	372	79	33	14	13	4	2	0	0	0
11	450	450	399	202	47	22	4	0	0	0	0	0	0
12	238	238	222	113	35	3	0	0	0	0	0	0	0
13	74	74	71	57	15	0	0	0	0	0	0	0	0

In Table 5.3, the k column describes the specificity of the contrast sets, and the

MSSRT column describes the MSSRT values. Each entry at a given k represents the number of maximal valid k-specific contrast sets discovered at the corresponding MSSRT value. For instance, when the MSSRT is 0, COSINE did not discover any 1-specific contrast sets but it discovered 1084 9-specific contrast sets. A dotted line occurs at the specificity k which represents the maximum specificity of the maximal valid contrast sets for each MSSRT value. For example, when the MSSRT is 0, the maximum specificity is 13, however when the MSSRT is 0.99, the maximum specificity is 7. A dashed line occurs at the specificity k which represents the floor of the average specificity of the maximal valid contrast sets for each MSSRT value. For example, when the MSSRT is 0, the average specificity is 8, however when the MSSRT is 0.99, the average specificity is 2.

In Table 5.3, the maximum specificity of the maximal valid contrast sets discovered for MSSRT values from 0 to 0.2 is 13. The average specificity of the maximal valid contrast sets discovered for MSSRT values from 0 to 0.2 is 8. This means that increasing the MSSRT from 0 through to 0.2 did not affect the maximum depth of the search space to which COSINE could find maximal valid contrast sets. For MSSRT values above 0.2, the average and maximal specificities decline steadily, though when the MSSRT is 0.99, COSINE still discovers 7-specific contrast sets.

Similarly, in Tables 5.4 to 5.6, the average and maximal specificities of the maximal valid contrast sets are unchanged for MSSRT values form 0 to 0.5 for the Mushroom dataset, 0 to 0.3 for the Waveform dataset, and 0 to 0.05 for the Spambase dataset. These results demonstrate that using lower values for the MSSRT as a constraint on the search process does not reduce the maximal or average specificity of the maximal valid contrast sets discovered.

Table 5.4: Summary of Contrast Sets discovered - Mushroom

	MSSRT												
k	*0.00*	*0.01*	*0.05*	*0.10*	*0.20*	*0.30*	*0.40*	*0.50*	*0.60*	*0.70*	*0.80*	*0.90*	*0.99*
1	0	0	0	0	0	1	1	1	5	5	6	6	6
2	0	0	0	0	1	0	0	0	3	3	2	1	2
3	0	0	0	1	2	2	2	2	1	3	3	4	6
4	0	0	0	1	1	0	0	0	1	0	2	7	5
5	1	1	2	3	5	7	8	8	9	15	15	10	10
6	13	13	14	10	8	18	17	17	17	16	16	14	6
7	68	68	69	50	30	32	36	36	36	34	40	19	10
8	131	131	129	99	54	47	37	37	34	28	21	14	3
9	228	228	228	197	118	79	59	59	28	25	12	6	3
10	176	176	176	157	99	73	67	67	46	15	13	6	3
11	291	290	289	242	158	92	60	60	14	16	8	8	4
12	243	241	240	158	92	47	32	32	10	6	2	0	0
13	157	157	159	130	76	60	25	25	8	5	2	2	2
14	80	80	78	64	35	20	9	9	0	0	0	0	0
15	44	43	42	27	10	4	3	3	0	0	0	0	0
16	43	43	43	29	19	9	1	1	0	0	0	0	0

Table 5.5: Summary of Contrast Sets discovered - Waveform

	MSSRT												
k	*0.00*	*0.01*	*0.05*	*0.10*	*0.20*	*0.30*	*0.40*	*0.50*	*0.60*	*0.70*	*0.80*	*0.90*	*0.99*
1	2	2	5	5	6	6	7	13	46	82	100	104	104
2	379	379	309	237	214	176	125	131	67	32	14	12	12
3	11890	11885	11881	11664	10277	8139	1918	159	62	3	2	2	2
4	20009	20009	20009	20001	19819	18876	3565	10	3	0	0	0	0
5	562	562	562	562	562	559	208	0	0	0	0	0	0

Table 5.6: Summary of Contrast Sets Discovered - Spambase

	MSSRT												
k	*0.00*	*0.01*	*0.05*	*0.10*	*0.20*	*0.30*	*0.40*	*0.50*	*0.60*	*0.70*	*0.80*	*0.90*	*0.99*
1	1	1	0	0	0	1	1	1	1	2	2	2	21
2	91	91	95	249	404	512	521	567	620	736	738	758	574
3	483	483	243	300	447	432	426	401	343	131	140	111	0
4	1045	1045	879	790	571	200	129	68	60	34	11	0	0
5	2971	2971	2636	1160	182	8	1	0	0	0	0	0	0
6	4714	4714	5436	826	18	0	0	0	0	0	0	0	0
7	5506	5506	6491	443	1	0	0	0	0	0	0	0	0
8	5670	5670	6343	200	0	0	0	0	0	0	0	0	0
9	8278	8278	7374	32	0	0	0	0	0	0	0	0	0
10	8641	8641	6750	1	0	0	0	0	0	0	0	0	0
11	6332	6332	4347	0	0	0	0	0	0	0	0	0	0
12	2670	2443	2008	0	0	0	0	0	0	0	0	0	0
13	1181	1001	919	0	0	0	0	0	0	0	0	0	0
14	433	420	397	0	0	0	0	0	0	0	0	0	0
15	140	140	140	0	0	0	0	0	0	0	0	0	0
16	42	42	42	0	0	0	0	0	0	0	0	0	0
17	11	11	11	0	0	0	0	0	0	0	0	0	0

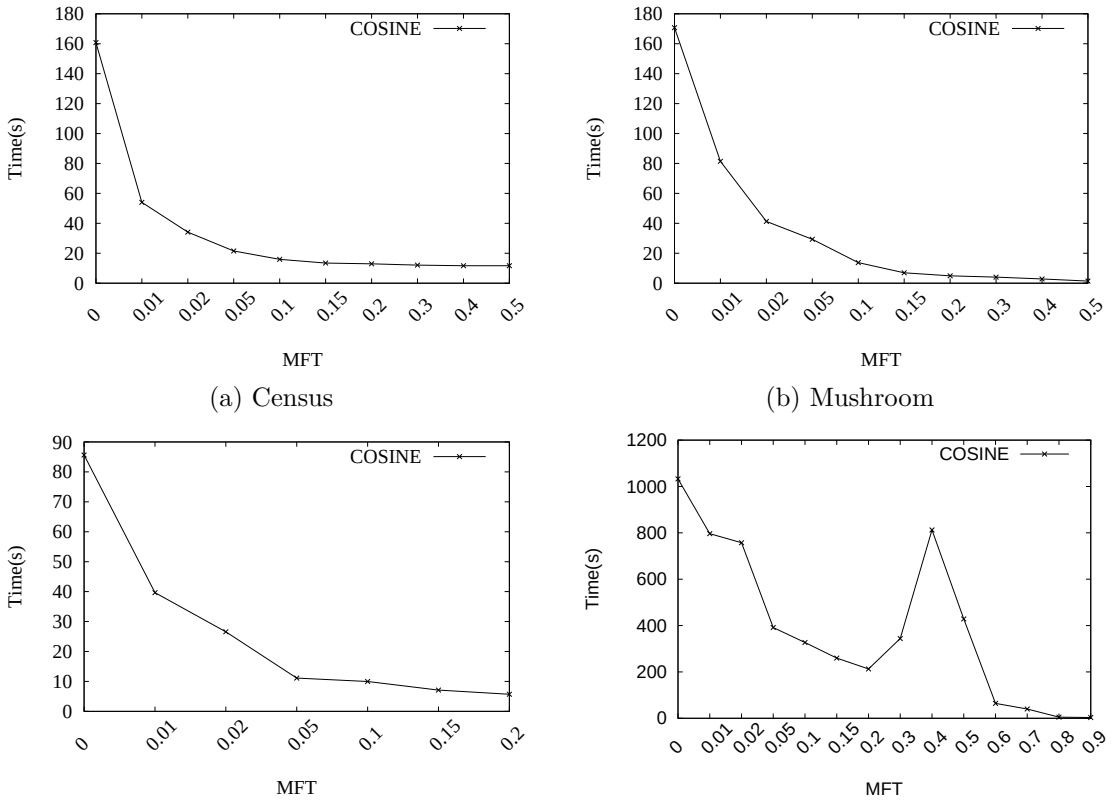
5.3.2 Minimum Frequency Threshold

We ran a series of discovery tasks on the Census, Mushroom, Waveform, and Spambase datasets to evaluate the effect of the MFT as a constraint on the COSINE algorithm for mining maximal contrast sets. For each dataset, we use an MSSRT of 0 and an MSDT of 0 for COSINE, we average the results over 10 consecutive runs, and we use a significance level of 0.95. In section 5.3.2.1, we examine the run time of COSINE as the MFT is varied. In Section 5.3.2.2, we examine the interestingness of the contrast sets discovered by COSINE as the MFT is varied. In Section 5.3.2.3, we examine the maximum and average specificity of the contrast sets discovered by COSINE as the MFT is varied.

5.3.2.1 Runtime

We evaluate the run time performance of COSINE when mining maximal valid contrast sets as the MFT is varied. Figures 5.10(a) and 5.10(b) show the average runtime for COSINE on the Census and Mushroom, respectively, as the MFT is varied from 0 to 0.5. Figure 5.10(c) shows the average runtime for COSINE on the Waveform dataset as the MFT is varied from 0 to 0.2, as values greater than 0.2 did not produce any contrast sets. Figure 5.10(d) shows the average runtime for COSINE on the Spambase dataset as the MFT is varied from 0 to 0.9, as values up to 0.9 also produced maximal valid contrast sets.

In Figures 5.10(a), 5.10(b), and 5.10(c) there is a sharp decrease in the run time of COSINE as the MFT increases from 0 to 0.01, declining more gradually thereafter, eventually plateauing for MFT values greater than 0.02 for the Census and Waveform datasets, and 0.05 on the Mushroom dataset. For the Spambase dataset, there is a sharp decrease in the run time as the MFT is increased from 0 to 0.05, decreasing more gradually as the MFT is increased from 0.05 to 0.2. The run time increases again as the MFT is increased from 0.2 to 0.4, declining sharply thereafter. Generally, a decrease



(a) Census (b) Mushroom

(c) Waveform (d) Spambase

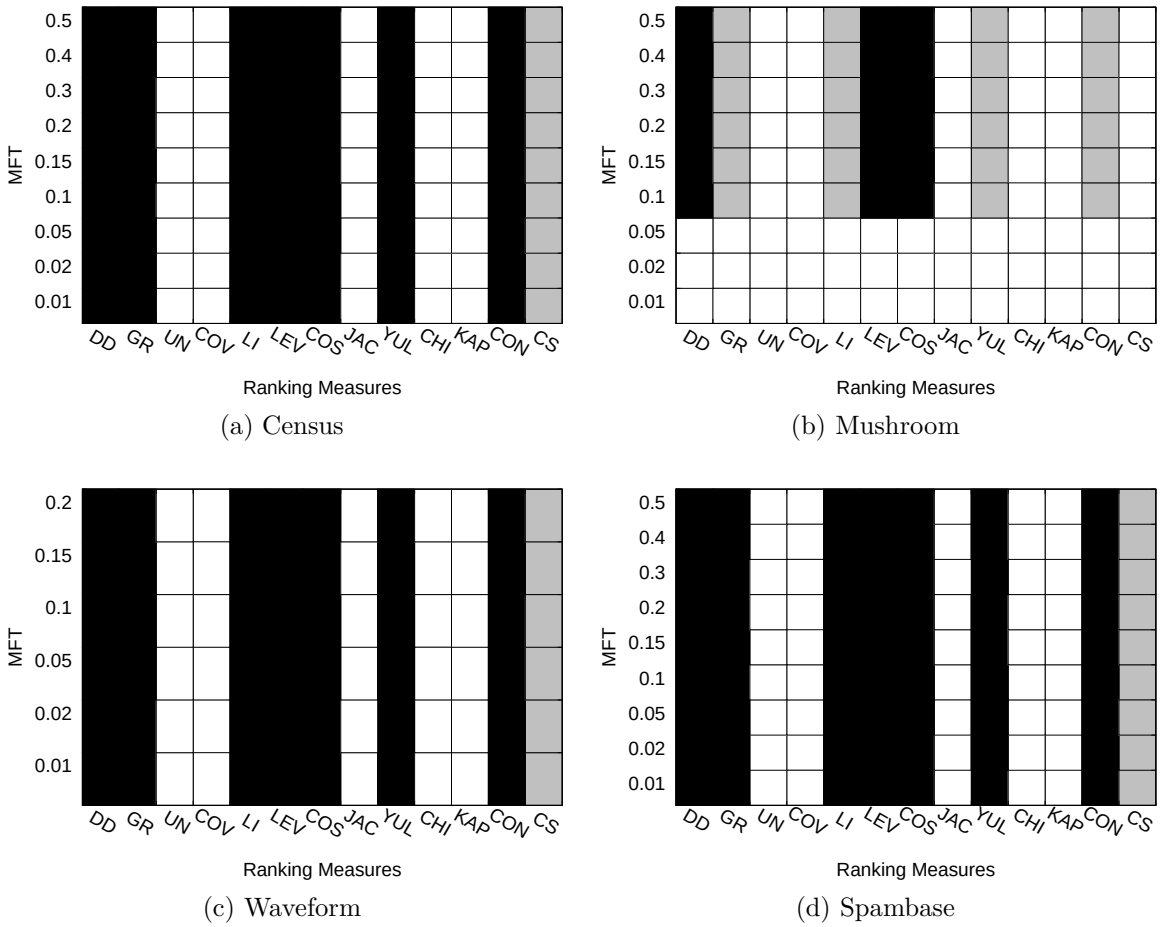
Figure 5.10: Run time vs. MFT

in run time as the MFT increases occurs because the MFT serves as a constraint on the depth of the search space and as its value increases, less contrast sets will meet this threshold and therefore COSINE will take less time to complete the search. Additionally, as the depth of the search space is constrained, branches of the search space which would have otherwise been pruned away because a more specific superset that subsumes would have been found, would also have to be explored, possibly increasing the running time.

5.3.2.2 Interestingness

We evaluate the interestingness of the maximal valid contrast sets discovered by COSINE on the Census, Mushroom, Waveform, and Spambase datasets using thirteen ranking measures: distribution difference (DD), growth rate (GR), unusualness (UN), coverage (COV), lift (LI), leverage (LEV), change of support (COS), Jaccard coefficient (JAC),

Yule's Q coefficient (YUL), ϕ-coefficient (CHI), κ-coefficient (KAP), conviction (CON), and collective strength (CS). Once again we vary the MFT from 0 to 0.5 for the Census, and Mushroom datasets, 0 to 0.2 for the Waveform dataset, and from 0 to 0.9 for the Spambase dataset. We again use a standard T-test at a confidence level of 95% to determine whether there is a statistically significant difference between the average interestingness values obtained by COSINE at an MFT of 0 and those obtained by COSINE when the MFT is greater than 0.



(a) Census (b) Mushroom

(c) Waveform (d) Spambase

Figure 5.11: Significance Values vs. MFT

The results are shown in Figures 5.11a to 5.11d, where statistical significance of the interestingness values obtained by COSINE when the MFT is 0 versus the interestingness values obtained by COSINE when the MFT is greater than 0 are plotted against

the MFT for each of the four datasets. In the graphs shown, the statistical significance is described by a three-step scale: white, black, and grey indicate that the interestingness values obtained by COSINE when the MFT is greater than 0 are significantly higher, significantly lower, and not significantly different, respectively, compared to those obtained by COSINE when the MFT is 0. For example, in Figure 5.11(a), the average interestingness of the results obtained by COSINE when the MFT is 0.01 is significantly higher than those obtained by COSINE when the MFT is 0 for UN, COV, JAC, CHI, and KAP, significantly lower for DD, GR, LI, LEV, COS, YUL, and CON, and not significantly different for CS.

In Figures 5.11(a), 5.11(c), and 5.11(d), the average interestingness of the results obtained by COSINE when the MFT is greater than 0 is significantly higher than those obtained by COSINE when the MFT is 0 for UN, COV, JAC, CHI, and KAP, significantly lower for DD, GR, LI, LEV, COS, YUL, and CON, and not significantly different for CS.

In Figure 5.11(b), the average interestingness of the results obtained by COSINE for MFT values from 0.01 to 0.05 is significantly higher than those obtained by COSINE when the MFT is 0 for all measures. However, the average interestingness of the results obtained by COSINE when the MFT is greater than 0.05 is significantly higher than those obtained by COSINE when the MFT is 0 for UN, COV, JAC, CHI, KAP, and CS, significantly lower for DD, LEV, and COS, and not significantly different for GR, LI, YUL, and CON.

For three of the four datasets, the average interestingness of the results obtained by COSINE for MFT values greater than 0 is significantly higher than the interestingness values obtained by COSINE when the MFT is 0 for only 5 of the 13 ranking measures. These results demonstrate that selecting contrast sets with high support does not necessarily produce a subset of contrast sets that is on average more interesting.

5.3.2.3 Specificity

We evaluate the specificity of the maximal valid contrast sets discovered by COSINE within the context of average and maximum specificities as the MFT is varied as described in the previous two sections. Tables 5.7 and 5.8 show the number of maximal valid contrast sets discovered by COSINE as the MFT is varied from 0 to 0.5. Table 5.9 shows the number of maximal valid contrast sets discovered by COSINE as the MFT is varied from 0 to 0.9. Table 5.10 shows the number of maximal valid contrast sets discovered by COSINE as the MFT is varied from 0 to 0.2. The columns have the same meaning as Table 5.3.

Table 5.7: Summary of Contrast Sets discovered - Census

	MFT									
k	*0.00*	*0.01*	*0.02*	*0.05*	*0.10*	*0.15*	*0.20*	*0.30*	*0.40*	*0.50*
1	0	0	0	0	0	0	0	0	0	0
2	1	6	3	9	3	3	2	2	1	1
3	12	60	35	15	25	10	1	6	5	2
4	75	155	87	47	23	9	12	1	5	1
5	277	271	168	70	24	8	8	4	2	0
6	690	468	220	63	20	14	6	1	0	0
7	1052	434	202	54	8	2	0	1	0	0
8	1290	308	144	26	2	2	1	0	0	0
9	1084	173	43	8	2	0	0	0	0	0
10	766	56	12	2	0	0	0	0	0	0
11	450	8	1	0	0	0	0	0	0	0
12	238	0	0	0	0	0	0	0	0	0
13	74	0	0	0	0	0	0	0	0	0

In Tables 5.7, 5.8, and 5.10 the average and maximal specificities of the maximal valid contrast sets discovered by COSINE decrease steadily as the MFT increases from 0 to 0.5, 0 to 0.2, respectively. In Table 5.9, the maximal specificity of the maximal valid contrast sets discovered by COSINE initially remains constant at 17 for MFT values from 0.01 to 0.4, but it decreases sharply thereafter. The average specificity of the maximal valid contrast sets discovered by COSINE initially never decreases below 8 for MFT values from 0 to 0.4, but it decreases thereafter.

These results demonstrate that using the MFT as a constraint on the search process reduces the maximal or average specificity of the maximal valid contrast sets discovered.

Table 5.8: Summary of Contrast Sets discovered - Mushroom

	MFT									
k	*0.00*	*0.01*	*0.02*	*0.05*	*0.10*	*0.15*	*0.20*	*0.30*	*0.40*	*0.50*
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	1	1	0
3	0	0	0	2	0	1	6	2	3	3
4	0	1	1	22	10	12	14	9	9	4
5	1	17	45	36	50	24	11	14	16	7
6	13	66	69	51	104	45	36	20	8	1
7	68	130	182	72	162	60	37	9	2	0
8	131	266	215	91	126	30	25	6	0	0
9	228	206	301	86	174	15	3	2	0	0
10	176	452	271	24	113	17	4	0	0	0
11	291	249	142	18	40	1	1	0	0	0
12	243	130	84	2	30	1	1	0	0	0
13	157	104	46	1	6	0	0	0	0	0
14	80	49	15	0	0	0	0	0	0	0
15	44	10	0	0	0	0	0	0	0	0
16	43	1	0	0	0	0	0	0	0	0

Table 5.9: Summary of Contrast Sets Discovered - Spambase

	MFT													
k	*0.00*	*0.01*	*0.02*	*0.05*	*0.10*	*0.15*	*0.20*	*0.30*	*0.40*	*0.50*	*0.60*	*0.70*	*0.80*	*0.90*
1	1	0	1	1	1	1	1	1	0	1	0	0	2	4
2	91	95	87	29	17	19	18	19	1	19	15	34	32	26
3	483	243	134	66	20	25	16	15	19	20	623	658	280	4
4	1045	879	306	175	71	112	53	19	15	127	4636	2700	254	0
5	2971	2636	721	531	219	294	297	11	20	678	10331	2839	13	0
6	4714	5436	1814	956	566	582	894	13	71	1923	7645	675	0	0
7	5506	6491	2649	1217	1283	625	592	57	608	1973	1288	11	0	0
8	5670	6343	3741	1220	2881	2286	55	69	1942	5346	16	0	0	0
9	8278	7374	5708	2437	4317	3121	76	268	4292	12168	0	0	0	0
10	8641	6750	7132	2912	3480	1075	171	817	6613	8958	0	0	0	0
11	6332	4347	6448	1794	1123	230	213	947	7692	2037	0	0	0	0
12	2670	2008	2331	760	347	259	259	1133	5823	381	0	0	0	0
13	1181	919	786	473	219	219	219	661	3092	353	0	0	0	0
14	433	397	345	323	194	194	194	310	1362	90	0	0	0	0
15	140	140	128	126	121	121	121	98	334	1	0	0	0	0
16	42	42	42	42	42	42	42	34	104	0	0	0	0	0
17	11	11	11	11	11	11	11	11	2	0	0	0	0	0

Table 5.10: Summary of Contrast Sets discovered - Waveform

	MFT						
k	*0.00*	*0.01*	*0.02*	*0.05*	*0.10*	*0.15*	*0.20*
1	2	11	13	41	82	58	32
2	379	1340	2230	316	0	0	0
3	11890	8034	171	0	0	0	0
4	20009	12	0	0	0	0	0
5	562	0	0	0	0	0	0

5.4 Evaluation of GENCCS

GENCCS utilizes the minimum mutual information threshold (MMIT) and the minimum all-confidence threshold (MACT) to prune attributes and the combine sets of 1-specific contrast sets, respectively, so that only the attribute-interval pairs that are highly correlated are combined. In this section, we examine the effect of both these constraints on the run time, interestingness, and specificity of the maximal correlated valid contrast sets using four variants of our algorithm which vary the MMIT: GENCCS (sets MMIT to half of the average mutual information), GENCCS-0 (sets the MMIT to 0), GENCCS-1 (sets the MMIT to the average mutual information), and GENCCS-2 (sets the MMIT to double the average mutual information).

We use the same datasets to evaluate GENCCS as were used for the evaluation of COSINE: namely Census, Mushroom, Waveform, and Spambase. The objective is to evaluate the performance of GENCCS when mining maximal correlated valid contrast sets. For each dataset, we set the minimum subset support ratio threshold (MSSRT) to 0, the minimum frequency threshold (MFT) to 0 and the minimum support difference threshold (MSDT) to 0. We average the results over 10 consecutive runs, using a significance level of 0.95. In section 5.4.1, we examine the run time of GENCCS as the MACT is varied. In Section 5.4.2, we examine the interestingness of the correlated contrast sets discovered by GENCCS as the MACT is varied. In Section 5.4.3, we examine the maximum and average specificity of the correlated contrast sets discovered by GENCCS as the MACT is varied. In Section 5.4.4, we examine the cardinality of the longest combine sets of the contrast sets discovered by GENCCS as the MACT is varied.

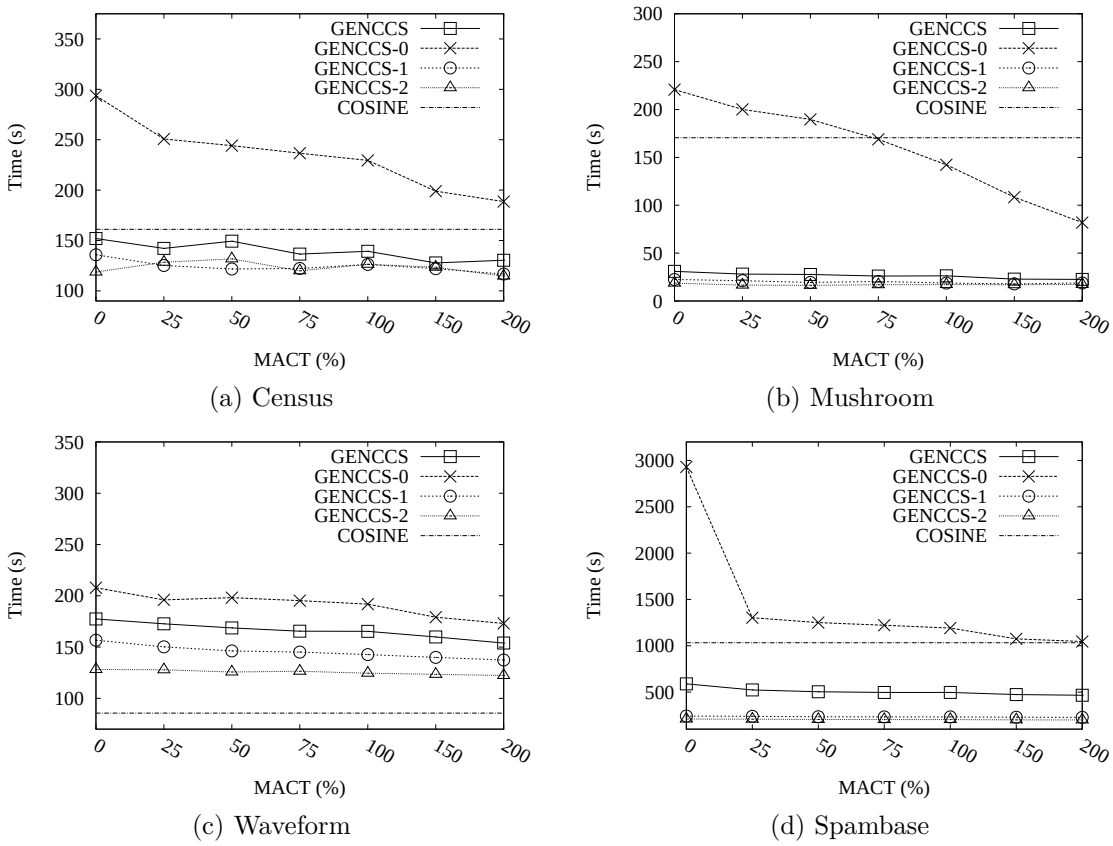
From our experiments, we draw four main conclusions. First, the run times of GENCCS, GENCCS-1, and GENCCS-2 are significantly lower than that of COSINE. Second, the average interestingness values of the maximal correlated valid contrast sets discovered by GENCCS are significantly higher than that of those discovered by COSINE.

Third, the maximum specificity of the maximal correlated valid contrast sets discovered decreases as the MMIT is increased but is constant as the MACT is increased. Fourth, the average specificity of the maximal correlated valid contrast sets discovered by GENCCS, GENCCS-0, GENCCS-1, and GENCCS-2 at higher MACT values is significantly higher than when the MACT is 0.

