INFORMATION-THEORETICAL AND COMBINATORIAL METHODS IN DATA-MINING

A Dissertation Presented

by

Szymon Jaroszewicz

Submitted to the Office of Graduate Studies, University of Massachusetts Boston, in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2003

Computer Science Program

INFORMATION-THEORETICAL AND COMBINATORIAL METHODS IN DATA-MINING

A Dissertation Presented

by

Szymon Jaroszewicz

Approved as to style and content by:

Dan A. Simovici, Professor
Chairperson of Committee

William Campbell, Associate Professor
Member

Peter Fejer, Professor
Member

Lucila Ohno-Machado, Professor
Member

Dan A. Simovici, Program Director
Computer Science Program

Peter Fejer, Chairperson
Computer Science Department

ABSTRACT

INFORMATION-THEORETICAL AND COMBINATORIAL METHODS IN DATA-MINING

December 2003

Szymon Jaroszewicz,
M.Sc., Technical University of Szczecin
Ph.D., University of Massachusetts Boston

Directed by Professor Dan A. Simovici

Various applications of information theoretical and combinatorial methods in data mining are presented.

An axiomatization has been introduced for a family of entropies including both Shannon entropy and the Gini index as special cases. These entropies, and distances based on them, were then applied to decision tree construction. It has been shown experimentally that trees using distances based on generalized entropies as splitting criteria are smaller than those constructed using other criteria without significant loss in accuracy.

One of the major problems in association rule mining is the huge number of rules produced. This work contains contributions to two principal methods of addressing the problem: sorting rules based on some interestingness measure, and rule pruning. A new measure of rule interestingness is introduced generalizing three well-known measures: chi-squared, entropy gain and Gini gain, which moreover gives a whole family of intermediate measures with interesting properties. Also,

a method of pruning association rules using the Maximum Entropy Principle has been introduced. The usefulness of both methods is shown experimentally. It is worth mentioning that Maximum Entropy pruning gives a high reduction in the number of rules while retaining most of the interesting ones.

An idea has been suggested by H. Mannila to use supports of itemsets discovered with a data mining algorithm to obtain the size of arbitrary database queries. Here a solution to the problem is presented using a variation of the so called Bonferroni inequalities.

Modifications of Bonferroni inequalities have been developed which allow for obtaining bounds on sizes of arbitrary database queries based on supports of frequent itemsets. Special cases like estimating support of an unknown itemset, or of an itemset with negated attributes are also considered. Experiments show that useful bounds can be obtained from the inequalities in many significant cases.

To my parents.

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Dan Simovici for all the time spent with me preparing this dissertation, and for all the interesting research I did under his direction.

I wish to thank Professor Ethan Bolker for his kind help during my first semester of teaching, and Professors Peter Fejer, Bill Campbell, Dan Simovici, Carl Offner and George Lukas whose courses I had the pleasure to take.

I also wish to acknowledge the support received from the J. William Fulbright Scholarship Board as a visiting researcher at the University of Massachusetts Boston (grant no. IIE#:15984114-ICB).

I would also like to thank Prof. Jeffrey Burr from the University of Massachusetts at Boston Gerontology Center for providing us with the elderly people census data.

TABLE OF CONTENTS

1 INTRODUCTION **1**

1.1 The Need for Data Mining 1

1.2 Basic Notation 2

1.3 Typical Data Mining Tasks 3

1.3.1 Classification, Decision Trees 3

1.3.2 Association rules 4

1.4 Overview of the Dissertation 6

1.4.1 Generalized Entropy Distances with Application to Decision Tree Construction 7

1.4.2 Association rule pruning and interestingness 7

1.4.3 Bonferroni inequalities 8

2 AXIOMATIZATION OF GENERALIZED ENTROPY OF PARTITIONS WITH APPLICATION TO DECISION TREE INDUCTION **10**

2.1 Introduction and Basic Notations 10

2.2 An Axiomatization of Generalized Entropy 13

2.3 Axiomatization of non-Shannon Entropies 19

2.4 Axiomatization of Shannon Entropy 23

2.5 Conditional Entropy 23

2.6 Metrics on Partitions Induced by Generalized Entropies 25

2.7 Generalized Gain as a Selection Criterion for Splitting Attributes in Decision Trees . . . 32
2.8 Experimental Results . . . 34

3 ASSOCIATION RULE PRUNING AND INTERESTINGNESS 37

3.1 Introduction . . . 37
3.2 Basic notations . . . 38
3.3 Interestingness of Rules . . . 42
3.4 A General Measure of Rule Interestingness . . . 43
3.5 Properties of the General Measure of Interestingness . . . 45
3.6 Experimental results . . . 52
3.6.1 Synthetic data . . . 53
3.6.2 The mushroom database . . . 54
3.7 Pruning Redundant Association Rules Using Maximum Entropy Principle . . . 57
3.8 Interestingness of A Rule with Respect to A Set of Constraints . . . 60
3.9 The pruning algorithm . . . 61
3.10 Experimental Evaluation of the Pruning Algorithm . . . 68

4 MEASURES ON BOOLEAN POLYNOMIALS AND THEIR APPLICATIONS IN DATA MINING . . . 73

4.1 Introduction . . . 73
4.2 A Representation Result for MFBAs . . . 76
4.3 An Inclusion-Exclusion Principle for MFBAs . . . 78

4.4 Applications in Data Mining and Database Query Optimization . . 83
4.4.1 Accuracy of Inclusion-Exclusion Principle 83
4.4.2 Support in tables with missing values 86
4.5 Approximating Supports of Itemsets Using Bonferroni-type Inequalities 89
4.5.1 A Recursive Procedure for Computing Bonferroni Bounds from Frequent Itemsets 91
4.6 The Estimation Algorithm 95
4.6.1 Experimental results 95

5 CONCLUSIONS 103

5.1 Generalized Entropy Distances with Applications to Decision Tree Construction 103
5.2 Association rule pruning and interestingness 104
5.3 Bonferroni inequalities 105

REFERENCES 106

LIST OF TABLES

2.1 Accuracy Results 36

3.1 Rules on synthetic data ordered by $\Upsilon_{D_{\chi^2,a}}$ for different values of a. . 55

3.2 Rules on mushroom dataset ordered by $\Upsilon_{D_{\chi^2,a}}$ for different values of a. 56

3.3 Rules manually selected from the `lenses` database 69

3.4 Rules selected from the `lenses` database 70

3.5 Top 12 rules involving `urban` attribute generated from the elderly people census data 71

3.6 Numbers of rules and computation times for various datasets 72

4.1 Results for the `census` dataset 100

4.2 Influence of the order of inequalities on the bounds 102

4.3 Estimates for itemsets with negations 102

LIST OF FIGURES

1.1 Pseudocode of a general decision tree construction algorithm 4

2.1 Comparative Experimental Results 36

3.1 An lgorithm for finding l-nonredundant association rules. 63

4.1 Itemset support estimation algorithm. 96

4.2 Itemset support estimation algorithm, additional functions. 97

4.3 Discovered vs. total frequent itemsets for the `mushroom` dataset . . 99

4.4 Ratios of discovered to total frequent itemsets for the `census` data . 99

CHAPTER 1

INTRODUCTION

1.1 The Need for Data Mining

In recent years business and scientific research has seen an explosion in the amount of data collected. Supermarket chains routinely collect terabytes of data on purchases their customers make. Several scientific fields like astronomy or particle physics now have to deal with databases in the range of petabytes.

This data is useless unless it can be analyzed which, due to its huge size, is a very difficult task. It has been noted [FRB02] that as processor speed doubles every 18 month (according to Moore's law), the amount of data stored by companies doubles every year, so increase in computing power will not provide a solution and a new methodology is required.

This resulted in the creation of the data mining field. Early efforts in data mining concentrated on modifying Machine Learning algorithms to scale up better with the size of the datasets.

The first major algorithm developed specifically for large datasets came with the introduction in 1993 by Rakesh Agrawal et al. [AIS93] of the association rule mining problem. In [AIS93] the *Apriori* algorithm has been presented which is capable of finding all association rules satisfying certain criteria even in very large datasets. The paper has been followed by hundreds of publications further improving the

algorithm, and applying it in other areas such as clustering.

1.2 Basic Notation

We first introduce database related notation used throughout the dissertation. A database table is a triple $\tau = (T, H, \rho)$, where T is the name of the table, H is its heading, and ρ is its content. Elements of ρ will be called *tuples* or *records* of the table. Database tables will sometimes be called *datasets*.

Elements of H, the *attributes* of the table, will be denoted by uppercase letters $A, B, C, \ldots$ from the beginning of the alphabet, and occasionally with letters X and Y. Each attribute has associated with it a set $\mathrm{Dom}(A)$ called its *domain*. Sets of attributes $\{A_1, A_2, \ldots, A_k\}$ will often be written as $A_1A_2 \ldots A_k$ according to standard notation used in database literature. Subsets of H (*attribute sets*), will be denoted using uppercase Roman letters $I, J, K, L, \ldots$. The union of attribute sets I, J will usually be written as IJ. The domain of a set of attributes $A_1A_2 \ldots A_k$ is defined as

$$\mathrm{Dom}(A_1A_2 \ldots A_k) = \mathrm{Dom}(A_1) \times \mathrm{Dom}(A_2) \times \ldots \times \mathrm{Dom}(A_k). \tag{1.1}$$

The content of the table is a subset of the domain of its header: $\rho \subseteq \mathrm{Dom}(H)$. See [ST95] for more details on the terminology.

It is assumed that the tuples of table $\tau = (T, H, \rho)$ are samples taken from some probability distribution over $\mathrm{Dom}(H)$.

Following datamining terminology we will occasionally refer to binary attributes as *items* and to sets of items as *itemsets*.

1.3 Typical Data Mining Tasks

Typical data mining tasks relevant to this work will now be briefly characterized (see [WF00] for a thorough introduction).

1.3.1 Classification, Decision Trees

An important data mining task is *classification*, that is predicting the value of an attribute Y based on values of some other attributes $X_1, X_2, \ldots, X_n$. The task is achieved by first building a model for Y based on a dataset for which its values are known, the so called *training set*.

There are many known methods for classification, the most widely used in datamining are *decision trees* [Qui93]. In a decision tree each non-leaf node is labeled with a test, each branch with a possible test outcome, and leaf nodes are labeled with predicted outcomes.

A series of algorithms for building decision trees has been proposed by Quinlan [Qui93], the most important ones being ID3 and its successor C4.5. The main idea behind those algorithms is to recursively split the dataset based on an outcome of a newly chosen test. This is better described by the pseudocode in figure 1.1.

The algorithm is usually followed by a pruning stage which is beyond the scope of this introduction.

The two key details of the algorithm are on lines 2 (the *stopping criterion*) and 4 (the *splitting criterion*). More relevant from the point of view of this work is the splitting criterion. Typically splits are done based on a single attribute, and the attribute maximizing some information theoretical criterion is chosen. In [Qui93]

```
1. BuildTree(database τ):
2.    If in τ, Y is almost always equal to some y_0 ∈ Dom(Y)
3.      return a leaf labeled with y_0.
4.    Choose a test C
5.    Split τ into τ_1, τ_2, ..., τ_k based on the outcome of C.
6.    For every τ_i, let T_i = BuildTree(τ_i)
7.    Return a node labeled with C, with T_1, ..., T_k as subtrees.
```

Figure 1.1: Pseudocode of a general decision tree construction algorithm

the so called information gain is used:

$$\mathsf{gain}(Y, X) = H(Y) - H(Y|X),$$

where H denotes the Shannon entropy. Intuitively we select an attribute which gives us the most information about the predicted attribute Y (see further chapters or [Qui93] for details). To prevent attributes with a large number of values from being favored over those with small domains, a modification called *gain ratio* is often used:

$$\mathsf{gainratio}(Y, X) = \frac{\mathsf{gain}(Y, X)}{H(X)}.$$

1.3.2 Association rules

Another important data mining task relevant to this work is association rule mining. The method was originally designed to help analyze supermarket purchase data. Each item sold by the supermarket has a binary attribute assigned to it.

Each row in a table represents a transaction, with items purchased in the transaction having their corresponding attributes set to 1. We are then interested in finding facts like 'people who buy beer are also likely to buy diapers'. The methodology can of course be applied to other domains, and has been followed by hundreds of research papers extending it in various directions. Some more formal definitions are given below, full details can be found in [AIS93].

Let $\tau = (T, H, \rho)$ be a table whose heading H consists entirely of binary attributes. Each such attribute is called an *item*, subsets of H are called *itemsets*. The *support* of an itemset I is defined as

$$\text{supp}(I) = |\{t \in \rho \mid t_I = 1, 1, \ldots, 1\}|,$$

where t_I denotes the projection of t onto attribute set I. An itemset with support greater than some user specified threshold minsupp is called *frequent*.

An association rule is a pair of itemsets (I, J), usually denoted as $I \to J$. I is called the rule's *antecedent* and J its *consequent*. There are two important measures associated with association rules: *support* and *confidence* defined respectively as:

$$\text{supp}(I \to J) = \text{supp}(I \cup J),$$

and

$$\text{conf}(I \to J) = \frac{\text{supp}(I \cup J)}{\text{supp}(I)}.$$

The *association rule mining problem* is to find in a given table all association rules with support and confidence greater than some user specified thresholds minsupp and minconf.

A solution to the problem presented in [AIS93] consists of two stages:

1. Find all frequent itemsets (using minsupp as the minimum support)

2. Discover association rules with minimum confidence using frequent itemsets discovered in step 1.

Step 1 of the above procedure is much more difficult, and consequently has attracted much more attention in the literature. Given below is a sketch of the Apriori algorithm, one of the first solutions to the problem. See [AIS93] for a full description.

1. $C_1 = \{\{A\} \mid A \in H\}$
2. $k = 1$
3. Find frequent itemsets in C_k and place them in F_k.
4. C_{k+1} = $\{I \subseteq H \mid |I| = k+1$, all subsets of size k of I are in $F_k\}$
5. $k = k + 1$
6. goto 3

The main idea here is to search for frequent itemsets of increasing sizes, and pruning the infrequent itemsets using the fact that an itemset cannot be frequent unless all its subsets are frequent. The algorithm does at most $kmax$ passes over the database, where $kmax$ is the size of the largest frequent itemset.

1.4 Overview of the Dissertation

This work is concerned with information theoretical and combinatorial concepts applied to datamining. Contributions of this work are summarized below.

1.4.1 Generalized Entropy Distances with Application to Decision Tree Construction

In Chapter 2 we introduce an axiomatization of a family of entropies which includes both Shannon entropy and the Gini index as special cases. These entropies are then applied to decision tree construction.

As shown above, one of the criteria for splitting attribute selection in decision tree construction is Shannon entropy gain or the corresponding gain ratio. In [LGR93] it has been shown that the expression

$$H(Y|X) + H(X|Y), \tag{1.2}$$

where H is the Shannon Entropy is a distance between attribute sets. It has also been shown that using this distance as a splitting criterion during decision tree construction often leads to much smaller trees with almost no loss in accuracy.

In this work it is shown that formula (1.2) remains a distance if Shannon entropy is replaced with any of the generalized entropies. It is shown experimentally that generalized entropies are useful as a splitting criterion for building decision trees as well. Generalized entropies produced in many cases even smaller decision trees without significant loss in accuracy.

1.4.2 Association rule pruning and interestingness

One of the major problems in association rule mining is the huge number of rules produced, creating a secondary data mining problem. Indeed, even a toy `contact-lenses` database, consisting of 5 attributes and 24 rows, produces hundreds of association rules, most of them between independent attributes. There are two methods of dealing with this problem:

1. Sorting rules according to some interestingness measure. Of course, choosing the right interestingness measure is crucial for this method.

2. Pruning rules so that only those most interesting from the users' perspective are retained.

This work contains contributions in both directions. A new interestingness measure is given generalizing three important known measures: chi-squared, entropy gain and Gini gain. Second, a method of pruning association rules using the Maximum Entropy Principle is presented. Usefulness of both methods is shown experimentally. It is worth noting that Maximum Entropy pruning gives a high reduction in the number of rules while retaining most of the interesting ones.

1.4.3 Bonferroni inequalities

A series of seminal papers by H. Mannila et al. [MT96, Man01, PMS01] introduced the idea of using supports of itemsets discovered with a data mining algorithm to obtain the size of arbitrary database queries.

In this work a solution to the problem is presented using a variation of the so called Bonferroni inequalities [GS96]. Let $A_1, A_2, \ldots, A_n$ be events in some probability space. Bonferroni inequalities have the form

$$\sum_{t=0}^{2m+1} (-1)^t S_t \leq P(\neg A_1 \wedge \neg A_2 \wedge \ldots \wedge \neg A_n) \leq \sum_{t=0}^{2m} (-1)^t S_t,$$

for every $m \geq 0$, where

$$S_k = \sum P(A_{i_1} \wedge A_{i_2} \wedge \ldots \wedge A_{i_k}),$$

where summation is over all $1 \leq i_1 < i_2 < \ldots < i_k \leq n$, see [GS96] for details and applications.

Here variants of Bonferroni inequalities have been developed which allow for estimating bounds of arbitrary database queries based on supports of frequent itemsets. Special cases like estimating support of an itemset with an unknown support or of an itemset with negated attributes are also considered. Experiments show that useful bounds can be obtained from the inequalities in many important cases.

CHAPTER 2

AXIOMATIZATION OF GENERALIZED ENTROPY OF PARTITIONS WITH APPLICATION TO DECISION TREE INDUCTION

2.1 Introduction and Basic Notations

The notion of partition of a finite set is naturally linked to the notion of probability distribution. Namely, if A is a finite set and $\pi = \{B_1, \ldots, B_n\}$ is a partition of A, then the probability distribution attached to π is $(p_1, \ldots, p_n)$, where $p_i = \frac{|B_i|}{|A|}$ for $1 \leq i \leq n$. Thus, it is natural to consider the notion of entropy of a partition via the entropy of the corresponding probability distribution. Axiomatizations for entropy and entropy-like characteristics of probability distributions represent a problem with a rich history in information theory. Previous relevant work include the results of A.I. Khinchin [Khi57], D.K. Faddeev [Fad56], R.S. Ingarden and K. Urbanik [IU62] who investigated various axiomatizations of entropy, and Z. Daróczy who presented in [Dar70] an unified treatment of entropy-like characteristics of probability distributions using the notion of information function. An overview of axiomatizations of entropies of random variables can be found in [MR75].

The work is also related to [SR93, JS99], where we introduced an axiomatization for the notion of functional entropy. This numerical characteristic of functions is related to the complexity of circuits that realize the functions (cf.[CA90]) and serves as an estimate for power dissipation of a circuit realizing a function (cf.[HW97]). Information measures, especially conditional entropy of a logic function and its variables, have been used for minimization of logic functions (See [LGR93] and [CSS98]). It is also naturally related to the notion of entropy of partitions since every function $f : A \longrightarrow B$ defines a partition on its definition domain A whose blocks are $\{f^{-1}(b) \mid b \in \mathrm{Ran}(f)\}$.

In a different direction, starting from the notion of impurity of a set relative to a partition, a common generalization of Shannon entropy and of Gini index has been used in [D 00] for clustering of non-categorial data. P. A. Devijer used the Gini index in pattern recognition in [Dev74].

Partitions play a central role in classifications. Indeed, if a set of tuples T is described by attributes $a_1, \ldots, a_n$, then each set of attributes K defines a partition $\pi(K)$ of T, where two tuples belong to the same block of $\pi(K)$ if they have equal projections on K. Note that if $H \subseteq K$, then $\pi(K) \leq \pi(H)$ for any attribute sets H and K.

More on application of partitions to classification is presented in later sections of this chapter. First we discuss the notion of entropy of partitions.

All sets considered in this chapter are nonempty and finite unless stated explicitly otherwise. The sets $\mathbb{R}, \mathbb{R}_{\geq 0}, \mathbb{Q}, \mathbb{N}, \mathbb{N}_1$ denote the set of reals, the set of nonnegative reals, the set of rational numbers, the set of natural numbers, and the set $\{n \in \mathbb{N} \mid n \geq 1\}$, respectively. The domain and range of a function f are denoted by $\mathrm{Dom}(f)$ and $\mathrm{Ran}(f)$ respectively.

Definition 2.1.1 Let A be a set. A collection of nonempty, pairwise disjoint sets $\{B_1, B_2, \ldots, B_m\}$ such that

$$\bigcup_{i=1}^{m} B_i = A$$

is called a *partition* of A. ∎

Denote by $\mathsf{PART}(A)$ the set of partitions of the set A. The class of all partitions of finite sets is denoted by PART. The one-block partition of A is denoted by ω_A. The partition $\{\{a\} \mid a \in A\}$ is denoted by ι_A.

If $\pi, \pi' \in \mathsf{PART}(A)$, then $\pi \leq \pi'$ if every block of π is included in a block of π'. Clearly, for every $\pi \in \mathsf{PART}(A)$ we have $\iota_A \leq \pi \leq \omega_A$.

The partially ordered set $(\mathsf{PART}(A), \leq)$ is a lattice (see, for example a very lucid study of this lattice in [Ler81]). If $\sigma, \sigma' \in \mathsf{PART}(A)$, then σ' covers σ if $\sigma \leq \sigma'$ and, there is no partition $\sigma_1 \in \mathsf{PART}(A)$ such that $\sigma < \sigma_1 < \sigma'$. This is denoted by $\sigma \prec \sigma'$. It is easy to see that $\sigma \prec \sigma'$ if and only if σ' can be obtained from σ by fusing two of its blocks into a block of σ'.

If A, B are two disjoint sets, $\pi \in \mathsf{PART}(A)$, $\sigma \in \mathsf{PART}(B)$, where $\pi = \{A_1, \ldots, A_m\}$, $\sigma = \{B_1, \ldots, B_n\}$, then the partition $\pi + \sigma$ is the partition of $A \cup B$ given by

$$\pi + \sigma = \{A_1, \ldots, A_m, B_1, \ldots, B_n\}.$$

Whenever the "+" operation is defined it is easily seen to be associative. In other words, if A, B, C are pairwise disjoint sets, and $\pi \in \mathsf{PART}(A)$, $\sigma \in \mathsf{PART}(B)$, $\tau \in \mathsf{PART}(C)$, then $\pi + (\sigma + \tau) = (\pi + \sigma) + \tau$.