5.4.1 Runtime

We compare the run time performance of GENCCS, GENCCS-0, GENCCS-1, and GENCCS-2 to COSINE when mining maximal correlated valid contrast sets as the MACT is varied. Figures 5.12(a) to 5.12(d) show the average runtime for GENCCS, GENCCS-0, GENCCS-1, and GENCCS-2 on the Census, Mushroom, Waveform, and Spambase datasets, respectively, as the MACT is varied from 0% to 200% of the average all-confidence. For comparison, we also show the average runtime for COSINE, represented as a dashed line. The run time for COSINE is a constant value because the MACT is not a constraint on the COSINE algorithm, thus increasing its value has no effect on the time taken by COSINE to discover maximal valid contrast sets.

When the MACT is 0, the search space generated and subsequently the maximal contrast sets discovered by GENCCS-0 is identical to that of COSINE. However, the run times of GENCCS-0 and COSINE are different because of the additional time it takes to measure the mutual information and all confidence in GENCCS-0. The difference in run times between GENCCS and COSINE gives us a measure of the cost of determining the mutual information and all confidence. For example, the cost for the Spambase dataset is 1898 seconds (i.e., the difference between the run times of GENCCS-0 (2930) and COSINE (1032)). The cost is highest for the Spambase dataset at 1898 seconds (i.e, a 183% increase), followed by the Waveform dataset at 122 seconds (i.e., a 143% increase), then the Census dataset at 133 seconds (i.e., a 82% increase), and finally the Mushroom dataset at 50 seconds (i.e., a 30% increase).



(a) Census (b) Mushroom

(c) Waveform (d) Spambase

Figure 5.12: Run time vs. MACT

For the Census, Mushroom, and Spambase datasets, GENCCS, GENCCS-1, and GENCCS-2 have a significantly lower runtime than COSINE at all MACT values. GENCCS-0, however, has a higher runtime than COSINE for all MACT values on the Census dataset but the runtime declines to that of COSINE when the MACT is 150% on the Spambase dataset and below that of COSINE when the MACT is above 75% on the Mushroom dataset. The difference in run times is significant because measuring the mutual information for all pairs of attributes and measuring the all-confidence of all 1-specific contrast sets adds to the run time of the algorithm thus any improvements over COSINE means that the cost of measuring the mutual information and all-confidence are outweighed by the benefit of them reducing the search space. For the Waveform dataset, all four algorithms, GENCCS, GENCCS-0, GENCCS-1, and GENCCS-2 have a higher runtime than COSINE at all MACT values.

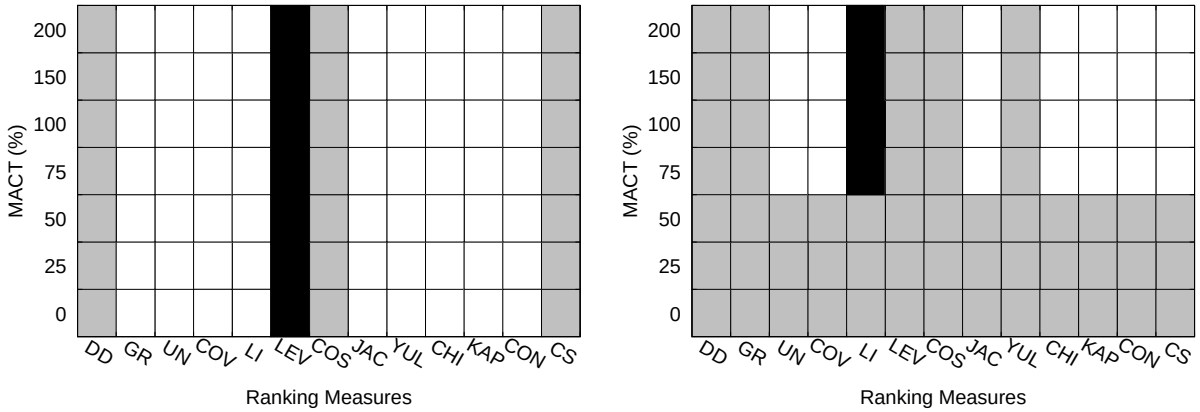
5.4.2 Interestingness

We compare the interestingness of the maximal valid contrast sets discovered by GENCCS, GENCCS-0, GENCCS-1, and GENCCS-2 to COSINE on the Census, Mushroom, Waveform, and Spambase datasets using thirteen ranking measures: distribution difference (DD), growth rate (GR), unusualness (UN), coverage (COV), lift (LI), leverage (LEV), change of support (COS), Jaccard coefficient (JAC), Yule's Q coefficient (YUL), ϕ-coefficient (CHI), κ-coefficient (KAP), conviction (CON), and collective strength (CS). Once again, we vary the MACT from 0% to 200% of the average all confidence. We use a standard T-test at a confidence level of 95% to determine whether there is a statistically significant difference between the average interestingness values obtained by GENCCS and COSINE, GENCCS-0 and COSINE, GENCCS-1 and COSINE, and GENCCS-2, respectively.

These results are shown in Figures 5.13 to 5.16. In Figures 5.13 to 5.16, the difference in the interestingness values obtained by GENCCS, GENCCS-0, GENCCS-1, and GENCCS-2, respectively versus COSINE are plotted against the MACT for each of the four datasets. Once again, white, black, and grey indicate that the interestingness values are significantly higher, significantly lower, and not significantly different, respectively.

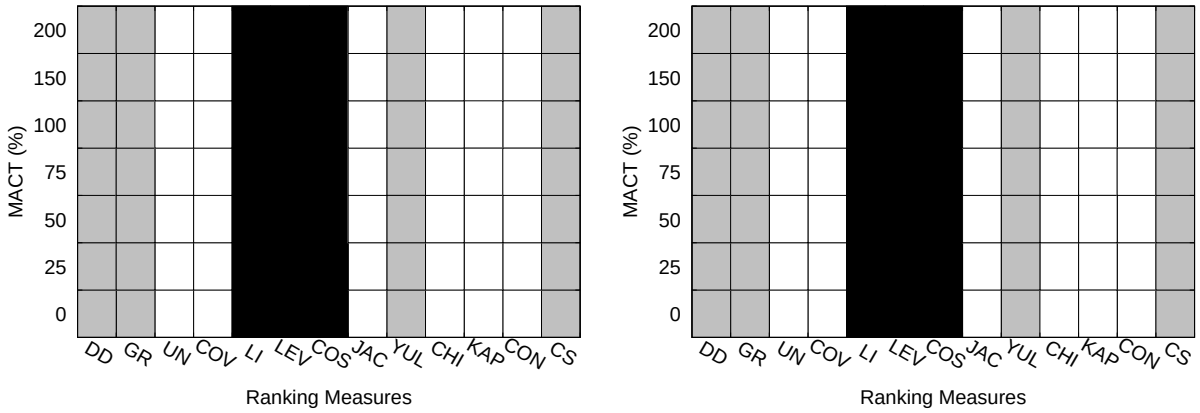
In Figure 5.13(a), the average interestingness of the results obtained by GENCCS is significantly higher than those obtained by COSINE for GR, UN, COV, LI, JAC, CHI, YUL, KAP, and CON, significantly lower for LEV, and not significantly different than COSINE for DD, COS, and CS, for all MACT values.

In Figure 5.13(b), the average interestingness of the results obtained by GENCCS-0 is not significantly different from those obtained by COSINE for all measures for MACT values from 0 to 50%. Above MACT values of 50%, the average interestingness of the results obtained by GENCCS are significantly higher than those obtained by COSINE for UN, COV, JAC, CHI, KAP, CON, and CS, significantly lower for LI, and not significantly



(a) GENCCS vs. COSINE

(b) GENCCS-0 vs. COSINE



(c) GENCCS-1 vs. COSINE

(d) GENCCS-2 vs. COSINE

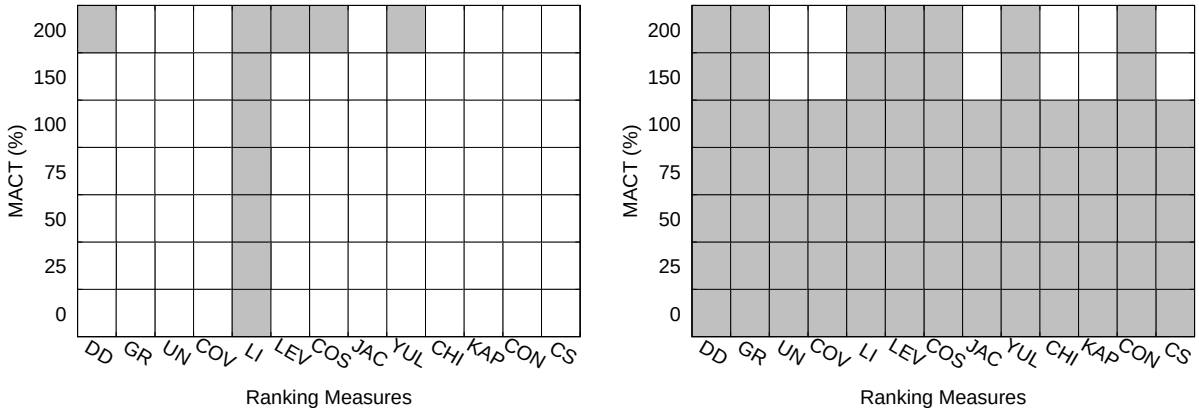
Figure 5.13: Significance Values vs. MACT - Census

different than COSINE for DD, GR, LEV, COS and YUL

In Figures 5.13(c) and 5.13(d), the average interestingness of the results obtained by GENCCS-1 and GENCCS-2 are significantly higher than those obtained by COSINE for UN, COV, JAC, CHI, KAP, and CON, significantly lower for LI, LEV, and COS, and not significantly different than COSINE for DD, GR, YUL, and CS, for all MACT values.

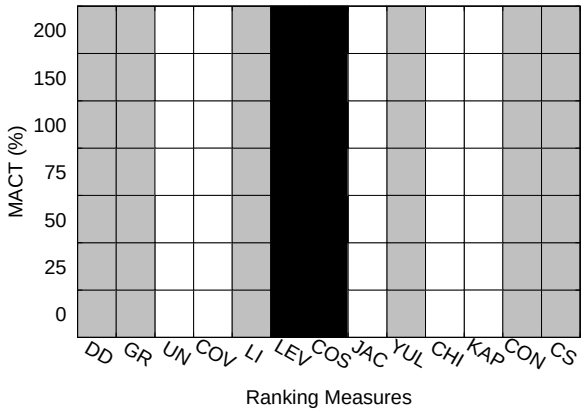
The average interestingness of the results obtained by each of GENCCS-0, GENCCS, GENCCS-1, and GENCCS-2 versus COSINE for the Mushroom, Waveform, and Spambase datasets can be interpreted in a similar manner, and is left as an exercise for the reader.

Our results demonstrate, on average, that the contrast sets found by GENCCS,

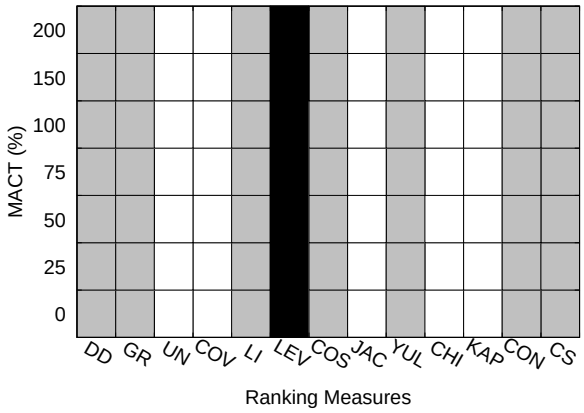


(a) GENCCS vs. COSINE

(b) GENCCS-0 vs. COSINE



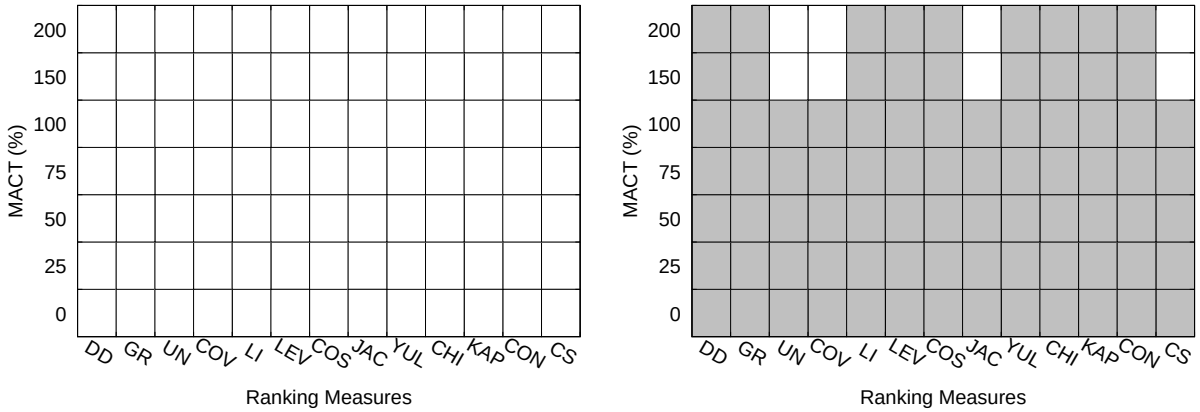
(c) GENCCS-1 vs. COSINE



(d) GENCCS-2 vs. COSINE

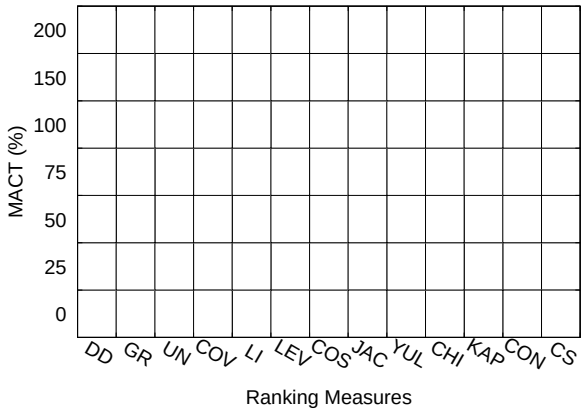
Figure 5.14: Significance Values vs. MACT - Mushroom

GENCCS-0, GENCCS-1, and GENCCS-2, are either significantly more interesting or not significantly different than the contrast sets found by COSINE. They also demonstrate that of the four variants, GENCCS most often produces the most interesting results. This is a very significant result because it demonstrates that combining only the attribute-interval pairs that are highly correlated produces a subset of contrast sets that is either significantly more interesting or not significantly different than the contrast sets obtained by combining all possible attribute-interval pairs. Having already demonstrated the improvements in run time, the results with interestingness values further validate the choice of GENCCS over COSINE.

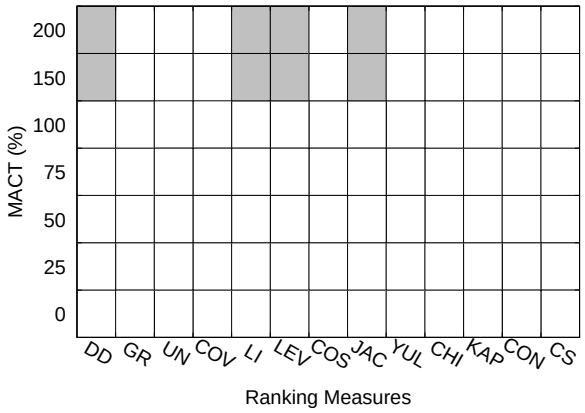


(a) GENCCS vs. COSINE

(b) GENCCS-0 vs. COSINE



(c) GENCCS-1 vs. COSINE



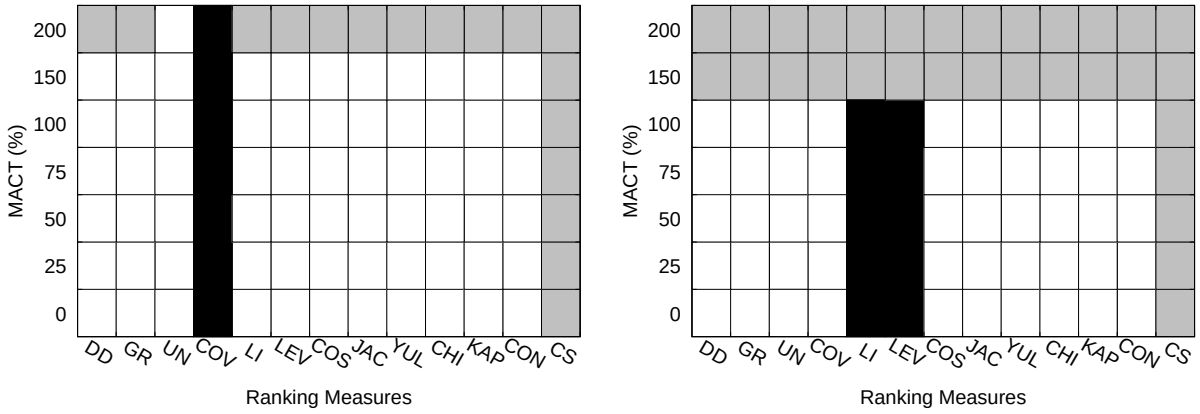
(d) GENCCS-2 vs. COSINE

Figure 5.15: Significance Values vs. MACT - Waveform

5.4.3 Specificity

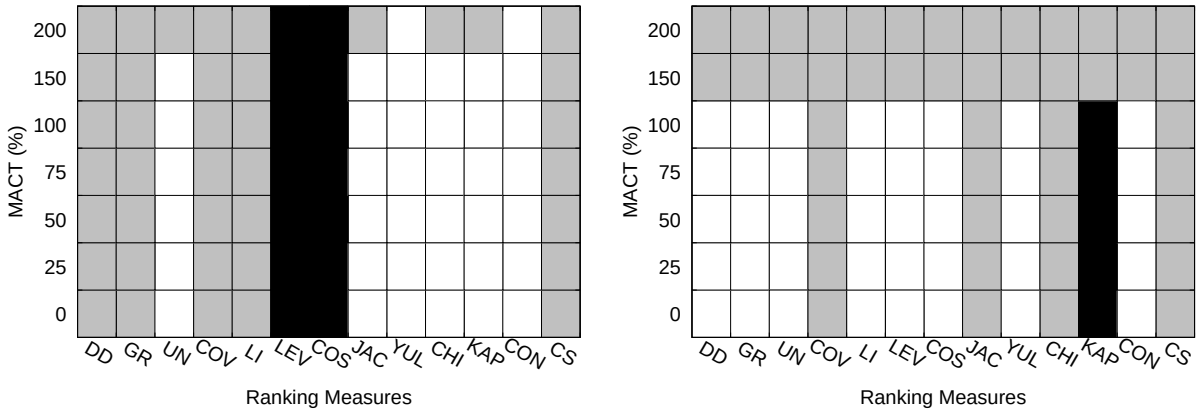
We compare the specificity of the maximal valid contrast sets discovered by GENCCS, GENCCS-0, GENCCS-1, and GENCCS-2 to COSINE within the context of average and maximum specificities as the MACT is varied. For the average specificity, we use a standard T-test at a confidence level of 95% to determine whether there is a significant difference between the average specificity values when the MACT is 0, and those obtained when the MACT is greater than 0. Once again, we use the white, black, and grey three-step scale to compare statistical significance.

The results for average specificity are shown in Figures 5.17(a) to 5.17(d) for the



(a) GENCCS vs. COSINE

(b) GENCCS-0 vs. COSINE



(c) GENCCS-1 vs. COSINE

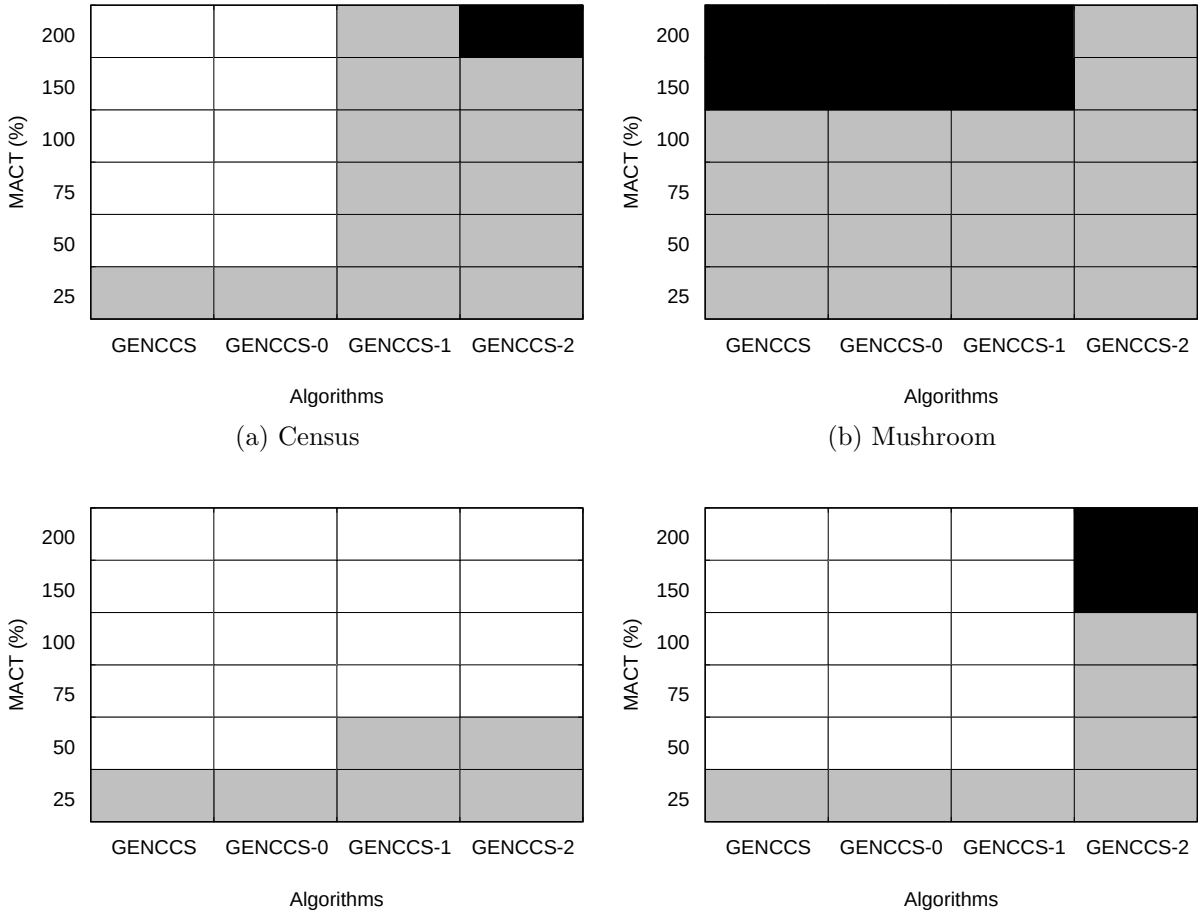
(d) GENCCS-2 vs. COSINE

Figure 5.16: Significance Values vs. MACT - Spambase

Census, Mushroom, Waveform, and Spambase datasets, respectively. The raw data from which these figures were generated can be found in Appendix B.

In Figures 5.17(a), 5.17(c) and 5.17(d), the average specificity of the results obtained by GENCCS and GENCCS-0 for MACT values from 50% to 200% is significantly higher than the results obtained when the MACT is 0. For the Mushroom dataset, in Figure 5.17(b), the average specificity of the results obtained by GENCCS and GENCCS-0 for MACT values from 25% to 100% is not significantly different than the results obtained by GENCCS and GENCCS-0 when the MACT is 0, but significantly lower for MACT values from 150 to 200%.

For the Waveform and Spambase datasets, the average specificity of the results ob-



(a) Census

(b) Mushroom

(c) Waveform

(d) Spambase

Figure 5.17: Comparison of Average Specificity

tained by GENCCS-1 for MACT values of 25% to 200% is generally significantly higher than the results obtained when the MACT is 0, except when the MACT is 50% for the Waveform dataset, where there is no significant difference. For the Census and Mushroom datasets, the average specificity of the results obtained by GENCCS-1 for MACT values of 25% to 200% is generally not significantly different than the results obtained when the MACT is 0, except when the MACT is above 100% for the Mushroom dataset, where it is significantly lower.

The average specificity of the results obtained by GENCCS-2 for MACT values form 25% to 200% is generally not significantly different than the results obtained when the MACT is 0 except for MACT values above 50% for the Waveform dataset where it

is significantly higher, and MACT values above 100% and 150% for the Census and Spambase datasets, respectively, where it is significantly lower.

Our results demonstrate that, in general, increasing the MACT significantly increases the average specificity of the contrast sets discovered for GENCCS and GENCCS-0.

Table 5.11 shows the maximum specificity of the maximal valid contrast sets discovered by GENCCS, GENCCS-0, GENCCS-1, and GENCCS-2 on the Census, Mushroom, Waveform, and Spambase datasets, respectively.

Table 5.11: Maximum Specificity

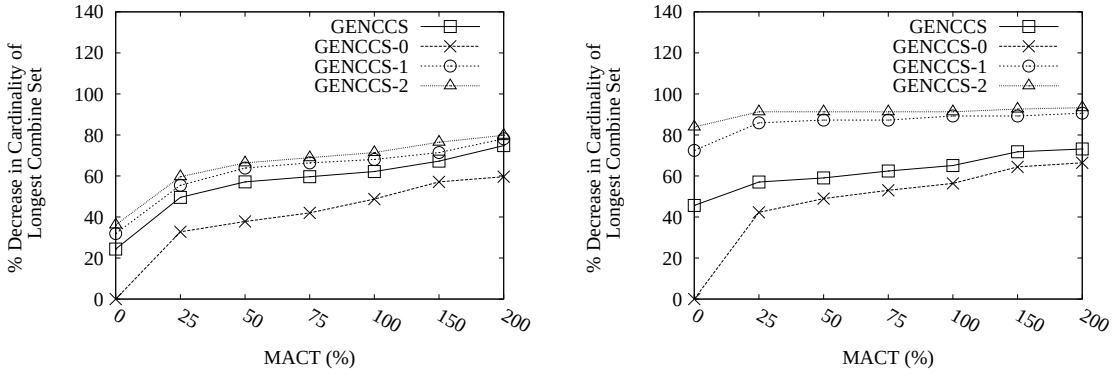
	Maximum Specificity				
Dataset	GENCCS	GENCCS-0	GENCCS-1	GENCCS-2	COSINE
Census	8	13	4	3	13
Mushroom	12	16	10	9	16
Waveform	5	5	5	5	5
Spambase	17	17	15	15	17

For all four datasets, the maximum specificity of the maximal correlated contrast sets discovered by GENCCS, GENCCS-0, GENCCS-1, and GENCCS-2 is unchanged when the MACT is increased. Also, the maximum specificity of the maximal correlated contrast sets discovered by GENCCS-0 is the same as that of COSINE.

For the Census dataset, the maximum specificity decreased from 13 with GENCCS-0 through to 8 with GENCCS, to 4 with GENCCS-1, then 3 with GENCCS-2. This represents a decline as the MMIT is increased. Similarly, on the Mushroom dataset, the maximum specificity decreased from 16 with GENCCS-0 through to 12 with GENCCS, to 10 with GENCCS-1, then 9 with GENCCS-2. However, on the Spambase dataset, the maximum specificity remained at 17 for both GENCCS-0 and GENCCS but declined to 15 for both GENCCS-1 and GENCCS-2, while on the Waveform dataset, the maximum specificity remains at 5 for each of GENCCS-0, GENCCS, GENCCS-1, and GENCCS-2. This means that increasing the MACT did not affect the depth of the search space to which either GENCCS, GENCCS-0, GENCCS-1, or GENCCS-2 could find maximal correlated valid contrast sets, while increasing the MMIT did.

5.4.4 Cardinality of the Longest Combine Set

We compare the percentage decrease in the cardinality of the longest combine sets for each of GENCCS, GENCCS-0, GENCCS-1, and GENCCS-2 to the longest combine sets for COSINE when mining maximal valid contrast sets as the MACT is varied. The results are shown in Figures 5.18(a) to 5.18(d) for the Census, Mushroom, Waveform, and Spambase datasets, respectively.



(a) Census (b) Mushroom

(c) Waveform (d) Spambase

Figure 5.18: % Decrease in the Cardinality of the Longest Combine Set vs. MACT

When the MACT is 0, GENCCS-0 is equivalent to COSINE, in terms of the search space generated, and in the cardinality of the longest combine set, thus the percentage decrease is 0. For each dataset, the cardinality of the longest combine set decreases as the MACT increases. The sharpest decline occurs with GENCCS-0 when the MACT is increased from 0 to 25%, where the cardinality of the longest combine set decreases

by 33%, 42%, 24%, and 67%, respectively, on the Census, Mushroom, Waveform, and Spambase datasets. When the MACT is 0, the size of the longest combine set also decreases as the MMIT is increased. The sharpest decline occurs when the MMIT is increased from 0 to 50% (i.e., from GENCCS-0 to GENCCS), where the cardinality of the longest combine set decreases by 24%, 46%, and 59% on the Census, Mushroom, and Waveform datasets. On the Spambase dataset, the sharpest decline occurs when the MMIT is increased from 50 to 100% (i.e., from GENCCS to GENCCS-1), where the cardinality of the longest combine set decreases by 41%.

The longest combine set can be found on the first level of the search space tree and it represents the largest possible subtree that can be generated. The bigger this combine set, the bigger the subtree will be and thus the longer GENCCS will take to find contrasts sets. Thus any reduction in the size of the longest combine set would reduce the run time of GENCCS.

Our results demonstrate that increasing both the MMIT and the MACT significantly decreases the cardinality of the longest combine set in the search space.

5.5 Minimum Support Ratio Threshold (MSRT)

We ran a series of discovery tasks on the Census, Mushroom, Waveform, and Spambase datasets to evaluate the effect of the MSRT as a constraint on the COSINE and GENCCS algorithms for mining maximal valid contrast sets and maximal correlated valid contrast sets, respectively. We use two variants of our algorithms: λ-COSINE, which is a variant of COSINE that uses the MSRT as a constraint, and $\lambda-$GENCCS, a variant of GENCCS that uses the MSRT as a constraint.

We use the same datasets to evaluate the MSRT as were used in the evaluation of COSINE and GENCCS: namely Census, Mushroom, Waveform, and Spambase. For each dataset, we set the minimum subset support ratio threshold (MSSRT) to 0, the

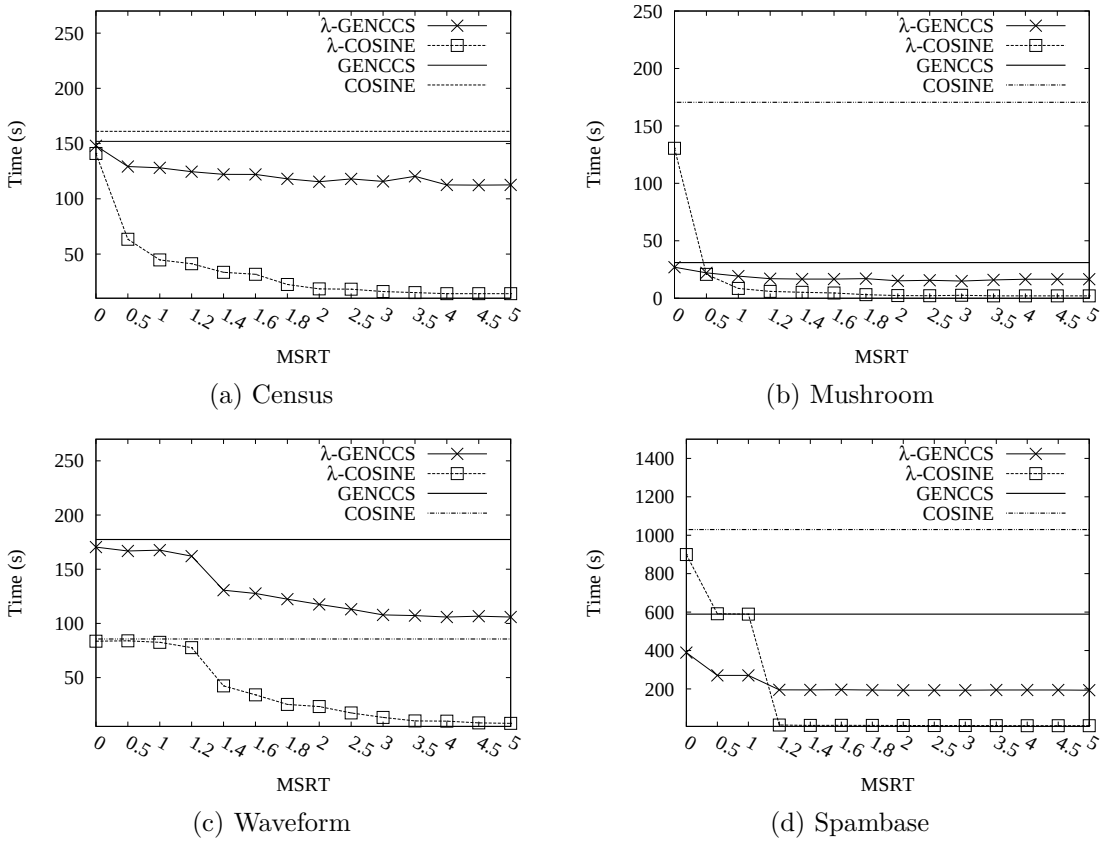
minimum frequency threshold (MFT) to 0, the minimum support difference threshold (MSDT) to 0, the minimum mutual information threshold (MMIT) to half of the average mutual information, and the minimum all confidence threshold (MACT) to half of the average all confidence. We average the results over 10 consecutive runs, and we use a significance level of 0.95. In Section 5.5.1, we examine the run time of λ-COSINE and λ-GENCCS as the MSRT is varied. In Section 5.5.2, we examine the interestingness of the contrast sets discovered by λ-COSINE and λ-GENCCS as the MSRT is varied. In Section 5.5.3, we examine the maximum and average specificity of the contrast sets discovered by λ-COSINE and λ-GENCCS as the MSRT is varied. In Section 5.5.4, we examine the number of contrast sets discovered by λ-COSINE and λ-GENCCS as the MSRT is varied.

From our experiments, we draw three main conclusions. First, the run time of λ-COSINE is significantly lower than that of COSINE for all MSRT values. Second, the average interestingness of the maximal valid contrast sets and maximal correlated valid contrast sets discovered by λ-COSINE and λ-GENCCS, respectively, are significantly higher than that of those discovered by COSINE and GENCCS. Third, using lower values for the MSRT does not reduce the maximal or average specificities of the maximal valid contrast sets or the maximal correlated valid contrast sets discovered by λ-COSINE and λ-GENCCS from that of COSINE and GENCCS, respectively.

5.5.1 Runtime

We compare the run time performance of λ-COSINE and λ-GENCCS to COSINE and GENCCS when mining maximal valid contrast sets and maximal correlated valid contrast sets, respectively, as the MSRT is varied. Figures 5.19(a) to 5.19(d) show the average runtime for λ-COSINE and λ-GENCCS on the Census, Mushroom, Waveform, and Spambase datasets, respectively, as the MSRT is varied from 0 to 5. For comparison, we also show the average runtime for COSINE, represented as a straight line, and for

GENCCS, represented as a dashed line. The run times for COSINE and GENCCS are constant values because the MSRT is not a constraint on the COSINE and GENCCS algorithms, thus increasing its value has no effect on the time taken by COSINE and GENCCS to discover maximal valid contrast sets and maximal correlated valid contrast sets, respectively.



(a) Census (b) Mushroom

(c) Waveform (d) Spambase

Figure 5.19: Run time vs. MSRT

In Figures 5.19(a), 5.19(b), and 5.19(d) there is a sharp decrease in the run time of λ-COSINE as the MSRT increases from 0 to 0.5, declining more gradually thereafter, eventually plateauing for MSRT values greater than 1 for the Mushroom and Spambase datasets, and 1.8 on the Census dataset. On the Spambase dataset, there is also a second sharper decline in the run time of λ-COSINE as the MSRT increases from 1 to 1.2. For the Waveform dataset, there is also a sharp decrease in the run time as the MSRT is increased from 1.2 to 1.4 though much less so than on the other datasets, then decreasing

more gradually thereafter, eventually plateauing for MSRT values greater than 2.

On all four datasets, there is a gradual decrease in the run time of λ-GENCCS as the MSRT is increased from 0 to 5, though on the Waveform and Spambase datasets, there is a period of sharper decline for MSRT values from 1.2 to 1.4 for Waveform, and 0 to 0.5 for Spambase.

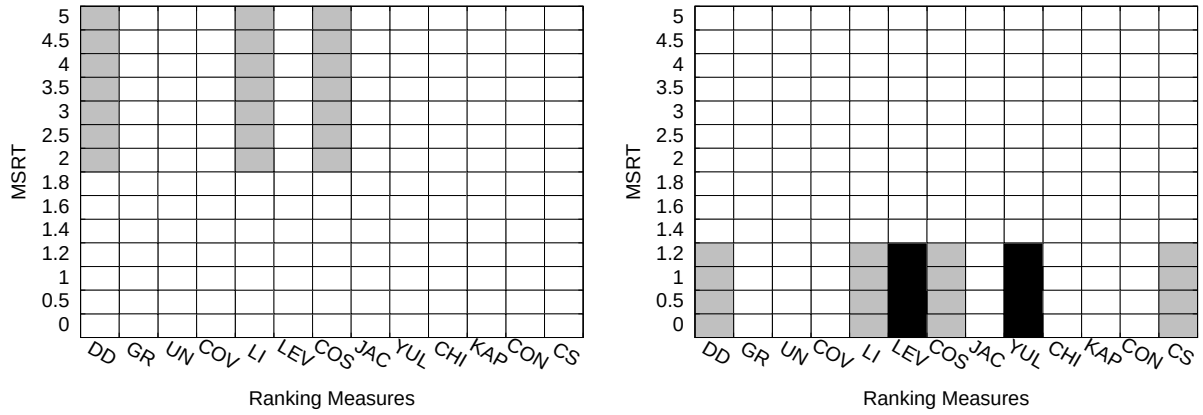
On the Mushroom and Spambase datasets, the run time of λ-COSINE is initially less than that of COSINE but greater than that of GENCCS and λ-GENCCS for MSRT values of 0 to 1 for Spambase and only 0 for Mushroom. On the Census and Waveform datasets, the run time of λ-COSINE is always less than that of COSINE, GENCCS, and λ-GENCCS for all MSRT values, though on the Waveform dataset, the run times of GENCCS, and λ-GENCCS are always higher than that of COSINE.

Generally, a decrease in run time as the MSRT increases occurs because the MSRT serves as a constraint on the depth of the search space and as its value increases, less contrast sets will meet this threshold and therefore λ-COSINE and λ-GENCCS will take less time to complete the search than COSINE and GENCCS, respectively. Additionally, any ∞-contrast set discovered allows the entire subtree made up of its combine set to be pruned away. Thus the run time for λ-COSINE and λ-GENCCS when the MSRT is 0 is less than that of COSINE and GENCCS, respectively.

5.5.2 Interestingness

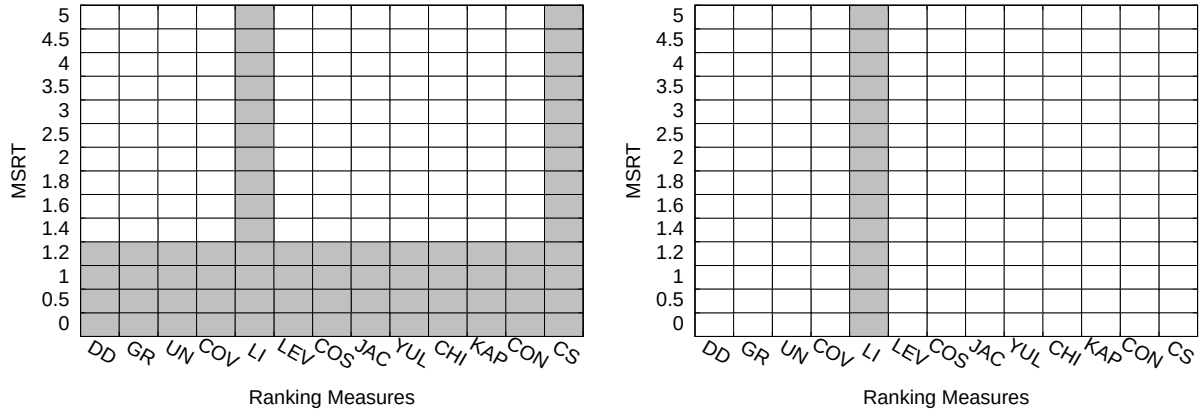
We compare the interestingness of the maximal valid contrast sets discovered by λ-COSINE and λ-GENCCS to COSINE and GENCCS, respectively, on the Census, Mushroom, Waveform, and Spambase datasets using thirteen ranking measures: distribution difference (DD), growth rate (GR), unusualness (UN), coverage (COV), lift (LI), leverage (LEV), change of support (COS), Jaccard coefficient (JAC), Yule's Q coefficient (YUL), ϕ-coefficient (CHI), κ-coefficient (KAP), conviction (CON), and collective strength (CS). Once again we vary the MSRT from 0 to 5. We use a standard T-test at a confidence

level of 95% to determine statistical significance.



(a) λ-COSINE vs. COSINE (b) λ-GENCCS vs. GENCCS

Figure 5.20: Significance Values vs. MSRT - Census

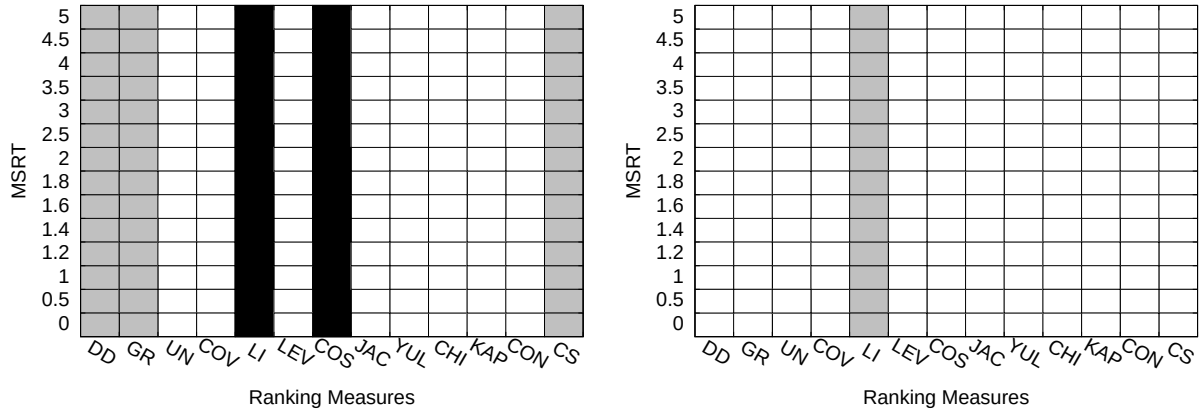


(a) λ-COSINE vs. COSINE (b) λ-GENCCS vs. GENCCS

Figure 5.21: Significance Values vs. MSRT - Mushroom

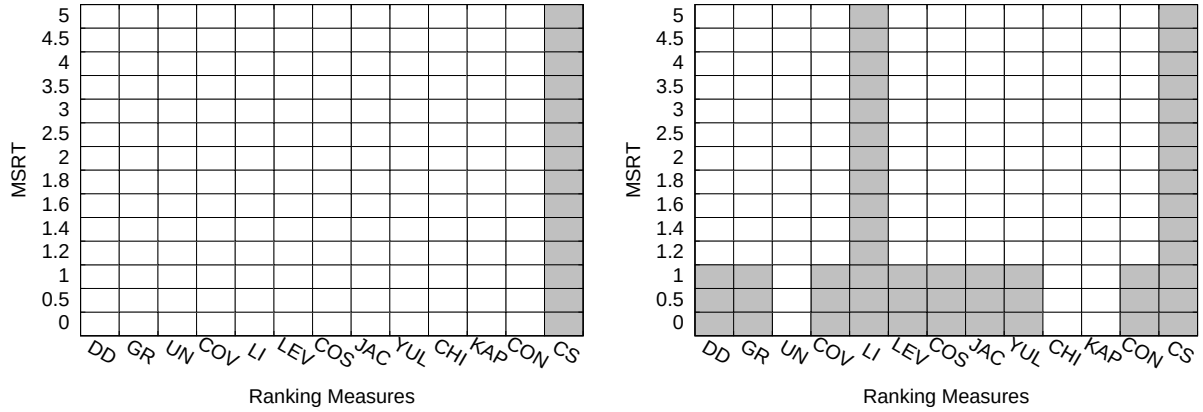
The results are shown in Figures 5.20 to 5.23. In Figures 5.20 to 5.23, the difference in the interestingness values obtained by λ-COSINE versus COSINE and λ-GENCCS versus GENCCS, respectively, are plotted against the MSRT for each of the four datasets. Once again, white, black, and grey indicate that the interestingness values are significantly higher, significantly lower, and not significantly different, respectively.

In Figure 5.20(a), the average interestingness of the results obtained by λ-COSINE is significantly higher than the results obtained by COSINE for MSRT values from 0 to 1.8



(a) λ-COSINE vs. COSINE (b) λ-GENCCS vs. GENCCS

Figure 5.22: Significance Values vs. MSRT - Waveform



(a) λ-COSINE vs. COSINE (b) λ-GENCCS vs. GENCCS

Figure 5.23: Significance Values vs. MSRT - Spambase

for all measures. For MSRT values above 1.8, the average interestingness of the results obtained by λ-COSINE is significantly higher than the results obtained by COSINE for all measures except DD, LI, and COS, where there is no significant difference.

In Figure 5.20(b), the average interestingness of the results obtained by λ-GENCCS is significantly higher than the results obtained by GENCCS for MSRT values from 0 to 1.2 for GR, UN, COV, JAC, CHI, KAP, and CON, significantly lower for LEV and YUL, and not significantly different for DD, LI, COS, and CS. For MSRT values above 1.2, the average interestingness of the results obtained by λ-GENCCS is significantly higher than

the results obtained by COSINE for all measures.

The average interestingness of the results obtained by λ-COSINE and λ-GENCCS versus the interestingness values obtained by COSINE and GENCCS, respectively for the Mushroom, Waveform, and Spambase datasets can be interpreted in a similar manner.

The average interestingness of the results obtained by λ-COSINE and λ-GENCCS is generally significantly higher than the interestingness values obtained by COSINE and GENCCS, respectively. These results demonstrate that contrast sets, both correlated and otherwise, where the maximum group support deviates the least from the minimum group support have lower interestingness values, thus removing them produces a subset of contrast sets that is on average more interesting.

5.5.3 Specificity

We compare the specificity of the maximal valid contrast sets discovered by λ-COSINE and λ-GENCCS to COSINE and GENCCS, respectively, within the context of average and maximum specificities as the MSRT is varied. For the average specificity, we use a standard T-test at a confidence level of 95% to determine whether there is a significant difference between the results obtained by λ-COSINE versus COSINE and λ-GENCCS vs GENCCS. Once again, we use the white, black, and grey three-step scale to indicate that the interestingness values obtained by λ-COSINE and λ-GENCCS are significantly higher, significantly lower, and not significantly different, respectively, compared to those obtained by COSINE and GENCCS.

The results are shown in Figure 5.24, where statistical significance of the average specificity values obtained by λ-COSINE versus COSINE and λ-GENCCS versus GENCCS, respectively, are plotted against the MSRT for each of the four datasets. The raw data from which this figure was generated can be found in Appendix C. In the graph, the statistical significance is described by a three-step scale: white, black, and grey indicate that the average specificity values obtained by λ-COSINE and λ-GENCCS are signifi-

cantly higher, significantly lower, and not significantly different, respectively, compared to those obtained by COSINE and GENCCS. For example, when the MSDT is 0, the average specificity of the results obtained by λ-COSINE and λ-GENCCS are not significantly different than those obtained by COSINE and GENCCS, respectively, for all datasets.

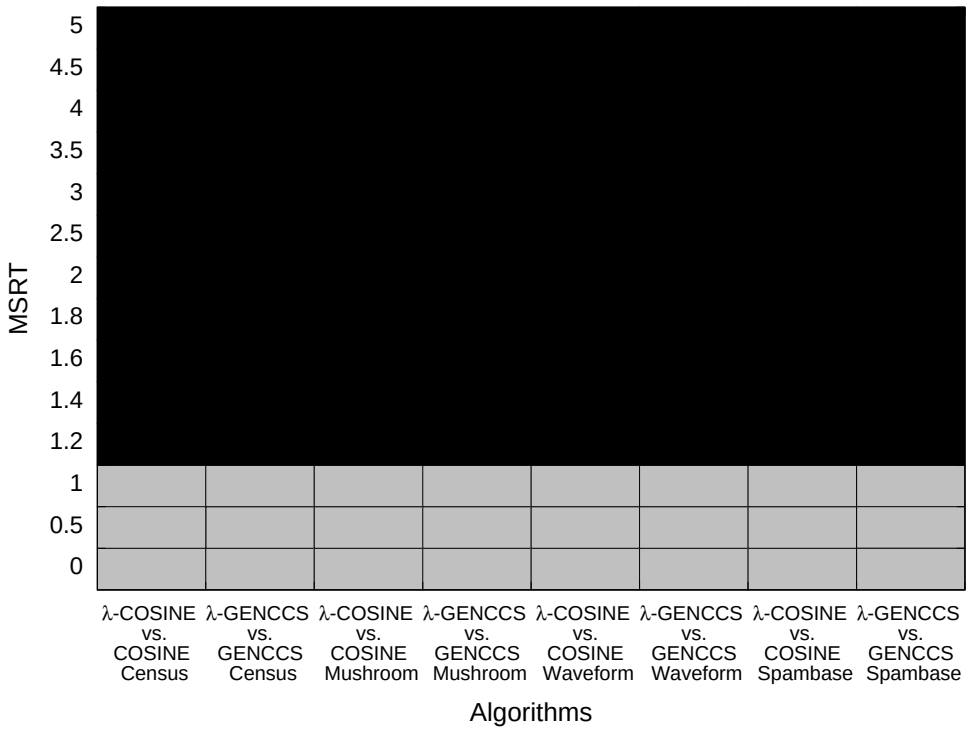
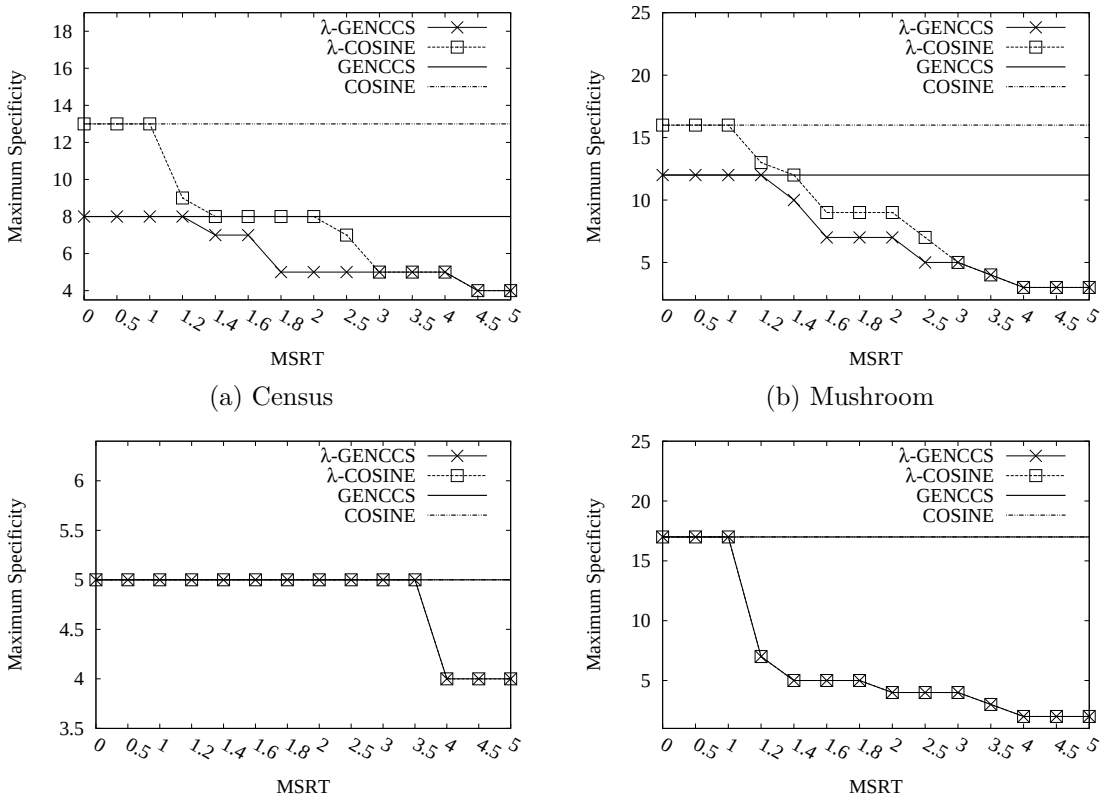


Figure 5.24: Comparison of Average Specificity

For all four datasets, the average specificity of the results obtained by λ-COSINE and λ-GENCCS are not significantly different than those obtained by COSINE and GENCCS, respectively, for MSRT values from 0 to 1. This means that increasing the MSRT from 0 through to 1 did not affect the average depth of the search space to which λ-COSINE and λ-GENCCS could discover maximal valid contrast sets, as compared to COSINE and GENCCS, respectively. For MSRT values from 1.2 to 5, the average specificity of the results obtained by λ-COSINE and λ-GENCCS are significantly lower than those obtained by COSINE and GENCCS, respectively, for all four datasets.

Figures 5.25(a) to 5.25(d) show the maximum specificity of the maximal valid contrast sets discovered by λ-COSINE and λ-GENCCS on the Census, Mushroom, Waveform, and Spambase datasets, respectively, as the MSRT is varied. For comparison, we also show the maximum specificity of the maximal valid contrast sets for COSINE, represented as a straight line, and the maximum specificity of the maximal correlated valid contrast sets for GENCCS, represented as a dashed line.



(a) Census (b) Mushroom

(c) Waveform (d) Spambase

Figure 5.25: Maximum Specificity vs. MSRT

For all four datasets, the maximum specificity of the maximal contrast sets discovered by λ-COSINE is equivalent to that of COSINE for MSRT values from 0 to 1. This means that increasing the MSRT from 0 through to 1 did not affect the maximum depth of the search space to which λ-COSINE could discover maximal valid contrast sets, as compared to COSINE. For MSRT values above 1, the maximum specificity of the maximal contrast sets discovered by λ-COSINE begins to decline until it becomes equivalent to that of

λ-GENCCS, for MSRT values above 2.5 on the Census and Mushroom datasets. For the Waveform and Spambase datasets, the maximum specificity of the maximal contrast sets discovered by λ-COSINE is equivalent to that of λ-GENCCS for all MSRT values.