Observe that if A and B are disjoint, then $\iota_A + \iota_B = \iota_{A \cup B}$. Also, $\omega_A + \omega_B$ is the partition $\{A, B\}$ of the set $A \cup B$.

If $\pi = \{A_1, \ldots, A_m\}$, $\sigma = \{B_1, \ldots, B_n\}$ are partitions of two arbitrary sets, then we denote the partition $\{A_i \times B_j \mid 1 \leq i \leq m, 1 \leq j \leq n\}$ of $A \times B$ by $\pi \times \sigma$.

Note that $\iota_A \times \iota_B = \iota_{A \times B}$ and $\omega_A \times \omega_B = \omega_{A \times B}$.

Traditionally, the notion of Shannon entropy is introduced for a discrete random variable's distribution

$$X : \begin{pmatrix} x_1 & \cdots & x_n \\ p_1 & \cdots & p_n \end{pmatrix}$$

as $\mathcal{H}(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$.

We define the Shannon entropy of $\pi = \{B_1, \ldots, B_m\} \in \mathsf{PART}(A)$ as the Shannon entropy of the probability distribution $(\frac{|B_1|}{|A|}, \ldots, \frac{|B_m}{|A|})$ induced by π.

Here we present an axiomatization leading to the above formula, as well as many of its generalizations, based solely on partitions. It was first presented in [SJ02], where we introduced an axiomatization of a general notion of entropy for partitions of finite sets. Our system of axioms shows the common nature of Shannon entropy and of other measures of distribution concentration such as the Gini index.

2.2 An Axiomatization of Generalized Entropy

We introduce below a system of four axioms satisfied by several types of entropy-like characteristics of probability distributions.

Definition 2.2.1 Let $\beta \in \mathbb{R}$, $\beta > 0$, and let $\Phi : \mathbb{R}^2_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ be a continuous function such that $\Phi(x, y) = \Phi(y, x)$, $\Phi(x, 0) = x$ for $x, y \in \mathbb{R}$.

A (Φ, β)-*system of axioms for a partition entropy* $\mathcal{H} : \mathsf{PART} \longrightarrow \mathbb{R}_{\geq 0}$ consists of the following axioms:

(P1) If $\pi, \pi' \in \mathsf{PART}(A)$ are such that $\pi \leq \pi'$, then $0 \leq \mathcal{H}(\pi') \leq \mathcal{H}(\pi)$.

(P2) If A, B are two finite sets such that $|A| \leq |B|$, then $\mathcal{H}(\iota_A) \leq \mathcal{H}(\iota_B)$.

(P3) For all disjoint sets A, B and partitions $\pi \in \mathsf{PART}(A)$, and $\sigma \in \mathsf{PART}(B)$ we have:

$$\begin{aligned}&\mathcal{H}(\pi + \sigma)\\ &= \left(\frac{|A|}{|A|+|B|}\right)^{\beta} \mathcal{H}(\pi) + \left(\frac{|B|}{|A|+|B|}\right)^{\beta} \mathcal{H}(\sigma)\\ &\quad +\mathcal{H}(\{A, B\}).\end{aligned}$$

(P4) If $\pi \in \mathsf{PART}(A)$ and $\sigma \in \mathsf{PART}(B)$, then

$$\mathcal{H}(\pi \times \sigma) = \Phi(\mathcal{H}(\pi), \mathcal{H}(\sigma)).$$

□

Lemma 2.2.2 *For every (Φ, β)-entropy $\mathcal{H}$ and set A we have $\mathcal{H}(\omega_A) = 0$.*

Proof. Let A, B be two disjoint sets that have the same cardinality, $|A| = |B|$. Since $\omega_A + \omega_B$ is the partition $\{A, B\}$ of the set $A \cup B$, by Axiom **(P3)** we have

$$\mathcal{H}(\omega_A + \omega_B) = \left(\frac{1}{2}\right)^{\beta} (\mathcal{H}(\omega_A) + \mathcal{H}(\omega_B)) + \mathcal{H}(\{A, B\}),$$

which implies $\mathcal{H}(\omega_A) + \mathcal{H}(\omega_B) = 0$. Since $\mathcal{H}(\omega_A) \geq 0$ and $\mathcal{H}(\omega_B)) \geq 0$ (Axiom **(P1)**) it follows that $\mathcal{H}(\omega_A) = \mathcal{H}(\omega_B) = 0$. ■

Lemma 2.2.3 *Let A, B be two disjoint sets and let $\pi, \pi' \in \mathsf{PART}(A \cup B)$ be defined by $\pi = \sigma + \iota_B$ and $\pi' = \sigma + \omega_B$, where $\sigma \in \mathsf{PART}(A)$. Then,*

$$\mathcal{H}(\pi) = \mathcal{H}(\pi') + \left(\frac{|B|}{|A|+|B|}\right)^{\beta} \mathcal{H}(\iota_B).$$

Proof. By Axiom **(P3)** we can write:

$$\begin{aligned}\mathcal{H}(\pi) &= \left(\frac{|A|}{|A|+|B|}\right)^{\beta}\mathcal{H}(\sigma)\\ &\quad+\left(\frac{|B|}{|A|+|B|}\right)^{\beta}\mathcal{H}(\iota_B)+\mathcal{H}(\{A,B\}),\end{aligned}$$

and

$$\begin{aligned}\mathcal{H}(\pi') &= \left(\frac{|A|}{|A|+|B|}\right)^{\beta}\mathcal{H}(\sigma)\\ &\quad+\left(\frac{|B|}{|A|+|B|}\right)^{\beta}\mathcal{H}(\omega_B)+\mathcal{H}(\{A,B\})\\ &= \left(\frac{|A|}{|A|+|B|}\right)^{\beta}\mathcal{H}(\sigma)+\mathcal{H}(\{A,B\})\\ &\quad\text{(by Lemma 2.2.2).}\end{aligned}$$

The above equalities imply immediately the equality of the lemma. ∎

Theorem 2.2.4 *For every (Φ,β)-entropy, and partition $\pi=\{A_1,\ldots,A_n\}\in \mathsf{PART}(A)$ we have:*

$$\mathcal{H}(\pi)=\mathcal{H}(\iota_A)-\sum_{j=1}^{n}\left(\frac{|A_j|}{|A|}\right)^{\beta}\mathcal{H}(\iota_{A_j}).$$

Proof. Starting from the partition π consider the following sequence of partitions in $\mathsf{PART}(A)$:

$$\begin{aligned}\pi_0 &= \omega_{A_1}+\omega_{A_2}+\omega_{A_3}+\cdots+\omega_{A_n}\\ \pi_1 &= \iota_{A_1}+\omega_{A_2}+\omega_{A_3}+\cdots+\omega_{A_n}\\ \pi_2 &= \iota_{A_1}+\iota_{A_2}+\omega_{A_3}+\cdots+\omega_{A_n}\\ &\vdots\\ \pi_n &= \iota_{A_1}+\iota_{A_2}+\iota_{A_3}+\cdots+\iota_{A_n}.\end{aligned}$$

Let $\sigma_j = \iota_{A_1} + \cdots + \iota_{A_j} + \omega_{A_{j+2}} + \cdots + \omega_{A_n}$. Then, $\pi_j = \sigma_j + \omega_{A_{j+1}}$ and $\pi_{j+1} = \sigma_j + \iota_{A_{j+1}}$; therefore, by Lemma 2.2.3, we have

$$\mathcal{H}(\pi_{j+1}) = \mathcal{H}(\pi_j) + \left(\frac{|A_{j+1}|}{|A|}\right)^{\beta} \mathcal{H}(\iota_{A_{j+1}})$$

for $0 \leq j \leq n-1$.

A repeated application of this equality yields:

$$\mathcal{H}(\pi_n) = \mathcal{H}(\pi_0) + \sum_{j=0}^{n-1} \left(\frac{|A_{j+1}|}{|A|}\right)^{\beta} \mathcal{H}(\iota_{A_{j+1}}).$$

Observe that $\pi_0 = \pi$ and $\pi_n = \iota_A$. Consequently,

$$\mathcal{H}(\pi) = \mathcal{H}(\iota_A) - \sum_{j=1}^{n} \left(\frac{|A_j|}{|A|}\right)^{\beta} \mathcal{H}(\iota_{A_j}).$$

∎

Note that if A, B are two sets such that $|A| = |B| > 0$, then, by Axiom **(P2)**, we have $\mathcal{H}(\iota_A) = \mathcal{H}(\iota_B)$. Therefore, the value of $\mathcal{H}(\iota_A)$ depends only on the cardinality of A, and there exists a function $\mu : \mathbb{N}_1 \longrightarrow \mathbb{R}$ such that $\mathcal{H}(\iota_A) = \mu(|A|)$ for every nonempty set A. Axiom **(P2)** also implies that μ is an increasing function. We will refer to μ as the *core* of the (Φ, β)-system of axioms.

Corollary 2.2.5 *Let $\mathcal{H}$ be a (Φ, β)-entropy. For the function μ defined in Axiom* **(P2)** *and every partition $\pi = \{A_1, \ldots, A_n\} \in \mathsf{PART}(A)$ we have:*

$$\mathcal{H}(\pi) = \mu(|A|) - \sum_{j=1}^{n} \left(\frac{|A_j|}{|A|}\right)^{\beta} \mu(|A_j|). \tag{2.1}$$

Proof. The statement is an immediate consequence of Theorems 2.2.4. ∎

Theorem 2.2.6 *Let $\pi = \{B_1, \ldots, B_n\}$ be a partition of the set A. Define the partition π' obtained by fusing the blocks B_1 and B_2 of π as $\pi' = \{B_1 \cup B_2, B_3, \ldots, B_n\}$*

of the same set. Then

$$\mathcal{H}(\pi) = \mathcal{H}(\pi') + \left(\frac{|B_1 \cup B_2|}{|A|}\right)^{\beta} \mathcal{H}(\{B_1, B_2\}).$$

Proof. A double application of Corollary 2.2.5 yields:

$$\begin{aligned}\mathcal{H}(\pi') &= \mu(|A|) - \left(\frac{|B_1 \cup B_2|}{|A|}\right)^{\beta} \mu(|B_1 \cup B_2|) \\ &\quad - \sum_{i>2}^{n} \left(\frac{|B_i|}{|A|}\right)^{\beta} \mu(|B_i|)\end{aligned}$$

and

$$\begin{aligned}\mathcal{H}(\{B_1, B_2\}) &= \mu(|B_1 \cup B_2|) - \left(\frac{|B_1|}{|B_1 \cup B_2|}\right)^{\beta} \mu(|B_1|) \\ &\quad - \left(\frac{|B_2|}{|B_1 \cup B_2|}\right)^{\beta} \mu(|B_2|).\end{aligned}$$

Substituting the above expressions in

$$\mathcal{H}(\pi') + \left(\frac{|B_1 \cup B_2|}{|A|}\right)^{\beta} \mathcal{H}(\{B_1, B_2\})$$

we obtain $\mathcal{H}(\pi)$. ∎

Theorem 2.2.6 allows us to extend Axiom **(P3)**:

Corollary 2.2.7 *Let $A_1, \ldots, A_m$ be nonempty, disjoint sets and let $\pi_i \in \mathsf{PART}(A_i)$ for every $1 \leq i \leq m$. We have*

$$\begin{aligned}\mathcal{H}(\pi_1 + \cdots + \pi_m) &= \sum_{i=1}^{m} \left(\frac{|A_i|}{|A|}\right)^{\beta} \mathcal{H}(\pi_i) \\ &\quad + \mathcal{H}(\{A_1, \ldots, A_m\}),\end{aligned}$$

where $A = A_1 \cup \cdots \cup A_m$.

Proof. The argument is by induction on $m \geq 2$. The basis step, $m = 2$, is Axiom **(P3)**. Suppose that the statement holds for m and let $A_1, \ldots, A_m, A_{m+1}$

be $m+1$ disjoint sets. Further, suppose that $\pi_1, \ldots, \pi_m, \pi_{m+1}$ are partitions of these sets, respectively. Then, $\pi_m + \pi_{m+1}$ is a partition of the set $A_m \cup A_{m+1}$. By the inductive hypothesis we have

$$\begin{aligned}
&\mathcal{H}(\pi_1 + \cdots + (\pi_m + \pi_{m+1})) = \textstyle\sum_{i=1}^{m-1} \left(\frac{|A_i|}{|A|}\right)^{\beta} \mathcal{H}(\pi_i) \\
&+ \left(\tfrac{|A_m|+|A_{m+1}|}{|A|}\right)^{\beta} \mathcal{H}(\pi_m + \pi_{m+1}) \\
&+\mathcal{H}(\{A_1, \ldots, (A_m \cup A_{m+1})\}),
\end{aligned}$$

where $A = A_1 \cup \cdots \cup A_m \cup A_{m+1}$.

Axiom **(P3)** implies:

$$\begin{aligned}
&\mathcal{H}(\pi_1 + \cdots + (\pi_m + \pi_{m+1})) = \textstyle\sum_{i=1}^{m-1} \left(\frac{|A_i|}{|A|}\right)^{\beta} \mathcal{H}(\pi_i) \\
&+ \left(\tfrac{|A_m|}{|A|}\right)^{\beta} \mathcal{H}(\pi_m) + \left(\tfrac{|A_{m+1}|}{|A|}\right)^{\beta} \mathcal{H}(\pi_{m+1}) \\
&+ \left(\tfrac{|A_m|+|A_{m+1}|}{|A|}\right)^{\beta} \mathcal{H}\{A_m, A_{m+1}\} \\
&+\mathcal{H}(\{A_1, \ldots, (A_m \cup A_{m+1})\}).
\end{aligned}$$

Finally, an application of Theorem 2.2.6 gives the desired equality. ∎

Theorem 2.2.8 *Let μ be the core of a (Φ, β)-system. If $a, b \in \mathbb{N}_1$, then*

$$\mu(ab) - \frac{\mu(a)}{b^{\beta-1}} = \mu(b).$$

Proof. Let $A = \{x_1, \ldots, x_a\}$ and $B = \{y_1, \ldots, y_b\}$ be two nonempty sets. Note that $\omega_A \times \iota_B$ consists of b blocks of size a: $A \times \{y_1\}, \ldots, A \times \{y_b\}$. By Axiom **(P4)**,

$$\mathcal{H}(\omega_A \times \iota_B) = \Phi(\mathcal{H}(\omega_A), \mathcal{H}(\iota_B)) = \Phi(0, \mathcal{H}(\iota_B)) = \mathcal{H}(\iota_B).$$

On the other hand,

$$\begin{aligned}
\mathcal{H}(\omega_A \times \iota_B) &= \mathcal{H}(\iota_{A\times B}) - \sum_{i=1}^{b} \left(\frac{1}{b}\right)^{\beta} \mathcal{H}(\iota_{A\times\{y_i\}}) \\
&= \mu(ab) - \frac{1}{b^{\beta}} b \cdot \mu(a),
\end{aligned}$$

which gives the needed equality. ∎

An entropy is said to be *non-Shannon* if it defined by a (Φ, β)-system of axioms such that $\beta \neq 1$; otherwise, that is if $\beta = 1$, the entropy will be referred to as a *Shannon* entropy.

2.3 Axiomatization of non-Shannon Entropies

In this section we use axiomatization of generalized entropy to obtain an axiomatization for non-Shannon entropy.

Theorem 2.3.1 *Let $\mathcal{H}$ be a non-Shannon entropy defined by a (Φ, β)-system of axioms and let μ be the core of this system of axioms.*

There is a constant $c \in \mathbb{R}$ such that $c \cdot (\beta - 1) \geq 0$ and

$$\mu(a) = c \cdot \left(1 - \frac{1}{a^{\beta-1}}\right)$$

for $a \in \mathbb{N}_1$.

Proof. Theorem 2.2.8 implies that

$$\mu(ab) = \frac{\mu(a)}{b^{\beta-1}} + \mu(b) = \frac{\mu(b)}{a^{\beta-1}} + \mu(a),$$

for every $a, b \in \mathbb{N}_1$. Consequently,

$$\frac{\mu(a)}{1 - \frac{1}{a^{\beta-1}}} = \frac{\mu(b)}{1 - \frac{1}{b^{\beta-1}}},$$

for all $a, b \in \mathbb{N}_1$, which gives the desired equality. ∎

Note that for $\beta \neq 1$ we have:

$$c = \begin{cases} \lim_{a\to\infty} \mu(a) \text{ if } \beta > 1 \\ \lim_{a\to 0} \mu(a) \text{ if } \beta < 1. \end{cases} \tag{2.2}$$

Corollary 2.3.2 *If $\mathcal{H}$ is a non-Shannon entropy defined by a (Φ, β)-system of axioms, then there exists a constant $c \in \mathbb{R}$ such that for all $\pi \in \mathsf{PART}(A)$, where $\pi = \{A_1, \ldots, A_n\}$ the following equality holds*

$$\mathcal{H}(\pi) = c \cdot \left(1 - \sum_{j=1}^{n} \left(\frac{|A_j|}{|A|}\right)^{\beta}\right). \tag{2.3}$$

Proof. By Corollary 2.2.5 and by Theorem 2.3.1 we have

$$\begin{aligned}
&\mathcal{H}(\pi)\\
&= \mu(|A|) - \sum_{j=1}^{n} \left(\frac{|A_j|}{|A|}\right)^{\beta} \mu(|A_j|)\\
&= c \cdot \left(1 - \frac{1}{|A|^{\beta-1}}\right) - c \cdot \sum_{j=1}^{n} \left(\frac{|A_j|}{|A|}\right)^{\beta} \cdot \left(1 - \frac{1}{|A_j|^{\beta-1}}\right)\\
&= c \cdot \left(1 - \frac{1}{|A|^{\beta-1}}\right) - c \cdot \sum_{j=1}^{n} \left(\frac{|A_j|}{|A|}\right)^{\beta} + c \cdot \sum_{j=1}^{n} \frac{|A_j|}{|A|^{\beta}}\\
&= c \cdot \left(1 - \sum_{j=1}^{n} \left(\frac{|A_j|}{|A|}\right)^{\beta}\right).
\end{aligned}$$

∎

Theorem 2.3.3 *Let $\mathcal{H}$ be the non-Shannon entropy defined by a (Φ, β)-system and let c be as defined by Equality (2.2), where μ is the core of the (Φ, β)-system of axioms. The function Φ introduced by Axiom* **(P4)** *is given by $\Phi(x, y) = x + y - \frac{1}{c}xy$ for $x, y \in [0, c]$ when $\beta > 1$ and for $x, y \in [0, \infty)$ when $\beta < 1$.*

Proof. Let $\pi = \{A_1, \ldots, A_n\} \in \mathsf{PART}(A)$ and $\sigma = \{B_1, \ldots, B_m\} \in \mathsf{PART}(B)$ be two partitions. Since

$$\begin{aligned}
\sum_{j=1}^{n} \left(\frac{|A_j|}{|A|}\right)^{\beta} &= 1 - \frac{1}{c}\mathcal{H}(\pi)\\
\sum_{k=1}^{m} \left(\frac{|B_k|}{|B|}\right)^{\beta} &= 1 - \frac{1}{c}\mathcal{H}(\sigma)
\end{aligned}$$

we can write:

$$\begin{aligned}\mathcal{H}(\pi\times\sigma) &= c\left(1-\sum_{j=1}^{n}\sum_{k=1}^{m}\left(\frac{|A_j||B_k|}{|A||B|}\right)^{\beta}\right)\\ &= c\left(1-\left(1-\frac{1}{c}\mathcal{H}(\pi)\right)\left(1-\frac{1}{c}\mathcal{H}(\sigma)\right)\right)\\ &= \mathcal{H}(\pi)+\mathcal{H}(\sigma)-\frac{1}{c}\mathcal{H}(\pi)\mathcal{H}(\sigma).\end{aligned}$$

Let us now show that the set of values of entropies is dense in the required intervals. For a given n take rational n-dimensional vectors $s=(1,0,\ldots,0)$ and $t=(\frac{1}{n},\ldots,\frac{1}{n})$. Let $\alpha\in\mathbb{Q}$, $0\le\alpha\le 1$. Denote $r_\alpha=(1-\alpha)s+\alpha t=\left(1-\alpha+\frac{\alpha}{n},\frac{\alpha}{n},\ldots,\frac{\alpha}{n}\right)$. Consider a function

$$f(\alpha)=c\left[1-\sum_{i=0}^{n}r_{\alpha,i}^{\beta}\right]=c\left[1-\left(1-\alpha+\frac{\alpha}{n}\right)^{\beta}-\frac{n-1}{n^{\beta}}\right],$$

where $r_{\alpha,i}$ is the i-th coordinate of r_α. It is easy to see that it is a continuous function of α, and its range is $[0,c(1-n^{\beta-1})]$. Since the set of rational numbers is dense and $f(\alpha)$ is continuous in α, values of entropies are dense in $[0,c(1-n^{1-\beta})]$. Since n can be chosen arbitrarily large, the values of entropies are dense in the interval $[0,c)$ when $\beta>1$ and in the interval $[0,\infty)$ when $\beta<1$. The density of the set of values of entropies and the continuity of Φ implies the desired form of Φ on the required intervals. ∎

Choosing $c=\frac{1}{\beta-1}$ in the equality (2.3) we obtain the Havrda-Charvát entropy (see [KK92]):

$$\mathcal{H}_\beta(\pi)=\frac{1}{\beta-1}\cdot\left(1-\sum_{j=1}^{n}\left(\frac{|A_j|}{|A|}\right)^{\beta}\right).$$

The limit case, $\lim_{\beta\to 1}\mathcal{H}_\beta(\pi)$ yields the Shannon entropy. The case $\beta=1$ is considered independently in the next section.

If $\beta = 2$ we obtain the Gini index,

$$\mathcal{H}_2(\pi) = 1 - \sum_{j=1}^{n} \left(\frac{|A_j|}{|A|} \right)^2,$$

which is widely used in machine learning and data mining.

Theorem 2.3.4 *If $c(\beta - 1) \geq 0$, then for all $\beta \neq 1$, the expression (2.3) satisfies the Axioms* **(P1), (P2), (P3), (P4)**.