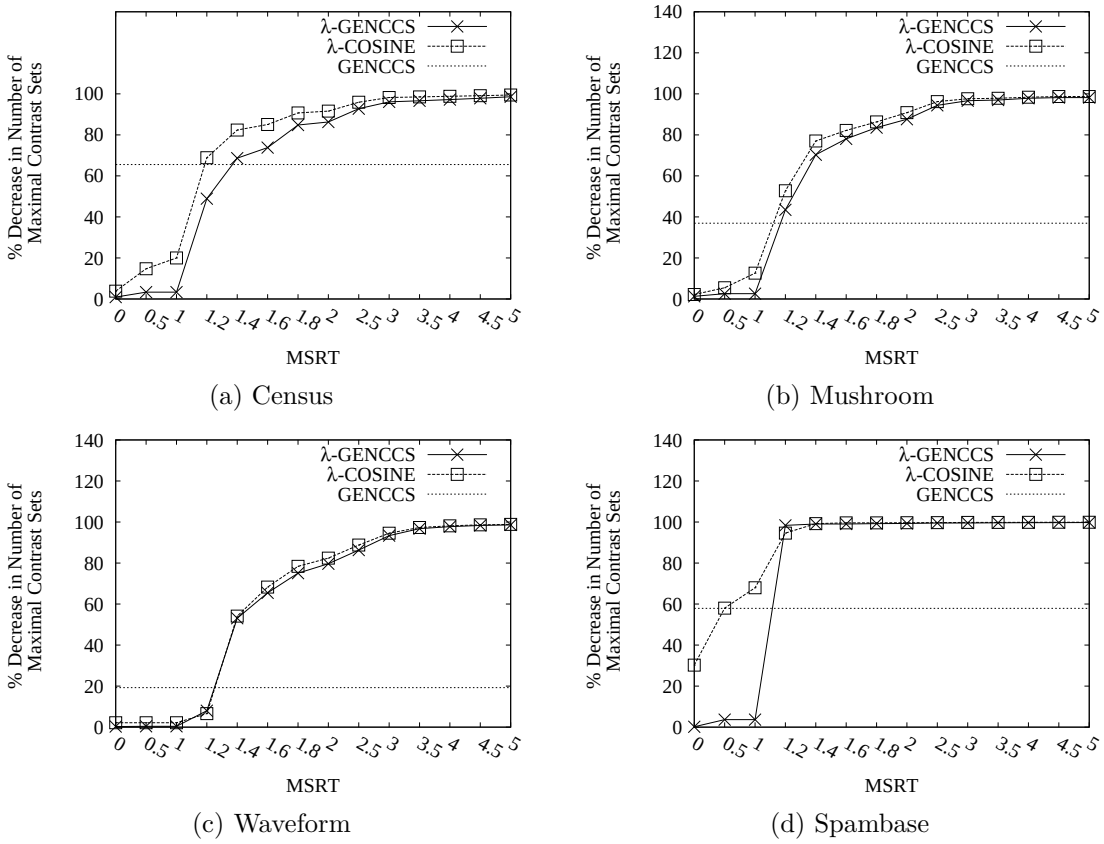
For the Census and Mushroom datasets, the maximum specificity of the maximal contrast sets discovered by λ-GENCCS is equivalent to that of GENCCS for MSRT values from 0 to 1.2, declining thereafter until it becomes equivalent to that of λ-COSINE. Here again, this means that increasing the MSRT from 0 through to 1.2 did not affect the maximum depth of the search space to which λ-GENCCS could discover maximal correlated valid contrast sets, as compared to GENCCS.

Our results demonstrate that, in general, using lower values for the MSRT does not reduce the maximal or average specificities of the maximal valid contrast sets and the maximal correlated valid contrast sets discovered by λ-COSINE and λ-GENCCS, respectively, from that of COSINE and GENCCS.

5.5.4 Number of Contrast Sets

We compare the percentage decrease in the number of maximal valid contrast sets and the number of maximal correlated valid contrast sets discovered by λ-COSINE and λ-GENCCS to that of those discovered by COSINE and GENCCS, respectively as the MSRT is varied. The results are shown in Figures 5.26(a) to 5.26(d) for the Census, Mushroom, Waveform, and Spambase datasets, respectively, as the MSRT is varied. For comparison, we also show the percentage decrease in the number of maximal contrast sets discovered by GENCCS to that of those discovered by COSINE, represented as a straight line.

The number of contrast sets discovered by λ-COSINE and λ-GENCCS sharply decreases from those discovered by COSINE, and GENCCS, respectively, as the MSRT increases from 1 to 1.2 on the Census, Mushroom, and Spambase datasets, and from 1.2 to 1.4 on the Waveform dataset. The sharpest decline occurs with λ-GENCCS over



(a) Census (b) Mushroom

(c) Waveform (d) Spambase

Figure 5.26: % Decrease in the Number of Maximal Contrast Sets vs. MSRT

GENCCS on the Spambase dataset, where the percentage decrease rises from 4% to 98% as the MSRT increases from 1 to 1.2.

The percentage decrease in the number of maximal contrast sets discovered by λ-COSINE over COSINE and λ-GENCCS over GENCCS, respectively, is higher than the percentage decrease in the number of maximal contrast sets discovered by GENCCS over COSINE for MSRT values above 1.2 on the Census and Waveform datasets, and for MSRT values above 1 on the Mushroom and Spambase datasets.

The number of contrast sets discovered can pose a challenge for an end-user in doing further analysis and making decisions from the analyses. Thus any reduction in the number of contrast sets discovered would produce a more manageable subset of results for an end-user to work with.

Our results demonstrate that increasing the MSRT significantly decreases the number

of contrasts sets discovered by both COSINE and GENCCS.

5.6 Evaluation of Discretize

We ran a series of discovery tasks to compare the results obtained by COSINE when the *Discretize* algorithm is used for discretization with those obtained when equal width and equal frequency binning are used for discretization. For each dataset, we use a minimum subset support ratio (MSSRT) of 0, a minimum support difference threshold (MSDT) of 0, and a significance level of 0.95 in statistical tests. We use a three-tier scale (i.e., H=High, M=Moderate, L=Low) for setting the minimum frequency threshold (MFT). Low, Moderate, and High correspond to MFT thresholds of 0.1%, 5%, and 20%, respectively. At the first level of the search space, we measure the number of candidate contrast sets that contain a discretized interval and the percentage of these contrast sets that satisfy our frequency criteria. We measure only at the first level of the search space, because those frequencies are solely dependent on the discretized interval, and not any combinations with other values. For the equal width and equal frequency binning, we also use a three-tier scale (i.e., H=High, M=Moderate, L=Low) for the number of bins to be created for each quantitative attribute. Low, Moderate, and High correspond to 10, 40, and 80 bins, respectively. We use the WEKA [44] implementation of the equal width and equal frequency binning algorithms. Table 5.12 shows the results obtained for the seven variants of the algorithms on the Census, Waveform, and Spambase datasets. There are no results for the Mushroom dataset because it has no quantitative attributes.

The *Number of Candidates* column describes the number of contrast sets that contain a discretized interval. The *Number of Successful Candidates* column describes the number of contrast sets that contain a discretized interval that satisfy the frequency criteria. The *Success Rate* column describes the percentage of candidate contrast sets that satisfy the frequency criteria. For example, Table 5.12 shows that for low, moderate, and high MFT

Table 5.12: Success Rate for Discretization Algorithms - Census

		MFT = L		*MFT = M*		*MFT = H*	
Algorithm	*Number of Candidates*	*Number of Successful Candidates*	*Success Rate*	*Number of Successful Candidates*	*Success Rate*	*Number of Successful Candidates*	*Success Rate*
Discretize	159	122	76.7%	47	29.6%	19	11.9%
EW-L	60	34	56.7%	17	28.3%	6	10%
EW-M	120	58	48.3%	22	18.3%	6	5%
EW-H	240	84	33.3%	19	7.9%	5	2.1%
EF-L	60	46	76.7%	27	45%	5	8.3%
EF-M	111	65	58.6%	24	21.6%	5	4.5%
EF-H	209	98	47.3%	9	4.3%	5	2.4%

thresholds, Discretize has a success rate of 76.7%, 29.6, and 11.9%, respectively.

Table 5.13: Success Rate for Discretization Algorithms - Waveform

		MFT = L		*MFT = M*		*MFT = H*	
Algorithm	*Number of Candidates*	*Number of Successful Candidates*	*Success Rate*	*Number of Successful Candidates*	*Success Rate*	*Number of Successful Candidates*	*Success Rate*
Discretize	268	115	42.9%	74	27.6%	39	14.6%
EW-L	400	125	31.3%	99	24.6%	31	7.8%
EW-M	800	224	28%	157	19.6%	0	0%
EW-H	1600	382	23.9%	87	5.4%	0	0%
EF-L	400	164	41%	64	16%	0	0%
EF-M	800	299	37.4%	169	21.1%	0	0%
EF-H	1600	523	32.7%	0	0%	0	0%

Table 5.14: Success Rate for Discretization Algorithms - Spambase

		MFT = L		*MFT = M*		*MFT = H*	
Algorithm	*Number of Candidates*	*Number of Successful Candidates*	*Success Rate*	*Number of Successful Candidates*	*Success Rate*	*Number of Successful Candidates*	*Success Rate*
Discretize	1351	1013	74.9%	328	24.2%	128	11.1%
EW-L	570	135	23.7%	66	11.6%	55	9.6%
EW-M	1140	205	18%	76	6.7%	54	4.7%
EW-H	2280	314	13.7%	81	3.4%	55	2.4%
EF-L	570	423	74.2%	113	19.8%	51	8.9%
EF-M	1140	661	58%	67	5.9%	51	4.5%
EF-H	2276	906	39.8%	54	2.4%	51	2.2%

For all three datasets, Discretize achieves the highest success rate for low, moderate, and high MFT thresholds. At low MFT thresholds, the difference in success rates is highest for EW-L and lowest for EF-L. The difference in success rates is important because it demonstrates that using the distribution of the values to create the width of the intervals produces more intervals which can be used in the discovery process than just arbitrarily picking a width. It also demonstrates that equally weighting the intervals can be a good substitute for Discretize.

Chapter 6

Case Studies

In this section we present two case studies demonstrating the use of COSINE and GENCCS in two real-world research projects. In the first case study, COSINE was used to evaluate factors that influenced ethical decision-making behaviour of physicians in a cross-cultural health care study. In the second case study, COSINE was used to study the impact of user created interfaces in the online game World of Warcraft. In Section 6.1, we present the first case study. In Section 6.2, we present the second case study.

6.1 Case Study I: Ethical Decision-Making

A cross cultural study of health decision-making was conducted to better understand factors that influence the ethical decision-making behaviour of physicians. There were four specific objectives. First, to determine if ethical ideologies differ among physicians in six culturally distinct nations. Second, to explore the extent to which the religiosity of physicians varies depending on culture. Third, to determine the extent to which ethical ideology and religiosity impact perceptions of ethical dilemmas involving seniors. Fourth, to study the behavioural intentions of physicians from disparate countries in response to ethical dilemmas.

A questionnaire was utilized to collect data from the 1255 physicians who volunteered to participate from 6 nations: Canada, Ireland, Japan, India, China, and Thailand. The collected data contains demographic information on each physician (such as gender, age, and religious affiliation), information on decision making issues influencing each physician (such as university training, family views, and peer attitudes),other general information relating to ethical ideology and religiosity, and information on how physicians respond to ethical dilemmas.

Initially, our analysis was focused on the first two of the four previously mentioned objectives. To identify the differences between the ethical ideologies of the physicians from the six nations, the six nations were used as the groups and the responses to the first five of the 20 statements in the Ethical Ideology section of the questionnaire and the first of 7 questions in the Decision Making Influences section of the questionnaire were used as the contrast sets. In the Ethical Ideology section, there were 9 responses to each statement listed on a continuum: completely disagree, largely disagree, moderately disagree, slightly disagree, neither agree or disagree, slightly agree, moderately agree, largely agree, and completely agree. In the Decision Making Influences section, there were 7 responses to each statement listed on a continuum such that 1 was not important and 7 was extremely important. Figures D.1 and D.2 in Appendix 5 shows 7 statements in the Decision Making Influences section and the 20 statements in the Ethical Ideology section of the questionnaire, respectively. Table 6.1 shows the data representation used.

Table 6.1: Data Representation

D1	E1	E2	E3	E4	E5
$\{1, 2, \ldots, 7\}$	$\{1, 2, \ldots, 7\}$	$\{1, 2, \ldots, 7\}$	$\{1, 2, \ldots, 7\}$	$\{1, 2, \ldots, 7\}$	$\{1, 2, \ldots, 7\}$

The columns labelled D1, E1, E2, E3, E4, and E5 describe attributes representing the first question in the Decision Making Influences section, and the first five statements in the Ethical Ideology section. For each attribute, the values that it can take on are also listed, and correspond to the specific question or statement on the questionnaire. For

example, the values 1 through 7 for D1 correspond to the 7 responses for the first question in the Decision Making Influences section of the questionnaire.

We ran COSINE at the 95% significance level with the following parameters, $\epsilon = 0$, $\sigma = 0.10$, $\kappa = 0$, $\omega = 0$, $m =$ Coverage, and $t =$ ALL. These parameters were chosen because the end-user wanted the discovered contrast sets to have high support.

COSINE generated a total of 67 valid contrast sets. We restrict our discussion to the five top-ranked contrast sets. The meaning of each contrast set is translated into an English representation for the reader's convenience.

Contrast Set 1: {E3:[9,9] & E4:[9,9] & E5:[9,9] (0.106)} → {Country:[Ireland, Ireland] (0.177) ≫ ¬Country:[Ireland, Ireland] (0.1)}, {¬Country:[Thailand, Thailand] (0.123) ≫ Country:[Thailand, Thailand] (0.06)}, {Country:[China, China] (0.287) ≫ ¬Country:[China, China] (0.08)}

English Representation: Physicians who completely agree that the existence of potential harm to others is always wrong, irrespective of the benefits to be gained and who completely agree that one should never psychologically or physically harm another person and who completely agree that one should not perform an action which might in any way threaten the dignity and welfare of another individual are more likely to come from China and Ireland by 72.1% and 43.5%, respectively, and less likely to come from Thailand by 51.2%.

Contrast Set 2: {E4:[9,9] & E5:[9,9](0.149)} → {Country:[India, India](0.361) ≫ ¬Country:[India, India] (0.118)}, {¬Country:[China, China] (0.319) ≫ Country:[China, China] (0.123)}

English Representation: Physicians who completely agree that one should never psychologically or physically harm another person and who completely agree that one should not perform an action which might in any way threaten the dignity and welfare of another individual are more likely to come from India by 67.3% and less likely to come from China by 61.5%.

Contrast Set 3: {E2:[8,8] (0.349)} → {¬Country:Japan, Japan] (0.467) ≫ Country:Japan, Japan] (0.202)}, {Country:[China, China] (0.301) ≫ ¬Country:[China, China] (0.127)}

English Representation: Physicians who largely agree that risks to another should never be tolerated, irrespective of how small the risks might be are less likely to come from Japan by 57% and more likely to come from China by 57.7%.

Contrast Set 4: {E4:[8,8] (0.390)} → {¬Country:[India, India] (0.161) ≫ Country:[India, India] (0.137)}

English Representation: Physicians who largely agree that one should never psychologically or physically harm another person are less likely to come from India by 64.9%.

Contrast Set 5: {D1:[7,7] & E2:[8,8](0.14)} → {¬Country:Japan, Japan](0.361) ≫ Country:Japan, Japan](0.071)}, {¬Country:Canada, Canada] (0.132) ≫ Country:[Canada, Canada] (0.077)}, {Country:[Thailand, Thailand] (0.197) ≫ ¬Country:[Thailand, Thailand] (0.077)}

English Representation: Physicians who largely agree that risks to another should never be tolerated, irrespective of how small the risks are and whose code of ethics is very important are less likely to come from Canada and Japan by 42% and 49.4% respectively and more likely to come from Thailand by 60.9%.

To identify the differences in religiosity between the physicians of the six nations, the six nations were used as the groups and the responses to the first 4 of the 11 questions in the Salience in Religious Commitment Scale section with the sixth of 7 questions in the Background Information section, and the first and fifth of 7 questions in the Decision Making Influences section the questionnaire were used as the contrast sets. In the Salience in Religious Commitment Scale section, there were 3 or 4 responses to the first three questions, and the user-provided response for the fourth. Figures D.1 and D.5 shows the relevant sections of the questionnaire. Table 6.2 shows the data representation used.

Table 6.2: Data Representation

B6	D1	D5	S1	S2	S3	S4
$\{1, 2, \ldots, 22\}$	$\{1, 2, \ldots, 7\}$	$\{1, 2, \ldots, 7\}$	$\{1, 2, 3\}$	$\{1, 2, 3, 4\}$	$\{1, 2, 3, 4\}$	$\{1, 2, \ldots, 43\}$

The columns labelled B6, D1, D5, S1, S2, S3, and S4 describe attributes representing the sixth question in the Background Information section, the first and fifth questions in the Decision Making Influences section, and the first four questions in the Salience in Religious Commitment Scale section.

We again ran COSINE at the 95% significance level with the following parameters, $\epsilon = 0$, $\sigma = 0.15$, $\kappa = 0$, $\omega = 0$, $m =$ Coverage, and $t =$ ALL. COSINE generated a total of 49 valid contrast sets. We again restrict our discussion to the five top-ranked contrast sets.

Contrast Set 1: {D1:[1,1] & S1:[2,3] (0.173)} → {Country:[Thailand, Thailand] (0.257) ≫ ¬Country:[Thailand, Thailand] (0.144)}

English Representation: Physicians who indicate that their professional code of ethics is important and whose religious faith is only of minor importance compared to other aspects of their life are more likely from Thailand by 43.9%.

Contrast Set 2: {S2:[1,1] & S3:[1,1] & S4:[0,0] (0.12)} → {Country:[China, China] (0.322) ≫ ¬Country:[China, China] (0.093)}, {¬Country:[Thailand, Thailand](0.156) ≫ Country:[Thailand, Thailand] (0.013)}

English Representation: Physicians who seldom base their decisions on their religious faith and who strongly disagree on whether their life would not have much meaning without their religious faith and who have not attended religious services during the past year are less likely to come from Thailand by 91.7% and more likely to come from China by 71.1%.

Contrast Set 3: {D1:[1,1] & S4:[0,0] (0.12)} → {Country:[China, China] (0.417) ≫ ¬Country:[China, China] (0.178)}, {¬Country:[Thailand, Thailand] (0.236) ≫ Country:[Thailand, Thailand] (0.116)}

English Representation: Physicians who indicate that their professional code of ethics is not important and who have not attended religious services during the past year are less likely to come from country Thailand 50.8% and more likely from China by 57.3%.

Contrast Set 4: {S1:[2,2] & S2:[1,1] & S4:[0,0] (0.11)} → {Country:Japan, Japan] (0.179) ≫ ¬Country:Japan, Japan] (0.1)}, {¬Country:[Canada, Canada] (0.04) ≫ Country:[Canada, Canada] (0.119)}, {¬Country:[Thailand, Thailand] (0.147) ≫ Country:[Thailand, Thailand] (0.003)}, {Country:[China, China] (0.417) ≫ ¬Country:[China, China] (0.07)}

English Representation: Physicians who indicate that their religious faith is only of minor importance for their life compared to other aspects of their life and who seldom make decisions based on religious faith and who have not attended religious services during the past year are less likely to come from Canada and Thailand by 66.4% and 98%, respectively, and more likely from countries Japan and China by 44.1% and 83.2%, respectively.

Contrast Set 5: {S2:[2,2] & S4:[43,43] & D5:[0,0] (0.105)} → {¬Country:[India, India] (0.117) ≫ Country:[India, India] (0.013)}, {¬Country:[Ireland, Ireland] (0.112) ≫ Country:[Ireland, Ireland] (0.04)}

English Representation: Physicians who sometimes base their decisions on their religious faith and who have not attended religious services during the past year and who did not specify a religion are less likely to come from countries Ireland and India by 64.3% and 88.9%, respectively..

We asked that our results be evaluated and feedback provided as an informal assessment of the usefulness of the results and general methodology to our end-user. The response was as follows: "..This is certainly an interesting and exhaustive procedure you have developed! The endorsements for the individual items by country are helpful in getting a general sense of our data. There are two aspects to the procedure that are particularly interesting. First, the specific group differences analysis (explained in per-

centages) is interesting and presents a way of looking at our data that we have used yet. Second and related, are the analyses where you examine different combinations of items. The combination analyses coupled with the group difference analyses can be very helpful in exploring and understanding our data....."

6.2 Case Study II: World of Warcraft User Interface Study

The World of Warcraft User Interface Study was conducted in order to examine the the World of Warcraft video game and the effect that user created interfaces have had on it and its community. The survey explored why users modify the interface, what benefits and drawbacks are present in user-created interface content, and why the community donates its time. World of Warcraft (WoW) is a very popular massively multiplayer online role-playing game (MMORPG). It has spawned a large user interface modding (modification) community that has created countless interface tweaks and mods. The mods range from simple button or interface location changes, to entire reworks of the interface with new themes and functionality added.

A questionnaire was utilized to collect data from 280 players who volunteered to participate. The collected data contains demographic information on each player (such as gender and age), information on game playing preferences, modification strategies, and developments methods. There were 141 questions, of which 106 had responses to choose from, and the remaining ones, respondents supplied their own.

Initially, our end-user was unsure of what specific groups or variables to be used in the analysis. We decided to perform two analyses. The first analysis was to identify the differences between the types of players. There were three types of players, beginner, intermediate, and advanced. Due to the small size of the dataset, and the small number of beginners, just 5, we combined the beginners and intermediates to get a group size of

94 which we call non-expert, with the advanced having 186. To identify the differences between the expert and non-expert groups, we used questions 6 through 11, each with a minimum of 2 and a maximum of 7 responses, as the contrast sets. Figure show the relevant sections of the questionnaire. Table 6.3 shows the data representation used.

Table 6.3: Data Representation

Q6	Q7	Q8	Q9	Q10	Q11
$\{1, 2, \ldots, 7\}$	$\{1, 2\}$	$\{1, 2, \ldots, 7\}$	$\{1, 2\}$	$\{1, 2\}$	$\{1, 2\}$

The columns labelled Q5, Q6, Q7, Q8, Q9, Q10, and Q11 describe questions 5, 6, 7, 8, 9, 10, and 11, respectively in the questionnaire. For each attribute, the values that it can take on are also listed, and correspond to the responses for that question on the questionnaire. For example, the values 1 through 7 for Q5 correspond to the 7 responses for Question 5 in the questionnaire.

We ran COSINE at the 90% significance level with the following parameters, $\epsilon = 0$, $\sigma = 0$, $\kappa = 0$, $\omega = 0$, $m =$ Coverage, and $t =$ ALL. We used these parameters because we wanted to obtain the largest subset of results possible which we hoped would guide the user in making further decisions about other questions to be explored.

COSINE generated a total of 3 valid contrast sets. The meaning of each contrast set is translated into an English representation for the reader's convenience.

Contrast Set 1: $\{$`Q6:[3,3]` $(0.236)\} \rightarrow \{$`Player:[Non-expert, Non-expert]` $(0.255) \gg$ `Player:[Expert, Expert]` $(0.113)\}$

English Representation: Players who play an average of 11 - 15 hours per week are more likely by 55.7% to be a non-expert player than an expert player.

Contrast Set 2: $\{$`Q8:[3,3]` $(0.161)\} \rightarrow \{$`Player:[Expert, Expert]` $(0.296) \gg$ `Player:[Non-expert, Non-expert]` $(0.117)\}$

English Representation: Players who have played massively multiplayer online (MMO)'s for 2 - 3 years are more likely by 60.5% to be an expert player than a non-expert player.

Contrast Set 3: $\{$`Q8:[3,3]` & `Q9:[2,2]` $(0.143)\} \rightarrow \{$`Player:[Expert, Expert]` $(0.285) \gg$

`Player:[Non-expert, Non-expert]` (0.106)}

English Representation: Players who have played MMO's for 2 - 3 years and who currently use interface mods are more likely by 62.8% to be an expert player than a non-expert player.

We asked that our results be evaluated and feedback provided as an informal assessment of the usefulness of the results and general methodology to our end-user. The response to these results was as follows: "....So for the first couple of results (having to do with expert versus non expert), that is not something i would have expected. I am curious why someone who play 11-15 hours a week (which seems like a lot) was more likely to not be an expert...."

Similarly, to determine the differences between players who had played a MMO game before WoW and those who had not. There were 116 players who had not, and 164 who had. These were our groups. We worked with only five questions, each with a minimum of 2 responses and a maximum of 7 responses.

The second analysis was to identify the differences between players who had played an MMO game before WoW and those who had not. There were three types of players, beginner, intermediate, and advanced. There were 116 players who had not, and 164 who had. These were our groups. We worked with questions 8, 9, 10, 11, and 12, each with a minimum of 2 and a maximum of 7 responses. Figure show the relevant sections of the questionnaire. Table 6.4 shows the data representation used.

Table 6.4: Data Representation

`Q8`	`Q9`	`Q10`	`Q11`	`Q12`
$\{1, 2, \ldots, 7\}$	$\{1, 2\}$	$\{1, 2\}$	$\{1, 2\}$	$\{1, 2, \ldots, 6\}$

The columns labelled `Q8`, `Q9`, `Q10`, `Q11`, and `Q12` describe questions 8, 9, 10, 11, and 12, respectively in the questionnaire. For each attribute, the values that it can take on are also listed, and correspond to the responses for that question on the questionnaire.

We again ran COSINE with the following parameters, $\epsilon = 0$, $\sigma = 0$, $\kappa = 0$, $\omega = 0$,

m = Coverage, and t = ALL.

COSINE generated a total of 24 valid contrast sets. Again, the meaning of each contrast set is translated into an English representation for the reader's convenience. Contrast sets 1 to 4 show the four that were top-ranked.

Contrast Set 1: {`Q11:[1,1]` (0.692)} → {`Player:[MMO_no, MMO_no]` (0.828) ≫ `Player:[MMO_yes, MMO_yes]` (0.592)}

English Representation: Players who have not used interface mods for a previous game are more likely by 39.9% to have not played a MMO game before WoW.

Contrast Set 2: {`Q11:[2,2]` (0.308)} → {`Player:[MMO_yes, MMO_yes]` (0.402) ≫ `Player:[MMO_no, MMO_no]` (0.356)}

English Representation: Players who have used interface mods for a previous game are more likely by 11.5% to have played a MMO game before WoW.

Contrast Set 3: {`Q8:[2,2]` & `Q11:[2,2]` (0.158)} → {`Player:[MMO_yes, MMO_yes]` (0.483) ≫ `Player:[MMO_no, MMO_no]` (0.061)}

English Representation: Players who have played MMO games for 1 - 2 years and have used interface mods for a previous game are more likely by 87.4% to have played a MMO game before WoW.

Contrast Set 4: {`Q8:[2,2]` & `Q10:[2,2]` (0.126)} → {`Player:[MMO_yes, MMO_yes]` (0.552) ≫ `Player:[MMO_no, MMO_no]` (0.061)}

English Representation: Players who have played MMO games for 1 - 2 years and have used interface mods in the past and have used interface mods for a previous game are more likely by 88.9% to have played a MMO game before WoW.

The response to these results was as follows: "....As for the last part "Players who have used interface mods for a previous game are more likely by 11.5% to have played a massively multiplayer online (MMO) game before WoW." I don't really know how relavent that is. It seems to me that this would be a pretty obvious conclusion because mmo's are pretty much the only games out there where you would experience a modified

interface, so to say that they are more likely to have played an previous mmo if they had also experienced ui addons before does not seem to hold a great deal of value....."