Proof. Let us first prove that for two partitions π and π', $\pi \leq \pi'$ implies $\mathcal{H}(\pi') \leq \mathcal{H}(\pi)$. It suffices to show the result for $\pi \prec \pi'$. Without loss of generality assume that $\pi = \{A_1, A_2, \ldots, A_k\}$, $\pi' = \{A_1, A_2, \ldots, A_{k-1} \cup A_k\}$. We have

$$\begin{aligned}
\mathcal{H}(\pi) - \mathcal{H}(\pi') &= c \left[\left(\frac{|A_{k-1}| + |A_k|}{|A|} \right)^\beta - \left(\frac{|A_{k-1}|}{|A|} \right)^\beta - \left(\frac{|A_k|}{|A|} \right)^\beta \right] \\
&= c \left(\frac{|A_{k-1}| + |A_k|}{|A|} \right)^\beta \left[\left(\frac{|A_{k-1}| + |A_k|}{|A_{k-1}| + |A_k|} \right)^\beta \right. \\
&\quad \left. - \left(\frac{|A_{k-1}|}{|A_{k-1}| + |A_k|} \right)^\beta - \left(\frac{|A_k|}{|A_{k-1}| + |A_k|} \right)^\beta \right].
\end{aligned}$$

The non-negativity of the above expression follows from concavity of cx^β when $c(\beta - 1) \geq 0$. It is easy to see that $\mathcal{H}(\omega_A) = 0$ which implies non-negativity of $\mathcal{H}$ and proves that expression (2.3) satisfies **(P1)**.

The case of **(P2)** follows from the fact that $\mathcal{H}(\iota_A) = c(1 - |A|^{1-\beta})$ is an increasing function of $|A|$ when $c(\beta - 1) \geq 0$, $\beta \neq 1$.

(P3) can easily be verified using elementary algebraical transformations, and **(P4)** follows from the equality $\mathcal{H}(\pi \times \sigma) = \mathcal{H}(\pi) + \mathcal{H}(\sigma) - \frac{1}{c}\mathcal{H}(\pi)\mathcal{H}(\sigma)$ proved earlier. ∎

2.4 Axiomatization of Shannon Entropy

When $\beta = 1$, by Theorem 2.2.8, we have

$$\mu(ab) = \mu(a) + \mu(b)$$

for $a, b \in \mathbb{N}_1$. If $\eta : \mathbb{N}_1 \longrightarrow \mathbb{R}$ is the function defined by $\eta(a) = a\mu(a)$ for $a \in \mathbb{N}_1$, then η is clearly an increasing function and we have

$$\eta(ab) = ab\mu(ab) = b\eta(a) + a\eta(b)$$

for $a, b \in \mathbb{N}_1$. By Theorem A.6 from [SR93], there exists a constant $c \in \mathbb{R}$ such that $\eta(a) = ca \log_2 a$ for $a \in \mathbb{N}_1$, so $\mu(a) = c \log_2(a)$. Then, equation (2.1) implies:

$$\mathcal{H}(\pi) = c \cdot \sum_{i=1}^{n} \frac{a_i}{a} \log_2 \frac{a_i}{a},$$

for every partition $\pi = \{A_1, \ldots, A_n\}$ of a set A, where $|A_i| = a_i$ for $1 \leq i \leq n$, and $|A| = a$. Taking $c = -1$ gives exactly the expression for Shannon's entropy.

The continuous function Φ is determined, as in the previous case. If $\pi = \{A_1, \ldots, A_n\} \in \mathsf{PART}(A)$ and $\sigma = \{B_1, \ldots, B_m\} \in \mathsf{PART}(B)$, then we must have

$$\mathcal{H}(\pi \times \sigma) = \mathcal{H}(\pi) + \mathcal{H}(\sigma) = \Phi(\mathcal{H}(\pi), \mathcal{H}(\sigma)).$$

An argument similar to the proof of Theorem 2.3.3 shows that the set of values of Shannon entropies is dense in the interval $[0, \infty)$, which combined with the continuity of Φ implies $\Phi(x, y) = x + y$.

2.5 Conditional Entropy

The entropies previously introduced generate corresponding conditional entropies.

Let $\pi \in \mathsf{PART}(A)$ and let $C \subseteq A$. Denote by π_C the "trace" of π on C given by

$$\pi_C = \{B \cap C | B \in \pi \text{ such that } B \cap C \neq \emptyset\}$$

Clearly, $\pi_C \in \mathsf{PART}(C)$; also, if C is a block of π, then $\pi_C = \omega_C$.

Definition 2.5.1 The *conditional entropy* defined by the (Φ, β)-entropy $\mathcal{H}$ is the function $\mathcal{C} : \mathsf{PART}^2 \longrightarrow \mathbb{R}_{\geq 0}$ given by:

$$\mathcal{C}(\pi, \sigma) = \sum_{j=1}^{n} \frac{|C_j|}{|A|} \cdot \mathcal{H}(\pi_{C_j}),$$

where $\pi, \sigma \in \mathsf{PART}(A)$, $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$. □

We denote the value of $\mathcal{C}(\pi, \sigma)$ by $\mathcal{H}(\pi|\sigma)$. Note that $\mathcal{H}(\pi|\omega_A) = \mathcal{H}(\pi)$.

The partition $\pi \wedge \sigma$ whose blocks consist of the nonempty intersections of the blocks of π and σ can be written as

$$\pi \wedge \sigma = \pi_{C_1} + \cdots + \pi_{C_n} = \sigma_{B_1} + \cdots + \sigma_{B_m}.$$

Notice that $\pi \wedge \sigma$ is the infimum of partitions π and σ in the lattice $(\mathsf{PART}(A), \leq)$.

Therefore, by Corollary 2.2.7, we have:

$$\mathcal{H}(\pi \wedge \sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|A|}\right)^{\beta} \mathcal{H}(\pi_{C_j}) + \mathcal{H}(\sigma).$$

For those entropies with $\beta > 1$ we have

$$\mathcal{H}(\pi \wedge \sigma) \leq \mathcal{H}(\pi|\sigma) + \mathcal{H}(\sigma), \tag{2.4}$$

while for those having $\beta < 1$, the reverse inequality holds. In the case of Shannon entropy, $\beta = 1$ and

$$\begin{aligned} \mathcal{H}(\pi \wedge \sigma) &= \mathcal{H}(\pi|\sigma) + \mathcal{H}(\sigma) && (2.5) \\ &= \mathcal{H}(\sigma|\pi) + \mathcal{H}(\pi). && (2.6) \end{aligned}$$

If $\mathcal{H}$ is a (Φ, β)-entropy, $\pi, \sigma \in \mathsf{PART}(A)$ are such that $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$, then the conditional entropy $\mathcal{H}(\pi|\sigma)$ is given by:

$$\begin{aligned}\mathcal{H}(\pi|\sigma) &= \sum_{j=1}^{n} \frac{|C_j|}{|A|}\mu(|C_j|) \\ &\quad - \sum_{j=1}^{n}\sum_{i=1}^{m} \frac{|B_i \cap C_j|^\beta}{|A| \cdot |C_j|^{\beta-1}}\mu(|B_i \cap C_j|).\end{aligned}$$

This equality follows immediately from Corollary 2.2.5.

In the case of Shannon entropy, taking $\beta = 1$ and $\mu(n) = \log_2 n$ we obtain the well-known expression of conditional entropy:

$$\mathcal{H}(\pi|\sigma) = -\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{|B_i \cap C_j|}{|A|} \log_2 \frac{|B_i \cap C_j|}{|C_j|}.$$

In the case of the Gini index we have $\beta = 2$ and $\mu(a) = c\left(1 - \frac{1}{a}\right)$ for $a \in \mathbb{N}_1$. Consequently, after some elementary transformations, the conditional Gini index is:

$$\mathcal{H}(\pi|\sigma) = 1 - \sum_{j=1}^{n}\sum_{i=1}^{m} \frac{|B_i \cap C_j|^2}{|A| \cdot |C_j|}.$$

2.6 Metrics on Partitions Induced by Generalized Entropies

Using the axioms above and choosing $c = \frac{1}{\beta-1}$ we obtain a family of entropies

$$\mathcal{H}_\beta(\pi) = \frac{1}{\beta - 1}\left(1 - \sum_{j=1}^{n}\left(\frac{|A_j|}{|A|}\right)^\beta\right),$$

for $\beta \neq 1$, and Shannon entropy for $\beta = 1$ (limiting case). In this section we prove that the family can be used to introduce a metric on the partitions of a set.

A direct consequence of the Axioms is that $\mathcal{H}(\omega_A) = 0$ for any set A (Lemma 2.2.2). The following reciprocal result also holds:

Lemma 2.6.1 *Let A be a finite set and let $\pi \in \mathsf{PART}(A)$ such that $\mathcal{H}_\beta(\pi) = 0$. Then, $\pi = \omega_A$.*

Proof. Suppose that $\mathcal{H}_\beta(\pi) = 0$ but $\pi < \omega_A$. Then, there exists a block C of π such that $\emptyset \subset C \subset A$. If $\theta = \{C, A - C\}$, then clearly we have $\pi \leq \theta$, so $0 \leq \mathcal{H}_\beta(\theta) \leq \mathcal{H}_\beta(\pi)$, which implies $\mathcal{H}_\beta(\theta) = 0$. If $\beta > 1$, then

$$\mathcal{H}_\beta(\theta) = c\left(1 - \left(\frac{|C|}{|A|}\right)^\beta - \left(\frac{|A-C|}{|A|}\right)^\beta\right) = 0.$$

The concavity of the function $f(x) = x^\beta + (1-x)^\beta$ on $[0,1]$ (when $\beta > 1$) implies either $C = A$ or $C = \emptyset$, which is a contradiction. Thus, $\pi = \omega_A$. A similar argument works for the other cases. ∎

Theorem 2.6.2 *Let A be a finite set and let $\pi, \sigma \in \mathsf{PART}(A)$. We have $\mathcal{H}_\beta(\pi|\sigma) = 0$ if and only if $\sigma \leq \pi$.*

Proof. Suppose that $\sigma = \{C_1, \ldots, C_n\}$. If $\sigma \leq \pi$, then $\pi_{C_j} = \omega_{C_j}$ for $1 \leq j \leq n$, so $\mathcal{H}_\beta(\pi|\sigma) = 0$. Conversely, suppose that

$$\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{n} \frac{|C_j|}{|A|} \cdot \mathcal{H}_\beta(\pi_{C_j}) = 0.$$

This implies $\mathcal{H}_\beta(\pi_{C_j}) = 0$ for $1 \leq j \leq n$, so $\pi_{C_j} = \omega_{C_j}$ for $1 \leq j \leq n$ by Lemma 2.6.1. This means that every block C_j of σ is included in a block of π, which implies $\sigma \leq \pi$. ∎

Lemma 2.6.3 *Let $a, b \in [0,1]$ such that $a + b = 1$. Then, for $\beta \geq 1$ we have:*

$$\sum_{i=1}^{n} (ax_i + by_i)^\beta \leq a \sum_{i=1}^{n} x_i^\beta + b \sum_{i=1}^{n} y_i^\beta,$$

for every $x_1, \ldots, x_n, y_1, \ldots, y_n \in [0,1]$. For $\beta \leq 1$, the reverse inequality holds.

Proof. The statement follows immediately from concavity of the function $f(x) = x^\beta$ for $\beta > 1$ on the interval $[0, 1]$. ∎

Theorems 2.6.4 and 2.6.7 extend well-known monotonicity properties of Shannon entropy.

Theorem 2.6.4 *If π, σ, σ' are partitions of the finite set A such that $\sigma \leq \sigma'$, then $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi|\sigma')$ for $\beta > 0$.*

Proof. To prove this statement it suffices to consider only the case when $\sigma \prec \sigma'$. Suppose initially that $\beta > 1$.

Let $\sigma, \sigma' \in \mathsf{PART}(A)$ such that $\sigma \prec \sigma'$. Suppose that D, E are blocks of σ such that $C = D \cup E$, where C is a block of σ'; the partition π is $\{B_1, \ldots, B_n\}$.

Define $x_i = \frac{|B_i \cap D|}{|D|}$ and $y_i = \frac{|B_i \cap E|}{|E|}$ for $1 \leq i \leq n$. If we choose $a = \frac{|D|}{|C|}$ and $b = \frac{|E|}{|C|}$, then

$$|C| \sum_{i=1}^{n} \frac{|B_i \cap C|^\beta}{|C|^\beta} \leq |D| \sum_{i=1}^{n} \frac{|B_i \cap D|^\beta}{|D|^\beta} + |E| \sum_{i=1}^{n} \frac{|B_i \cap E|^\beta}{|E|^\beta},$$

by Lemma 2.6.3. Consequently, we can write:

$$\begin{aligned}
\mathcal{H}_\beta(\pi|\sigma) &= \cdots + \frac{|D|}{|A|}\mathcal{H}_\beta(\pi_D) + \frac{|E|}{|A|}\mathcal{H}_\beta(\pi_E) + \cdots \\
&= \cdots + \frac{|D|}{|A|}\left(1 - \sum_{i=1}^{n} \frac{|B_i \cap D|^\beta}{|D|^\beta}\right) + \frac{|E|}{|A|}\left(1 - \sum_{i=1}^{n} \frac{|B_i \cap E|^\beta}{|E|^\beta}\right) + \cdots \\
&\leq \cdots + \frac{|C|}{|A|}\left(1 - \sum_{i=1}^{n} \frac{|B_i \cap C|^\beta}{|C|^\beta}\right) + \cdots = \mathcal{H}_\beta(\pi|\sigma').
\end{aligned}$$

For $\beta < 1$ we have

$$|C| \sum_{i=1}^{n} \frac{|B_i \cap C|^\beta}{|C|^\beta} \geq |D| \sum_{i=1}^{n} \frac{|B_i \cap D|^\beta}{|D|^\beta} + |E| \sum_{i=1}^{n} \frac{|B_i \cap E|^\beta}{|E|^\beta},$$

by the second part of Lemma 2.6.3. Thus,

$$\begin{aligned}\mathcal{H}_\beta(\pi|\sigma) &= \cdots + \frac{|D|}{|A|}\mathcal{H}_\beta(\pi_D) + \frac{|E|}{|A|}\mathcal{H}_\beta(\pi_E) + \cdots \\ &= \cdots + \frac{|D|}{|A|}\left(\sum_{i=1}^{n}\frac{|B_i \cap D|^\beta}{|D|^\beta} - 1\right) + \frac{|E|}{|A|}\left(\sum_{i=1}^{n}\frac{|B_i \cap E|^\beta}{|E|^\beta} - 1\right) + \cdots \\ &\leq \cdots + \frac{|C|}{|A|}\left(\sum_{i=1}^{n}\frac{|B_i \cap C|^\beta}{|C|^\beta} - 1\right) + \cdots = \mathcal{H}_\beta(\pi|\sigma').\end{aligned}$$

For $\beta = 1$ the inequality is a well-known property of Shannon entropy. ∎

Corollary 2.6.5 *For every $\pi, \sigma \in \mathsf{PART}(A)$ and $\beta > 0$, we have $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi)$.*

Proof. Since $\sigma \leq \omega_A$, by Theorem 2.6.4 we have $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi|\omega_A) = \mathcal{H}_\beta(\pi)$. ∎

Corollary 2.6.6 *Let A be a finite set. For $\beta \geq 1$ we have $\mathcal{H}_\beta(\pi \wedge \sigma) \leq \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma)$ for every $\pi, \sigma \in \mathsf{PART}(A)$.*

Proof. By Inequality (2.4) and by Corollary 2.6.5 we have

$$\mathcal{H}_\beta(\pi \wedge \sigma) \leq \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma) \leq \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma).$$

∎

Theorem 2.6.7 *If π, π', σ are partitions of the finite set A such that $\pi \leq \pi'$, then $\mathcal{H}_\beta(\pi|\sigma) \geq \mathcal{H}_\beta(\pi'|\sigma)$.*

Proof. Suppose that $\sigma = \{C_1, \ldots, C_n\}$. Then, it is clear that $\pi_{C_j} \leq \pi'_{C_j}$ for $1 \leq j \leq n$. Therefore, $\mathcal{H}_\beta(\pi_{C_j}) \geq \mathcal{H}_\beta(\pi'_{C_j})$ by Axiom **(P1)**, which implies immediately the desired inequality. ∎

Lemma 2.6.8 *Let A be a nonempty set and let $\{A', A''\}$ be a two-block partition of A. If $\pi \in \mathsf{PART}(A)$, $\sigma' \in \mathsf{PART}(A')$, and $\sigma'' \in \mathsf{PART}(A'')$, then*

$$\mathcal{H}_\beta(\pi|\sigma' + \sigma'') = \frac{|A'|}{|A|}\mathcal{H}_\beta(\pi'|\sigma') + \frac{|A''|}{|A|}\mathcal{H}_\beta(\pi''|\sigma''),$$

where $\pi' = \pi_{A'}$ and $\pi'' = \pi_{A''}$.

Proof. Note that $\sigma' + \sigma''$ is a partition of A. The lemma follows immediately from the definition of conditional entropy. ∎

Theorem 2.6.9 *Let A be a nonempty set and let $\{A_1, \ldots, A_\ell\}$ be a partition of A. If $\pi \in \mathsf{PART}(A)$, $\sigma_k \in \mathsf{PART}(A_k)$ for $1 \leq k \leq \ell$, then*

$$\mathcal{H}_\beta(\pi|\sigma_1 + \cdots + \sigma_\ell) = \sum_{k=1}^{\ell} \frac{|A_k|}{|A|}\mathcal{H}_\beta(\pi_k|\sigma_k)$$

where $\pi_k = \pi_{A_k}$ for $1 \leq k \leq \ell$.

Proof. The result follows immediately from Lemma 2.6.8 due to the associativity of the partial operation "+". ∎

Theorem 2.6.10 *If $\beta > 1$, then for every three partitions π, σ, τ of a finite set A we have*

$$\mathcal{H}_\beta(\pi|\sigma \wedge \tau) + \mathcal{H}_\beta(\sigma|\tau) \geq \mathcal{H}_\beta(\pi \wedge \sigma|\tau).$$

If $\beta < 1$ we have the reverse inequality, and for $\beta = 1$ we have the equality

$$\mathcal{H}_\beta(\pi|\sigma \wedge \tau) + \mathcal{H}_\beta(\sigma|\tau) = \mathcal{H}_\beta(\pi \wedge \sigma|\tau).$$

Proof. Suppose that $\pi = \{B_1, \ldots, B_m\}$, $\sigma = \{C_1, \ldots, C_n\}$, and $\tau = \{D_1, \ldots, D_\ell\}$. We noted already that $\sigma \wedge \tau = \sigma_{D_1} + \cdots + \sigma_{D_\ell} = \tau_{C_1} + \cdots + \tau_{C_n}$. Consequently,

by Theorem 2.6.9 we have $\mathcal{H}_\beta(\sigma \wedge \pi) = \sum_{k=1}^{\ell} \frac{|D_k|}{|A|} \mathcal{H}_\beta(\pi_{D_k}|\sigma_{D_k})$. Also, we have $\mathcal{H}_\beta(\sigma|\tau) = \sum_{k=1}^{\ell} \frac{|D_k|}{|A|} \mathcal{H}_\beta(\sigma_{D_k})$.

If $\beta > 1$ we saw that $\mathcal{H}_\beta(\pi_{D_k} \wedge \sigma_{D_k}) \leq \mathcal{H}_\beta(\pi_{D_k}|\sigma_{D_k}) + \mathcal{H}_\beta(\sigma_{D_k})$, for every k, $1 \leq k \leq \ell$, which implies

$$\begin{aligned} \mathcal{H}_\beta(\sigma \wedge \pi) + \mathcal{H}_\beta(\sigma|\tau) &\geq \sum_{k=1}^{\ell} \frac{|D_k|}{|A|} \mathcal{H}_\beta(\pi_{D_k} \wedge \sigma_{D_k}) \\ &= \sum_{k=1}^{\ell} \frac{|D_k|}{|A|} \mathcal{H}_\beta((\pi \wedge \sigma)_{D_k}) \\ &= \mathcal{H}_\beta(\pi \wedge \sigma|\tau). \end{aligned}$$

Using a similar argument we obtain the second inequality of the theorem. The equality for the Shannon case was obtained in [Man91]. ∎

Corollary 2.6.11 *Let A be a finite set. For $\beta \geq 1$ and for $\pi, \sigma, \tau \in \mathsf{PART}(A)$ we have the inequality: $\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\tau) \geq \mathcal{H}_\beta(\pi|\tau)$.*

Proof. Note that by Theorem 2.6.4 we have: $\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\tau) \geq \mathcal{H}_\beta(\pi|\sigma \wedge \tau) + \mathcal{H}_\beta(\sigma|\tau)$. Therefore, for $\beta \geq 1$, by Theorems 2.6.10 and 2.6.7 we obtain $\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\tau) \geq \mathcal{H}_\beta(\pi \wedge \sigma|\tau) \geq \mathcal{H}_\beta(\pi|\tau)$. ∎

Definition 2.6.12 Let $\beta > 1$. The mapping $d_\beta : \mathsf{PART}(A)^2 \longrightarrow \mathbb{R}_{\geq 0}$ is defined by $d_\beta(\pi, \sigma) = \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi)$ for $\pi, \sigma \in \mathsf{PART}(A)$. □

The following result generalizes a result of López de Mántaras ([Man91]):

Corollary 2.6.13 *d_β is a metric on $\mathsf{PART}(A)$.*

Proof. If $d_\beta(\pi, \sigma) = 0$, then $\mathcal{H}_\beta(\pi|\sigma) = \mathcal{H}_\beta(\sigma|\pi) = 0$. Therefore, by Theorem 2.6.2 we have $\sigma \leq \pi$ and $\pi \leq \sigma$, so $\pi = \sigma$. The symmetry of d_β is immediate. The triangular property is a direct consequence of Corollary 2.6.11. ∎

In [Man91] it is shown that the mapping $e_1 : \mathsf{PART}(A)^2 \longrightarrow \mathbb{R}_{\geq 0}$ that corresponds to Shannon entropy, defined by

$$e_1(\pi, \sigma) = \frac{d_1(\pi, \sigma)}{\mathcal{H}_1(\pi \wedge \sigma)}$$

for $\pi, \sigma \in \mathsf{PART}(A)$ is also a metric on $\mathsf{PART}(A)$. This result is extended next.

Theorem 2.6.14 *Let A be a finite, non-empty set. For $\beta \geq 1$, the mapping $e_\beta : \mathsf{PART}(A)^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by*

$$e_\beta(\pi, \sigma) = \frac{2d_\beta(\pi, \sigma)}{d_\beta(\pi, \sigma) + \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma)}$$

for $\pi, \sigma \in \mathsf{PART}(A)$ is a metric on $\mathsf{PART}(A)$ such that $0 \leq e_\beta(\pi, \sigma) \leq 1$.