Chapter 7

Conclusion

The objective of this thesis was to develop and evaluate a technique for discovering contrast sets. Five goals were realized in obtaining this objective:

- Discretize, a novel, yet simple, discretization approach akin to equal-width binning which uses the mean and standard deviation to determine the intervals for continuous values, was introduced.

- COSINE, a vertical mining technique for discovering valid contrast sets on large datasets with categorical and quantitative attributes, where the association rules generated can have multiple discrete and continuous values in both the antecedent and consequent, was introduced and evaluated.

- GENCCS, a vertical mining technique, which builds on COSINE, that uses mutual information and all confidence to select only the attributes and attribute-values that are highly correlated for discovering contrast sets was introduced and evaluated.

- λ-contrast set and ∞-contrast set, novel contrast sets, were introduced.

- The use of objective measures as measures of interestingness for contrast sets generated from datasets was introduced and evaluated.

This chapter is organized as follows. In Section 7.1, we summarize the original contributions made by this thesis. In Section 7.2, we summarize the experimental results. In Section 7.3, we conclude with suggestions for future research.

7.1 Original Contributions

This thesis makes five primary contributions to data mining. First, we introduced and evaluated the Discretize algorithm. An efficient version of the algorithm is described which uses the mean and standard deviation to create intervals for continuous-valued attributes. Second, we introduced and evaluated the COSINE algorithm. An efficient version of the algorithm is described which traverses the search space generating all valid contrast sets and maximal valid contrast sets. Third, we introduced and evaluated the GENCCS algorithm. An efficient version of the algorithm is described which traverses the search of only attribute-interval pairs that are highly correlated, generating all correlated valid contrast sets, and maximal correlated valid contrast sets. Fourth, we introduced and evaluated 14 objective measures of interestingness for ranking the interestingness of the contrast sets generated using the COSINE and GENCCS algorithms. Finally, we introduced λ-contrast sets and ∞-contrast sets, and modified the COSINE and GENCCS algorithms to discover λ-contrast sets and correlated λ-contrast sets, respectively.

7.2 Summary

In Chapter 1, we introduced the problem of mining contrast sets from datasets. We also introduced the problem of ranking the interestingness of the contrast sets generated using objective measures of interestingness.

In Chapter 2, we presented a detailed survey of existing contrast set mining techniques and algorithms, and describe their most important features. We also present a survey of interestingness measures from four data mining tasks which are closely related to contrast

set mining.

In Chapter 3, we define valid contrast sets, λ-contrast sets, and ∞-contrast sets, and describe the COSINE and GENCCS algorithms. The operation of the COSINE algorithm was thoroughly described via detailed walkthroughs of three versions of the algorithm for discovering all valid contrast sets, maximal valid contrast sets, and λ valid contrast sets, respectively. Generally, the COSINE algorithm uses a vertical data format, diffsets, which not only allows simultaneous generation and support counting of the contrast sets, but also an efficient method of storing the instances in the dataset which can then be easily propagated throughout the search space. It also uses a backtracking search paradigm in order to enumerate all possible configurations of contrast sets in the search space. Both techniques allow for efficient pruning of the search space. GENCCS builds on COSINE in that instead of enumerating the entire search space in order to identify valid contrast sets, it utilizes mutual information and all confidence to select only the attributes and attribute-values that are highly correlated, resulting in a smaller search space.

In Chapter 4, we describe 14 objective measures drawn from various areas in data mining, that we considered for use as measures of interestingness. We introduced the distribution difference measure and the interestingness factor measure for ranking contrast sets. We adapted the growth rate and unusualness measures from emerging patterns and subgroup discovery, respectively, for contrast set mining. Finally, we also adapted coverage, lift, leverage, change of support, the Jaccard coefficient, ϕ coefficient, Yule's Q coefficient, conviction, and collective strength, 10 probability based measures, from association rule mining for contrast set mining. A detailed example of ranking a contrast set was presented for each measure.

In Chapter 5, the experimental results showed that COSINE is an effective technique for efficient generation of all valid contrast sets, and maximal valid contrast sets. It was shown that when the minimum support difference is not a constraint on the search space, COSINE discovered more interesting contrast sets more efficiently than STUCCO

and CIGAR. While COSINE discovered less maximal contrast sets than STUCCO and CIGAR, the average specificities of those discovered are significantly higher and their maximal specificities are at least that of CIGAR.

Experimental results also showed that GENCCS is an effective technique for efficient generation of all valid correlated contrast sets, and maximal correlated valid contrast sets. It was shown that when the attributes and attribute-values that are highly correlated are selected, GENCCS discovered more interesting contrast sets more efficiently than COSINE. While GENCCS discovered less maximal contrast sets than COSINE, the average and maximal specificities of those discovered are at least that of COSINE.

The experimental results show that the minimum support ratio threshold is an effective constraint on the search for valid contrast sets and correlated valid contrast sets as discovered by COSINE and GENCCS, respectively. While the number of contrast sets discovered by COSINE and GENCCS when the minimum support ratio threshold was used as a constraint is significantly reduced than when it is not, on average the contrast sets discovered were more interesting.

We have 13 interestingness measures for ranking contrast sets to choose from (distribution difference, growth rate, unusualness, coverage, lift, leverage, change of support, the Jaccard coefficient, ϕ coefficient, Yule's Q coefficient, conviction, and collective strength). Since each measure determines interestingness differently, the choice of which measure to use may make a difference. Unfortunately, determining the most suitable measure is a matter of trial-and-error and may be based on the needs of the end-user. We recommend using multiple measures to rank the contrast sets then using the interestingness factor to get an overall rank order for the contrast sets. The interestingness factor allows the measures to be weighted equally or otherwise.

Finally, given the experimental results, we have two algorithms for mining contrast sets to choose from (COSINE and GENCCS) and a constraint which can be applied to either to choose from (MSRT). Since each algorithm discovers a different subset of

contrast sets, the choice of which algorithm to use may make a difference. We recommend using COSINE to find the largest subset of contrast sets where every attribute in the dataset must be used. These contrast sets will have the highest specificity, average and maximal. We recommend using GENCCS to find a smaller subset of contrast sets where the attributes and attribute values are targeted based on their correlation with each other. These contrast sets will have an average and maximal specificities that is comparable to that of obtained by COSINE in some cases, and on average will be more interesting than that of those obtained by COSINE. We recommend using the MSRT as a constraint on either COSINE or GENCCS to find the smallest subset of contrast sets either from all attributes if applied to COSINE or a smaller subset if applied to GENCCS. These contrast sets will have average and maximal specificities significantly lower than that of COSINE or GENCCS, but on average will be more interesting than that of those obtained by COSINE and GENCCS.

7.3 Areas for Future Research

Considerable research remains to be done in the discovery of contrast sets and in the application of interestingness measures to the problem of ranking the interestingness of the contrast sets discovered. We see five major areas for future research.

First, other objective interestingness measures need to be evaluated to determine their suitability for ranking the interestingness of contrast sets generated from datasets. There are many possible candidates in the literature. Possible measures include *Laplace Correction* [51], *Gini Index* [38], *Goodman and Kruskal* [40], and *Information Gain* [67].

Second, subjective measures which are based on user beliefs or biases regarding relationships in the data [39], need also be evaluated to determine their suitability for ranking the interestingness of contrast sets generated from datasets. Possible measures include *Unexpectedness* [111] [122], and *Novelty* [81] [82]. Consider unexpectedness which

is related to a belief system. Given evidence E (contrast sets) the degree of belief in a contrast set X is given by:

$$P(X|E,\mathcal{D}) = \frac{P(E|X,\mathcal{D})\ P(X|\mathcal{D})}{P(E|X,\mathcal{D})\ P(X|\mathcal{D}) + P(E|\neg X,\mathcal{D})\ P(\neg X|\mathcal{D})},$$

where $\mathcal{D}$ is the context representing the previous evidence supporting X.

The interestingness measure for a contrast set Y, relative to a belief system B, is defined as the relative difference by the prior and posterior probabilities, and is given by

$$I(Y,B) = \sum_{X \in B} \frac{P(Y|X,\mathcal{D}) - P(X|\mathcal{D})}{P(X|\mathcal{D})}.$$

Consider the contrast set $X = \{A : [0,0], B : [0,0]\ 0.3\} \rightarrow \{E : [0,0]\ 0.1 \gg \neg E : [0,0]\ 0.2\}$. Assume that the user's specifies the degree of belief in $P(X|\mathcal{D})$ as 0.3 based on a dataset where 3 out of 10 transactions support this belief. Suppose a valid contrast set, $Y = \{A : [0,0]\ 0.5\} \rightarrow \{Y : [0,0]\ 0.2 \gg \neg Y : [0,0]\ 0.3\}$, is discovered. Also assume that $P(Y|X,\mathcal{D})$ is 1 and the confidence of $P(Y|\neg X,\mathcal{D})$ is 0.33, then $P(X|E,\mathcal{D})$ is calculated as $\frac{1\times 0.3}{1\times 0.3 + 0.33\times 0.7} = 0.56$ and the unexpectedness for Y is calculated as $\frac{|0.56-0.3|}{0.3} = 0.86$.

Third, other methods for describing contrast sets to end users need to be developed and evaluated. While there has been no published efforts specifically for contrast sets, several methods for subgroup visualization have been developed by [8], [37], and [62]. Based on the compatibility of contrast set mining and subgroup discovery established in [89], visualization methods developed for subgroup discovery can be adapted to contrast set mining. One simple method is to use a bar chart to visualize the contrast sets.

Fourth, it would be interesting to develop and evaluate parallel versions of COSINE and GENCCS. From the first level of the search space, each element in the combine set is valid, and each combination of the elements with the prefix is independent of each other. Consider the first level of a search tree with a prefix set $\{\texttt{A} : [0,0]\}$ and associated combine set $\{\texttt{B} : [0,0]$, $\texttt{B} : [1,1]$, $\texttt{C} : [0,0]$, $\texttt{C} : [1,1]$, $\texttt{D} : [20, 22.9)$, $\texttt{D} : [22.9, 26]\}$. In the search for

all valid contrast sets, all combinations of the prefix with each element in the combine set is independent of each other and each resulting sub-tree can further be traversed independently of the other sub-trees. Thus this traversals can be done in parallel on the same processor or distributed to multiple processors. While parallel algorithms for contrast set mining are not found in the literature, there is a plethora of research done in the closely related area of the parallel and distributed mining of frequent itemsets and maximal frequent itemsets [22] [91] [125] [141].

Finally, other measures need to be evaluated to determine their suitability for measuring the relationships between attributes and attribute-values in the dataset. These relationships can be utilized to generate a smaller search space for discovering valid contrast sets. There is certainly no shortage of possible candidates in the literature. Possible measures include *Pearson Correlation* [103], *Brownian Covariance* [14], *Odds Ratio* [87], *Information Gain* [67], *Spearman's rho* [21], and *Kendall's Tau* [58].

Bibliography

[1] Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. Depth First Generation of Long Patterns. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '00, pages 108–118, New York, NY, USA, 2000. ACM. ISBN 1-58113-233-6. doi: http://doi.acm.org/10.1145/347090.347114.

[2] Charu C. Aggarwal and Philip S. Yu. A New Framework for Itemset Generation. In *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 18–24, New York, NY, USA, 1998. ACM. ISBN 0-89791-996-3. doi: http://doi.acm.org/10.1145/275487.275490.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1-55860-153-8.

[4] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining Association Rules between Sets of Items in Large Databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, New York, NY, USA, 1993. ACM. ISBN 0-89791-592-5. doi: http://doi.acm.org/10.1145/170035.170072.

[5] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 1st edition, 1990.

[6] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[7] Martin Atzmller and Frank Puppe. SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery. In Johannes Frnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *PKDD*, volume 4213 of *Lecture Notes in Computer Science*, pages 6–17. Springer, 2006. ISBN 3-540-45374-1. doi: 10.1007/118716376.

[8] Martin Atzmueller and Frank Puppe. Semi-Automatic Visual Subgroup Mining using VIKAMINE. *J. UCS*, 11(11):1752–1765, nov 2005.

[9] Martin Atzmueller, Frank Puppe, and Hans-Peter Buscher. Profiling Examiners using Intelligent Subgroup Mining. In *Proc. 10th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2005)*, pages 46–51, Aberdeen, Scotland, 2005.

[10] Yonatan Aumann and Yehuda Lindell. A Statistical Theory for Quantitative Association Rules. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 261–270, New York, NY, USA, 1999. ACM. ISBN 1-58113-143-7. doi: 10.1145/312129.312243.

[11] Stephen D. Bay and Michael J. Pazzani. Detecting Change in Categorical Data: Mining Contrast Sets. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 302–306, New York, NY, USA, 1999. ACM. ISBN 1-58113-143-7. doi: http://doi.acm.org/10.1145/312129.312263.

[12] Stephen D. Bay and Michael J. Pazzani. Detecting Group Differences: Mining

Contrast Sets. *Data Min. Knowl. Discov.*, 5(3):213–246, 2001. ISSN 1384-5810. doi: http://dx.doi.org/10.1023/A:1011429418057.

[13] Roberto J. Bayardo. Efficiently Mining Long Patterns from Databases. In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 85–93, New York, NY, USA, 1998. ACM. ISBN 0-89791-995-5. doi: http://doi.acm.org/10.1145/276304.276313.

[14] Peter J. Bickel and Ying Xu. Discussion of: Brownian Distance Covariance. *Annals of Applied Statistics*, 3:1266–1269, 2009. doi: 10.1214/09-AOAS312A.

[15] Diane M. Billings, Diane J. Skiba, and Helen R. Connors. Best Practices in Web-based Courses: Generational Differences Across Undergraduate and Graduate Nursing Students. *Journal of Professional Nursing*, 21(2):126 – 133, 2005. ISSN 8755-7223. doi: 10.1016/j.profnurs.2005.01.002.

[16] C. Borgelt. Efficient Implementations of Apriori and Eclat. In *FIMI*, volume 90 of *CEUR Workshop Proc.* CEUR-WS.org, 2003.

[17] Anne-Laure Boulesteix, Gerhard Tutz, and Korbinian Strimmer. A CART-based Approach to Discover Emerging Patterns in Microarray Data. *Bioinformatics*, 19 (18):2465–2472, 2003. doi: 10.1093/bioinformatics/btg361.

[18] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255–264, Tucson, Arizona, USA, May 1997.

[19] Doug Burdick, Manuel Calimlim, Jason Flannick, Johannes Gehrke, and Tomi Yiu. MAFIA: A Maximal Frequent Itemset Algorithm. *IEEE Trans. on Knowl. and Data Eng.*, 17:1490–1504, November 2005. ISSN 1041-4347. doi: http://dx.doi.org/10.1109/TKDE.2005.183.

[20] Martha Carr, Hillary Hettinger Steiner, Brandon Kyser, and Barry Biddlecomb. A Comparison of Predictors of Early Emerging Gender Differences in Mathematics Competency. *Learning and Individual Differences*, 18(1):61 – 75, 2008. ISSN 1041-6080. doi: 10.1016/j.lindif.2007.04.005.

[21] John C. Caruso and Norman Cliff. Empirical Size, Coverage, and Power of Confidence Intervals for Spearman's Rho. *Educational and Psychological Measurement*, 57(4):637–654, 1997. doi: 10.1177/0013164497057004009.

[22] S.M. Chung and Congnan Luo. Parallel Mining of Maximal Frequent Itemsets from Databases. In *Tools with Artificial Intelligence, 2003. Proceedings. 15th IEEE International Conference on*, pages 134 – 139, nov. 2003. doi: 10.1109/TAI.2003.1250181.

[23] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

[24] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 2006. ISBN 0-471-24195-9.

[25] Harald Cramer. *Mathematical Methods of Statistics*. Princeton University Press, Princeton NJ, 1946.

[26] Ruth Defrin, Libby Shramm, and Ilana Eli. Gender Role Expectations of Pain is Associated with Pain Tolerance Limit but not with Pain Threshold. *PAIN*, 145 (1-2):230 – 236, 2009. ISSN 0304-3959. doi: 10.1016/j.pain.2009.06.028.

[27] Guozhu Dong and Jinyan Li. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 43–52, New York, NY, USA, 1999. ACM. ISBN 1-58113-143-7. doi: http://doi.acm.org/10.1145/312129.312191.

[28] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. In *ICML*, pages 194–202, 1995.

[29] B Dunkel and N Soparkar. Data Organization and Access for Efficient Data Mining. In *Proceedings of the 15th International Conference on Data Engineering*, ICDE '99, pages 522–, Washington, DC, USA, 1999. IEEE Computer Society. ISBN 0-7695-0071-4.

[30] Hongjian Fan and Kotagiri Ramamohanarao. Efficiently Mining Interesting Emerging Patterns. In *Advances in Web-Age Information Management*, pages 189–201, 2003.

[31] Hongjian Fan, Ming Fan, Kotagiri Ramamohanarao, and Mengxu Liu. Further Improving Emerging Pattern Based Classifiers Via Bagging. In *Proceedings of the 10th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining*, PAKDD'06, pages 91–96, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-33206-5, 978-3-540-33206-0.

[32] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery: an overview. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in knowledge discovery and data mining*, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996. ISBN 0-262-56097-6.

[33] Flavia Franconi, Sandra Brunelleschi, Luca Steardo, and Vincenzo Cuomo. Gender Differences in Drug Responses. *Pharmacological Research*, 55(2):81 – 95, 2007. ISSN 1043-6618. doi: 10.1016/j.phrs.2006.11.001.

[34] William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus. Knowledge Discovery in Databases: An Overview. In *Knowledge Discovery in Databases*, pages 1–30. AAAI/MIT Press, 1991. ISBN 0-262-62080-4.

[35] Takeshi Fukuda, Yasuhido Morimoto, Shinichi Morishita, and Takeshi Tokuyama. Mining Optimized Association Rules for Numeric Attributes. In *PODS '96: Proceedings of the fifteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 182–191, New York, NY, USA, 1996. ACM. ISBN 0-89791-781-2. doi: http://doi.acm.org/10.1145/237661.237708.

[36] Dragan Gamberger and Nada Lavrac. Expert-Guided Subgroup Discovery: Methodology and Application. *J. Artif. Intell. Res. (JAIR)*, 17:501–527, 2002.

[37] Dragan Gamberger, Nada Lavrač, and Dietrich Wettschereck. Subgroup Visualization: A Method and Application to Population Screening. In Marko Bohanec, Branko Kavšek, Nada Lavrač, and Dunja Mladenic, editors, *In Proceedings of the International Workshop on intelligent Data Analysis in Medicine and Pharmacology, IDAMAP*, pages 35–40. Helsinki University Printing House, August 2002. ISBN 952-10-0634-X.

[38] Joseph L. Gastwirth. The Estimation of the Lorenz Curve and Gini Index. *Review of Economics and Statistics*, 54(3):306–316, 1972. ISSN 00346535. doi: 10.2307/1937992.

[39] Liqiang Geng and Howard J. Hamilton. Interestingness Measures for Data Mining: A Survey. *ACM Comput. Surv.*, 38(3):9, 2006. ISSN 0360-0300. doi: http://doi.acm.org/10.1145/1132960.1132963.

[40] Leo A Goodman and William H Kruskal. Measures of Association for Cross Classifications III: Approximate Sampling Theory. *Journal of the American Statistical Association*, 58(302):310–364, 1963. URL `http://www.jstor.org/stable/2283271`.

[41] Karam Gouda and Mohammed J. Zaki. GenMax: An Efficient Algorithm for Mining Maximal Frequent Itemsets. *Data Min. Knowl. Discov.*, 11(3):223–242, 2005. ISSN 1384-5810. doi: http://dx.doi.org/10.1007/s10618-005-0002-x.

[42] Sylvie Guillaume. Discovery of Ordinal Association Rules. In *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '02, pages 322–327, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-43704-5.

[43] Dimitrios Gunopulos, Heikki Mannila, and Sanjeev Saluja. Discovering All Most Specific Sentences by Randomized Algorithms. In *Proceedings of the 6th International Conference on Database Theory*, pages 215–229, London, UK, 1997. Springer-Verlag. ISBN 3-540-62222-5.

[44] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: AnUupdate. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278.

[45] Jiawei Han. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005. ISBN 1558609016.

[46] Jiawei Han, Jian Pei, and Yiwen Yin. Mining Frequent Patterns without Candidate Generation. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, SIGMOD '00, pages 1–12, New York, NY, USA, 2000. ACM. ISBN 1-58113-217-4. doi: http://doi.acm.org/10.1145/342009.335372.

[47] R.J. Hilderman and T. Peckham. A Statistically Sound Alternative Approach to Mining Contrast Sets. *Proceedings of the 4th Australasian Data Mining Conference (AusDM'05)*, pages 157–172, Dec. 2005.

[48] Robert Hilderman and Howard Hamilton. Heuristic Measures of Interestingness. In Jan Zytkow and Jan Rauch, editors, *Principles of Data Mining and Knowledge Discovery*, volume 1704 of *Lecture Notes in Computer Science*, pages 232–241. Springer Berlin / Heidelberg, 1999. ISBN 978-3-540-66490-1.

[49] Laura C. Holdcraft and William G. Iacono. Cross-generational Effects on Gender Differences in Psychoactive Drug Abuse and Dependence. *Drug and Alcohol Dependence*, 74(2):147 – 158, 2004. ISSN 0376-8716. doi: 10.1016/j.drugalcdep.2003.11.016.

[50] S Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979. ISSN 0303-6898.

[51] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, April 2003. ISBN 0521592712.

[52] Yanming Jiang, Kazim Sheikh, and Claudia Bullock. Is There a Sex or Race Difference in Stroke Mortality? *Journal of Stroke and Cerebrovascular Diseases*, 15(5): 179 – 186, 2006. ISSN 1052-3057. doi: 10.1016/j.jstrokecerebrovasdis.2006.05.007.

[53] Hui jing Huang, Yongsong Qin, Xiaofeng Zhu, Jilian Zhang, and Shichao Zhang. Difference Detection Between Two Contrast Sets. *Data Warehousing and Knowledge Discovery*, pages 481–490, 2006.

[54] Branko Kavsek and Nada Lavrac. APRIORI-SD: Adapting Association Rule Learning to Subgroup Discovery. *Applied Artificial Intelligence*, 20(7):543–583, 2006.

[55] Yiping Ke, James Cheng, and Wilfred Ng. An Information-theoretic Approach to Quantitative Association Rule Mining. *Knowl. Inf. Syst.*, 16(2):213–244, 2008. ISSN 0219-1377. doi: http://dx.doi.org/10.1007/s10115-007-0104-4.

[56] Yiping Ke, James Cheng, and Wilfred Ng. Correlated Pattern Mining in Quantitative Databases. *ACM Trans. Database Syst.*, 33:14:1–14:45, September 2008. ISSN 0362-5915. doi: http://doi.acm.org/10.1145/1386118.1386120.

[57] Micere Keels. Ethnic group differences in early head start parents' parenting beliefs and practices and links to children's early cognitive development. *Early*

Childhood Research Quarterly, 24(4):381 – 397, 2009. ISSN 0885-2006. doi: 10.1016/j.ecresq.2009.08.002.

[58] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):pp. 81–93, 1938. ISSN 00063444. URL `http://www.jstor.org/stable/2332226`.

[59] Sunghan Kim, Gwenden L Dueker, Lynn Hasher, and David Goldstein. Children's time of day preference: age, gender and ethnic differences. *Personality and Individual Differences*, 33(7):1083 – 1090, 2002. ISSN 0191-8869. doi: 10.1016/S0191-8869(01)00214-8.

[60] Won-Young Kim, Young-Koo Lee, and Jiawei Han. CCMine: Efficient Mining of Confidence-Closed Correlated Patterns. In Honghua Dai, Ramakrishnan Srikant, and Chengqi Zhang, editors, *Advances in Knowledge Discovery and Data Mining*, volume 3056 of *Lecture Notes in Computer Science*, pages 569–579. Springer Berlin Heidelberg, 2004. ISBN 978-3-540-22064-0.

[61] Willi Klösgen. Explora: a multipattern and multistrategy discovery assistant. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 249–271. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996. ISBN 0-262-56097-6.

[62] Willi Klösgen and Stephan R. W. Lauer. Presentation and Visualization: Visualization of Data Mining Results. In Willi Klösgen and Jan M. Zytkow, editors, *Handbook of Data Mining and Knowledge Discovery*, pages 509–515. Oxford University Press, Inc., New York, NY, USA, 2002. ISBN 0-19-511831-6.

[63] Willi Klösgen, Michael May, and Jim Petch. Mining Census Data for SpatialEeffects on Mortality. *Intell. Data Anal.*, 7(6):521–540, December 2003. ISSN 1088-467X.

[64] Petra Kralj, N. Lavrac, D. Gamberger, and A. Krstacic. Supporting Factors to Improve the Explanatory Potential of Contrast Set Mining: Analyzing Brain Ischaemia Data. In Tomaz Jarm, Peter Kramar, and Anze Zupanic, editors, *11th Mediterranean Conference on Medical and Biomedical Engineering and Computing 2007*, volume 16 of *IFMBE Proceedings*, pages 157–161. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-73043-9.

[65] Petra Kralj, Nada Lavrac, Dragan Gamberger, and Antonija Krstacic. Contrast Set Mining for Distinguishing Between Similar Diseases. In Riccardo Bellazzi, Ameen Abu-Hanna, and Jim Hunter, editors, *AIME*, volume 4594 of *Lecture Notes in Computer Science*, pages 109–118. Springer, 2007. ISBN 978-3-540-73598-4.

[66] Petra Kralj Novak, Nada Lavrač, Dragan Gamberger, and Antonija Krstačić. CSM-SD: Methodology for Contrast Set Mining through Subgroup Discovery. *J. of Biomedical Informatics*, 42(1):113–122, 2009. ISSN 1532-0464. doi: http://dx.doi.org/10.1016/j.jbi.2008.08.007.

[67] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):pp. 79–86, 1951. ISSN 00034851. doi: doi:10.1214/aoms/1177729694.

[68] Tarald O. Kvalseth. Entropy and Correlation: Some Comments. *Systems, Man and Cybernetics, IEEE Transactions on*, 17(3):517 –519, May 1987. ISSN 0018-9472. doi: 10.1109/TSMC.1987.4309069.

[69] Nada Lavrac, Peter A. Flach, and Blaz Zupan. Rule Evaluation Measures: A Unifying View. In *Proceedings of the 9th International Workshop on Inductive Logic Programming*, ILP '99, pages 174–185, London, UK, UK, 1999. Springer-Verlag. ISBN 3-540-66109-3.

[70] Nada Lavrac, Branko Kavsek, Peter A. Flach, and Ljupco Todorovski. Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

[71] Y.-K. Lee, W.-Y. Kim, D. Cai, and J. Han. CoMine: Efficient Mining of Correlated Patterns. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 581 – 584, nov. 2003. doi: 10.1109/ICDM.2003.1250982.

[72] Jinyan Li, Guozhu Dong, and Kotagiri Ramamohanarao. Instance-Based Classification by Emerging Patterns. In *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 191–200, London, UK, 2000. Springer-Verlag. ISBN 3-540-41066-X.

[73] Jinyan Li, Guozhu Dong, and Kotagiri Ramamohanarao. Making Use of the Most Expressive Jumping Emerging Patterns for Classification. *Knowl. Inf. Syst.*, 3(2): 1–29, 2001. ISSN 0219-1377.

[74] Jinyan Li, Huiqing Liu, James R. Downing, Allen Eng-Juh Yeoh, and Limsoon Wong. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. *Bioinformatics*, 19(1):71–78, 2003. doi: 10.1093/bioinformatics/19.1.71.