Proof. It easy to see that $0 \leq e_\beta(\pi, \sigma) \leq 1$ since, by Corollary 2.6.5, $\mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma) \geq \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) = d_\beta(\pi, \sigma)$. We need to show only that the triangular inequality is satisfied by e_β for $\beta > 1$. We can write:

$$\begin{aligned}
&e_\beta(\pi, \sigma) + e_\beta(\sigma, \tau) = \\
&2 \cdot \tfrac{\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi)}{\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma)} + 2 \cdot \tfrac{\mathcal{H}_\beta(\sigma|\tau) + \mathcal{H}_\beta(\tau|\sigma)}{\mathcal{H}_\beta(\sigma|\tau) + \mathcal{H}_\beta(\tau|\sigma) + \mathcal{H}_\beta(\sigma) + \mathcal{H}_\beta(\tau)}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma) \leq \\
&\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\sigma|\tau) + \mathcal{H}_\beta(\tau|\sigma) + \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\tau)
\end{aligned}$$

because $\mathcal{H}_\beta(\sigma) \leq \mathcal{H}_\beta(\sigma|\tau) + \mathcal{H}_\beta(\tau)$ by Inequality (2.4) and Axiom **(P1)**. Similarly,

$$\begin{aligned}
&\mathcal{H}_\beta(\sigma|\tau) + \mathcal{H}_\beta(\tau|\sigma) + \mathcal{H}_\beta(\sigma) + \mathcal{H}_\beta(\tau) \leq \\
&\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\sigma|\tau) + \mathcal{H}_\beta(\tau|\sigma) + \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\tau)
\end{aligned}$$

because $\mathcal{H}_\beta(\sigma) \leq \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\pi)$. This yields the inequality:

$$
\begin{array}{l}
e_\beta(\pi,\sigma) + e_\beta(\sigma,\tau) \geq \\
2 \cdot \frac{\mathcal{H}_\beta(\pi|\sigma)+\mathcal{H}_\beta(\sigma|\pi)+\mathcal{H}_\beta(\sigma|\tau)+\mathcal{H}_\beta(\tau|\sigma)}{\mathcal{H}_\beta(\pi|\sigma)+\mathcal{H}_\beta(\sigma|\pi)+\mathcal{H}_\beta(\sigma|\tau)+\mathcal{H}_\beta(\tau|\sigma)+\mathcal{H}_\beta(\pi)+\mathcal{H}_\beta(\tau)} = \\
2 \cdot \frac{1}{1+\frac{\mathcal{H}_\beta(\pi)+\mathcal{H}_\beta(\tau)}{\mathcal{H}_\beta(\pi|\sigma)+\mathcal{H}_\beta(\sigma|\pi)+\mathcal{H}_\beta(\sigma|\tau)+\mathcal{H}_\beta(\tau|\sigma)}} \geq \\
2 \cdot \frac{1}{1+\frac{\mathcal{H}_\beta(\pi)+\mathcal{H}_\beta(\tau)}{\mathcal{H}_\beta(\pi|\tau)+\mathcal{H}_\beta(\tau|\pi)}} = e_\beta(\pi,\tau).
\end{array}
,
$$

∎

For $\beta = 1$, $e_1(\pi,\sigma) = \frac{d_1(\pi,\sigma)}{\mathcal{H}_1(\pi\wedge\sigma)}$, due to equality (2.5), which coincides with the expression obtained in [Man91] for the normalized distance.

2.7 Generalized Gain as a Selection Criterion for Splitting Attributes in Decision Trees

The standard selection criterion for splitting attributes is the information gain used by Quinlan [Qui93] in the classical C4.5 algorithm.

In [Weh96] generalized entropies have been mentioned in the context of splitting attribute selection, however only special cases of Shannon entropy and Gini index have been discussed and used in experiments.

In [Man91] it has been showed that using an entropy based distance d instead of the entropy gain as a splitting criterion gives smaller decision trees with little impact on accuracy. We generalized this work to distances based on generalized entropies.

We show that choosing the splitting attribute A based on the least value of $d_\beta(\pi(A),\pi)$, where π is the partition of the training set that corresponds to the target attribute of the classification generates smaller trees with comparable de-

grees of accuracy, and that in many cases best results are obtained for β different from 1 (Shannon entropy) or 2 (Gini index) thus proving the usefulness of generalized entropies.

Let $\pi, \sigma \in \mathsf{PART}(A)$. The *$\beta$-gain* of σ relative to π is the expression $G_\beta(\pi, \sigma) = \mathcal{H}_\beta(\pi) - \mathcal{H}_\beta(\pi|\sigma)$. The gain ratio is given by $R_\beta(\pi, \sigma) = \frac{G_\beta(\pi,\sigma)}{\mathcal{H}_\beta(\sigma)}$. For $\beta = 1$ we obtain Quinlan's gain defined through Shannon's entropy.

The next theorem establishes a monotonicity property of the distance d_β.

Theorem 2.7.1 *Let A be a finite sets and let $\pi, \pi', \sigma \in \mathsf{PART}(A)$ be three partitions such that π' is covered by π. In other words, $\pi = \{B_1, \ldots, B_m\}$ and $\pi' = \{B_1, \ldots, B'_m, B''_m\}$, where $B_m = B'_m \cup B''_m$. Suppose also that there exists a block C of σ such that $B_m \subseteq C$. Then, if $\beta \geq 1$, we have $d_\beta(\pi, \sigma) \leq d_\beta(\pi', \sigma)$ and $e_\beta(\pi, \sigma) \leq e_\beta(\pi', \sigma)$.*

Proof. For the case of Shannon's entropy, $\beta = 1$, the inequalities were proven in [Man91]. Therefore, we can assume that $\beta > 1$.

We claim that under the hypothesis of the theorem we have $\mathcal{H}_\beta(\sigma|\pi) = \mathcal{H}(\sigma|\pi')$. Note that $\sigma_{B_m} = \omega_{B_m}$, $\sigma_{B'_m} = \omega_{B'_m}$, and $\sigma_{B''_m} = \omega_{B''_m}$, since $B'_m, B''_m \subseteq B_m \subseteq C$. Therefore, $\mathcal{H}(\sigma_{B_m}) = \mathcal{H}(\sigma_{B'_m}) = \mathcal{H}(\sigma_{B''_m}) = 0$, hence

$$\mathcal{H}_\beta(\sigma|\pi) = \sum_{i=1}^{m} \frac{|B_i|}{|A|} \mathcal{H}_\beta(\sigma_{B_i}) = \sum_{i=1}^{m-1} \frac{|B_i|}{|A|} \mathcal{H}_\beta(\sigma_{B_i}) = \mathcal{H}_\beta(\sigma|\pi').$$

Theorem 2.6.7 implies $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi'|\sigma)$, which gives the first inequality.

Note that the second equality of the theorem:

$$\begin{aligned} e_\beta(\pi|\sigma) &= 2 \cdot \frac{\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi)}{\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma)} \\ &\leq 2 \cdot \frac{\mathcal{H}_\beta(\pi'|\sigma) + \mathcal{H}_\beta(\sigma|\pi')}{\mathcal{H}_\beta(\pi'|\sigma) + \mathcal{H}_\beta(\sigma|\pi') + \mathcal{H}_\beta(\pi') + \mathcal{H}_\beta(\sigma)} = e_\beta(\pi', \sigma) \end{aligned}$$

is equivalent to

$$\frac{\mathcal{H}_\beta(\sigma)+\mathcal{H}_\beta(\pi)}{\mathcal{H}_\beta(\pi|\sigma)+\mathcal{H}_\beta(\sigma|\pi)} \geq \frac{\mathcal{H}_\beta(\sigma)+\mathcal{H}_\beta(\pi')}{\mathcal{H}_\beta(\pi'|\sigma)+\mathcal{H}_\beta(\sigma|\pi')}. \tag{2.7}$$

Applying the definition of conditional entropy we can write:

$$\mathcal{H}_\beta(\pi|\sigma)-\mathcal{H}_\beta(\pi'|\sigma)=\frac{|C|}{|A|}\left[\frac{|B'_m|^\beta}{|C|^\beta}+\frac{|B''_m|^\beta}{|C|^\beta}-\frac{|B_m|^\beta}{|C|^\beta}\right]$$

and

$$\mathcal{H}_\beta(\pi)-\mathcal{H}_\beta(\pi')=\frac{|B'_m|^\beta+|B''_m|^\beta-|B_m|^\beta}{|A|^\beta},$$

which implies

$$\mathcal{H}_\beta(\pi|\sigma)-\mathcal{H}_\beta(\pi'|\sigma)=\left(\frac{|A|}{|C|}\right)^{\beta-1}\left[\mathcal{H}_\beta(\pi)-\mathcal{H}_\beta(\pi')\right]. \tag{2.8}$$

Thus, we obtain:

$$\mathcal{H}_\beta(\pi'|\sigma)-\mathcal{H}_\beta(\pi|\sigma)\geq\mathcal{H}_\beta(\pi')-\mathcal{H}_\beta(\pi). \tag{2.9}$$

By denoting $a=\mathcal{H}_\beta(\sigma)$ and $b=\mathcal{H}_\beta(\sigma|\pi)=\mathcal{H}_\beta(\sigma|\pi')$, the Inequality (2.7) can be written as:

$$\frac{a+\mathcal{H}_\beta(\pi)}{\mathcal{H}_\beta(\pi|\sigma)+b}\geq\frac{a+\mathcal{H}_\beta(\pi')}{\mathcal{H}_\beta(\pi'|\sigma)+b},$$

Elementary transformations yield: $\mathcal{H}_\beta(\pi'|\sigma)-\mathcal{H}_\beta(\pi|\sigma)\geq\frac{b+\mathcal{H}_\beta(\pi|\sigma)}{a+\mathcal{H}_\beta(\pi)}(\mathcal{H}_\beta(\pi')-\mathcal{H}_\beta(\pi))$, which is implied by Inequality (2.9) because $\frac{b+\mathcal{H}_\beta(\pi|\sigma)}{a+\mathcal{H}_\beta(\pi)}\leq 1$. This proves the second inequality of the theorem. ∎

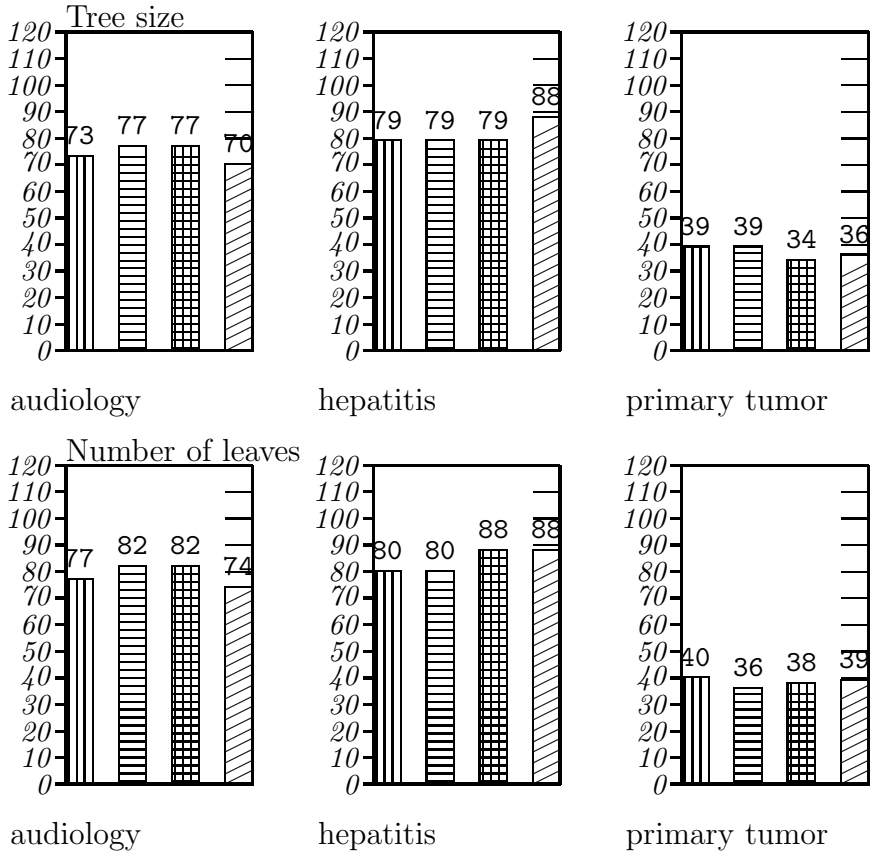
2.8 Experimental Results

The experiments have been conducted on 33 datasets from the UCI Machine Learning Repository. The `J48` tree builder from the `Weka` package [WF00] was used, in its original form as well as modified to support generalized entropy distances for

different values of the β parameter. Each experiment used 5-fold crossvalidation, average has been taken of the outcomes of the 5 runs and was performed with the tree pruning unchanged from the original version.

The tree size and the number of leaves diminish for 20 of the 33 databases analysed and grow for the remaining 13. The best reduction in size was achieved for the `primary-tumor` database, where the size of the tree was reduced to 37% for $\beta = 2.5$ and the number of leaves was reduced to 38.8% compared to the standard J48 algorithm that makes use of the gain ratio. On another hand, the largest increase in size and number of leaves was recorded for the `pima-diabetes` database, where for $\beta = 1$, we has an increase to 260% in size and to 256% in the number of leaves, though such an increase occurs rarely among the 13 databases where increases occur.

In Figure 2.1 we show the comparative performance of the distance d_β approach compared to the standard gain ratio for the databases which yielded the best results (`audiology`, `hepatitis`, and `primary-tumor`), in the case of the prunned trees. The 100% level refers in each case to the gain-ratio algorithm. It is interesting to observe that the accurracy diminishes slightly (by % for audiology database) or improves slightly, as shown in table 2.1, thus confirming previous results [Man91, BFO98, Min89] that accuracy is not affected substantially by the method used for tree construction.



The β factor:

$\beta = 1$ $\beta = 1.5$ $\beta = 2$ $\beta = 2.5$

Figure 2.1: Comparative Experimental Results

Database	J48	$\beta = 1$	$\beta = 1.5$	$\beta = 2$	$\beta = 2.5$
audiology	78.76%	73.42%	73.86%	73.86%	71.64%
hepatitis	78.06%	83.22%	83.22%	83.87%	83.87%
primary-tumor	40.99%	43.34%	41.87%	43.34%	43.05%

Table 2.1: Accuracy Results

CHAPTER 3

ASSOCIATION RULE PRUNING AND INTERESTINGNESS

3.1 Introduction

Many data mining algorithms produce huge sets of rules, practically impossible to analyze manually. Our goal is to identify a reasonably small, nonredundant set of interesting association rules describing well the most important relationships within the data.

The two main methods of achieving this goal known in the literature are sorting the rules according to some interestingness measure and rule pruning. In the first approach the (supposedly) most interesting rules, from users point of view, will be placed near the top of the sorted list. Of course choosing the right interestingness measure is crucial for effectiveness of this method.

Another approach is to remove or *prune* rules which are not interesting. Typically, rule sets are highly redundant and, so, it is usually possible to remove redundancies and leave a small nonredundant set of rules which describes well the key relationships between attributes and is small enough to be easily analyzed by the user.

This chapter contains contributions to both those methods. First, a new in-

terestingness measure generalizing many other important measures is introduced [JS01]; later, a method of association rule pruning using the Maximum Entropy Principle is described [JS02].

3.2 Basic notations

Now we introduce notation used throughout this chapter. See the Introduction for the description of database related notation used.

Values from domains of sets of attributes will be denoted by corresponding boldface lowercase letters, e.g. $\mathbf{i} \in \mathrm{Dom}(I)$. For $\mathbf{h} \in \mathrm{Dom}(H)$ and $I \subseteq H$, we denote the projection of $\mathbf{h}$ on I by $\mathbf{h}_I$. Sums $\sum_{\mathbf{i} \in \mathrm{Dom}(I)}$ will be abbreviated as $\sum_{\mathbf{i}}$.

Probability distributions will be denoted using letters P and Q. For a probability distribution P on $\mathrm{Dom}(H)$, and for $I \subseteq H$, let P_I be the marginal probability distribution on $\mathrm{Dom}(I)$ obtained by marginalizing the distribution P. In other words, we have

$$P_I(\mathbf{i}) = \sum \{P(\mathbf{h}) : \mathbf{h}_I = \mathbf{i}\}$$

for $\mathbf{i} \in \mathrm{Dom}(I)$. The joint distribution of H estimated from the data will be denoted by $\hat{P}$.

A uniform distribution over the domain of an attribute set I will be denoted by $\mathcal{U}_I$.

Let I, I' be attribute sets, and Q and Q' distributions over $\mathrm{Dom}(I)$ and $\mathrm{Dom}(I')$ respectively. The product of the distributions Q, Q' is the distribution $Q \times Q'$ over $K = I \cup I'$ such that

$$(Q \times Q')(\mathbf{k}) = Q(\mathbf{k}_I) \cdot Q'(\mathbf{k}_J),$$

for all $\mathbf{k} \in \mathrm{Dom}(K)$. The attribute sets I, J are *independent* in a distribution P if

$P_{IJ} = P_I \times P_J$, where IJ is an abbreviation for $I \cup J$.

Take an attribute set I and a distribution P_I over $\mathrm{Dom}(I)$. In order to simplify the formulas, and to make them more similar to standard textbook version, let us denote $p_{\mathbf{i}} = P_I(\mathbf{i})$. The Shannon entropy of I (or equivalently P_I) is defined as:

$$\mathcal{H}(I) = \mathcal{H}(P_I) = -\sum_{\mathbf{i}} p_{\mathbf{i}} \log p_{\mathbf{i}}.$$

The Gini index of I (or P_I) is defined as

$$\mathrm{gini}(I) = \mathrm{gini}(P_I) = 1 - \sum_{\mathbf{i}} p_{\mathbf{i}}^2.$$

Both of them are special cases of the *Havrda-Charvát α-entropy* of I (see [KK92]) defined as:

$$\mathcal{H}_\alpha(I) = \frac{1}{1-\alpha}\left(\sum_{\mathbf{i}} p_{\mathbf{i}}^\alpha - 1\right).$$

The limit case, when α tends towards 1 yields the Shannon entropy (with natural logarithm base), and for $\alpha = 2$ we obtain the Gini index.

If I, J are two sets of attributes. Let us denote $p_{\mathbf{i}} = P_I(\mathbf{i})$, $q_{\mathbf{j}} = P_J(\mathbf{j})$, and $p_{\mathbf{ij}} = \sum\{P_{IJ}(\mathbf{k}) \mid \mathbf{k} \in \mathrm{Dom}(IJ), \mathbf{k}_I = \mathbf{i}, \mathbf{k}_J = \mathbf{j}\}$. Notice that $p_{\mathbf{ij}}$ is the probability of $\mathbf{i}$ and $\mathbf{j}$ occurring simultaneously.

The conditional Shannon entropy of I conditioned upon J is given by

$$\mathcal{H}(I|J) = -\sum_{\mathbf{i}}\sum_{\mathbf{j}} p_{\mathbf{ij}} \log \frac{p_{\mathbf{ij}}}{q_{\mathbf{j}}},$$

Similarly, the Gini conditional index of these distributions is:

$$\mathrm{gini}(I|J) = 1 - \sum_{\mathbf{i}}\sum_{\mathbf{j}} \frac{p_{\mathbf{ij}}^2}{q_{\mathbf{j}}}.$$

These definitions allow us to introduce the Shannon gain (called entropy gain in

literature [Qui93]) and the Gini gain defined as:

$$\begin{aligned} \mathsf{gain}_{\text{gini}}(I, J) &= \text{gini}(I) - \text{gini}(I|J), \\ \mathsf{gain}_{\text{shannon}}(I, J) &= \mathcal{H}(I) - \mathcal{H}(I|J) \\ &= \mathcal{H}(I) + \mathcal{H}(J) - \mathcal{H}(I \cup J), \end{aligned} \tag{3.1}$$

respectively.

Notice that the Shannon gain is identical to the *mutual information* between attribute sets I and J [McE77]. For the Gini gain we can write:

$$\mathsf{gain}_{\text{gini}}(I, J) = \sum_{\mathbf{i}} \sum_{\mathbf{j}} \frac{p_{\mathbf{ij}}^2}{q_{\mathbf{j}}} - \sum_{\mathbf{i}} p_{\mathbf{i}}^2 \tag{3.2}$$

The notion of *distribution divergence* is central to the rest of the chapter.

Definition 3.2.1 Let $\mathcal{D}_I$ be a class of distributions over $\text{Dom}(I)$. A *distribution divergence* is a function $D : \mathcal{D}_I \times \mathcal{D}_I \longrightarrow \mathbb{R}$ such that:

1. $D(P, Q) \geq 0$ and $D(P, Q) = 0$ if and only if $P = Q$ for every $P, Q \in \mathcal{D}_I$.
2. When Q is fixed, $D(P, Q)$ is a convex function of P; in other words, if $P = a_1 P_1 + \cdots + a_k P_k$, where $a_1 + \ldots + a_k = 1$, then
$$D(P, Q) \geq \sum_{i=1}^{k} a_i D(P_i, Q).$$

□

Intuitively, the divergence represents how much distribution P differs from Q.

Two most commonly used divergencies are the *Kullback-Leibler divergence* (also known as *crossentropy*) and the *chi-squared divergence* [KK92] defined for distributions P, Q over the domain of an attribute set I as:

$$D_{KL}(P : Q) = \sum_{\mathbf{i}} P(\mathbf{i}) \log \frac{P(\mathbf{i})}{Q(\mathbf{i})},$$

$$D_{\chi^2}(P:Q) \;=\; \sum_{\mathbf{i}} \frac{(P(\mathbf{i})-Q(\mathbf{i}))^2}{Q(\mathbf{i})} = \sum_{\mathbf{i}} \frac{P(\mathbf{i})^2}{Q(\mathbf{i})} - 1,$$

respectively. Note that $|\rho| D_{\chi^2}$ equals the χ^2 dependency measure, well known from statistics [BA99].