[75] Jinyan Li, Huiqing Liu, See-Kiong Ng, and Limsoon Wong. Discovery of significant rules for classifying cancer diagnosis data. *Bioinformatics*, 19(2):93–102, 2003. doi: 10.1093/bioinformatics/btg1066.

[76] Jersey Liang, Joan M. Bennett, Hidehiro Sugisawa, Erika Kobayashi, and Taro Fukaya. Gender differences in old age mortality: Roles of health behavior and baseline health status. *Journal of Clinical Epidemiology*, 56(6):572–582, 2003. ISSN 0895-4356. doi: 10.1016/S0895-4356(03)00060-X.

[77] Dao-I Lin and Zvi M. Kedem. Pincer Search: A New Algorithm for Discovering the Maximum Frequent Set. In *Proceedings of the 6th International Conference*

on Extending Database Technology: Advances in Database Technology, EDBT '98, pages 105–119, London, UK, UK, 1998. Springer-Verlag. ISBN 3-540-64264-1.

[78] Jessica Lin and Eamonn Keogh. Group SAX: Extending the Notion of Contrast Sets to Time Series and Multimedia Data. In *Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases*, PKDD'06, pages 284–296, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-45374-1, 978-3-540-45374-1.

[79] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In *DMKD '03: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11, New York, NY, USA, 2003. ACM. doi: http://doi.acm.org/10.1145/882082.882086.

[80] Jun-Lin Lin and M.H. Dunham. Mining Association Rules: Anti-Skew Algorithms. In *Data Engineering, 1998. Proceedings., 14th International Conference on*, pages 486 –493, feb 1998. doi: 10.1109/ICDE.1998.655811.

[81] Bing Liu, Wynne Hsu, and Shu Chen. Using General Impressions to Analyze Discovered Classification Rules. In David Heckerman, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy, editors, *AAAI*, pages 31–36. AAAI Press, 1997.

[82] Bing Liu, Wynne Hsu, Lai-Fun Mun, and Hing-Yan Lee. Finding Interesting Patterns Using User Expectations. *IEEE Trans. on Knowl. and Data Eng.*, 11(6): 817–832, November 1999. ISSN 1041-4347. doi: 10.1109/69.824588.

[83] Huan Liu, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash. Discretization: An Enabling Technique. *Data Min. Knowl. Discov.*, 6(4):393–423, 2002. ISSN 1384-5810. doi: http://dx.doi.org/10.1023/A:1016304305535.

[84] Sheng Ma and Joseph L. Hellerstein. Mining Mutually Dependent Patterns. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 409–416, Washington, DC, USA, 2001. IEEE Computer Society. ISBN 0-7695-1119-8.

[85] R. J. Miller and Y. Yang. Association Rules over Interval Data. In *SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 452–461, New York, NY, USA, 1997. ACM. ISBN 0-89791-911-4. doi: http://doi.acm.org/10.1145/253260.253361.

[86] B. Minaei-Bidgoli, Pang-Ning Tan, and W.F. Punch. Mining Interesting Contrast Rules for a Web-based Educational System. *Machine Learning and Applications, 2004. Proceedings. 2004 International Conference on*, pages 320–327, December, 2004.

[87] Frederick Mosteller. Association and Estimation in Contingency Tables. *Journal of the American Statistical Association*, 63(321):1–28, 1968. ISSN 01621459. URL `http://www.jstor.org/stable/2283825`.

[88] Zohreh Nazeri, Daniel Barbara, Kenneth Jong, George Donohue, and Lance Sherry. Contrast-Set Mining of Aircraft Accidents and Incidents. In *Proceedings of the 8th industrial conference on Advances in Data Mining: Medical Applications, E-Commerce, Marketing, and Theoretical Aspects*, ICDM '08, pages 313–322, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-70717-2.

[89] Petra Kralj Novak, Nada Lavrač, and Geoffrey I. Webb. Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. *J. Mach. Learn. Res.*, 10:377–403, 2009. ISSN 1533-7928.

[90] Edward R. Omiecinski. Alternative Interest Measures for Mining Associations in

Databases. *IEEE Trans. on Knowl. and Data Eng.*, 15:57–69, January 2003. ISSN 1041-4347. doi: http://dx.doi.org/10.1109/TKDE.2003.1161582.

[91] M.E. Otey, S. Parthasarathy, Chao Wang, A. Veloso, and Jr. Meira, W. Parallel and Distributed Methods for Incremental Frequent Itemset Mining. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 34(6):2439 –2450, December 2004. ISSN 1083-4419. doi: 10.1109/TSMCB.2004.836887.

[92] Jong Soo Park, Ming-Syan Chen, and Philip S. Yu. An Effective Hash-based Algorithm for Mining Association Rules. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, SIGMOD '95, pages 175–186, New York, NY, USA, 1995. ACM. ISBN 0-89791-731-6. doi: http://doi.acm.org/10.1145/223784.223813.

[93] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering Frequent Closed Itemsets for Association Rules. In *Proceedings of the 7th International Conference on Database Theory*, ICDT '99, pages 398–416, London, UK, UK, 1999. Springer-Verlag. ISBN 3-540-65452-6.

[94] Jian Pei, Jiawei Han, and Runying Mao. CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 21–30, 2000.

[95] G. Piatetsky-Shapiro. Discovery, Analysis, and Presentation of Strong Rules. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery in Databases*. AAAI/MIT Press, Cambridge, MA, 1991.

[96] Gregory J. Pool, Andria F. Schwegler, Brian R. Theodore, and Perry N. Fuchs. Role of gender norms and group identification on hypothetical and experimental pain tolerance. *Pain*, 129(1-2):122 – 129, 2007. ISSN 0304-3959.

[97] James L. Powell. Estimation of Semiparametric Models. In R. F. Engle and D. McFadden, editors, *Handbook of Econometrics*, volume 4, chapter 41, pages 2443–2521. Elsevier, 1 edition, 1986.

[98] Jing Qin and Jerry Lawless. Empirical Likelihood and General Estimating Equations. *Annals of Statistics*, 22(1):300–325, 1994.

[99] Yongsong Qin and Shichao Zhang. EmpiricalLlikelihood Confidence Intervals for Differences between Two Datasets with Missing Data. *Pattern Recogn. Lett.*, 29 (6):803–812, April 2008. ISSN 0167-8655. doi: 10.1016/j.patrec.2007.12.010.

[100] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. ISBN 1-55860-238-0.

[101] J. N. K. Rao. On Variance Estimation With Imputed Survey Data. *Journal of the American Statistical Association*, 91(434):499–506, 1996.

[102] R. Rastogi and Kyuseok Shim. Mining Optimized Association Rules with Categorical and Numeric Attributes. *Knowledge and Data Engineering, IEEE Transactions on*, 14(1):29 –50, January 2002. ISSN 1041-4347. doi: 10.1109/69.979971.

[103] Joseph Lee Rodgers and W. Alan Nicewander. Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1):59–66, 1988. ISSN 00031305. URL `http://www.jstor.org/stable/2685263`.

[104] Geraldine Rouxel. Cognitive-affective determinants of performance in mathematics and verbal domains: Gender differences. *Learning and Individual Differences*, 12 (3):287 – 310, 2000. ISSN 1041-6080. doi: 10.1016/S1041-6080(01)00041-3.

[105] Ulrich Ruckert, Lothar Richter, and Stefan Kramer. Quantitative Association Rules Based on Half-Spaces: An Optimization Approach. In *Proceedings of the Fourth*

IEEE International Conference on Data Mining, ICDM '04, pages 507–510, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2142-8.

[106] Ron Rymon. Search through Systematic Set Enumeration. In Bernhard Nebel, Charles Rich, and William R. Swartout, editors, *KR*, pages 539–550. Morgan Kaufmann, 1992. ISBN 1-55860-262-3.

[107] Florin D. Salajan, Dieter J. Schnwetter, and Blaine M. Cleghorn. Student and faculty inter-generational digital divide: Fact or fiction? *Computers & Education*, 55(3):1393 – 1403, 2010. ISSN 0360-1315. doi: 10.1016/j.compedu.2010.06.017.

[108] Sunita Sarawagi, Shiby Thomas, and Rakesh Agrawal. Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications. *Data Min. Knowl. Discov.*, 4:89–125, July 2000. ISSN 1384-5810. doi: http://dx.doi.org/10.1023/A:1009887712954.

[109] Ashoka Savasere, Edward Omiecinski, and Shamkant B. Navathe. An Efficient Algorithm for Mining Association Rules in Large Databases. In *Proceedings of the 21th International Conference on Very Large Data Bases*, VLDB '95, pages 432–444, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-379-4.

[110] Pradeep Shenoy, Jayant R. Haritsa, S. Sudarshan, Gaurav Bhalotia, Mayank Bawa, and Devavrat Shah. Turbo-charging Vertical Mining of Large Databases. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, SIGMOD '00, pages 22–33, New York, NY, USA, 2000. ACM. ISBN 1-58113-217-4. doi: http://doi.acm.org/10.1145/342009.335376.

[111] Avi Silberschatz and Alexander Tuzhilin. What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Trans. on Knowl. and Data Eng.*, 8(6):970–974, December 1996. ISSN 1041-4347. doi: 10.1109/69.553165.

[112] Mondelle Simeon and Robert Hilderman. GENCCS: A Correlated Group Difference Approach to Contrast Set Mining. In *Proceedings of the 7th International Conference on Machine Learning and Data Mining in Pattern Recognition*, MLDM'11, pages 140–154, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-23198-8.

[113] Mondelle Simeon and Robert Hilderman. COSINE: A Vertical Group Difference Approach to Contrast Set Mining. In *Proceedings of the 24th Canadian conference on Advances in Artificial Intelligence*, Canadian AI'11, pages 359–371, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-21042-6.

[114] Mondelle Simeon and Robert J. Hilderman. Exploratory Quantitative Contrast Set Mining: A Discretization Approach. In *ICTAI '07: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Vol.2 (ICTAI 2007)*, pages 124–131, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-3015-X. doi: http://dx.doi.org/10.1109/ICTAI.2007.99.

[115] Mondelle Simeon, Robert J. Hilderman, and Howard J. Hamilton. Mining Interesting Contrast Sets. In Pascal Lorenz and Petre Dini, editors, *INTENSIVE 2012, The Fourth International Conference on Resource Intensive Applications and Services*, 2012.

[116] Hee Seok Song, Jae kyeong Kim, and Soung Hie Kim. Mining the change of customer behavior in an internet shopping mall. *Expert Systems with Applications*, 21(3):157 – 168, 2001. ISSN 0957-4174. doi: 10.1016/S0957-4174(01)00037-9.

[117] Arnaud Soulet, Bruno Cremilleux, and Francois Rioult. Condensed Representation of Emerging Patterns. In Honghua Dai, Ramakrishnan Srikant, and Chengqi Zhang, editors, *Advances in Knowledge Discovery and Data Mining*, volume 3056 of *Lecture*

Notes in Computer Science, pages 127–132. Springer Berlin / Heidelberg, 2004. ISBN 978-3-540-22064-0.

[118] Ramakrishnan Srikant and Rakesh Agrawal. Mining Quantitative Association Rules in Large Relational Tables. *SIGMOD Rec.*, 25(2):1–12, 1996. ISSN 0163-5808. doi: http://doi.acm.org/10.1145/235968.233311.

[119] Jonatan Taminau, Ruben Hillewaere, Stijn Meganck, Darrell Conklin, Ann Nowé, and Bernard Manderick. Applying Subgroup Discovery for the Analysis of String Quartet Movements. In *Proceedings of 3rd international workshop on Machine learning and music*, MML '10, pages 29–32, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0161-9. doi: 10.1145/1878003.1878014.

[120] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the Right Objective Measure for Association Analysis. *Inf. Syst.*, 29(4):293–313, 2004. ISSN 0306-4379. doi: http://dx.doi.org/10.1016/S0306-4379(03)00072-3.

[121] Ljupco Todorovski, Peter A. Flach, and Nada Lavrac. Predictive Performance of Weghted Relative Accuracy. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '00, pages 255–264, London, UK, UK, 2000. Springer-Verlag. ISBN 3-540-41066-X.

[122] Alexander Tuzhilin. On Subjective Measures ofIinterestingness in Knowledge Discovery. In *In Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pages 275–281. AAAI Press, 1995.

[123] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979. ISBN 0-408-70929-4.

[124] Christine A. Vaughan, William P. Sacco, and Jason W. Beckstead. Racial/ethnic Differences in Body Mass Index: The Roles of Beliefs about Thinness and Di-

etary Restriction. *Body Image*, 5(3):291 – 298, 2008. ISSN 1740-1445. doi: 10.1016/j.bodyim.2008.02.004.

[125] Adriano Veloso, Matthew Otey, Srinivasan Parthasarathy, and Wagner Meira. Parallel and Distributed Frequent Itemset Mining on Dynamic Datasets. In Timothy Pinkston and Viktor Prasanna, editors, *High Performance Computing - HiPC 2003*, volume 2913 of *Lecture Notes in Computer Science*, pages 184–193. Springer Berlin / Heidelberg, 2003. ISBN 978-3-540-20626-2.

[126] Geoffrey I. Webb. Magnum Opus: Setting New Standards in Association Rule Discovery, 2001. URL `www.rulequest.com`.

[127] Geoffrey I. Webb. Discovering Associations with Numeric Variables. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 383–388, New York, NY, USA, 2001. ACM. ISBN 1-58113-391-X. doi: 10.1145/502512.502569.

[128] Geoffrey I. Webb. Discovering Significant Patterns. *Mach. Learn.*, 68(1):1–33, 2007. ISSN 0885-6125. doi: http://dx.doi.org/10.1007/s10994-007-5006-x.

[129] Geoffrey I. Webb. Layered Critical Calues: A Powerful Direct-Adjustment Approach to Discovering Significant Patterns. *Mach. Learn.*, 71(2-3):307–323, 2008. ISSN 0885-6125. doi: http://dx.doi.org/10.1007/s10994-008-5046-x.

[130] Geoffrey I. Webb, Shane Butler, and Douglas Newlands. On Detecting Differences between Groups. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 256–265, New York, NY, USA, 2003. ACM. ISBN 1-58113-737-0. doi: http://doi.acm.org/10.1145/956750.956781.

[131] Tzu-Tsung Wong and Kuo-Lung Tseng. Mining Negative Contrast Sets from Data with Discrete Attributes. *Expert Syst. Appl.*, 29(2):401–407, 2005.

[132] Stefan Wrobel. An Algorithm for Multi-relational Discovery of Subgroups. In *PKDD '97: Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 78–87, London, UK, 1997. Springer-Verlag. ISBN 3-540-63223-9.

[133] Stefan Wrobel. *Inductive Logic Programming for Knowledge Discovery in Databases*, volume XIX, pages 74–101. Springer-Verlag, 2001.

[134] Hui Xiong, Pang-Ning Tan, and Vipin Kumar. Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 387 – 394, nov. 2003. doi: 10.1109/ICDM.2003.1250944.

[135] Hui Xiong, Pang-Ning Tan, and Vipin Kumar. Hyperclique Pattern Discovery. *Data Min. Knowl. Discov.*, 13:219–242, September 2006. ISSN 1384-5810. doi: 10.1007/s10618-006-0043-9.

[136] Y. Y. Yao, S. K. Michael Wong, and Cory J. Butz. On Information-Theoretic Measures of Attribute Importance. In *Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining*, PAKDD '99, pages 133–137, London, UK, 1999. Springer-Verlag. ISBN 3-540-65866-1.

[137] Raymond W. Yeung. *A First Course in Information Theory (Information Technology: Transmission, Processing and Storage)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0306467917.

[138] Masaharu Yoshioka. Analyzing Multiple News Sites by Contrasting Articles. In *Proceedings of the 2008 IEEE International Conference on Signal Image Technology and Internet Based Systems*, SITIS '08, pages 45–51, Washington, DC, USA, 2008. IEEE Computer Society. ISBN 978-0-7695-3493-0. doi: 10.1109/SITIS.2008.42.

[139] Larry T. H. Yu, Fu-lai Chung, Stephen C. F. Chan, and Simon M. C. Yuen. Using Emerging Pattern Based Projected Clustering and Gene Expression Data for Cancer Detection. In *Proceedings of the second conference on Asia-Pacific bioinformatics - Volume 29*, APBC '04, pages 75–84, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.

[140] G. Udny Yule. On the Association of Attributes in Statistics: With Illustrations from the Material of the Childhood Society. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 194(252-261):257–319, 1900. doi: 10.1098/rsta.1900.0019.

[141] Mohammed J. Zaki. Parallel and Distributed Association Mining: A Survey. *IEEE Concurrency*, 7(4):14–25, October 1999. ISSN 1092-3063. doi: 10.1109/4434.806975.

[142] Mohammed J. Zaki. Scalable Algorithms for Association Mining. *IEEE Trans. on Knowl. and Data Eng.*, 12:372–390, May 2000. ISSN 1041-4347. doi: 10.1109/69.846291.

[143] Mohammed J. Zaki and Karam Gouda. Fast Vertical Mining using Diffsets. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 326–335, New York, NY, USA, 2003. ACM. ISBN 1-58113-737-0. doi: http://doi.acm.org/10.1145/956750.956788.

[144] Mohammed J. Zaki and Ching-Jui Hsiao. Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure. *IEEE Trans. on Knowl. and Data Eng.*, 17:462–478, April 2005. ISSN 1041-4347. doi: http://dx.doi.org/10.1109/TKDE.2005.60.

[145] Mohammed Javeed Zaki and Ching-Jiu Hsiao. CHARM: An Efficient Algorithm for Closed Itemset Mining. In *SDM*, 2002.

[146] Jian Zhang and Joan Feigenbaum. Finding Highly Correlated Pairs Efficiently with Powerful Pruning. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 152–161, New York, NY, USA, 2006. ACM. ISBN 1-59593-433-2. doi: http://doi.acm.org/10.1145/1183614.1183640.

[147] Xiuzhen Zhang, Guozu Dong, and Ramamohanarao Kotagiri. Exploring Constraints to Efficiently Mine Emerging Patterns from Large High-dimensional Datasets. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 310–314, New York, NY, USA, 2000. ACM. ISBN 1-58113-233-6. doi: http://doi.acm.org/10.1145/347090.347158.

[148] Z Zhang, Y Lu, and B Zhang. An Effective Partitioning-Combining Algorithm for Discovering Quantitative Association Rules. In *Proc. of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 241–251, 1997.

Appendices

Appendix A

COSINE Data

Table A.1: Summary of Specificity Results - Mushroom

k	Alg.	MSDT														
		0	**0.01**	**0.03**	**0.05**	**0.1**	**0.5**	**0.75**	**1**	**2**	**5**	**10**	**15**	**20**	**25**	**30**
1	COSINE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	0	0	0	0	0	0	0	0	0	0	0	0	0	67	76
2	COSINE	0	0	0	0	0	0	0	0	0	0	5	5	2	3	4
	STUCCO	70	70	70	70	70	70	70	56	56	56	56	48	48	48	31
	CIGAR	0	0	0	0	0	0	0	0	0	0	0	0	0	57	67
3	COSINE	0	0	0	0	0	0	0	0	1	8	12	16	6	3	1
	STUCCO	674	674	674	674	674	674	674	621	621	320	289	239	194	132	98
	CIGAR	10	10	10	10	10	10	29	29	29	29	29	247	133	67	57
4	COSINE	0	1	1	1	1	2	3	9	28	40	24	17	13	5	1
	STUCCO	2197	2197	2197	2197	2197	2197	2197	1427	1427	916	818	731	545	348	20
	CIGAR	143	143	143	143	143	143	133	133	67	257	922	200	114	57	48
5	COSINE	1	13	13	13	13	28	41	48	48	69	47	15	7	4	0
	STUCCO	2490	2490	2490	2490	2490	2490	2490	719	719	306	194	143	56	22	0
	CIGAR	722	722	722	722	722	627	551	551	314	941	618	67	86	48	29
6	COSINE	13	68	68	68	68	95	94	107	113	91	20	15	7	0	1
	STUCCO	703	703	703	703	703	230	118	98	62	39	11	0	0	0	0
	CIGAR	2014	2014	2014	1758	1530	1330	1159	1159	656	722	447	57	48	48	10
7	COSINE	68	131	131	131	131	186	190	183	185	65	7	2	0	0	0
	STUCCO	98	98	98	98	98	0	0	0	0	0	0	0	0	0	0
	CIGAR	3696	3696	3325	2888	2518	2185	1900	1900	1083	513	342	29	10	10	0
8	COSINE	131	228	228	228	228	228	209	205	205	39	0	2	2	1	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	4864	4864	4066	3544	3078	2679	2328	2204	1330	76	48	29	0	0	0
9	COSINE	228	176	176	176	176	219	331	339	264	20	5	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	3487	3487	2584	2242	1957	1701	1482	532	473	5	0	0	0	0	0
10	COSINE	176	291	291	291	291	394	336	318	114	8	0	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	1881	1881	1311	1140	698	565	432	203	181	0	0	0	0	0	0
11	COSINE	291	243	243	243	243	239	176	116	97	0	0	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	722	722	618	532	247	179	151	119	9	0	0	0	0	0	0
12	COSINE	243	157	157	157	157	121	117	84	68	0	0	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	171	171	124	119	107	99	91	87	0	0	0	0	0	0	0
13	COSINE	157	80	80	80	80	90	74	73	14	0	0	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	86	86	95	87	79	11	4	1	0	0	0	0	0	0	0
14	COSINE	80	44	44	44	44	65	19	11	0	0	0	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	48	48	37	29	19	0	0	0	0	0	0	0	0	0	0
15	COSINE	44	43	43	43	43	7	1	1	0	0	0	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	COSINE	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.2: Summary of Statistical Measures for Specificity - Mushroom

Alg.	Meas.	MSDT 0	0.01	0.03	0.05	0.1	0.5	0.75	1	2	5	10	15	20	25	30
COSINE	Total	1475	1475	1475	1475	1475	1674	1591	1494	1137	340	120	72	37	16	8
	Mean	10.88	10.34	10.34	10.34	10.34	9.64	9.25	9.03	8.57	6.18	5.34	4.46	4.51	3.94	3.63
	SD	2.19	1.68	1.68	1.68	1.68	2.07	1.97	1.98	1.83	1.55	1.39	1.43	1.41	1.53	1.77
STUCCO	Total	6232	6232	6232	6232	6232	5661	5549	2921	2885	1637	1368	1161	843	550	149
	Mean	4.54	4.54	4.54	4.54	4.54	4.38	4.34	4.06	4.04	3.97	3.86	3.83	3.72	3.63	2.93
	SD	0.92	0.92	0.92	0.92	0.92	0.79	0.77	0.82	0.79	0.78	0.73	0.68	0.67	0.7	0.58
CIGAR	Total	17844	17844	15049	13214	11108	9529	8260	6918	4142	2543	2406	629	391	354	287
	Mean	8.02	8.02	7.87	7.85	7.7	7.63	7.59	7.29	7.3	5.93	5.3	4.63	4.64	3.89	2.71
	SD	1.59	1.59	1.62	1.64	1.59	1.51	1.52	1.46	1.35	0.65	1.05	1.43	1.05	1.93	1.45

Table A.3: Summary of Specificity Results - Census

k	Alg.	MSDT 0	0.01	0.03	0.05	0.1	0.5	0.75	1	2	5	10	15	20	25	30
1	COSINE	0	0	0	0	0	0	0	0	0	0	4	1	3	1	1
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	54	43	18
	CIGAR	0	0	0	0	0	0	0	0	0	0	0	0	4	4	16
2	COSINE	1	1	1	0	1	1	1	4	0	3	3	4	0	1	1
	STUCCO	19	19	19	19	19	19	22	37	66	243	187	151	49	21	21
	CIGAR	29	29	29	29	29	29	29	51	51	51	165	165	94	80	37
3	COSINE	12	12	12	11	22	18	19	12	16	8	8	0	0	0	0
	STUCCO	46	46	46	46	41	78	62	228	438	147	87	49	0	0	0
	CIGAR	233	233	233	233	233	233	233	200	200	200	141	98	15	0	0
4	COSINE	75	75	75	83	98	90	90	59	52	35	10	1	1	0	0
	STUCCO	196	196	173	771	764	543	448	411	398	99	43	8	0	0	0
	CIGAR	965	965	965	965	965	965	965	1079	979	875	78	24	5	0	0
5	COSINE	277	277	277	294	345	265	178	178	114	49	3	1	0	0	0
	STUCCO	723	723	642	642	572	1339	1454	1183	267	0	0	0	0	0	0
	CIGAR	2254	2254	2254	2254	2005	1785	1589	1040	492	101	18	2	0	0	0
6	COSINE	690	690	690	709	786	475	365	333	159	37	0	0	0	0	0
	STUCCO	3366	3366	2996	2721	2567	1703	854	132	19	0	0	0	0	0	0
	CIGAR	3348	3348	3348	3348	2978	2144	1460	934	123	97	0	0	0	0	0
7	COSINE	1052	1052	1052	1079	1201	621	467	422	181	28	0	0	0	0	0
	STUCCO	2744	2744	2443	2102	1926	184	121	0	0	0	0	0	0	0	0
	CIGAR	3337	3337	3337	3337	2970	1750	1152	265	87	43	0	0	0	0	0
8	COSINE	1290	1290	1290	1330	1307	575	499	345	137	23	0	0	0	0	0
	STUCCO	1801	1801	1602	927	732	0	0	0	0	0	0	0	0	0	0
	CIGAR	2299	2299	2299	2299	2046	408	159	99	73	29	0	0	0	0	0
9	COSINE	1084	1084	1084	1107	1173	446	335	248	94	1	0	0	0	0	0
	STUCCO	282	282	172	31	11	0	0	0	0	0	0	0	0	0	0
	CIGAR	1079	1079	1079	1079	959	188	121	88	37	3	0	0	0	0	0
10	COSINE	766	766	766	801	769	269	156	90	41	0	0	0	0	0	0
	STUCCO	172	172	47	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	328	328	328	328	229	52	47	39	4	0	0	0	0	0	0
11	COSINE	450	450	450	456	393	90	44	26	4	0	0	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	61	61	61	61	59	18	12	2	0	0	0	0	0	0	0
12	COSINE	238	238	238	222	171	30	9	5	0	0	0	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	27	27	27	27	25	0	0	0	0	0	0	0	0	0	0
13	COSINE	74	74	74	69	38	0	0	0	0	0	0	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.4: Summary of Statistical Measures for Specificity - Census

Alg.	Meas.	MSDT 0	0.01	0.03	0.05	0.1	0.5	0.75	1	2	5	10	15	20	25	30
COSINE	Total	6009	6009	6009	6161	6304	2880	2163	1722	798	184	28	7	4	2	2
	Mean	8.28	8.28	8.28	8.25	8.07	7.5	7.36	7.18	6.82	5.58	3.18	2.57	1.75	1.5	1.5
	SD	1.88	1.88	1.88	1.87	1.83	1.76	1.68	1.62	1.68	1.49	1.22	1.4	1.5	0.71	0.71
STUCCO	Total	9349	9349	8140	7259	6632	3866	2961	1991	1188	489	317	208	103	64	39
	Mean	6.7	6.7	6.63	6.23	6.17	5.34	5.15	4.58	3.78	2.71	2.55	2.31	1.48	1.33	1.54
	SD	1.17	1.17	1.11	1.18	1.17	0.88	0.85	0.85	0.91	0.78	0.72	0.54	0.5	0.47	0.51
CIGAR	Total	13960	13960	13960	13960	12498	7572	5767	3797	2046	1399	402	289	118	84	53
	Mean	6.57	6.57	6.57	6.57	6.52	5.85	5.61	5.18	4.59	4.18	2.87	2.53	2.18	1.95	1.7
	SD	1.56	1.56	1.56	1.56	1.57	1.38	1.38	1.41	1.35	1.13	0.88	0.68	0.55	0.21	0.46