An important, general class of distribution divergences was obtained by Cziszar in [Czi72] as:

$$D_\phi(P,Q) = \sum_{\mathbf{i}} Q(\mathbf{i}) \phi\left(\frac{P(\mathbf{i})}{Q(\mathbf{i})}\right),$$

where P and Q are two distributions over $\mathrm{Dom}(I)$ and $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ is a twice differentiable convex function such that $\phi(1) = 0$. We will also make an additional assumption that $0 \cdot \phi(\frac{0}{0}) = 0$ to handle the case when for some $\mathbf{i}$ both $P(\mathbf{i})$ and $Q(\mathbf{i})$ are zero. If for some $\mathbf{i}$, $P(\mathbf{i}) > 0$, and $Q(\mathbf{i}) = 0$ the value of $D_\phi(P,Q)$ is undefined.

The Cziszar divergence satisfies properties (1) and (2) given above (see [KK92]).

Both the Kullback-Leibler and the chi-squared divergencies are special cases of Cziszar divergence obtained by choosing $\phi(x) = x \log x$ and $\phi(x) = x^2 - x$ respectively.

An important subclass of Cziszar divergencies is the so called *Havrda-Charvát divergence* $D_{\mathcal{H}_\alpha}$ generated by $\phi(x) = \frac{x^\alpha - x}{\alpha - 1}$ [KK92]. D_{KL} and D_{χ^2} are special cases of $D_{\mathcal{H}_\alpha}$ for $\alpha \to 1$ and $\alpha = 2$ respectively.

The following result shows the invariance of Cziszar divergence with respect to distribution product:

Theorem 3.2.2 *For any two distributions P, P' over $\mathrm{Dom}(I)$, any distribution Q over $\mathrm{Dom}(J)$ and any Cziszar divergence measure D_ϕ we have $D_\phi(Q \times P, Q \times P') = D_\phi(P, P')$.*

Proof. The definition of Cziszar divergence implies

$$\begin{aligned} D_\phi(Q \times P, Q \times P') &= \sum_{\mathbf{j}} \sum_{\mathbf{i}} P'(\mathbf{i}) Q(\mathbf{j}) \phi\left(\frac{P(\mathbf{i})Q(\mathbf{j})}{P'(\mathbf{i})Q(\mathbf{j})}\right), \\ &= \sum_{\mathbf{j}} Q(\mathbf{j}) \sum_{\mathbf{i}} P'(\mathbf{i}) \phi\left(\frac{P(\mathbf{i})}{P'(\mathbf{i})}\right), \\ &= D_\phi(P, P'), \end{aligned}$$

which is the desired equality. ∎

Definition 3.2.3 A *rule* is a pair of attribute sets (I, J), if $I, J \subseteq H$.

If (I, J) is a rule, then we refer to I as the *antecedent* and to J as the *consequent* of the rule. A rule (I, J) will be denoted, following the prevalent convention in the literature, by $I \rightarrow J$. □

3.3 Interestingness of Rules

Determining the interestingness of rules is an important data mining problem. Many such measures have been proposed, and used in literature (see [BA99] for a survey). Another survey and a method of selecting the right measure based on experts opppinions has been presented in [TKS02]. In this chapter I concentrate on measures that assess how much knowledge we gain on the joint distribution of a set of attributes Q from the knowing the joint distribution of some set of attributes P.

Examples of such measures are *entropy gain*, *mutual information*, *Gini gain*, χ^2 [McE77, Mit97, SBM98, BA99, Mor98, MFM98]. The rules considered here are thus different from *association rules* studied in data mining, since we consider full joint distributions of both antecedent and consequent, while association rules consider only the probability of all attributes having some specified value. This

approach has the advantage of natural applicability to multivalued (not binary) attributes.

In this chapter it is demonstrated that all the above mentioned measures are special cases of a more general parametric measure of interestingness, and by choosing two numerical parameters a continuum of measures can be obtained containing several well-known interesting measures as special cases.

3.4 A General Measure of Rule Interestingness

Let $r = I \rightarrow J$ be a rule. Let P be a probability distribution over H. We will refer to P (and to its marginalizations) as the *observed distribution* (possibly equal to the distribution $\hat{P}$ estimated from data but derivations below make sense for any P no matter how it was obtained).

To construct an interestingness measure we will use a Bayesian approach, in that we will consider an assumed apriori distribution Θ of the consequent set of attributes J. This could be an *apriori* distribution of J, the observed distribution P_J, or some combination of these distributions. To define an interestingness measure of r we will be guided by two main considerations:

- The more the observed joint distribution of IJ diverges from the product distribution of I and the assumed distribution Θ of J the more interesting the rule is. Note that $P_{IJ} = P_I \times \Theta$ corresponds to the situation when I and J are independent and the observed distribution of J follows the assumed apriori distribution.

- The rule is not interesting if I, J are independent. Therefore, we need to consider a correcting term in the definition of an interestingness measure that

will decrease its value when P_J is different from the assumed distribution.

The choice of the distribution Θ of the consequent J of rules of the form $I \to J$ can be made starting either from the observed distribution, that is, adopting P_J for Θ, or from some exterior information. For example, if J is the sex attribute for a table that contains data concerning some experiment subjects, we can adopt as the assumed distribution either

$$P_{sex} = \begin{pmatrix} \text{'F'} & \text{'M'} \\ 0.45 & 0.55 \end{pmatrix},$$

assuming that 45% of the individuals involved are female, or the distribution

$$P_{gen\_pop} = \begin{pmatrix} \text{'F'} & \text{'M'} \\ 0.51 & 0.49 \end{pmatrix},$$

consistent with the general distribution of the sexes in the general population. Moreover, we can contemplate a convex combination of distributions of the form

$$\Theta_a = aP_J + (1-a)\Theta_0,$$

where P_J is the observed distribution of J and Θ_0 is a distribution that is based on some prior knowledge. The number a reflects the degree of confidence in the observed distribution; the closer this number is to 1, the higher the confidence, and the more preponderant P_J is in the assumed true distribution.

Definition 3.4.1 Let $r : I \to J$ be a rule, D be some measure of divergence between distributions, and let Θ be a distribution over $\text{Dom}(J)$.

The Υ *measure of interestingness* generated by D and Θ is defined by

$$\Upsilon_{D,\Theta}(r) = D(P_{IJ}, P_I \times \Theta) - D(P_J, \Theta).$$

□

In the above definition Θ represents the assumed distribution of J while P_J is the observed distribution of J. The term $D(P_J, \Theta)$ measures the degree to which P_J diverges from the prior distribution Θ, and $D(P_{IJ}, P_I \times \Theta)$ measures how far P_{IJ} diverges from the joint distribution of I and J in case they were independent, and J was distributed according to Θ.

The justification for the correcting term $D(P_J, \Theta)$ is given in the following theorem:

Theorem 3.4.2 *If I and J are independent, and D is a Cziszar measure of divergence then $\Upsilon_{D,\Theta}(I \rightarrow J) = 0$.*

Proof. In this case $P_{IJ} = P_I \times P_J$, and by Theorem 3.2.2 we have $\Upsilon_{D,\Theta}(I \rightarrow J) = D(P_I \times P_J, P_I \times \Theta) - D(P_J, \Theta) = D(P_J, \Theta) - D(P_J, \Theta) = 0$. ∎

Note that if D is a Cziszar divergence $D = D_\phi$, then the invariance of these divergences implies:

$$\Upsilon_{D_\phi,\Theta}(I \rightarrow J) = D_\phi(P_{IJ}, P_I \times \Theta) - D_\phi(P_I \times P_J, P_I \times \Theta).$$

3.5 Properties of the General Measure of Interestingness

Initially, we discuss several basic properties of the proposed measure.

Theorem 3.5.1 *If D is a Cziszar divergence, then*

$$\Upsilon_{D,P_J}(I \rightarrow J) = \Upsilon_{D,P_I}(J \rightarrow I)$$

Proof. We have $\Upsilon_{D,P_J}(I \rightarrow J) = D(P_{IJ}, P_I \times P_J)$, $\Upsilon_{D,P_I}(J \rightarrow I) = D(P_{JI}, P_J \times P_I)$, and the proof follows from the permutational symmetry of Cziszar's divergence [KK92]. ∎

The above property means that when the assumed distribution of the consequent is kept equal to the distribution observed from data, then the measure is symmetric with respect to the direction of the rule, i.e. exchanging the antecedent and consequent does not change the interestingness.

Theorem 3.5.2 *Let D be a Cziszar divergence. If K is a set of attributes independent of I, and jointly of IJ, then, for any Θ*

$$\Upsilon_{D,\Theta}(KI \to J) = \Upsilon_{D,\Theta}(I \to J).$$

If K is a set of attributes independent of J, and jointly of IJ, then

$$\Upsilon_{D,P_{KJ}}(I \to KJ) = \Upsilon_{D,P_J}(I \to J).$$

Proof.

$$\begin{aligned}
\Upsilon_{D,\Theta}(KI \to J) &= D(P_{KIJ}, P_{KI} \times \Theta) - D(P_J, \Theta) \\
&= D(P_K \times P_{IJ}, P_K \times P_I \times \Theta) - D(P_J, \Theta) \\
&= D(P_{IJ}, P_I \times \Theta) - D(P_J, \Theta) = \Upsilon_{D,\Theta}(I \to J) \\
&\quad \text{(by Theorem 3.2.2)}
\end{aligned}$$

For the second part, it follows from theorem 3.5.1 and from the first part of this theorem that

$$\begin{aligned}
\Upsilon_{D,P_{KJ}}(I \to KJ) &= \Upsilon_{D,P_I}(KJ \to I) = \Upsilon_{D,P_I}(J \to I) \\
&= \Upsilon_{D,P_J}(I \to J).
\end{aligned}$$

∎

The previous result gives a desirable property of $\Upsilon_{D,\Theta}$ since adding independent attributes should not affect rules interestingness.

Note that if $\Theta = P_J$, that is, when Θ equals the observed distribution of the consequent, then Υ becomes symmetric and is not affected by adding independent attributes to either the antecedent or the consequent.

Next, we consider several important special cases of the interestingness measure.

If the divergence D and the assumed distribution used in the definition of the interestingness measure are chosen appropriately, then the interestingness $\Upsilon_{D,\Theta}(I \to J)$ is proportional to a gain of the set of attributes of the consequent J of the rule relative to the antecedent I. Both the Gini gain, $\mathsf{gain}_{\text{gini}}(J, I)$, and the entropy gain, $\mathsf{gain}_{\text{shannon}}(J, I)$, can be obtained by appropriate choice of D. Moreover a measure proportional to the χ^2 statistic can be obtained in that way.

Let I, J be two sets of attributes; denote $p_{\mathbf{i}} = P_I(\mathbf{i})$, $q_{\mathbf{j}} = P_J(\mathbf{j})$, and $p_{\mathbf{ij}} = \sum\{P_{IJ}(\mathbf{k}) \mid \mathbf{k} \in \mathrm{Dom}(IJ), \mathbf{k}_I = \mathbf{i}, \mathbf{k}_J = \mathbf{j}\}$.

Theorem 3.5.3 *Let $I \longrightarrow J$ be a rule. If $D = D_{\mathrm{KL}}$ then*

$$\Upsilon_{D,\Theta}(I \to J) = \mathsf{gain}_{\text{shannon}}(J, I),$$

regardless of the choice of Θ.

Proof. The definition of the Kullback-Leibler divergence allows us to write:

$$\begin{aligned}
\Upsilon_{D_{\mathrm{KL}},\Theta}(I \to J) &= D_{\mathrm{KL}}\left(P_{IJ}, P_I \times \Theta\right) - D_{\mathrm{KL}}\left(P_J, \Theta\right) \\
&= \sum_{\mathbf{i}} \sum_{\mathbf{j}} p_{\mathbf{ij}} \log \frac{p_{\mathbf{ij}}}{p_{\mathbf{i}} \theta_{\mathbf{j}}} - \sum_{\mathbf{j}} q_{\mathbf{j}} \log \frac{q_{\mathbf{j}}}{\theta_{\mathbf{j}}} \\
&= \sum_{\mathbf{i}} \sum_{\mathbf{j}} p_{\mathbf{ij}} \log p_{\mathbf{ij}} - \sum_{\mathbf{i}} \sum_{\mathbf{j}} p_{\mathbf{ij}} \log p_{\mathbf{i}} \\
&\quad - \sum_{\mathbf{i}} \sum_{\mathbf{j}} p_{\mathbf{ij}} \log \theta_{\mathbf{j}} - \sum_{\mathbf{j}} q_{\mathbf{j}} \log q_{\mathbf{j}} + \sum_{\mathbf{j}} q_{\mathbf{j}} \log \theta_{\mathbf{j}} \\
&= \mathcal{H}(I) + \mathcal{H}(J) - \mathcal{H}(IJ) = \mathsf{gain}_{\text{shannon}}(J, I), \\
&\qquad \text{(by Equality (3.1))}
\end{aligned}$$

which completes the proof. ∎

The above theorem means that for the case D_{KL} the family of measures generated by Θ reduces to a single measure: the Shannon gain (mutual information). This is not the case for other divergences.

Theorem 3.5.4 *Let $I \longrightarrow J$ be a rule. If $D = D_{\chi^2}$ and $\Theta = \mathcal{U}_J$, $n = |\operatorname{Dom}(J)|$, then*

$$\Upsilon_{D,\Theta}(I \to J) = n \cdot \mathsf{gain}_{\mathrm{gini}}(J, I).$$

Proof. We have

$$\begin{aligned}
\Upsilon_{\chi^2,\mathcal{U}_J}(I \to J) &= D_{\chi^2}\left(P_{IJ}, P_I \times \mathcal{U}_J\right) - D_{\chi^2}\left(P_J, \mathcal{U}_J\right) \\
&= \sum_{\mathbf{i}} \sum_{\mathbf{j}} \frac{p_{\mathbf{ij}}^2}{\frac{p_{\mathbf{i}}}{n}} - \sum_{\mathbf{j}} \frac{q_{\mathbf{j}}^2}{\frac{1}{n}} \\
&= n \left(\sum_{\mathbf{i}} \sum_{\mathbf{j}} \frac{p_{\mathbf{ij}}^2}{p_{\mathbf{i}}} - \sum_{\mathbf{j}} q_{\mathbf{j}}^2 \right) \\
&= n \cdot \mathsf{gain}_{\mathrm{gini}}(J, I), \\
&\qquad \text{(by Equality (3.2))}
\end{aligned}$$

which is the desired equality. ∎

Theorem 3.5.5 *We have $\Upsilon_{D_{\chi^2},P_J}(I \longrightarrow J)$ is proportional to $\chi^2(I, J)$, the chi-squared statistics [BA99] for attribute sets I, J.*

Proof.

$$\begin{aligned}
\Upsilon_{D_{\chi^2},P_J}(I \longrightarrow J) &= D_{\chi^2}\left(P_{IJ}, P_I \times P_J\right) - D_{\chi^2}\left(P_J, P_J\right) \\
&= D_{\chi^2}\left(P_{IJ}, P_I \times P_J\right) \\
&= \frac{\chi^2(I, J)}{|\rho|},
\end{aligned}$$

where $|\rho|$ is the number of rows in the database table from which the observed distribution is estimated. ∎

Note that above we treat attribute sets, for example $I = \{A_1, \ldots, A_r\}$ and $J = \{B_1, \ldots, B_s\}$, as single attributes with the domains given by (1.1). This is appropriate, since we are interested in how one *set* of attributes I influences the joint distribution of another *set* of attributes J. Another way, used in [SBM98], is to compute $\chi^2(A_1, \ldots, A_r, B_1, \ldots, B_s)$; however this is not what we want.

The case $D = D_{\chi^2}$ is of practical interest since it includes two widely used measures (χ^2 and $\mathsf{gain}_{\text{gini}}$) as special cases, and allows for obtaining a continuum of measures "in between" the two.

Theorem 3.5.6 proven below shows that the generalized measure interestingness $\Upsilon_{D,\Theta}(I \to J)$ is minimal when I and J are independent and thus, it justifies our definition of this measure through variational considerations. We begin with a technical result.

Theorem 3.5.6 *Let $\Upsilon_{D,\Theta}$ be the measure of interestingness generated by the assumed distribution Θ and the Kullback-Leibler divergence, or the χ^2-divergence and let $I \to J$ be a rule. For any fixed distribution P over H and a fixed distribution Θ, the value of $\Upsilon_{D,\Theta}(I \to J)$ is minimal (and equal to 0) if only if $P_{IJ} = P_I \times P_J$, i.e., when I and J are independent.*

Proof. It is clear that if I and J are independent, then we have in both cases $\Upsilon_{D,\Theta}(I \to J) = 0$.

When $D = D_{\text{KL}}$ the result follows from the properties of Shannon gain/mutual information [McE77].

We need to prove the result for $D = D_{\chi^2}$. It has been noted in Section 3.2 that D_{χ^2} is a special case of Cziszar divergence for $\phi(x) = x^2 - x$. From the conditions

on ϕ it follows that for all Cziszar's divergences the respective functions ϕ have the property that the inverses of their first derivatives are monotonic functions and therefore can be inverted. Indeed, in the case of D_{χ^2} we have $\phi(x) = x^2 - x$, and $(\phi')^{-1}(x) = x/2 + 1/2$.

We will use *Lagrange multipliers* method to find the minimum of $D_{\chi^2}(P_{IJ}, P_I \times \Theta)$ subject to the following set of constraints:

$$\sum_{\mathbf{i}} \sum_{\mathbf{j}} p_{\mathbf{ij}} = 1 \tag{3.3}$$

$$\sum_{\mathbf{j}} p_{\mathbf{ij}} = p_{\mathbf{i}} \tag{3.4}$$

$$\sum_{\mathbf{i}} p_{\mathbf{ij}} = q_{\mathbf{j}} \tag{3.5}$$

The Lagrangian is

$$\begin{aligned} L &= \sum_{\mathbf{i}} \sum_{\mathbf{j}} p_{\mathbf{i}} \theta_{\mathbf{j}} \phi \left(\frac{p_{\mathbf{ij}}}{p_{\mathbf{i}} \theta_{\mathbf{j}}} \right) + \lambda \left(\sum_{\mathbf{i}} \sum_{\mathbf{j}} p_{\mathbf{ij}} - 1 \right) \\ &+ \sum_{\mathbf{i}} \lambda_{\mathbf{i}} \left(\sum_{\mathbf{j}} p_{\mathbf{ij}} - p_{\mathbf{i}} \right) + \sum_{\mathbf{j}} \mu_{\mathbf{j}} \left(\sum_{\mathbf{i}} p_{\mathbf{ij}} - q_{\mathbf{j}} \right), \end{aligned}$$

and

$$\frac{\partial L}{\partial p_{\mathbf{ij}}} = \phi' \left(\frac{p_{\mathbf{ij}}}{p_{\mathbf{i}} \theta_{\mathbf{j}}} \right) + \lambda + \lambda_{\mathbf{i}} + \mu_{\mathbf{j}}. \tag{3.6}$$

By equating (3.6) to zero we get:

$$\phi' \left(\frac{p_{\mathbf{ij}}}{p_{\mathbf{i}} \theta_{\mathbf{j}}} \right) = -(\lambda + \lambda_{\mathbf{i}} + \mu_{\mathbf{j}}).$$

In the case of the D_{χ^2} measure of divergence we have $(\phi')^{-1}(x) = \frac{x}{2} + \frac{1}{2}$. Therefore, $p_{\mathbf{ij}}$ can be written as

$$p_{\mathbf{ij}} = p_{\mathbf{i}} \theta_{\mathbf{j}} \left[\frac{-\lambda - \lambda_{\mathbf{i}} - \mu_{\mathbf{j}}}{2} + \frac{1}{2} \right] = \frac{1}{2} p_{\mathbf{i}} \theta_{\mathbf{j}} (1 - \lambda - \lambda_{\mathbf{i}} - \mu_{\mathbf{j}}).$$

Substituting into (3.4) we get $\sum_{\mathbf{j}} p_{\mathbf{ij}} = \frac{1}{2} \sum_{\mathbf{j}} p_{\mathbf{i}} \theta_{\mathbf{j}} (1 - \lambda - \lambda_{\mathbf{i}} - \mu_{\mathbf{j}}) = p_{\mathbf{i}}$, and

$$\sum_{\mathbf{j}} \theta_{\mathbf{j}} (1 - \lambda - \lambda_{\mathbf{i}} - \mu_{\mathbf{j}}) = 2.$$

After splitting the sum we get $1 - \lambda - \lambda_{\mathbf{i}} - \sum_{\mathbf{j}} \theta_{\mathbf{j}}\mu_{\mathbf{j}} = 2$, and

$$\lambda_{\mathbf{i}} = -\lambda - \sum_{\mathbf{j}} \theta_{\mathbf{j}}\mu_{\mathbf{j}} - 1 = c_\alpha.$$

Similarly, substituting into (3.5) we get $\sum_{\mathbf{i}} p_{\mathbf{ij}} = \frac{1}{2}\sum_{\mathbf{i}} p_{\mathbf{i}}\theta_{\mathbf{j}}(1 - \lambda - \lambda_{\mathbf{i}} - \mu_{\mathbf{j}}) = q_{\mathbf{j}}$, and

$$\sum_{\mathbf{i}} p_{\mathbf{i}}(1 - \lambda - \lambda_{\mathbf{i}} - \mu_{\mathbf{j}}) = 2\frac{q_{\mathbf{j}}}{\theta_{\mathbf{j}}}.$$

After splitting the sum we get $1 - \lambda - \mu_{\mathbf{j}} - \sum_{\mathbf{i}} p_{\mathbf{i}}\lambda_{\mathbf{i}} = 2\frac{q_{\mathbf{j}}}{\theta_{\mathbf{j}}}$, and

$$\mu_{\mathbf{j}} = 1 - \lambda - \sum_{\mathbf{i}} p_{\mathbf{i}}\lambda_{\mathbf{i}} - 2\frac{q_{\mathbf{j}}}{\theta_{\mathbf{j}}} = c_\beta - 2\frac{q_{\mathbf{j}}}{\theta_{\mathbf{j}}}.$$

Thus,

$$p_{\mathbf{ij}} = \frac{1}{2}p_{\mathbf{i}}\theta_{\mathbf{j}}(1 - \lambda - c_\alpha - c_\beta + 2\frac{q_{\mathbf{j}}}{\theta_{\mathbf{j}}}) = \frac{1}{2}p_{\mathbf{i}}\theta_{\mathbf{j}}(c_\gamma + 2\frac{q_{\mathbf{j}}}{\theta_{\mathbf{j}}}),$$

for some constant c_γ. By using 3.3 we get $c_\gamma = 0$, and $p_{\mathbf{ij}} = p_{\mathbf{i}} \cdot p_{\mathbf{j}}$. ∎

We proved that $\mathsf{gain}_{\text{shannon}}$ and $\mathsf{gain}_{\text{gini}}$ are equivalent to $\Upsilon_{D_{\text{KL}},\mathcal{U}_J}$ and $\Upsilon_{D_{\chi^2},\mathcal{U}_J}$, respectively. It is thus natural to define a notion of *gain* for any divergence D as

$$\mathsf{gain}_D(I \to J) = \Upsilon_{D,\mathcal{U}_J}(I \to J).$$

Let $P_J|\mathbf{i}$ denote the probability distribution of J conditioned on $I = \mathbf{i}$. For any Cziszar measure D_ϕ we have:

$$\begin{aligned}
\mathsf{gain}_{D_\phi}(I \to J) &= D_\phi(P_{IJ}, P_I \times \mathcal{U}_J) - D_\phi(P_J, \mathcal{U}_J) \\
&= \sum_{\mathbf{i}} p_{\mathbf{i}} \sum_{\mathbf{j}} \frac{1}{n}\phi\left(\frac{p_{\mathbf{ij}}}{p_{\mathbf{i}} \cdot \frac{1}{n}}\right) - D_\phi(P_J, \mathcal{U}_J) \\
&= -\left[D_\phi(P_J, \mathcal{U}_J) - \sum_{\mathbf{i}} p_{\mathbf{i}} D_\phi(P_J|\mathbf{i}, \mathcal{U}_J)\right].
\end{aligned}$$

As special cases we have $\mathsf{gain}_{\text{gini}} \equiv \mathsf{gain}_{\chi^2}$, and $\mathsf{gain}_{\text{shannon}} \equiv \mathsf{gain}_{\text{KL}}$.