Table A.5: Summary of Specificity Results- Waveform

k	Alg.	MSDT														
		0	**0.01**	**0.03**	**0.05**	**0.1**	**0.5**	**0.75**	**1**	**2**	**5**	**10**	**15**	**20**	**25**	**30**
1	COSINE	2	2	2	2	2	3	7	7	10	15	55	63	41	17	1
	STUCCO	9	9	9	9	16	103	109	216	287	290	741	445	198	132	77
	CIGAR	9	9	9	9	9	9	9	9	9	1412	645	541	265	206	102
2	COSINE	379	379	379	379	379	395	550	675	1219	1255	55	0	0	0	0
	STUCCO	30959	30959	30959	30959	23743	21898	17312	15239	5341	2105	713	0	0	0	0
	CIGAR	460	460	460	460	363	287	1076	7481	4605	2477	1105	17	15	0	0
3	COSINE	11890	11890	11890	11890	11890	12581	16445	16577	5348	7	0	0	0	0	0
	STUCCO	13349	13349	13349	13349	11943	10992	8015	5689	1543	12	0	0	0	0	0
	CIGAR	38901	38901	38901	38901	30732	26277	19181	14397	2687	9	9	0	0	0	0
4	COSINE	20009	20009	20009	20009	20009	19490	8199	2853	10	0	0	0	0	0	0
	STUCCO	917	917	917	917	615	411	276	88	50	0	0	0	0	0	0
	CIGAR	10793	10793	10793	10793	9873	7157	6284	882	55	0	0	0	0	0	0
5	COSINE	562	562	562	562	562	402	36	3	0	0	0	0	0	0	0
	STUCCO	288	288	288	288	193	129	87	58	0	0	0	0	0	0	0
	CIGAR	106	106	106	106	83	67	19	0	0	0	0	0	0	0	0
6	COSINE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	5	5	5	5	0	0	0	0	0	0	0	0	0	0	0

Table A.6: Summary of Statistical Measures for Specificity - Waveform

Alg.	Meas.	MSDT														
		0	**0.01**	**0.03**	**0.05**	**0.1**	**0.5**	**0.75**	**1**	**2**	**5**	**10**	**15**	**20**	**25**	**30**
COSINE	Total	32842	32842	32842	32842	32842	32871	25237	20115	6587	1277	110	63	41	17	1
	Mean	3.63	3.63	3.63	3.63	3.63	3.61	3.31	3.11	2.81	1.99	1.5	1	1	1	1
	SD	0.54	0.54	0.54	0.54	0.54	0.54	0.51	0.41	0.4	0.13	0.5	0	0	0	0
STUCCO	Total	45522	45522	45522	45522	36510	33533	25799	21290	7221	2407	1454	445	198	132	77
	Mean	2.35	2.35	2.35	2.35	2.38	2.36	2.34	2.27	2.19	1.88	1.49	1	1	1	1
	SD	0.55	0.55	0.55	0.55	0.55	0.53	0.52	0.49	0.5	0.33	0.5	0	0	0	0
CIGAR	Total	50274	50274	50274	50274	41060	33797	26569	22769	7356	3898	1759	558	280	206	102
	Mean	3.21	3.21	3.21	3.21	3.24	3.21	3.2	2.71	2.38	1.64	1.64	1.03	1.05	1	1
	SD	0.44	0.44	0.44	0.44	0.45	0.43	0.49	0.53	0.5	0.48	0.49	0.17	0.23	0	0

Table A.7: Summary of Specificity Results - Spambase

k	Alg.	MSDT											
		0	**0.01**	**0.03**	**0.05**	**0.1**	**0.5**	**0.75**	**1**	**2**	**5**	**10**	**15**
1	COSINE	1	0	0	0	0	93	88	85	75	8	1	1
	STUCCO	2	2	2	2	2	2	2	2	2	2	2	2
	CIGAR	7	7	7	7	5	7	10	12	7	12	7	7
2	COSINE	91	95	95	95	95	241	234	233	188	43	3	5
	STUCCO	443	443	443	443	443	443	443	443	443	172	68	25
	CIGAR	283	283	252	224	131	79	45	26	21	19	143	10
3	COSINE	483	243	243	243	243	879	844	800	715	131	16	0
	STUCCO	681	681	681	681	681	681	681	1457	568	221	120	59
	CIGAR	1502	1338	1190	1059	626	369	219	129	76	31	137	76
4	COSINE	1045	879	879	879	879	2634	2401	2238	984	347	23	4
	STUCCO	1556	1556	1556	1556	1556	1556	1556	1590	620	242	95	0
	CIGAR	26863	23909	21280	18938	11174	6593	3889	2294	1354	528	79	29
5	COSINE	2971	2636	2636	2636	2636	5435	4633	3942	1629	571	2	0
	STUCCO	3666	3666	3666	3666	10732	6332	3737	3657	1090	424	166	0
	CIGAR	25735	22903	20385	18143	10705	6317	3727	2199	1297	507	62	0
6	COSINE	4714	5436	5436	5436	5436	6491	5403	4762	2058	713	1	0
	STUCCO	36267	32271	32271	32271	26939	15894	9378	1190	463	181	32	0
	CIGAR	19685	17519	15594	13878	8187	4831	2849	1680	454	123	33	0
7	COSINE	5506	6491	6491	6491	6491	6343	5165	4915	3013	597	0	0
	STUCCO	32789	22619	22619	22619	11713	6910	4078	607	237	69	0	0
	CIGAR	17626	15687	13963	12426	7333	4327	2554	1507	407	110	23	0
8	COSINE	5670	6343	6343	6343	6343	7374	6077	5914	3524	261	3	0
	STUCCO	11854	10897	10897	10897	10911	6437	387	281	56	0	0	0
	CIGAR	17117	15234	13559	12067	7119	4201	2478	1165	315	85	0	0
9	COSINE	8278	7374	7374	7374	7374	6750	5521	5533	2834	102	0	0
	STUCCO	3997	3417	3417	3417	6523	3848	281	39	3	0	0	0
	CIGAR	14516	13042	11607	10332	6095	3596	2877	1352	365	79	0	0
10	COSINE	8641	6750	6750	6750	6750	4347	3615	3517	1194	26	0	0
	STUCCO	1754	1653	1653	1653	2878	1247	51	0	0	0	0	0
	CIGAR	9237	8221	7316	6512	3841	1805	848	399	108	29	0	0
11	COSINE	6332	4347	4347	4347	4347	2008	1690	1660	524	20	0	0
	STUCCO	651	443	443	443	524	30	5	0	0	0	0	0
	CIGAR	4486	3991	3553	3163	2530	1189	559	263	71	19	0	0
12	COSINE	2670	2008	2008	2008	2008	919	707	702	221	28	0	0
	STUCCO	213	193	193	76	5	0	0	0	0	0	0	0
	CIGAR	3123	2498	1998	1598	1278	601	282	133	3	2	0	0
13	COSINE	1181	919	919	919	919	397	333	330	193	39	0	0
	STUCCO	95	44	44	0	0	0	0	0	0	0	0	0
	CIGAR	231	185	148	118	94	44	11	5	0	0	0	0
14	COSINE	433	397	397	397	397	140	108	108	35	19	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	121	97	78	62	50	24	0	0	0	0	0	0
15	COSINE	140	140	140	140	140	42	31	31	15	2	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	26	11	3	0	0	0	0	0	0	0	0	0
16	COSINE	42	42	42	42	42	11	11	11	4	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	0	0	0	0	0	0	0	0	0	0	0	0
17	COSINE	11	11	11	11	11	0	0	0	0	0	0	0
	STUCCO	0	0	0	0	0	0	0	0	0	0	0	0
	CIGAR	0	0	0	0	0	0	0	0	0	0	0	0

Table A.8: Summary of Statistical Measures for Specificity - Spambase

Alg.	Meas.	MSDT											
		0	**0.01**	**0.03**	**0.05**	**0.1**	**0.5**	**0.75**	**1**	**2**	**5**	**10**	**15**
COSINE	Total	48209	44111	44111	44111	44111	44104	36861	34781	17206	2907	49	10
	Mean	8.71	8.45	8.45	8.45	8.45	7.45	7.39	7.48	7.29	6.13	3.82	2.7
	SD	2.31	2.23	2.23	2.23	2.23	2.23	2.26	2.26	2.21	2.01	1.36	1.16
STUCCO	Total	93968	77885	77885	77724	72907	53380	38699	36199	18485	3311	343	36
	Mean	6.75	6.7	6.7	6.69	6.68	6.43	6.17	5.75	5.77	4.58	3.7	2.47
	SD	1.24	1.29	1.29	1.26	1.53	1.43	1.36	1.3	1.21	1.44	1.26	0.61
CIGAR	Total	140558	124925	110938	98527	67168	55183	41548	39890	20178	5064	272	68
	Mean	6.65	6.64	6.62	6.61	6.81	6.61	6.45	6.43	5.97	5.31	3.08	2.78
	SD	2.31	2.29	2.27	2.25	2.26	2.04	1.92	1.86	1.81	1.34	1.24	0.81

Appendix B

GENCCS Data

Table B.1: Summary of Specificity Results - Mushroom

k	Alg.	MACT						
		0	**0.25**	**0.5**	**0.75**	**1**	**1.5**	**2**
1	GENCCS-0	0	0	1	1	1	1	3
	GENCCS	1	1	2	3	4	4	5
	GENCCS-1	9	9	10	12	13	14	15
	GENCCS-2	17	17	18	20	21	23	23
2	GENCCS-0	0	0	0	2	3	2	4
	GENCCS	1	1	1	2	2	1	4
	GENCCS-1	13	12	15	11	10	10	11
	GENCCS-2	21	21	20	19	16	16	12
3	GENCCS-0	0	1	0	0	1	5	4
	GENCCS	18	19	18	16	16	21	18
	GENCCS-1	10	11	9	11	12	23	13
	GENCCS-2	39	39	35	36	34	28	27
4	GENCCS-0	0	0	2	2	0	0	6
	GENCCS	52	51	51	51	49	51	37
	GENCCS-1	45	43	45	47	46	41	36
	GENCCS-2	42	42	41	34	34	34	22
5	GENCCS-0	1	1	2	1	3	9	10
	GENCCS	152	154	151	145	134	126	103
	GENCCS-1	65	65	58	53	55	45	28
	GENCCS-2	33	37	38	40	33	24	14
6	GENCCS-0	13	14	15	15	14	18	19
	GENCCS	156	153	152	148	132	108	97
	GENCCS-1	81	83	77	72	62	53	46
	GENCCS-2	27	24	27	23	20	12	16
7	GENCCS-0	68	67	70	69	68	74	73
	GENCCS	215	218	209	206	200	181	148
	GENCCS-1	79	79	79	79	76	61	52
	GENCCS-2	8	8	8	7	7	3	1
8	GENCCS-0	131	131	128	131	139	148	164
	GENCCS	146	145	138	130	124	121	110
	GENCCS-1	66	66	65	65	52	36	26
	GENCCS-2	0	0	0	0	0	0	0
9	GENCCS-0	228	227	227	224	192	177	174
	GENCCS	95	95	93	92	79	64	34
	GENCCS-1	22	23	23	19	14	14	5
	GENCCS-2	12	12	11	9	9	8	8
10	GENCCS-0	176	176	176	167	161	145	130
	GENCCS	69	69	68	70	59	43	14
	GENCCS-1	14	12	11	9	9	6	6
	GENCCS-2	0	0	0	0	0	0	0
11	GENCCS-0	291	292	291	293	286	271	195
	GENCCS	20	21	20	14	10	10	8
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
12	GENCCS-0	243	244	233	231	240	200	136
	GENCCS	5	3	3	3	3	1	1
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
13	GENCCS-0	157	157	165	166	151	104	76
	GENCCS	0	0	0	0	0	0	0
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
14	GENCCS-0	80	80	70	61	53	53	38
	GENCCS	0	0	0	0	0	0	0
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
15	GENCCS-0	44	42	39	41	37	34	19
	GENCCS	0	0	0	0	0	0	0
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
16	GENCCS-0	43	42	41	29	15	3	2
	GENCCS	0	0	0	0	0	0	0
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0

Table B.2: Summary of Statistical Measures for Specificity - Mushroom

Alg.	Meas.	MACT						
		0	**0.25**	**0.5**	**0.75**	**1**	**1.5**	**2**
GENCCS-0	Total	1475	1474	1460	1433	1364	1244	1053
	Mean	10.88	10.86	10.81	10.75	10.68	10.41	9.98
	SD	2.19	2.19	2.21	2.19	2.15	2.2	2.29
GENCCS	Total	930	930	906	880	812	731	579
	Mean	6.95	6.94	6.92	6.9	6.84	6.72	6.52
	SD	1.85	1.84	1.85	1.85	1.84	1.84	1.76
GENCCS-1	Total	404	403	392	378	349	303	238
	Mean	6.13	6.13	6.09	6.05	5.91	5.65	5.52
	SD	1.95	1.92	1.98	1.97	1.99	2.09	2.14
GENCCS-2	Total	199	200	198	188	174	148	123
	Mean	4.2	4.19	4.21	4.1	4.06	3.78	3.75
	SD	1.99	1.97	1.97	1.95	1.99	2.01	2.14

Table B.3: Summary of Specificity Results - Census

k	Alg.	MACT						
		0	**0.25**	**0.5**	**0.75**	**1**	**1.5**	**2**
1	GENCCS-0	0	8	10	14	17	20	22
	GENCCS	11	17	18	21	22	25	26
	GENCCS-1	14	20	21	24	24	25	27
	GENCCS-2	37	37	38	38	38	38	39
2	GENCCS-0	1	0	5	8	10	8	14
	GENCCS	43	35	33	27	22	15	16
	GENCCS-1	70	64	57	57	51	46	47
	GENCCS-2	61	62	59	60	57	52	52
3	GENCCS-0	12	7	10	26	22	24	33
	GENCCS	237	230	216	204	175	145	122
	GENCCS-1	226	223	213	204	178	152	121
	GENCCS-2	139	137	130	124	112	97	78
4	GENCCS-0	75	68	55	45	36	34	32
	GENCCS	503	499	481	455	408	335	246
	GENCCS-1	162	162	154	152	141	129	117
	GENCCS-2	0	0	0	0	0	0	0
5	GENCCS-0	277	252	223	192	156	115	83
	GENCCS	532	530	520	494	452	374	332
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
6	GENCCS-0	690	652	558	491	431	333	242
	GENCCS	409	409	398	385	366	337	317
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
7	GENCCS-0	1052	1035	965	883	789	646	458
	GENCCS	227	227	227	222	215	188	185
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
8	GENCCS-0	1290	1276	1233	1146	1039	845	638
	GENCCS	110	110	110	110	103	95	95
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
9	GENCCS-0	1084	1081	1062	1034	962	816	700
	GENCCS	0	0	0	0	0	0	0
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
10	GENCCS-0	766	760	748	742	691	587	503
	GENCCS	0	0	0	0	0	0	0
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
11	GENCCS-0	450	450	447	431	404	353	318
	GENCCS	0	0	0	0	0	0	0
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
12	GENCCS-0	238	238	235	231	225	194	179
	GENCCS	0	0	0	0	0	0	0
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
13	GENCCS-0	74	74	74	74	70	68	68
	GENCCS	0	0	0	0	0	0	0
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0

Table B.4: Summary of Statistical Measures for Specificity - Census

Alg.	Meas.	MACT						
		0	**0.25**	**0.5**	**0.75**	**1**	**1.5**	**2**
GENCCS-0	Total	6009	5901	5625	5317	4852	4043	3290
	Mean	8.28	8.31	8.37	8.42	8.46	8.51	8.61
	SD	1.88	1.88	1.88	1.91	1.92	1.96	2.06
GENCCS	Total	2072	2057	2003	1918	1763	1514	1339
	Mean	5.02	5.03	5.05	5.07	5.11	5.15	5.25
	SD	1.44	1.45	1.45	1.46	1.47	1.49	1.53
GENCCS-1	Total	472	469	445	437	394	352	312
	Mean	3.14	3.12	3.12	3.11	3.11	3.09	3.05
	SD	0.77	0.8	0.81	0.83	0.85	0.88	0.93
GENCCS-2	Total	237	236	227	222	207	187	169
	Mean	2.43	2.42	2.41	2.39	2.36	2.32	2.23
	SD	0.75	0.75	0.76	0.76	0.77	0.79	0.8

Table B.5: Summary of Specificity Results - Waveform

k	Alg.	MACT						
		0	**0.25**	**0.5**	**0.75**	**1**	**1.5**	**2**
1	GENCCS-0	2	4	9	14	15	22	24
	GENCCS	4	6	11	15	16	23	25
	GENCCS-1	5	7	11	17	17	23	25
	GENCCS-2	18	19	24	27	27	31	34
2	GENCCS-0	379	370	251	119	42	7	9
	GENCCS	328	319	222	111	45	14	12
	GENCCS-1	292	285	207	109	55	22	15
	GENCCS-2	210	204	147	73	32	9	10
3	GENCCS-0	11890	11890	11882	11767	11508	10259	7722
	GENCCS	9431	9431	9427	9332	9113	8037	6083
	GENCCS-1	6696	6696	6694	6613	6444	5680	4394
	GENCCS-2	4360	4360	4358	4305	4190	3670	2887
4	GENCCS-0	20009	20009	20009	20009	20009	19952	19458
	GENCCS	16248	16248	16248	16248	16248	16199	15777
	GENCCS-1	10976	10976	10976	10976	10976	10944	10676
	GENCCS-2	5829	5829	5829	5829	5829	5815	5701
5	GENCCS-0	562	562	562	562	562	562	562
	GENCCS	500	500	500	500	500	500	500
	GENCCS-1	369	369	369	369	369	369	369
	GENCCS-2	252	252	252	252	252	252	252

Table B.6: Summary of Statistical Measures for Specificity - Waveform

Alg.	Meas.	MACT						
		0	**0.25**	**0.5**	**0.75**	**1**	**1.5**	**2**
GENCCS-0	Total	32842	32835	32713	32471	32136	30802	27775
	Mean	3.63	3.63	3.64	3.65	3.66	3.68	3.74
	SD	0.54	0.54	0.53	0.52	0.52	0.51	0.49
GENCCS	Total	26511	26504	26408	26206	25922	24773	22397
	Mean	3.64	3.64	3.64	3.65	3.66	3.69	3.75
	SD	0.54	0.54	0.54	0.53	0.52	0.51	0.49
GENCCS-1	Total	18338	18333	18257	18084	17861	17038	15479
	Mean	3.62	3.62	3.63	3.64	3.65	3.68	3.73
	SD	0.56	0.56	0.55	0.54	0.53	0.52	0.5
GENCCS-2	Total	10669	10664	10610	10486	10330	9777	8884
	Mean	3.57	3.57	3.58	3.59	3.6	3.64	3.69
	SD	0.58	0.58	0.58	0.56	0.56	0.55	0.54

Table B.7: Summary of Specificity Results - Spambase

k	Alg.	MACT						
		0	**0.25**	**0.5**	**0.75**	**1**	**1.5**	**2**
1	GENCCS-0	1	3	15	22	25	29	39
	GENCCS	6	9	20	27	32	37	46
	GENCCS-1	14	17	26	36	39	43	50
	GENCCS-2	21	22	33	40	46	50	56
2	GENCCS-0	91	86	47	39	36	36	26
	GENCCS	63	58	31	31	27	34	23
	GENCCS-1	50	44	27	24	19	26	16
	GENCCS-2	17	17	15	15	15	18	13
3	GENCCS-0	483	443	261	178	161	113	43
	GENCCS	218	206	132	101	92	57	37
	GENCCS-1	210	202	143	109	88	59	44
	GENCCS-2	26	24	16	19	14	10	9
4	GENCCS-0	1045	994	741	579	482	366	283
	GENCCS	589	572	442	374	312	246	210
	GENCCS-1	510	499	421	382	343	251	216
	GENCCS-2	65	63	53	43	37	33	29
5	GENCCS-0	2971	2957	2263	2066	1586	1044	719
	GENCCS	1565	1555	1282	1243	1038	728	607
	GENCCS-1	848	847	768	752	678	451	419
	GENCCS-2	97	97	76	66	67	44	42
6	GENCCS-0	4714	4714	4072	3859	3038	2090	1721
	GENCCS	2134	2134	2002	1921	1726	1246	1148
	GENCCS-1	1060	1060	1028	1003	949	819	780
	GENCCS-2	129	129	123	121	116	100	94
7	GENCCS-0	5506	5506	5283	5158	4896	3209	2715
	GENCCS	2206	2206	2128	2079	2010	1678	1520
	GENCCS-1	1026	1026	1022	1001	964	921	874
	GENCCS-2	132	129	111	104	102	85	78
8	GENCCS-0	5670	5670	5625	5540	5441	4238	3818
	GENCCS	2683	2683	2683	2601	2591	2313	2100
	GENCCS-1	834	834	834	827	820	774	756
	GENCCS-2	82	82	78	76	70	63	62
9	GENCCS-0	8278	8278	8260	8216	8168	7275	6931
	GENCCS	3335	3335	3335	3293	3282	2924	2793
	GENCCS-1	717	717	717	716	716	691	677
	GENCCS-2	45	45	40	38	36	29	30
10	GENCCS-0	8641	8641	8641	8635	8514	7932	7505
	GENCCS	3454	3454	3454	3446	3392	3217	3003
	GENCCS-1	396	396	396	396	395	392	392
	GENCCS-2	50	50	49	49	47	37	34
11	GENCCS-0	6332	6332	6332	6332	6284	5904	5759
	GENCCS	2487	2487	2487	2487	2486	2450	2441
	GENCCS-1	181	181	181	181	181	181	181
	GENCCS-2	6	6	5	5	5	3	4
12	GENCCS-0	2670	2670	2670	2670	2669	2466	2466
	GENCCS	994	994	994	994	994	980	980
	GENCCS-1	66	66	66	66	66	66	66
	GENCCS-2	13	13	13	13	12	0	0
13	GENCCS-0	1181	1181	1181	1181	1179	1036	1036
	GENCCS	318	318	318	318	318	318	318
	GENCCS-1	5	5	5	5	5	5	5
	GENCCS-2	5	5	4	2	2	1	0
14	GENCCS-0	433	433	433	433	431	397	397
	GENCCS	186	186	186	186	186	186	186
	GENCCS-1	1	1	1	1	1	1	1
	GENCCS-2	3	3	2	0	0	0	0
15	GENCCS-0	140	140	140	140	140	140	140
	GENCCS	37	37	37	37	37	37	37
	GENCCS-1	5	5	5	5	5	5	5
	GENCCS-2	5	5	5	5	5	5	4
16	GENCCS-0	42	42	42	42	42	42	42
	GENCCS	14	14	14	14	14	14	14
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0
17	GENCCS-0	11	11	11	11	11	11	11
	GENCCS	4	4	4	4	4	4	4
	GENCCS-1	0	0	0	0	0	0	0
	GENCCS-2	0	0	0	0	0	0	0

Table B.8: Summary of Statistical Measures for Specificity - Spambase

Alg.	Meas.	MACT						
		0	**0.25**	**0.5**	**0.75**	**1**	**1.5**	**2**
GENCCS-0	Total	48209	48101	46017	45101	43103	36328	33651
	Mean	8.71	8.72	8.87	8.94	9.06	9.29	9.43
	SD	2.31	2.3	2.21	2.18	2.12	2.06	2
GENCCS	Total	20293	20252	19549	19156	18541	16469	15467
	Mean	8.48	8.49	8.62	8.66	8.75	8.95	9.04
	SD	2.34	2.33	2.26	2.24	2.21	2.18	2.16
GENCCS-1	Total	5923	5900	5640	5504	5269	4685	4482
	Mean	6.85	6.87	6.98	7.03	7.1	7.29	7.36
	SD	2.11	2.1	2.07	2.06	2.06	2.06	2.05
GENCCS-2	Total	696	690	623	596	574	478	455
	Mean	6.52	6.53	6.51	6.43	6.39	6.11	6.05
	SD	2.49	2.5	2.61	2.65	2.7	2.72	2.74

Appendix C

MSRT Data

Table C.1: Summary of Specificity Results - Mushroom

k	Alg.	MSRT													
		0	**0.5**	**1**	**1.2**	**1.4**	**1.6**	**1.8**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
1	λ-COSINE	1	1	1	0	0	0	0	0	0	0	0	1	1	2
	λ-GENCCS	1	2	2	1	1	1	2	2	1	1	1	2	2	2
2	λ-COSINE	2	3	2	0	0	3	3	2	6	7	9	16	15	14
	λ-GENCCS	1	1	1	2	6	9	7	5	8	5	7	15	14	14
3	λ-COSINE	0	1	5	3	14	17	21	22	13	16	15	7	3	3
	λ-GENCCS	18	18	18	13	11	22	27	19	14	14	14	3	1	1
4	λ-COSINE	2	0	0	19	17	37	51	35	19	7	7	0	0	0
	λ-GENCCS	52	51	51	41	21	34	43	36	16	9	6	0	0	0
5	λ-COSINE	1	3	9	33	30	95	62	32	10	5	1	0	0	0
	λ-GENCCS	152	151	151	81	42	64	37	25	13	2	0	0	0	0
6	λ-COSINE	15	14	18	91	32	58	27	14	7	0	0	0	0	0
	λ-GENCCS	154	152	152	111	47	56	27	18	0	0	0	0	0	0
7	λ-COSINE	69	68	74	145	117	35	24	22	1	0	0	0	0	0
	λ-GENCCS	211	209	209	128	75	18	10	10	0	0	0	0	0	0
8	λ-COSINE	131	139	148	157	72	12	11	6	0	0	0	0	0	0
	λ-GENCCS	140	138	138	80	52	0	0	0	0	0	0	0	0	0
9	λ-COSINE	224	192	177	99	24	6	3	2	0	0	0	0	0	0
	λ-GENCCS	95	93	93	39	11	0	0	0	0	0	0	0	0	0
10	λ-COSINE	167	161	145	55	23	0	0	0	0	0	0	0	0	0
	λ-GENCCS	69	68	68	21	10	0	0	0	0	0	0	0	0	0
11	λ-COSINE	293	286	271	52	9	0	0	0	0	0	0	0	0	0
	λ-GENCCS	20	20	20	8	0	0	0	0	0	0	0	0	0	0
12	λ-COSINE	231	240	200	17	1	0	0	0	0	0	0	0	0	0
	λ-GENCCS	5	3	3	1	0	0	0	0	0	0	0	0	0	0
13	λ-COSINE	166	151	104	26	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	λ-COSINE	61	57	57	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	λ-COSINE	42	41	41	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	λ-COSINE	38	37	37	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table C.2: Summary of Statistical Measures for Specificity - Mushroom

Alg.	Meas.	MSRT													
		0	**0.5**	**1**	**1.2**	**1.4**	**1.6**	**1.8**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
λ-COSINE	Total	1443	1394	1289	697	339	263	202	135	56	35	32	24	19	19
	Mean	10.79	10.78	10.59	8.07	7.09	5.41	5.09	4.99	4.04	3.29	3	2.25	2.11	2.05
	SD	2.22	2.24	2.37	2.04	1.78	1.37	1.46	1.57	1.24	0.96	0.8	0.53	0.46	0.52
λ-GENCCS	Total	918	906	906	526	276	204	153	115	52	31	28	20	17	17
	Mean	6.94	6.92	6.92	6.61	6.38	4.92	4.48	4.49	3.62	3.19	2.89	2.05	1.94	1.94
	SD	1.86	1.85	1.85	1.74	1.76	1.3	1.34	1.38	1.09	0.91	0.79	0.51	0.43	0.43