A parameterized version of Υ that takes into account the degree of confidence in the distribution of the consequent as it results from the data is introduced next.

Let us define the probability distribution $\Theta_a, a \in [0,1]$ by

$$\Theta_a = aP_J + (1-a)\mathcal{U}_J.$$

The value of a expresses the amount of confidence we have in P_J estimated from the data, assuming a uniform prior distribution, so Θ_a is the *a posteriori* distribution for J.

We can now define

$$\Upsilon_{D,a} = \Upsilon_{D,\Theta_a}.$$

Note that when $D = D_{\chi^2}$, we have (up to a constant factor) both $\chi^2(I \to J)$ and $\text{gini}_{\mathsf{gain}}(I \to J)$ as special cases of $\Upsilon_{D_{\chi^2},a}$. Moreover by taking different values of parameter a we can obtain a continuum of measures in between the two.

As noted before, both D_{χ^2} and D_{KL} divergence measures are special cases of Havrda-Charvát divergence $D_{\mathcal{H}_\alpha}$ for $\alpha \to 1$, and $\alpha = 2$ respectively. We can thus introduce $\Upsilon_{\alpha,a} = \Upsilon_{D_{\mathcal{H}_\alpha},\Theta_a}$, which allows us to obtain a family of interestingness measures, including (up to a constant factor) all three measures given in Section 3.5 as special cases, by simply changing two real valued parameters α and a.

Also note that for $a = 0$, we obtain a family of gains (as defined in chapter 3.5) for all the Havrda-Charvát divergences.

3.6 Experimental results

We evaluated the new measure on a simple synthetic dataset and on data from the UCI machine learning repository [BM98]. We concentrated on the case $D = D_{\chi^2}$, as potentially the most useful in practice, and found interestingness of rules for different values of parameter a (see chapter 3.5)

3.6.1 Synthetic data

To ensure measures throughout the family handle obvious cases correctly, and to make it easy to observe properties of the measure for different values of parameter a we first evaluated the rules on a synthetic dataset with 3 attributes A, B, C and with known probabilistic dependencies between them.

Values of attributes A and B have been generated from known probability distributions:

$$P_A = \begin{pmatrix} 0 & 1 & 2 \\ 0.1 & 0.5 & 0.4 \end{pmatrix}, P_B = \begin{pmatrix} 0 & 1 \\ 0.2 & 0.8 \end{pmatrix}.$$

Attribute C depends on attribute A. Denote $P_C|i$ the distribution of C conditioned upon $A = i$. We used

$$P_C|0 = \begin{pmatrix} 0 & 1 \\ 0.2 & 0.8 \end{pmatrix}, P_C|1 = \begin{pmatrix} 0 & 1 \\ 0.5 & 0.5 \end{pmatrix}, P_C|2 = \begin{pmatrix} 0 & 1 \\ 0.7 & 0.3 \end{pmatrix},$$

One million data points have been generated according to this distribution, for a few values of a we sorted all possible rules based on their $\Upsilon_{D_{\chi^2},a}$ interestingness values. Results are given in Table 3.1.

Discussion

1. Attribute B is totally independent of both A and C, so any rule containing only B as the antecedent or consequent should have interestingness 0. The experiments confirm this, for all values of parameter a such rules have interestingness close to zero, significantly lower than the interestingness of any other rules.

2. For $a = 0$ (the first quarter of the table) Υ becomes the Gini gain, a measure that is strongly asymmetric (and could thus suggest the direction of the

dependence) and strongly affected by adding extra independent attributes to the consequent (which is undesirable).

3. For $a = 1$ (the last quarter of the table) Υ becomes (up to a constant factor) the χ^2 measure of dependence. This measure is totally symmetric and not affected by presence of independent attributes in either antecedent or consequent. Indeed, it can be seen that all rules involving A and C have the same interestingness regardless of the presence of B in the antecedent or consequent.

4. As a varies from 0 to 1 the intermediate measures can be seen to become more and more symmetric. Measures for a being close to but less than 1 could be of practical interest since they seem to 'combine the best of the two worlds', that is, are still asymmetric and pretty insensitive to presence of independent attributes in the consequent. E.g. for $a = 0.9$ all rules having A in the antecedent and C in the consequent have interestingness close to 0.09, while rules having C in the antecedent and A in the consequent have all interestingness close to 0.082 regardless of the presence or absence of B in the consequents. So for $a = 0.9$ the intermediate measure correctly ranked the rules indicating the true direction of the relationship.

3.6.2 The mushroom database

We repeated the above experiment on data from the UCI machine learning repository [BM98]. Here we present results for the *agaricus-lepiota* database containing data on North American Mushrooms. To make the ruleset size manageable we restrict ourselves to rules involving the *class* attribute indicating whether the mushroom is edible or poisonous.

rule	$\Upsilon_{D_{\chi^2},0}$	rule	$\Upsilon_{D_{\chi^2},0.5}$
$A \rightarrow BC$	0.122061	$A \rightarrow BC$	0.0989161
$C \rightarrow AB$	0.0896776	$AB \rightarrow C$	0.0898611
$AB \rightarrow C$	0.0896287	$A \rightarrow C$	0.089861
$A \rightarrow C$	0.0896287	$C \rightarrow AB$	0.0769886
$BC \rightarrow A$	0.065851	$BC \rightarrow A$	0.0683164
$C \rightarrow A$	0.0658484	$C \rightarrow A$	0.0683142
$B \rightarrow AC$	3.16585e-06	$B \rightarrow AC$	2.50502e-06
$B \rightarrow A$	2.7369e-06	$B \rightarrow A$	2.35091e-06
$AC \rightarrow B$	1.37659e-06	$AC \rightarrow B$	1.51849e-06
$A \rightarrow B$	1.32828e-06	$A \rightarrow B$	1.46355e-06
$B \rightarrow C$	1.70346e-07	$B \rightarrow C$	1.72781e-07
$C \rightarrow B$	1.10069e-07	$C \rightarrow B$	1.22814e-07
rule	$\Upsilon_{D_{\chi^2},0.9}$	rule	$\Upsilon_{D_{\chi^2},1}$
$A \rightarrow BC$	0.0908769	$BC \rightarrow A$	0.0905673
$AB \rightarrow C$	0.0903859	$A \rightarrow BC$	0.0905673
$A \rightarrow C$	0.0903859	$C \rightarrow AB$	0.0905654
$C \rightarrow AB$	0.0834734	$AB \rightarrow C$	0.0905654
$BC \rightarrow A$	0.082009	$A \rightarrow C$	0.0905653
$C \rightarrow A$	0.082007	$C \rightarrow A$	0.0905653
$B \rightarrow AC$	2.19739e-06	$AC \rightarrow B$	2.15872e-06
$B \rightarrow A$	2.12646e-06	$B \rightarrow AC$	2.15872e-06
$AC \rightarrow B$	1.95101e-06	$A \rightarrow B$	2.08117e-06
$A \rightarrow B$	1.87986e-06	$B \rightarrow A$	2.08017e-06
$B \rightarrow C$	1.73782e-07	$C \rightarrow B$	1.74126e-07
$C \rightarrow B$	1.57306e-07	$B \rightarrow C$	1.74126e-07

Table 3.1: Rules on synthetic data ordered by $\Upsilon_{D_{\chi^2},a}$ for different values of a.

rule	$\Upsilon_{D_{\chi^2},0}$
class→odor ring-type	9.84024
class→odor spore-print-color	9.16709
class→odor veil-color	8.22064
class→odor gill-attachment	8.2026
class→gill-color spore-print-color	7.82161
class→ring-type spore-print-color	7.62564
class→odor stalk-root	7.60198
class→gill-color ring-type	7.28972
class→odor stalk-color-above-ring	7.19584
class→odor stalk-color-below-ring	7.14197

rule	$\Upsilon_{D_{\chi^2},0.9}$
odor→class stalk-root	3.61877
class stalk-root→odor	3.2782
odor→class cap-color	2.59777
odor→class ring-type	2.54896
odor→class spore-print-color	2.54864
stalk-color-above-ring→class stalk-color-below-ring	2.47669
class cap-color→odor	2.46105
odor→class gill-color	2.45027
stalk-color-below-ring→class stalk-color-above-ring	2.38593
class spore-print-color→odor	2.35384

rule	$\Upsilon_{D_{\chi^2},1}$
class stalk-root→odor	4.11701
class stalk-color-below-ring→stalk-color-above-ring	3.38287
stalk-color-below-ring→class stalk-color-above-ring	3.37968
class ring-type→odor	2.98764
class cap-color→odor	2.85308
odor→class gill-color	2.82423
odor→class spore-print-color	2.56331
odor→class stalk-color-below-ring	2.44004
class stalk-color-above-ring→odor	2.42725
class gill-color→spore-print-color	2.42224

Table 3.2: Rules on mushroom dataset ordered by $\Upsilon_{D_{\chi^2},a}$ for different values of a.

In the experiment we enumerated all rules involving up to 3 attributes and ranked them by interestingness for different values of parameter a. Top ten rules for each value of a are shown in Table 3.2. For $a = 1$ the symmetric rules were removed.

We noticed that for any value of a most of the rules involve the *odor* attribute. Indeed the inspection of data revealed that knowing the mushroom's odor allows for identifying its class with 98.5% accuracy, far better than for any other attribute.

We note also that similar rules are ranked close to the top for all values of a, which proves that measures thoughout the family identify dependencies correctly. From data omitted in the tables it can be observed that, as in the case of synthetic data, when a approaches 1 the measures become more and more symmetric and unaffected by independent attributes in the consequent.

3.7 Pruning Redundant Association Rules Using Maximum Entropy Principle

We begin by presenting a method of association rule pruning using Maximum Entropy approach. A subrule of an association rule $I \rightarrow J$ is a rule $K \rightarrow J$, such that $K \subset I$ (see [AIS93] or further sections for a detailed discussion of association rules). In [LHM99, AL99] a rule is considered not interesting if its confidence is close to that of one of its subrules. A similar approach (although in a slightly generalized setting) is used in [PT00] to prune the discovered rules. Also, in [PT00] a rule is considered interesting with respect to some set of beliefs if it contradicts at least one of the rules in the beliefs under the so-called monotonicity assumption. A detailed statistical analysis of interestingness of a rule with respect to a single subrule, and algorithms for finding rules interesting in this setting can be found

in [Suz97, SK98].

The current work on evaluation of interestingness considers the influence of each subrule separately, while in our approach we take into account the combined influence of all the subrules of a rule. Examples illustrating the advantages (in our opinion) of our approach are given in Section 3.9.

In [LHM99], apart from pruning, the authors also find so called *direction setting* rules which summarize the dataset. This procedure takes into account many subrules of a rule and is thus similar to our approach. However, our approach has the advantage of giving a more precise, probabilistic quantification of the influence of subrules on the interestingness of a rule.

Another approach to pruning discovered rules is based on selecting a minimal set of rules covering the dataset [TKR95, BVW00]. Again, those methods do not take into consideration probabilistic interactions between rules in the cover. Also, they may prune many interesting rules if they cover instances already covered by other rules.

A general study of measures of rule interestingness can be found in [BA99, JS01, HH99].

An overview of the interestingness of a rule with respect to a set of constraints can be found in [GHK94]. In [GHK94] the authors propose the method of *random worlds* and prove that in many important cases it is equivalent to the principle of maximum entropy.

Maximum entropy principle and other probability models have been also used in datamining in query selectivity estimation [PMS01]. There has also been work in applying MaxENT in speech processing [Rat96]

When $\mathrm{Dom}(A) = \{0, 1\}$ we say that A is a binary attribute. In this, and

the following sections of this chapter we use tables whose headings have the form $H = \{A_1, A_2, \ldots, A_m\}$ and consist of binary attributes.

Since attributes of H are assumed to be binary, subsets of H will be referred to as *itemsets*, according to standard datamining terminology [AIS93].

Since the choice of divergence is immaterial for the rest of the discussion we will simply denote the divergence by D meaning that either Kullback-Leibler or chi-squared divergence can be used.

A *constraint* C on the set of attributes H is a pair $\mathsf{C} = (I, p)$ where $I \subseteq H$, $p \in [0, 1]$. A probability distribution P *satisfies* a constraint $\mathsf{C} = (I, p)$ if $P_I(\mathbf{1}_I) = p$, where $\mathbf{1}_I = (1, 1, \ldots, 1) \in \mathrm{Dom}(I)$. Usually, the attribute set will be clear from context, so we will just write $\mathbf{1}$ instead of $\mathbf{1}_I$.

To remove redundancies in the rule set we need to define how interesting a rule is with respect to a set of constraints introduced by other rules.

Definition 3.7.1 A set of constraints $\mathcal{C}$ is *consistent* if there exists a joint probability distribution over H which satisfies all the constraints in $\mathcal{C}$. Otherwise, $\mathcal{C}$ is *inconsistent*. □

Here we will only be concerned with consistent sets of constraints. Dealing with inconsistent sets of constraints is an interesting topic of future research.

While determining interestingness of rules with respect to a consistent set of constraints $\mathcal{C}$ we will associate with $\mathcal{C}$ some joint probability distribution $P^{\mathcal{C}}$ over H.

Note that a set of constraints does not have to determine the joint probability distribution uniquely, and we have to choose one of the conforming distributions. The three main approaches to this problem are the maximum entropy principle

(MaxENT), the minimum interdependence principle, and the maximum likelihood (see [KK92, Adw97]). We use MaxENT, but it can be shown [KK92, Adw97], that in most cases all three approaches are equivalent. Philosophical justifications of the principles can be found in [KK92, GHK94].

Definition 3.7.2 Let $\mathcal{C}$ be a consistent set of constraints. A *probability distribution $P^{\mathcal{C}}$ over H is induced by $\mathcal{C}$* if it satisfies the following conditions:

1. $P^{\mathcal{C}}$ satisfies all the constraints in $\mathcal{C}$.
2. Of all probability distributions over H satisfying $\mathcal{C}$, $P^{\mathcal{C}}$ has the largest entropy.

□

It can be shown [Adw97] that $P^{\mathcal{C}}$ is unique.

3.8 Interestingness of A Rule with Respect to A Set of Constraints

We are now ready to define the interestingness of an association rule with respect to some set of constraints $\mathcal{C}$. For the definition of association rules see [AIS93].

The support of an itemset I is $\text{supp}(I) = \hat{P}_I(\mathbf{1})$. Rules with empty antecedents are allowed and the support and confidence of such rules are defined to be equal to the support of their consequents.

The set of constraints generated by an association rule $I \rightarrow J$ is defined as

$$\mathsf{C}(I \rightarrow J) = \{(I, \text{supp}(I)), (I \cup J, \text{supp}(I \cup J))\}.$$

We introduce two interestingness measures for association rules: the active and passive interestingness. The active interestingness reflects the impact of adding to the current set of constraints the set of constraints generated by the rule itself. The passive interestingness is determined by the difference between the confidence estimated from the data and the confidence estimated starting from the probability distribution induced by the constraints.

Definition 3.8.1 Let $\mathcal{C}$ be a consistent set of constraints, $I \to J$ be a rule and D some measure of distribution divergence. Denote by $Q^{\mathcal{K}}$ the probability distribution over $I \cup J$ induced by the set of constraints $\mathcal{K}$.

The *active interestingness* of $I \to J$ with respect to $\mathcal{C}$ is defined as:

$$\mathsf{I}^{\text{act}}(\mathcal{C}, I \to J) = D(Q^{\mathcal{C} \cup \mathsf{C}(I \to J)}, Q^{\mathcal{C}}).$$

The *passive interestingness* of $I \to J$ with respect to $\mathcal{C}$ is defined as:

$$\mathsf{I}^{\text{pass}}(\mathcal{C}, I \to J) = \left| \text{conf}(I \to J) - \frac{Q^{\mathcal{C}}(\mathbf{1})}{Q^{\mathcal{C}}_I(\mathbf{1})} \right|,$$

where $\text{conf}(I \to J)$ denotes the confidence of rule $I \to J$. □

Whenever we state facts that hold for either of these measures we simply talk about rule interestingness I.

3.9 The pruning algorithm

Definition 3.9.1 Let $\mathcal{R}$ be a set of association rules. Consider an association rule $I \to J$, where $I, J \subseteq H$. The rule $I \to J$ is I-*nonredundant with respect to* $\mathcal{R}$, if $I = \emptyset$ or $\mathsf{I}(\mathcal{C}^{I,J}(\mathcal{R}), I \to J)$ is significantly greater than 0, where $\mathcal{C}^{I,J}(\mathcal{R}) = \{\mathsf{C}(K \to J) : K \to J \in \mathcal{R}, K \subset I\}$. □

Note that we do not specify precisely what 'significantly greater' means. This may mean 'greater than some threshold' or 'statistically significant at some confidence level' or some combination of both.

A feature of our definition of redundancy is that it is not influenced by rules involving attributes not in $I \cup J$. For example, suppose that the joint distribution of attributes ABC is fully explained by rules $A \to B$ and $B \to C$. The rule $A \to C$ may still be considered I-nonredundant, even though it does not introduce any new information.

We believe this is the correct behavior. In general, if we have a long chain of rules $A \to B \to C \to \ldots \to Y \to Z$, the rule $A \to Z$ might not be easy to see and thus be interesting. Furthermore, the discovered rules do not necessarily correspond to true causality relations, and it might be better, at least until the user develops a better understanding of the dataset, to present him/her also rules indirectly implied by some other rules.

Another important advantage of our method is that single rules usually involve very few attributes, and thus local interestingness can be efficiently determined, even by direct application of the Generalized Iterative Scaling algorithm, as we show later.

An algorithm for producing a set of I-nonredundant rules with a single attribute in the consequent is given in Figure 3.1.

Examples below show how our method compares with other work in certain situations. Passive interestingness measure I^{pass} is used, but it is easy to see that the statements remain valid also for the active interestingness measure I^{act}. See discussion later in this section for details on how the maximum entropy distributions can be computed.

Input: A set $\mathcal{S}$ of association rules.

Output: Set $\mathcal{R}$ of l-nonredundant association rules of $\mathcal{S}$.

1. For each $A_i \in H$
2. $\quad \mathcal{R}_i = \{\emptyset \rightarrow A_i\}$
3. $\quad k = 1$
4. $\quad$ For each rule $I \rightarrow A_i \in \mathcal{S}$, $|I| = k$ do
5. $\quad\quad$ If $I \rightarrow A_i$ is l-nonredundant with respect to $\mathcal{R}_i$ then
6. $\quad\quad\quad$ Let $\mathcal{R}_i = \mathcal{R}_i \cup \{I \rightarrow A_i\}$
7. $\quad k = k + 1$
8. $\quad$ Goto 4
9. $\mathcal{R} = \bigcup_{A_i \in H} \mathcal{R}_i$

Figure 3.1: An lgorithm for finding l-nonredundant association rules.

Example 3.9.2 Let A, B, C be binary attributes, $P_A(1) = P_B(1) = 0.5$. The attribute C depends on A, B according to the following association rules:

assoc. rule	confidence
$\emptyset \to C$	0.5
$A \to C$	0.3
$B \to C$	0.7
$AB \to C$	0.3

Using the approach from [PT00, LHM99, AL99, SLR99] rules $\emptyset \to C$, $A \to C$ and $B \to C$ are interesting but $AB \to C$ is not, since it is explained by the rule $A \to C$. We claim however that the rule $AB \to C$ is interesting, since it tells us that when both A and B are 'present' it is A that influences C stronger.

Consider rules $\emptyset \to C$, $A \to C$, and $B \to C$. The set of constraints corresponding to them is $\mathcal{C} = \{(A, 0.5), (B, 0.5), (C, 0.5), (AC, 0.15), (BC, 0.35)\}$. The MaxENT distribution in this case is

$$P^{\mathcal{C}} = \begin{pmatrix} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ 0.105 & 0.105 & 0.045 & 0.245 & 0.245 & 0.045 & 0.105 & 0.105 \end{pmatrix},$$

and $P^{\mathcal{C}}_{ABC}(\mathbf{1})/P^{\mathcal{C}}_{AB}(\mathbf{1}) = 0.5$, different from $\text{conf}(AB \to C) = 0.3$, making the rule $AB \to C$ interesting. □

Example 3.9.3 Assume now that the confidences of the rules in the example above are

assoc. rule	confidence
$\emptyset \to C$	0.5
$A \to C$	0.3
$B \to C$	0.7
$AB \to C$	0.5

Using methods given in [LHM99, AL99] the rule $AB \to C$ is interesting, since its confidence differs from $\text{conf}(A \to C)$ and $\text{conf}(B \to C)$.