Table C.3: Summary of Specificity Results - Census

k	Alg.	MSRT													
		0	**0.5**	**1**	**1.2**	**1.4**	**1.6**	**1.8**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
1	λ-COSINE	8	10	14	2	3	3	3	3	4	6	5	6	6	6
	λ-GENCCS	11	18	18	16	16	16	16	15	14	13	11	10	10	9
2	λ-COSINE	0	5	8	43	61	61	60	59	34	41	31	17	15	8
	λ-GENCCS	43	33	33	28	36	34	30	30	15	26	19	11	11	7
3	λ-COSINE	7	10	26	276	240	204	157	145	61	50	44	39	24	14
	λ-GENCCS	237	216	216	228	209	167	115	102	60	35	33	31	18	10
4	λ-COSINE	68	55	45	581	293	256	171	154	84	10	10	9	8	4
	λ-GENCCS	493	481	481	387	177	139	86	77	44	6	6	5	5	2
5	λ-COSINE	252	223	192	620	261	209	97	80	44	1	1	1	1	1
	λ-GENCCS	531	520	520	365	110	98	42	37	17	1	1	1	1	1
6	λ-COSINE	659	561	491	500	125	102	47	46	18	0	0	0	0	0
	λ-GENCCS	401	398	398	165	55	42	13	12	0	0	0	0	0	0
7	λ-COSINE	989	883	946	252	73	57	23	18	1	0	0	0	0	0
	λ-GENCCS	227	227	227	138	42	42	12	10	0	0	0	0	0	0
8	λ-COSINE	1233	1101	987	174	5	5	1	1	0	0	0	0	0	0
	λ-GENCCS	110	110	110	22	5	5	1	1	0	0	0	0	0	0
9	λ-COSINE	1062	934	734	24	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	λ-COSINE	848	742	742	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	λ-COSINE	448	399	429	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	λ-COSINE	137	185	182	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	λ-COSINE	68	16	12	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table C.4: Summary of Statistical Measures for Specificity - Census

Alg.	Meas.	MSRT													
		0	**0.5**	**1**	**1.2**	**1.4**	**1.6**	**1.8**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
λ-COSINE	Total	5779	5124	4808	2472	1061	897	559	506	246	108	91	72	54	33
	Mean	8.27	8.27	8.28	5.14	4.36	4.3	3.96	3.92	3.76	2.62	2.68	2.75	2.69	2.58
	SD	1.84	1.84	1.9	1.49	1.33	1.34	1.29	1.29	1.19	0.77	0.79	0.83	0.93	1.03
λ-GENCCS	Total	2053	2003	2003	1349	650	543	315	284	150	81	70	58	45	29
	Mean	5.02	5.05	5.05	4.64	4.06	4.08	3.63	3.61	3.23	2.46	2.53	2.59	2.47	2.28
	SD	1.45	1.45	1.45	1.37	1.41	1.47	1.33	1.33	1.08	0.9	0.91	0.94	1.04	1.1

Table C.5: Summary of Specificity Results - Waveform

k	Alg.	MSRT													
		0	**0.5**	**1**	**1.2**	**1.4**	**1.6**	**1.8**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
1	λ-COSINE	15	15	15	2	2	2	2	2	2	2	3	5	6	4
	λ-GENCCS	4	11	11	11	10	10	10	10	9	9	9	10	9	8
2	λ-COSINE	42	42	42	378	347	345	331	322	274	238	208	184	175	169
	λ-GENCCS	259	222	222	225	225	231	230	231	203	201	179	164	158	156
3	λ-COSINE	11508	11508	11508	11563	7148	5625	3947	3205	2002	1184	564	390	260	206
	λ-GENCCS	9427	9427	9427	9132	5810	4865	3643	2982	1998	1183	564	390	260	206
4	λ-COSINE	20009	20009	20009	18251	7310	4315	2706	2159	1360	364	105	38	15	5
	λ-GENCCS	16248	16248	16248	14599	6148	3913	2614	2083	1360	364	105	38	15	5
5	λ-COSINE	562	562	562	484	257	137	111	92	53	1	1	1	0	0
	λ-GENCCS	500	500	500	423	248	137	111	92	53	1	1	1	0	0

Table C.6: Summary of Statistical Measures for Specificity - Waveform

Alg.	Meas.	MSRT													
		0	**0.5**	**1**	**1.2**	**1.4**	**1.6**	**1.8**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
λ-COSINE	Total	32136	32136	32136	30678	15064	10424	7097	5780	3691	1789	881	618	456	384
	Mean	3.66	3.66	3.66	3.61	3.5	3.41	3.37	3.35	3.32	3.07	2.88	2.75	2.62	2.55
	SD	0.52	0.52	0.52	0.54	0.58	0.58	0.6	0.61	0.63	0.58	0.6	0.58	0.57	0.54
λ-GENCCS	Total	26438	26408	26408	24390	12441	9156	6608	5398	3623	1758	858	603	442	375
	Mean	3.64	3.64	3.64	3.62	3.51	3.43	3.39	3.37	3.34	3.08	2.9	2.76	2.64	2.55
	SD	0.54	0.54	0.54	0.54	0.58	0.58	0.59	0.6	0.62	0.58	0.61	0.59	0.58	0.56

Table C.7: Summary of Specificity Results - Spambase

k	Alg.	MSRT													
		0	**0.5**	**1**	**1.2**	**1.4**	**1.6**	**1.8**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
1	λ-COSINE	39	9	46	1	3	9	9	12	34	42	42	41	46	46
	λ-GENCCS	6	20	20	17	19	19	21	35	42	42	41	46	46	47
2	λ-COSINE	26	58	23	712	120	74	78	86	56	40	38	36	21	15
	λ-GENCCS	63	31	31	37	68	70	74	55	38	36	35	20	15	6
3	λ-COSINE	43	206	37	1046	177	88	90	52	40	17	17	4	0	0
	λ-GENCCS	218	132	132	138	89	92	53	41	21	21	4	0	0	0
4	λ-COSINE	283	572	210	739	83	19	14	11	9	2	2	0	0	0
	λ-GENCCS	589	442	442	114	20	15	14	12	1	1	0	0	0	0
5	λ-COSINE	719	1555	607	104	2	1	1	1	1	0	0	0	0	0
	λ-GENCCS	1565	1282	1282	25	0	0	0	0	0	0	0	0	0	0
6	λ-COSINE	1721	2134	1148	9	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	2134	2002	2002	1	0	0	0	0	0	0	0	0	0	0
7	λ-COSINE	2715	2206	1520	4	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	2206	2128	2128	0	0	0	0	0	0	0	0	0	0	0
8	λ-COSINE	3818	2683	2100	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	2683	2683	2683	0	0	0	0	0	0	0	0	0	0	0
9	λ-COSINE	6931	3335	2793	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	3335	3335	3335	0	0	0	0	0	0	0	0	0	0	0
10	λ-COSINE	7505	3454	3003	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	3454	3454	3454	0	0	0	0	0	0	0	0	0	0	0
11	λ-COSINE	5759	2487	2441	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	2487	2487	2487	0	0	0	0	0	0	0	0	0	0	0
12	λ-COSINE	2466	994	980	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	994	994	994	0	0	0	0	0	0	0	0	0	0	0
13	λ-COSINE	1036	318	318	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	318	318	318	0	0	0	0	0	0	0	0	0	0	0
14	λ-COSINE	397	186	186	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	186	186	186	0	0	0	0	0	0	0	0	0	0	0
15	λ-COSINE	140	37	37	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	37	37	37	0	0	0	0	0	0	0	0	0	0	0
16	λ-COSINE	42	14	14	0	0	0	0	0	0	0	0	0	0	0
	λ-GENCCS	14	14	14	0	0	0	0	0	0	0	0	0	0	0
	λ-COSINE	11	4	4	0	0	0	0	0	0	0	0	0	0	0
17	λ-GENCCS	4	4	4	0	0	0	0	0	0	0	0	0	0	0
	λ-COSINE	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table C.8: Summary of Statistical Measures for Specificity - Spambase

Alg.	Meas.	MSRT													
		0	**0.5**	**1**	**1.2**	**1.4**	**1.6**	**1.8**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
λ-COSINE	Total	33640	20248	15463	2615	385	191	192	162	140	101	99	81	67	61
	Mean	9.42	8.49	9.03	3.11	2.9	2.63	2.58	2.4	2.19	1.79	1.79	1.54	1.31	1.25
	SD	2	2.33	2.16	0.87	0.76	0.75	0.72	0.75	0.9	0.79	0.8	0.59	0.47	0.43
λ-GENCCS	Total	20293	19549	19549	332	196	196	162	143	102	100	80	66	61	53
	Mean	8.48	8.61	8.61	3.29	2.56	2.53	2.37	2.21	1.81	1.81	1.54	1.3	1.25	1.11
	SD	2.34	2.26	2.26	0.95	0.8	0.77	0.82	0.91	0.79	0.8	0.59	0.46	0.43	0.32

Appendix D

Case Studies Questionnaires

Case Study I

I. Background Information:

a. What is your profession (including speciality if applicable)_____

b. Where do you work (check as many as applicable)

i. ____ Hospital

ii. ____ University

iii. ____ Private health clinic

iv. _____Other (please specify)

c. How many years have you been in practice?_____

d. Gender: ___Male ___Female

e. Age____

f. What is your religious affiliation? _____

g. Where did you complete your medical training?_____

II. Decision Making Influences

The following questions seek to provide us with a better understanding of what factors influence the ethical decision making behaviour of practitioners. Please indicate how influential is each of the following during those times that you are attempting to develop solutions to ethical dilemmas/problems:

1. **Professional codes of ethics**:
   1 2 3 4 5 6 7
   not important extremely important

2. **Family views**:
   1 2 3 4 5 6 7
   not important extremely important

3. **University training**:
   1 2 3 4 5 6 7
   not important extremely important

4. **Hospital value statements**:
   1 2 3 4 5 6 7
   not important extremely important

5. **Religious background**:
   1 2 3 4 5 6 7
   not important extremely important

6. **Staff views** (i.e., colleagues, co-workers, etc.):
   1 2 3 4 5 6 7
   not important extremely important

7. **Peer attitudes**:
   1 2 3 4 5 6 7
   not important extremely important

Figure D.1: Ethical Decision-Making Questionnaire Page 1

III. Ethical Ideology (*The Ethics Position Questionnaire*)

Instructions. You will find a series of general statements listed below. Each represents a commonly held opinion and there are no right or wrong answers. You will probably disagree with some items and agree with others. We are interested in the extent to which you agree or disagree with such matters of opinion.

Please read each statement carefully. Then indicate the extent to which you agree or disagree by placing in front of the statement the number corresponding to your feelings, where:

1 = Completely disagree	4 = Slightly disagree	7 = Moderately agree
2 = Largely disagree	5 = Neither agree nor disagree	8 = Largely agree
3 = Moderately disagree	6 = Slightly agree	9 = Completely agree

1. [ ] A person should make certain that their actions never intentionally harm another even to a small degree.
2. [ ] Risks to another should never be tolerated, irrespective of how small the risks might be.
3. [ ] The existence of potential harm to others is always wrong, irrespective of the benefits to
4. be gained.
5. [ ] One should never psychologically or physically harm another person.
6. [ ] One should not perform an action which might in any way threaten the dignity and
7. welfare of another individual.
8. [ ] If an action could harm an innocent other, then it should not be done.
9. [ ] Deciding whether or not to perform an act by balancing the positive consequences of the act against the negative consequences of the act is immoral.
10. [ ] The dignity and welfare of people should be the most important concern in any society.
11. [ ] It is never necessary to sacrifice the welfare of others.
12. [ ] Moral actions are those which closely match ideals of the most "perfect" action.
13. [ ] There are no ethical principles that are so important that they should be a part of any code of ethics.
14. [ ] What is ethical varies from one situation and society to another.
15. [ ] Moral standards should be seen as being individualistic; what one person considers to be moral may be judged to be immoral by another person.
16. [ ] Different types of moralities cannot be compared as to "rightness."
17. [ ] Questions of what is ethical for everyone can never be resolved since what is moral or immoral is up to the individual.
18. [ ] Moral standards are simply *personal* rules which indicate how a person should behave, and are not to be applied in making judgments of others.
19. [ ] Ethical considerations in interpersonal relations are so complex that individuals should be allowed to formulate their own individual codes.
20. [ ] Rigidly codifying an ethical position that prevents certain types of actions could stand in the way of better human relations and adjustment.
21. [ ] No rule concerning lying can be formulated; whether a lie is permissible to not permissible totally depends upon the situation.

20 [ ]. Whether a lie is judged to be moral or immoral depends upon the circumstances surrounding the action.

Figure D.2: Ethical Decision-Making Questionnaire Page 2

IV.Vignettes: Multidimensional Ethics Scale (Reidenbach & Robin, 1990)

Please read each of the vignettes and provide your beliefs about the action by placing an "X" between the pairs of items.

Vignette 1: Quality of Life – Level of Treatment

Mrs. X is an 85-year-old woman with terminal cancer and in significant pain. She is not expected to live more than 30 days. The attending physician realises that to provide her with adequate pain relief, the administration of an increased dose of morphine may result in further complications that may hasten her death. Without an increase in the dosage of her medication she will perhaps live an additional month yet suffer profoundly.

Action: The physician increases the dosage of morphine.

Your response to this action is that it is…

Just	1	2	3	4	5	6	7	Unjust
Fair	1	2	3	4	5	6	7	Unfair
Culturally Acceptable	1	2	3	4	5	6	7	Culturally Unacceptable
Individually Acceptable	1	2	3	4	5	6	7	Individually Unacceptable
Traditionally Acceptable	1	2	3	4	5	6	7	Traditionally Unacceptable
Acceptable to my Family	1	2	3	4	5	6	7	Unacceptable to my Family
Morally Right	1	2	3	4	5	6	7	Not Morally Right
Self Promoting	1	2	3	4	5	6	7	Not Self Promoting
Self Sacrificing	1	2	3	4	5	6	7	Not Self Sacrificing
Personally Satisfying	1	2	3	4	5	6	7	Not Personally Satisfying
Produces the Greatest Utility	1	2	3	4	5	6	7	Produces the Least Utility
Maximizes Benefits While Minimizes Harm	1	2	3	4	5	6	7	Minimizes Benefits While Maximizes Harm
Leads to the Greatest Good for the Greatest Number	1	2	3	4	5	6	7	Leads to the Least Good for the Greatest Number
Violates an Unwritten Contract	1	2	3	4	5	6	7	Does Not Violate an Unwritten Contract
Obligated to Act this Way	1	2	3	4	5	6	7	Not Obligated to Act this Way
Violates an Unspoken Promise	1	2	3	4	5	6	7	Does Not Violate an Unspoken Promise
The probability that I would undertake the same action is High	1	2	3	4	5	6	7	Low
The probability that my peers or colleagues would undertake the same action is High	1	2	3	4	5	6	7	Low
The action describe above is Ethical	1	2	3	4	5	6	7	Unethical

Figure D.3: Ethical Decision-Making Questionnaire Page 3

Vignette 2: Competence

Mr. O. is a 73-year-old retired executive who has begun to show signs of early onset of dementia. His wife reports that he is getting lost while driving his car in the city in which he has lived all his life and often loses his car after parking for short periods of time. Mr. O. vehemently denies that he has memory issues and diverts blame to improper placement of street signs. Following testing that confirms dementia, Dr. M. must decide whether or not Mr. O. is a driving hazard due to his declining mental faculties.

Action: Dr. M calls the motor vehicle department to rescind Mr. O's license to drive.

Your response to this action is that it is…

Just	1	2	3	4	5	6	7	Unjust
Fair	1	2	3	4	5	6	7	Unfair
Culturally Acceptable	1	2	3	4	5	6	7	Culturally Unacceptable
Individually Acceptable	1	2	3	4	5	6	7	Individually Unacceptable
Traditionally Acceptable	1	2	3	4	5	6	7	Traditionally Unacceptable
Acceptable to my Family	1	2	3	4	5	6	7	Unacceptable to my Family
Morally Right	1	2	3	4	5	6	7	Not Morally Right
Self Promoting	1	2	3	4	5	6	7	Not Self Promoting
Self Sacrificing	1	2	3	4	5	6	7	Not Self Sacrificing
Personally Satisfying	1	2	3	4	5	6	7	Not Personally Satisfying
Produces the Greatest Utility	1	2	3	4	5	6	7	Produces the Least Utility
Maximizes Benefits While Minimizes Harm	1	2	3	4	5	6	7	Minimizes Benefits While Maximizes Harm
Leads to the Greatest Good for the Greatest Number	1	2	3	4	5	6	7	Leads to the Least Good for the Greatest Number
Violates an Unwritten Contract	1	2	3	4	5	6	7	Does Not Violate an Unwritten Contract
Obligated to Act this Way	1	2	3	4	5	6	7	Not Obligated to Act this Way
Violates an Unspoken Promise	1	2	3	4	5	6	7	Does Not Violate an Unspoken Promise
The probability that I would undertake the same action is High	1	2	3	4	5	6	7	Low
The probability that my peers or colleagues would undertake the same action is High	1	2	3	4	5	6	7	Low
The action describe above is Ethical	1	2	3	4	5	6	7	Unethical

4

Figure D.4: Ethical Decision-Making Questionnaire Page 4

V. Salience in Religious Commitment Scale (Clark Roof & Perkins, 1975)

Please answer the following questions by *circling* the appropriate answer for the multiple-choice items and checking or providing the most accurate number for the fill-in-the-blank questions.

1. My religious faith is:
   ____ Important for my life, but no more important than certain other aspects of my life.
   ____ Only of minor importance for my life, compared to certain other aspects of my life.
   ____ Of central importance for my life, and would, if necessary come before all other aspects of my life.

2. Everyone must make many important life decisions, such as which occupation to pursue, what goals to strive for, whom to vote for, what to teach one's children, etc. When you have made, or do make decisions such as these, to what extent do you make the decisions on the basis of your religious faith?

   ____ I seldom if ever base such decisions on religious faith.
   ____ I sometimes base such decisions on my religious faith but definitely not most of the time.
   ____ I feel that most of my important decisions are based on my religious faith, but usually in a general, unconscious way.
   ____ I feel that most of my important decisions are based on my religious faith, and I usually attempt to make them so.

3. Without my religious faith, the rest of my life would not have much meaning to it.
   Strongly disagree disagree agree strongly agree

4. How many times have you attended religious services during the past year? ____ times

5. Which of the following best describes your practice of prayer or religious meditation?
   a. Prayer or meditation is a regular part of my daily life.
   b. I usually pray or meditate in times of stress or need but rarely at any other time.
   c. I pray or meditate only during formal ceremonies.
   d. I never pray or meditate.

6. When you have a serious personal problem, how often do you take religious advice or teaching into consideration?
   a. Almost always
   b. Usually
   c. Sometimes
   d. Never

7. How much influence would you say that religion has on the way that you choose to act and the way that you choose to spend your time each day?
   a. No influence
   b. A small influence
   c. Some influence
   d. A fair amount of influence
   e. A large influence

Figure D.5: Ethical Decision-Making Questionnaire Page 5

8. Which of the following statements comes closest to your belief about God?
    a. I am sure that God exists and that God is active in my life.
    b. Although I sometimes question God's existence, I do believe in God and believe God knows of me as a person.
    c. I don't know if there is a personal God, but I do believe in a higher power of some kind.
    d. I don't know if there is a personal God or a higher power of some kind, and I don't know if I ever will.
    e. I don't believe in a personal God or in a higher power.

9. Which one of the following statements comes closest to your belief about life after death (immortality)?
    a. I believe in a personal life after death, a soul existing as a specific individual spirit.
    b. I believe in a soul existing after death as a part of a universal spirit.
    c. I believe in a life after death of some kind, but I really don't know what it would be like.
    d. I don't know whether there is any kind of life after death, and I don't know if I will ever know.
    e. I don't believe in any kind of life after death.

10. During the past year, how often have you experienced a feeling of religious reverence or devotion?
    a. Almost daily
    b. Frequently
    c. Sometimes
    d. Rarely
    e. Never

11. Do you agree with the following statement? "Religion gives me a great amount of comfort and security in my life."
    a. Strongly disagree
    b. Disagree
    c. Uncertain
    d. Agree
    e. Strongly Agree

Figure D.6: Ethical Decision-Making Questionnaire Page 6

Case Study II

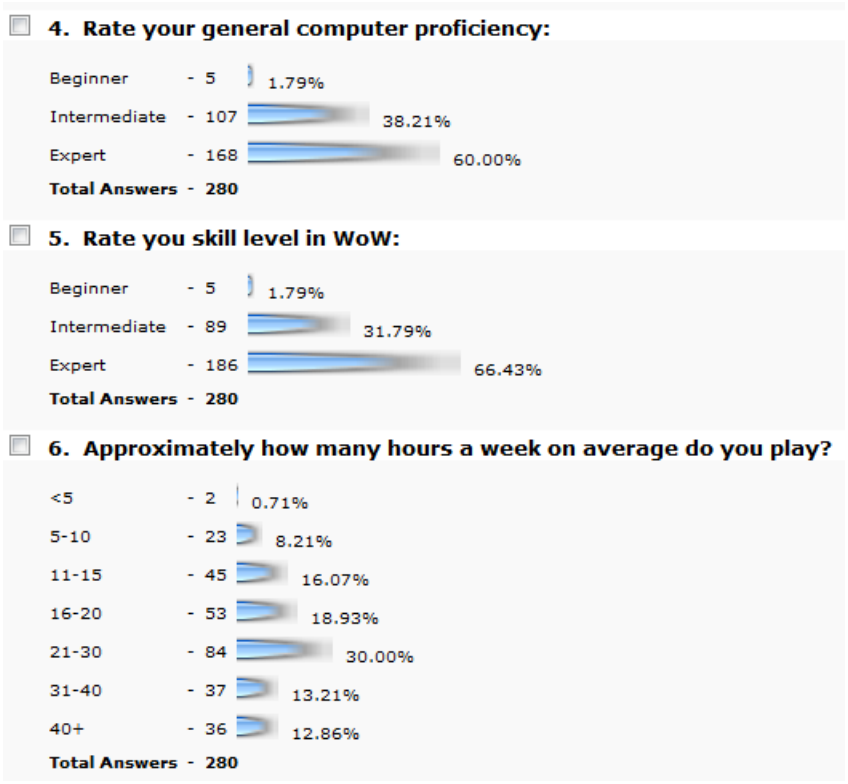


Figure D.7: WoW User Interface Study Questionnaire Questions 4 to 6

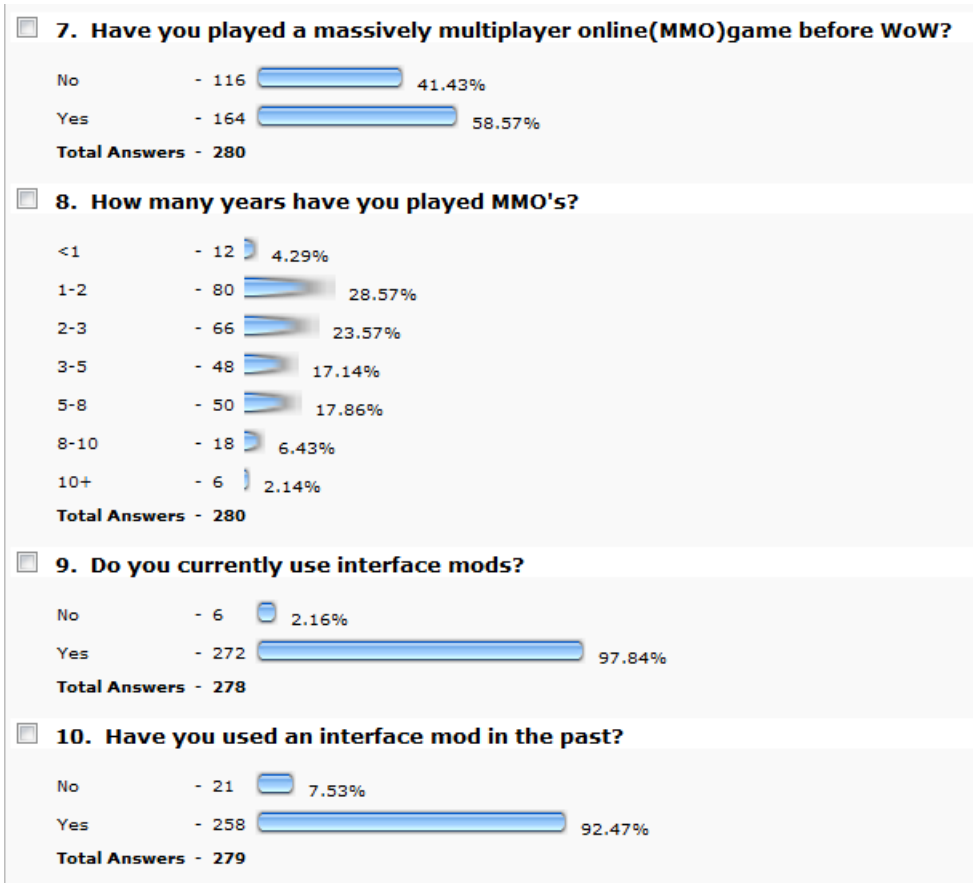


Figure D.8: WoW User Interface Study Questionnaire Questions 7 to 10

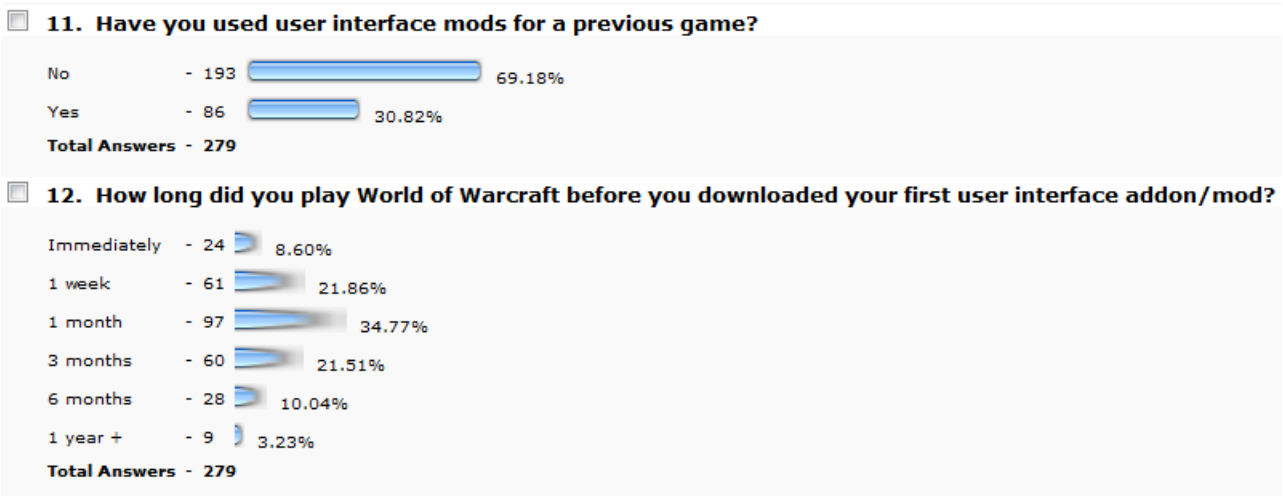


Figure D.9: WoW User Interface Study Questionnaire Questions 11 to 12