However, as seen above, the maximum entropy distribution induced by rules $\emptyset \to C$, $A \to C$ and $B \to C$ gives $P^{\mathcal{C}}_{ABC}(\mathbf{1})/P^{\mathcal{C}}_{AB}(\mathbf{1}) = 0.5$, and the rule $AB \to C$ is considered uninteresting. In other words, knowing the joint influence of AB on C does not give us any more information over what we have already know from other rules, since A and B are conditionally independent given C. The above result is intuitive since when both A and B influence C we would expect their joint influence to be an 'average' between the influences of A and B alone. □

Example 3.9.4 Suppose that attribute A is independent of B, C, and jointly of BC. Then, $P^{\mathcal{C}}_{ABC}(\mathbf{1})/P^{\mathcal{C}}_{AB}(\mathbf{1}) = P^{\mathcal{C}}_{BC}(\mathbf{1})/P^{\mathcal{C}}_{B}(\mathbf{1}) = \text{conf}(B \to C)$, and the rule $AB \to C$ is considered not interesting using our approach, but also using methods from [PT00, LHM99, SLR99, AL99] which explains their good behavior in practice. However as the examples above show, those methods can filter out certain interesting rules, and include some uninteresting ones. □

To compute the maximum entropy distribution we can use the Generalized Iterative Scaling (GIS) Algorithm [Adw97, Bad95, DR72, Csi89].

Let $\mathcal{C} = \{\mathsf{C}_1, \mathsf{C}_2, \ldots, \mathsf{C}_n\}$ be a set of constraints, where $\mathsf{C}_k = (I_k, p_k)$. GIS proceeds by assigning some initial values to each probability in $P^{\mathcal{C}}$, and iteratively updating them until all the constraints are satisfied. Let $P^{\mathcal{C}(i)}$ denote the distribution after i iterations. Updating in each iteration is performed according to the formula

$$P^{\mathcal{C}(i+1)}(\mathbf{h}) = P^{\mathcal{C}(i)}(\mathbf{h}) \prod_{\mathbf{h}_{I_k}=\mathbf{1}} \left[\frac{p_k}{P^{\mathcal{C}(i)}_{I_k}(\mathbf{h}_{I_k})} \right]^{\frac{1}{c}},$$

for every $\mathbf{h} \in \text{Dom}(H)$, assuming that $\frac{0}{0} = 0$. The algorithm is guaranteed to

converge if $\sum_{k=1}^{n} f_k(\mathbf{h}) = c$ is a constant independent of $\mathbf{h}$. In practice, this condition can always be satisfied by adding an additional constraint. See [Adw97, DR72, Csi89] for details and proof of convergence. The version of the algorithm presented in [Csi89] has the advantage of being able to cope with distributions with zero probabilities, and this is the one we use in our implementation.

The disadvantage of the GIS algorithm is its high computational cost caused by the necessity of computing the marginal probabilities, and in some cases by the large number of iterations required.

One of the main techniques for speeding up MaxENT computations is *decomposition* [Bad95, And74, DLS80]. However, in our case, we will only use maximum entropy distributions in few variables, and our experiments showed that decomposition does not give real improvement in efficiency. We noticed however that the number of rules considered interesting is small and thus constraints are usually simple. Closed form solutions are used for a few common cases; in every other situation we use the GIS algorithm.

We describe below the closed form solutions. For attribute set I denote $N_I = |\{\mathbf{x} \in \mathrm{Dom}(H) : \mathbf{x}_I = \mathbf{1}\}|$.

Theorem 3.9.5 *Let* $\mathcal{C} = \{(J, \hat{P}_J(\mathbf{1})), (K, \hat{P}_K(\mathbf{1})), (K \cup J, \hat{P}_{K\cup J}(\mathbf{1}))\}$, *where* $J, K \subset H, K \cap J = \emptyset$ *be a set of constraints. The MaxENT distribution induced by* $\mathcal{C}$ *is*

$$P^{\mathcal{C}}(\mathbf{x}) = \begin{cases} \frac{\hat{P}_{K\cup J}(\mathbf{1})}{N_{K\cup J}} & , \text{ if } \mathbf{x}_J = \mathbf{1} \wedge \mathbf{x}_K = \mathbf{1} \\ \frac{\hat{P}_J(\mathbf{1}) - \hat{P}_{K\cup J}(\mathbf{1})}{N_J - N_{K\cup J}} & , \text{ if } \mathbf{x}_J = \mathbf{1} \wedge \mathbf{x}_K \neq \mathbf{1} \\ \frac{\hat{P}_K(\mathbf{1}) - \hat{P}_{K\cup J}(\mathbf{1})}{N_K - N_{K\cup J}} & , \text{ if } \mathbf{x}_J \neq \mathbf{1} \wedge \mathbf{x}_K = \mathbf{1} \\ \frac{1 - \hat{P}_K(\mathbf{1}) - \hat{P}_J(\mathbf{1}) + \hat{P}_{K\cup J}(\mathbf{1})}{|\mathrm{Dom}(H)| - N_K - N_J + N_{K\cup J}} & , \text{ if } \mathbf{x}_J \neq \mathbf{1} \wedge \mathbf{x}_K \neq \mathbf{1}, \end{cases}$$

for $\mathbf{x} \in \mathrm{Dom}(H)$.

Proof. For every $R \subseteq \{K, J\}$ denote X_R the set of all $\mathbf{x} \in \mathrm{Dom}(H)$ such that $\mathbf{x}_I = \mathbf{1}$ if $I \in R$ and $\mathbf{x}_I \neq \mathbf{1}$ otherwise, for all $I \in \{J, K\}$. Note that $|X_{\{K,J\}}| = N_{K\cup J}$, $|X_{\{J\}}| = N_J - N_{K\cup J}$, $|X_{\{K\}}| = N_K - N_{K\cup J}$, and $|X_\emptyset| = |\mathrm{Dom}(H)| - N_K - N_J + N_{K\cup J}$. Also denote $P^*_R = \sum_{\mathbf{x}\in X_R} \hat{P}(\mathbf{x})$ for all $R \subseteq \{K, J\}$. Note that $P^*_{\{K,J\}} = \hat{P}_{K\cup J}(\mathbf{1})$, $P^*_{\{J\}} = \hat{P}_J(\mathbf{1}) - \hat{P}_{K\cup J}(\mathbf{1})$, $P^*_{\{K\}} = \hat{P}_K(\mathbf{1}) - \hat{P}_{K\cup J}(\mathbf{1})$, and $P^*_\emptyset = 1 - \hat{P}_K(\mathbf{1}) - \hat{P}_J(\mathbf{1}) + \hat{P}_{K\cup J}(\mathbf{1})$.

For a probability distribution P on H that satisfies the the set of constraints $\mathcal{C}$ we have:

$$\begin{aligned}\mathcal{H}(P) &= -\sum_{R\subseteq\{K,J\}} \sum_{\mathbf{x}\in X_R} P(\mathbf{x}) \log P(\mathbf{x}) \\ &= -\sum_{R\subseteq\{K,J\}} P^*_R \sum_{\mathbf{x}\in X_R} \frac{P(\mathbf{x})}{P^*_R} \log \frac{P(\mathbf{x})}{P^*_R} - \sum_{R\subseteq\{K,J\}} P^*_R \log P^*_R.\end{aligned}$$

It suffices to maximize the first term. Notice that for every $R \subseteq \{K, J\}$, we have $\sum_{\mathbf{x}\in X_R} \frac{P(\mathbf{x})}{P^*_R} = 1$, and thus P/P^*_R is a probability distribution over X_R, and its entropy $-\sum_{\mathbf{x}\in X_R} \frac{P(\mathbf{x})}{P^*_R} \log \frac{P(\mathbf{x})}{P^*_R}$ is maximized when $\frac{P(\mathbf{x})}{P^*_R} = 1/|X_R|$ for every $\mathbf{x} \in X_R$. This gives $P(\mathbf{x}) = P^*_R/|X_R|$ for every $\mathbf{x} \in X_R$ and completes the proof since every $\mathbf{x} \in \mathrm{Dom}(H)$ belongs to exactly one of the X_R's. ∎

Notice that when $J = K$, the above result reduces to

$$P^{\mathcal{C}}(\mathbf{x}) = \begin{cases} \frac{\hat{P}_J(\mathbf{1})}{N_J} & , \text{if } \mathbf{x}_J = \mathbf{1} \\ \frac{1-\hat{P}_J(\mathbf{1})}{|\mathrm{Dom}(H)|-N_J} & , \text{if } \mathbf{x}_J \neq \mathbf{1}, \end{cases} \tag{3.7}$$

for $\mathbf{x} \in \mathrm{Dom}(H)$.

Frequently, the only subrules of a rule $I \to J$ are $\emptyset \to J$, and $K \to J$, where $K \subset I$. In this case the MaxENT distribution induced by the subrules can be found by application of Theorem 3.9.5. If the only subrule is $\emptyset \to J$, then we can use Equality (3.7). Our experiments revealed that using the above theorem

reduces pruning time up to a factor of 10. See [Bad95, PMS01] for a more detailed discussion of methods of speeding up MaxENT computations.

3.10 Experimental Evaluation of the Pruning Algorithm

In this section we present an experimental evaluation of our pruning algorithm. We used passive interestingness I^{pass}, and considered a rule I^{pass}-nonredundant if its passive interestingness was greater than some threshold. Our experiments have shown that the passive measure of interestingness performed better than the active one I^{act}, which often pruned interesting rules with small support. The reason is that rules with small support usually have many attributes in the antecedent, and thus adding them as constraints affects only very few values in the joint probability distribution, while active interestingness depends on the whole distribution. Also, we did not use any minimum confidence threshold, because pruning provided a sufficient reduction in the number of rules, and setting a minimum confidence threshold occasionally pruned some of the interesting rules.

We first present the result of running the algorithm on the `lenses` database from the UCI machine learning archive [BM98]. The database has the advantage of being very small thus allowing manual selection of rules. Table 3.3 shows the rules having the `lenses` attribute as consequent, selected manually by the authors, providing a complete description of the dataset. Table 3.4 shows rules involving `lenses` attribute as consequent generated by the Apriori algorithm with minimum support 1 (1 record), no minimum confidence, post-processed with our pruning algorithm using passive interestingness with interestingness threshold 0.3. Negative values of interestingness mean that the presence of the antecedent decreases the probability of presence of the consequent.

antecedent→lenses	conf. [%]	supp. [%]
∅→soft	20.8	20.8
∅→hard	16.6	16.6
∅→none	62.5	62.5
tears=reduced→none	100	50
astigmatism=no,tears=normal→soft	83.3	20.8
astigmatism=yes,tears=normal→hard	66.6	16.6
age=pre-presbyopic,prescription=hypermetrope,astigmatism=yes→none	100	8.3
age=presbyopic,prescription=myope,astigmatism=no→none	100	8.3
age=presbyopic,prescription=hypermetrope,astigmatism=yes→none	100	8.3

Table 3.3: Rules manually selected from the `lenses` database

Rules have been sorted based on the product of support and interestingness, with an extra condition, that a rule cannot be printed until all its subrules have been printed. Also, note that the `lenses` dataset contains multivalued attributes. Since our method only handles boolean attributes we encode each original attribute with a number of boolean attributes, one for each possible value of the original attribute.

The Apriori algorithm produced 113 rules having `lenses` attribute as the consequent. After pruning, 16 nonredundant rules were left with a nonempty antecedent. This is a significant reduction.

When rules with all possible consequents are considered, our method outputs 40 rules out of 890 produced by Apriori. Also, note that all rules selected manually are also considered interesting by our pruning algorithm, and the top three rules are indeed identical in both cases, which suggests that really interesting rules are indeed retained by our algorithm.

antecedent→lenses	I^P [%]	conf. [%]	supp. [%]
∅→soft	0	20.8	20.8
∅→hard	0	16.6	16.6
∅→none	0	62.5	62.5
tears=reduced→none	37.5	100	50
astigmatism=no,tears=normal→soft	62.5	83.3	20.8
astigmatism=yes,tears=normal→hard	50	66.6	16.6
tears=normal→none	-37.5	25	12.5
prescription=myope,astigmatism=yes→hard	33.3	50	12.5
prescription=myope,tears=normal→hard	33.3	50	12.5
prescription=hypermetrope,astigmatism=yes,tears=normal→none	41.4	66.6	8.3
age=pre-presbyopic,prescription=hypermetrope,astigmatism=yes→none	37.5	100	8.3
age=presbyopic,prescription=myope,astigmatism=no→none	37.5	100	8.3
age=presbyopic,prescription=hypermetrope,astigmatism=yes→none	37.5	100	8.3
age=young,astigmatism=yes→hard	33.3	50	8.3
age=young,tears=normal→hard	33.3	50	8.3
age=presbyopic,astigmatism=no,tears=normal→soft	-32.9	50	4.1
age=presbyopic,prescription=hypermetrope,astigmatism=no,tears=normal→soft	49.3	100	4.1
prescription=hypermetrope,astigmatism=yes,tears=normal→hard	-32.9	33.3	4.1
age=young,prescription=hypermetrope,astigmatism=yes,tears=normal→hard	39.3	100	4.1

Table 3.4: Rules selected from the `lenses` database

We also applied our method to a dataset of census data of elderly people obtained from The University of Massachusetts at Boston Gerontology Center. The dataset contains about 330 thousand records, 11 attributes with up to five values, and is available at `http://www.cs.umb.edu/ ~sj /datasets /census.arff.gz`. We used 1% minimum support and no minimum confidence. The Apriori algorithm produced 247476 rules practically impossible to analyze by hand. After pruning with 10% interestingness threshold only 2056 were considered nonredundant, and after further restricting this set to rules with a given consequent attribute we were able to obtain easily manageable sets of interesting association rules. Some of them, concerning the `urban` (whether a person lives in a city or not) attribute are given in Table 3.5. Although the pruning time was quite long (over 4 hours on a 100MHz Pentium machine), it was still much easier to use our method than to handle hundreds of thousands of rules manually. See Table 3.6 for further details.

antecedent→consequent	I^P [%]	conf. [%]	supp. [%]
∅→urban=no	0	22.4	22.4
∅→urban=yes	0	77.5	77.5
immigr=no,region=south→urban=yes	-11.8	65.7	26.2
race=white→urban=yes	-10.6	66.8	22.5
region=west→urban=yes	12.8	90.3	16.9
race=hisp→urban=yes	12.4	89.9	15.4
region=south,race=black→urban=yes	-10.6	66.8	17.8
immigr=no,region=south→urban=no	11.8	34.2	13.7
alone=yes,region=south→urban=yes	-10.5	66.9	15
immigr=before75→urban=yes	15.9	93.4	9.7
region=neast,race=black→urban=yes	19.7	97.2	6.7
region=midw,race=black→urban=yes	18.9	96.5	6.9
age=below75,region=neast→urban=yes	10.5	88	11.3
race=white→urban=no	10.6	33.1	11.1

Table 3.5: Top 12 rules involving `urban` attribute generated from the elderly people census data

dataset	min. support	interestingness threshold	number of rules		pruning time [s]
			Apriori	after pruning	
lenses	1(4%)	0.3	890	40	1.3
mushroom*	500(16%)	0.2	164125	5141	418
breast-cancer	30(10%)	0.15	2128	74	2.8
primary-tumor*	30(9%)	0.3	43561	67	21.8
primary-tumor*	30(9%)	0.2	43561	432	24
car	10(0.5%)	0.3	20669	293	11.1
car	10(0.5%)	0.15	20669	580	30.2
splice*	300(9%)	0.5	4847	24	3.0
splice*	300(9%)	0.3	4847	95	5.6
splice*	300(9%)	0.15	4847	290	7.2
splice*	200(6%)	0.3	35705	463	33.8
census(elderly people)	3000(1%)	0.3	247476	194	4801
census(elderly people)	3000(1%)	0.2	247476	621	5683
census(elderly people)	3000(1%)	0.1	247476	2056	15480

* itemsets with up to 4 attributes

Table 3.6: Numbers of rules and computation times for various datasets

Table 3.6 shows the number of rules generated by Apriori compared with the number of rules considered interesting by our algorithm, as well as pruning time, for various datasets from the UCI Machine Learning Archive [BM98]. All datasets have been mined with 0 minimum confidence. The interestingness thresholds and minimum supports have been chosen manually by trial and error such that the unpruned rules provide a lot of interesting information while keeping their number reasonably small. For some datasets values for a few different thresholds are given for comparison. All experiments have been performed on a 100MHz Pentium machine with 64MB of memory.

CHAPTER 4

MEASURES ON BOOLEAN POLYNOMIALS AND THEIR APPLICATIONS IN DATA MINING

4.1 Introduction

The focus of this chapter is a study of measures on free Boolean algebras with a finite number of generators (abbreviated as MFBAs). As we shall see, these measures play an important role in query optimization in relational databases, and also, in the study of frequent sets in data mining. We obtain general Bonferroni-type inequalities for sizes of arbitrary Boolean queries. The origin of our investigation resides in a series of seminal papers by H. Mannila et al. ([MT96, Man01, PMS01]) in which the idea of using supports of attribute sets discovered with a data mining algorithm to obtain the size of a database query was introduced.

The same problem has been investigated in parallel in [CG02] using different methodology. The authors managed to obtain tight bounds for support of an itemset for an important case when supports of all its subsets are known, and proved that width of the bounds decrease exponentially with itemsets size. Tightness of their bound is an important advantage over our results, which are however more general. It is possible for example to directly obtain bounds for support of an

itemset when all its subsets of a given size are known (not necessarily all of its subsets). The method from [CG02] requires recursion in such a case which can be computationally expensive.

Similar problems have been addressed in the area of statistical data protection, where it is important to assure that inferences about individual cases cannot be made from marginal totals (see [Dob01, BG99] for an overview). Those methods concentrate on obtaining the most accurate bounds possible (in order to rule out information disclosure), computational efficiency being a secondary concern. Algorithms usually involve repeated iterations over full contingency tables [BG99], branch and bound search [Dob01] or numerous applications of linear programming.

Let $\mathcal{B} = (B, \mathbf{0}, \mathbf{1}, \bar{}\,, \vee, \wedge)$ be a Boolean algebra, where $\mathbf{0}, \mathbf{1} \in B$ are two distinguished elements of $\mathcal{B}$, $\bar{}$ is a unary operation, and $\vee, \wedge$ are two binary associative, commutative, and idempotent operations that satisfy the usual axioms of Boolean algebras (see, for example [Rud74]). Here $\mathbf{0}$ and $\mathbf{1}$ are the least and the largest element of the algebra, respectively.

We define $x^b = x$ if $b = 1$ and $x^b = \bar{x}$ if $b = 0$, for $x \in B$ and $b \in \{0, 1\}$.

It is a well-known fact that a Boolean algebra $\mathcal{B} = (B, \mathbf{0}, \mathbf{1}, \bar{}\,, \vee, \wedge)$ defines a Boolean ring structure, $\mathcal{B} = (B, \mathbf{0}, \mathbf{1}, \wedge, \oplus)$, where $\wedge$ plays the role of the multiplication, and $\oplus$ the role of addition, where

$$x \oplus y = (x \wedge \bar{y}) \vee (\bar{x} \wedge y)$$

for $x, y \in B$. This ring is unitary, commutative, and has characteristic 2 (since $x \oplus x = \mathbf{0}$ for every x). Also, $\mathbf{1} \oplus x = \bar{x}$.

Let $A = \{a_1, \ldots, a_n\}$ be a set of n variables. The set $\mathsf{pol}(A)$ of *Boolean polynomials of the n variables in A* is defined inductively by:

1. $\mathbf{0}$, $\mathbf{1}$, and each a_i belong to $\mathsf{pol}(A)$ for $1 \leq i \leq n$;

2. if p, q belong to $\mathsf{pol}(A)$, then $\bar{p}$, $(p \vee q)$, and $(p \wedge q)$ belong to $\mathsf{pol}(A)$.

If $p, q \in \mathsf{pol}(A)$, then we denote by $(p \oplus q)$ the polynomial $((p \wedge \bar{q}) \vee (\bar{p} \wedge q))$. A Boolean polynomial $(\cdots((p_1 \omega p_2)\omega p_3)\omega \cdots \omega p_n)$ is denoted by $(p_1 \omega p_2 \omega \cdots \omega p_n)$, where $\omega \in \{\vee, \wedge, \oplus\}$.

Let $\mathcal{B} = (B, \mathbf{0}, \mathbf{1}, \bar{}\,, \vee, \wedge)$ be a Boolean algebra and let $A = \{a_1, \ldots, a_n\}$ be a set of n variables. The n-ary function $f_p : B^n \longrightarrow B$ generated by a polynomial $p \in \mathsf{pol}(A)$ is defined in the usual way. We write $p = q$ for $p, q \in \mathsf{pol}(A)$ if $f_p = f_q$.

Let $\vec{b} = (b_1, \ldots, b_n)$ be a sequence of elements of the set $\{0, 1\}$. An A-minterm is a Boolean polynomial

$$p_{\vec{b}} = a_1^{b_1} \wedge \cdots \wedge a_n^{b_n},$$

The set of A-minterms is denoted by $\mathsf{mint}(A)$. Any Boolean polynomial in $\mathsf{pol}(A)$ can be uniquely written as a disjunction of some subset of A-minterms (up to the order of the disjuncts). This observation implies that the Boolean algebra $\mathsf{pol}(A)$ is isomorphic to the Boolean algebra of collections of subsets of the set A; thus, $\mathsf{pol}(A)$ has 2^{2^n} elements.

For a set of polynomials $M = \{p_1, \ldots, p_n\}$ and $J = \{j_1, \ldots, j_m\} \subseteq \{1, \ldots, n\}$ we denote by p_J the conjunction $p_{j_1} \wedge \cdots \wedge p_{j_m}$. For the special case, when $J = \emptyset$ we write $p_J = \mathbf{1}$.

A *measure* on a Boolean algebra $\mathcal{B} = (B, \mathbf{0}, \mathbf{1}, \bar{}\,, \vee, \wedge)$ is a non-negative, real-valued function $\mu : B \longrightarrow \mathbb{R}$ such that $\mu(x \vee y) = \mu(x) + \mu(y)$ for every $x, y \in B$ such that $x \wedge y = \mathbf{0}$.

4.2 A Representation Result for MFBAs

In this context, we find it convenient to use the relational database terminology. We modify slightly the database related notation of previous chapters to better fit in the Boolean algebra setting.

$A = \{a_1, \ldots, a_n\}$ be a set of variables. Members of A as *attributes*. We attach a set $\mathrm{Dom}(a_i)$ to each attribute a_i such that $|\mathrm{Dom}(a_i)| \geq 2$. The set $\mathrm{Dom}(a_i)$ is the *domain of* a_i.

A table is a triple $\tau = (T, A, \rho)$, where T is the name of the table, $A = \{a_1, \ldots, a_n\}$ is the heading of the table and $\rho = \{t_1, \ldots, t_m\}$ is a finite set of functions of the form $t_i : A \longrightarrow \bigcup_{a \in A} \mathrm{Dom}(a)$ such that $t_i(a) \in \mathrm{Dom}(a)$ for every $a \in A$. Following the relational database terminology we shall refer to these functions as A-tuples, or simply as tuples. If $\mathrm{Dom}(a_i) = \{0, 1\}$ for $1 \leq i \leq n$, then τ is a binary table.

Let $\tau = (T, A, \rho)$ be a binary table. A *query on the table* τ is a Boolean polynomial in $\mathsf{pol}(A)$. This definition of queries is a formalization of the usual notion of queries in databases.

Example 4.2.1 To retrieve in SQL all tuples t of τ such that at least two of $t(a_1), t(a_2)$ and $t(a_3)$ equal 1 we write the query as

select * **from** T **where** $(a_1 = 1$ **and** $a_2 = 1)$ **or** $(a_2 = 1$ **and** $a_3 = 1)$
or $(a_1 = 1$ **and** $a_3 = 1)$

The condition specified in this select corresponds to the polynomial $(a_1 \wedge a_2) \vee (a_2 \wedge a_3) \vee (a_1 \wedge a_3)$. □

A query p defines a table (T_p, A, ρ_p), where ρ_p is defined inductively as follows:

1. $\rho_{\mathbf{0}} = \emptyset$ and $\rho_{\mathbf{1}} = \rho$;

2. if $p = a_i$, then $\rho_p = \{t \in \rho | t(a_i) = 1\}$;

3. if $p = \bar{q}$, then $\rho_p = \rho - \rho_q$;

4. if $p = (q_1 \vee q_2)$, then $\rho_p = \rho_{p_1} \cup \rho_{p_2}$ and,

5. if $p = (q_1 \wedge q_2)$, then $\rho_p = \rho_{p_1} \cap \rho_{p_2}$.

It is easy to see that for a conjunction

$$p = a_{i_1}^{b_1} \wedge \cdots \wedge a_{i_m}^{b_m},$$

where $b_i \in \{0, 1\}$ for $1 \leq i \leq m$, the set ρ_p consists of those tuples t such that $t(a_{i_\ell}) = b_\ell$ for $1 \leq \ell \leq m$.

Theorem 4.2.2 *A function $\mu : \mathsf{pol}(A) \longrightarrow \mathbb{N}$ is a measure if and only if there exists a binary table $\tau = (T, A, \rho)$ such that $\mu(p) = |\rho_p|$ for all $p \in \mathsf{pol}(A)$.*

Proof. Suppose that $\tau = (T, A, \rho)$ is a table. Define the mapping $\mu_\tau : \mathsf{pol}(A) \longrightarrow \mathbb{R}$ by $\mu(p) = |\rho_p|$ for every $p \in \mathsf{pol}(A)$. Let p, q be two polynomials such that $(p \wedge q) = \mathbf{0}$. Then, $\mu_\tau(p \vee q) = |\rho_{p \vee q}| = |\rho_p \cup \rho_q|$. Since $p \wedge q = \mathbf{0}$ we have $\rho_p \cap \rho_q = \emptyset$, so $\mu_\tau(p \vee q) = \mu_\tau(p) + \mu_\tau(q)$. Thus, μ_τ is a measure on $\mathsf{pol}(A)$.

Conversely, let μ be a measure on $\mathsf{pol}(A)$, where $A = \{a_1, \ldots, a_n\}$. If $\vec{b} = (b_1, \ldots, b_n) \in \{0, 1\}^n$, $p_{\vec{b}} = a_1^{b_1} \wedge \cdots \wedge a_n^{b_n}$ is a minterm and $\mu(p_{\vec{b}}) = k$ consider a set $\sigma_{p_{\vec{b}}}$ of k tuples $t_{\vec{b}}^1, \ldots, t_{\vec{b}}^k$, where $t_{\vec{b}}^j(a_i) = b_i$ for every i, j, $1 \leq j \leq k$, and $1 \leq i \leq n$. Define the table $\tau_\mu = (T, A, \rho)$, where $\rho = \bigcup \{\sigma_{p_{\vec{b}}} | p_{\vec{b}} \in \mathsf{mint}(A)\}$.

We claim that $\mu(p) = |\rho_p|$ for every polynomial $p \in \mathsf{pol}(A)$. Suppose that p can be expressed as a disjunction of minterms $p = p_{\vec{b}_1} \vee \cdots \vee p_{\vec{b}_k}$, where $\vec{b}_1, \ldots, \vec{b}_k \in$

$\{\mathbf{0},\mathbf{1}\}^n$. Then, $\mu(p) = \sum_{j=1}^{k} \mu(p_{\vec{b}_j})$, because $p_{\vec{b}_l} \wedge p_{\vec{b}_h} = \mathbf{0}$ when $l \neq h$. On the other hand, $|\rho_p| = |\bigcup_{j=1}^{k} \rho_{p_{\vec{b}_j}}| = \sum_{j=1}^{k} |\rho_{p_{\vec{b}_j}}|$, so $\mu(p) = |\rho_p|$. ∎

We shall refer to μ_τ as the *measure induced by the table* τ on $\mathsf{pol}(A)$.

Measures induced by tables are generated by pseudo-Boolean functions which range over the set $\mathbb{N}$ (see [HR66]). Namely, let $A = \{a_1, \ldots, a_n\}$ be a set of n attributes. Define the pseudo-Boolean function $f : \{0,1\}^n \longrightarrow \mathbb{N}$ by $f(b_1, \ldots, b_n) = \mu_\tau(p_{b_1,\ldots,p_n})$. Then, it is easy to verify that for every polynomial $p \in \mathsf{pol}(A)$ we have

$$\mu_\tau(p) = \sum \{f(\vec{b}) \mid p_{\vec{b}} \in \mathsf{mint}(A) \text{ and } p_{\vec{b}} \leq p\}. \tag{4.1}$$

Conversely, if $f : \{0,1\}^n \longrightarrow \mathbb{N}$ is an integer-valued, non-negative pseudo-Boolean function, then the function μ defined as in Equality (4.1) is clearly a measure on $\mathsf{pol}(A)$.

In the next section we regard the set of minterms $\mathsf{mint}(A)$ as a sample space and each polynomial $p \in \mathsf{pol}(A)$ as an event on this sample space. The event p occurs in $p_{\vec{b}}$ if $p_{\vec{b}} \leq p$. Thus, if μ is a measure on $\mathsf{pol}(A)$, then the mapping $P_\mu : \mathsf{pol}(A) \longrightarrow \mathbb{R}$ given by $P_\mu(p) = \frac{\mu(p)}{\mu(\mathbf{1})}$ is a probability on $\mathsf{pol}(A)$.

4.3 An Inclusion-Exclusion Principle for MFBAs

Let p be a polynomial in $\mathsf{pol}(A)$. It is known that p can be uniquely written as

$$p = \sum_{(i_1,\ldots,i_m)}^{\oplus} c_{(i_1,\ldots,i_m)} \wedge a_{i_1} \wedge \cdots \wedge a_{i_m},$$

where the summation $\sum^{\oplus}$ involves the "exclusive or" operation $\oplus$ and is extended to all subsets $\{i_1, \ldots, i_m\}$ of $\{1, \ldots, n\}$. The coefficients $c_{(i_1,\ldots,i_m)}$ belong to the set $\{\mathbf{0},\mathbf{1}\}$. Thus, for a measure μ on $\mathsf{pol}(A)$ it is interesting to evaluate $\mu(p_1 \oplus p_2 \oplus \cdots \oplus p_m)$, where $p_1, \ldots, p_m$ are polynomials in $\mathsf{pol}(A)$.

The *indicator random variable* of a polynomial p (see [GS96]) is the variable I_p defined by

$$I_p(p_{\vec{b}}) = \begin{cases} 1 & \text{if } p_{\vec{b}} \leq p \\ 0 & \text{otherwise.} \end{cases}$$

for $p_{\vec{b}} \in \mathsf{mint}(A)$.

Note that the expected value $E[I_p]$ of I_p equals $P_\mu(p)$.

If $M = \{p_1, \ldots, p_n\}$ and $J = \{j_1, \ldots, j_m\} \subseteq \{1, \ldots, n\}$, then $p_{M,J} = p_{j_1} \wedge \ldots \wedge p_{j_m}$, and $I_{p_{M,J}} = I_{p_{j_1}} \cdots I_{p_{j_m}}$.

For a set of polynomials M denote by $S^\mu_{M,k}$ the probability that exactly k events in M hold:

$$S^\mu_{M,k} = \sum\{P_\mu(p_{M,K}) \mid |K| = k\}.$$

The number of k-subsets K of M such that $p_{M,K}$ holds is given by the random variable $\sum\{I_{p_{M,K}} \mid |K| = k\}$. By the previous observation

$$S^\mu_{M,k} = \sum\{E(I_{p_{M,K}}) \mid |K| = k\} = E\left[\sum\{I_{p_{M,K}} \mid |K| = k\}\right].$$

Let ν_M be the random variable on $\mathsf{mint}(A)$ such that $\nu_M(p_{\vec{b}}) = |\{p_i \in M \mid p_{\vec{b}} \leq p_i\}|$. Note that ν_M gives the number of events in M that hold and, therefore, the random variable $\binom{\nu_M}{k}$ gives the number of k-subsets Q of M such that $p_{M,Q}$ holds, which means that $\binom{\nu_M}{k} = \sum\{I_{p_{M,K}} \mid |K| = k\}$, and

$$S^\mu_{M,k} = E\left[\binom{\nu_M}{k}\right]. \tag{4.2}$$

The equality (4.2) is the basis of the method of indicators, that is a method of proving probabilistic identities by taking expectations of their non-probabilistic counterparts, see [GS96] for details.

Theorem 4.3.1 *Let $\mu : \mathsf{pol}(A) \longrightarrow \mathbb{R}$ be a measure on the free Boolean algebra $\mathsf{pol}(A)$, where $A = \{a_1, \ldots, a_n\}$. If $M = \{p_1, \ldots, p_m\}$ is a set of m polynomials of $\mathsf{pol}(A)$, then*

$$\mu(p_1 \oplus \cdots \oplus p_m) = \mu(\mathbf{1}) \cdot \sum_{k=1}^{m} (-2)^{k-1} \cdot S^{\mu}_{M,k}. \tag{4.3}$$

Proof. Let $a \in \mathbb{N}$, note that $(-1)^a = \sum_{k=0}^{a} (-2)^k \binom{a}{k}$, which yields, after elementary transformations:

$$\sum_{k=1}^{a} (-2)^{k-1} \binom{a}{k} = (-1)^a - 1 = \begin{cases} 0 & \text{if } a \text{ is even} \\ 1 & \text{if } a \text{ is odd.} \end{cases}$$

This implies

$$\sum_{k=1}^{\nu_M} (-2)^{k-1} \binom{\nu_M}{k} = \sum_{k=1}^{|M|} (-2)^{k-1} \binom{\nu_M}{k} = \begin{cases} 0 & \text{if } \nu_M \text{ is even} \\ 1 & \text{if } \nu_M \text{ is odd.} \end{cases}$$

By taking expectations of both sides, and using equality (4.2) we get

$$E\left[\sum_{k=1}^{|M|} (-2)^{k-1} \binom{\nu_M}{k}\right] = \sum_{k=1}^{|M|} (-2)^{k-1} S^{\mu}_{M,k} = P_{\mu}(\nu_M \text{ is odd}) = P_{\mu}(p_1 \oplus \ldots \oplus p_m).$$

which yields the desired equality. ∎

Corollary 4.3.2 *Let $\mu, \mu' : \mathsf{pol}(A) \longrightarrow \mathbb{R}$ be two measures on the free Boolean algebra $\mathsf{pol}(A)$, where $A = \{a_1, \ldots, a_n\}$. If $\mu(p) = \mu'(p)$ for every conjunction p of the form $p = a_{i_1} \wedge \cdots \wedge a_{i_m}$, then $\mu = \mu'$.*

Proof. The result follows immediately from Theorem 4.3.1. ∎

Example 4.3.3 Consider the "majority polynomial" $p_{maj} = (a_1 \wedge a_2) \vee (a_2 \wedge a_3) \vee (a_1 \wedge a_3)$. For $f_{p_{maj}}$ we have $f_{p_{maj}}(x_1, x_2, x_3) = 1$ if and only if at least two of its arguments are equal to 1. Note that

$$p_{maj} = (a_1 \wedge a_2) \oplus (a_2 \wedge a_3) \oplus (a_1 \wedge a_3).$$

Theorem 4.3.1 allows us to write

$$\begin{aligned}\mu(p_{maj}) &= \mu(a_1 \wedge a_2) + \mu(a_2 \wedge a_3) + \mu(a_1 \wedge a_3) \\ &\quad -2\mu((a_1 \wedge a_2) \wedge (a_2 \wedge a_3)) - 2\mu((a_1 \wedge a_2) \wedge (a_1 \wedge a_3)) \\ &\quad -2\mu((a_2 \wedge a_3) \wedge (a_1 \wedge a_3)) + 4\mu((a_1 \wedge a_2) \wedge (a_2 \wedge a_3) \wedge (a_1 \wedge a_2)) \\ &= \mu(a_1 \wedge a_2) + \mu(a_2 \wedge a_3) + \mu(a_1 \wedge a_3) - 2\mu(a_1 \wedge a_2 \wedge a_3).\end{aligned}$$

□

Corollary 4.3.2 shows that the values of a measure on $\mathsf{pol}(A)$ are completely determined by its values on positive conjunctions of the form a_I for $I \subseteq \{1, \ldots, n\}$. Note that the contribution of every tuple of a table $\tau = (T, A, \rho)$ of the form $(b_1, \ldots, b_n)$ to the value of $\mu_\tau(I)$ equals 1 for every set I such that $I \subseteq \{i \in \{1, \ldots n\} \mid b_i = 1\}$.

Next, we obtain Bonferroni-type inequalities [GS96] that give bounds on the value of $\mu(p_1 \oplus \ldots \oplus p_m)$. To this end we need the following technical result:

Define W_b^a for $a, b \in \mathbb{N}$ and $b \leq a$ as $W_b^a = \sum_{k=b}^{a} (-2)^{k-1} \binom{a}{k}$. Alternatively, W_b^a can be written as

$$W_b^a = (-2)^{b-1} \sum_{k=b}^{a} (-2)^{k-b} \binom{a}{k} = (-2)^{b-1} \sum_{\ell=0}^{a-b} (-2)^{\ell} \binom{a}{b+\ell}.$$

Lemma 4.3.4 *The signs of the members of the sequence $(W_b^a, W_{b+1}^a, \ldots, W_a^a)$ are alternating.*

Proof. Define

$$U_b^a = \sum_{\ell=0}^{a-b} (-2)^{\ell} \binom{a}{b+\ell}$$

for $a, b \in \mathbb{N}$ and $b \leq a$. Since $W_b^a = (-2)^{b-1} U_b^a$ it suffices to prove that the numbers U_b^a are have all the same sign.

Note that $U_b^b = 1$ for $b \in \mathbb{N}$. We can write:

$$\begin{aligned} U_b^a &= \sum_{\ell=0}^{a-b}(-2)^\ell \binom{a}{b+\ell} \\ &= \binom{a-1}{b} + \binom{a-1}{b-1} - 2\binom{a-1}{b+1} - 2\binom{a-1}{b} \\ &\quad +2^2\binom{a-1}{b+2} + 2^2\binom{a-1}{b+1} - 2^3\binom{a-1}{b+3} - 2^3\binom{a-1}{b+2} + \cdots \\ &\vdots \\ &\quad +(-2)^{a-b-1}\binom{a-1}{a-1} + (-2)^{a-b-1}\binom{a-1}{a-2} + (-2)^{a-b}\binom{a-1}{a-1} \\ &= \binom{a-1}{b-1} - U_b^{a-1}. \end{aligned}$$

Thus, we obtain

$$U_b^a = \binom{a-1}{b-1} - U_b^{a-1}. \tag{4.4}$$

We claim that $0 \leq U_b^a \leq \binom{a}{b-1}$ for $0 \leq b \leq a$. This can be shown by induction on $a \geq b$. The basis step $a = b$ is immediate. Suppose that the double inequality holds for $a-1$, that is, $0 \leq U_b^{a-1} \leq \binom{a-1}{b-1}$. Then, it is clear that $U_b^a \geq 0$. To show that $U_b^a \leq \binom{a}{b-1}$ we need to verify that $\binom{a-1}{b-1} - U_b^{a-1} \leq \binom{a}{b-1}$. Since $\binom{a}{b-1} = \binom{a-1}{b-1} + \binom{a-1}{b-2}$ for $b \geq 2$ the last inequality follows. ∎

Theorem 4.3.5 *For any $r, s \in \mathbb{N}$ we have:*

$$\mu(\mathbf{1}) \cdot \sum_{k=1}^{2r}(-2)^{k-1}S_k^\mu \leq \mu(p_1 \oplus \ldots \oplus p_m) \leq \mu(\mathbf{1}) \cdot \sum_{k=1}^{2s+1}(-2)^{k-1}S_k^\mu.$$

Proof. By equality (4.2) and Lemma 4.3.4 we get that for any $r, s \in \mathbb{N}$

$$\sum_{k=1}^{2r}(-2)^{k-1}\binom{a}{k} \leq \sum_{k=1}^{a}(-2)^{k-1}\binom{a}{k} \leq \sum_{k=1}^{2s+1}(-2)^{k-1}\binom{a}{k},$$

implying

$$\sum_{k=1}^{2r}(-2)^{k-1}\binom{\nu_M}{k} \leq \sum_{k=1}^{|M|}(-2)^{k-1}\binom{\nu_M}{k} \leq \sum_{k=1}^{2s+1}(-2)^{k-1}\binom{\nu_M}{k}.$$

By applying expectations and using equality (4.2) we get the desired result. ∎

Example 4.3.6 Consider the table $\tau = (T, a_1a_2a_3, \rho)$ given below, where $\mu(p) = |\rho_p|$:

T

a_1	a_2	a_3
0	0	0
0	1	0
1	0	0
0	0	0
0	1	0
1	0	1
0	1	1
1	1	0
1	1	0
0	1	1
1	0	1
1	1	0
1	1	1

and the majority polynomial p_{maj} from Example 4.3.3. We have $\mu(a_1 \wedge a_2) = 4$, $\mu(a_1 \wedge a_3) = 3$, $\mu(a_2 \wedge a_3) = 3$, giving $\mu(p_{maj}) \leq 10$. Also $\mu((a_1 \wedge a_2) \wedge (a_1 \wedge a_3)) = \mu((a_1 \wedge a_2) \wedge (a_2 \wedge a_3)) = \mu((a_1 \wedge a_3) \wedge (a_2 \wedge a_3)) = 1$ giving $\mu(p_{maj}) \geq 4$. The true value of $\mu(p_{maj})$ is 8. □

4.4 Applications in Data Mining and Database Query Optimization

4.4.1 Accuracy of Inclusion-Exclusion Principle

In database query optimization and in data mining, it is often necessary to estimate the number of rows in a database table satisfying a given query. Unfortunately,

in most cases, the exact number of rows satisfying a query cannot be computed exactly and has to be estimated (usually using the assumption of statistical independence between attributes).

Let $\tau = (T, A, \rho)$ be a binary table and let $K = \{a_{k_1}, \ldots, a_{k_m}\}$ be a set of attributes, $K \subseteq A$. The support of the set K relative to the table τ is the value of the probability $P_{\mu_{tau}}(a_{k_1} \wedge \cdots \wedge a_{k_m})$:

$$\mathrm{supp}_\tau(K) = \frac{|\{t \in \rho \mid t(a) = \mathbf{1} \text{ for all } a \in K\}|}{|\rho|}.$$

In other words, the support of an attribute set K in the table τ is defined by the value of the measure induced by the table on the Boolean polynomial that describes the attribute set. By extension, we can regard the number $\frac{\mu_\tau(q)}{\mu_\tau(\mathbf{1})}$ as the support of the query q and we denote this number by $\mathrm{supp}_\tau(q)$. Indeed, if $q \in \mathsf{pol}(A)$ is a query involving a table $\tau = (T, A, \rho)$ such that q can be written as

$$q = c \oplus \sum_{I \in \mathcal{I}}^{\oplus} a_I,$$

where $c \in \{\mathbf{0}, \mathbf{1}\}$ and $\mathcal{I}$ is a collection of subsets of $\{1, \ldots, n\}$, then $\mathrm{supp}_\tau(q)$ can be obtained from Theorem 4.3.1 using the numbers $\mathrm{supp}_\tau(a_I)$. Methods that obtain approximative estimations of query sizes been proposed [Man01], including the use of Maximum Entropy Principle. An open problem raised was estimating the quality of such an approximation.

The computation of the size of the query using Theorem 4.3.1 can be often simplified if there is a known maximal number of 1 components in the tuples of the table. For example, in a store that sells 1000 items (corresponding to 1000 attributes in a table that contains the records of purchases) it is often the case that we can use an empirical limit of, say, 8 items per tuple. In this case, conjunctions that contain more than 8 conjuncts can be discarded and the estimation

is considerably simplified. Even, if such an upper bound cannot be imposed apriori, it is often the case that we can discard large conjunctions (which have low support). However, there are some risks when approximations of this nature are performed due to the the large values of coefficients that multiply the supports for large conjunctions.

Indeed, consider the tables $\tau^n_{odd} = (T_o, A, \rho_{odd})$, $\tau^n_{even} = (T_e, A, \rho_{even})$, where

$$\rho_{odd} = \{t \in \mathrm{Dom}(A) \mid n_1(t) \text{ is odd}\} \text{ and } \rho_{even} = \{t \in \mathrm{Dom}(A) \mid n_1(t) \text{ is even}\},$$

where $n_1(t)$ denotes the number of attributes equal to 1 in tuple t and $|A| = n$.

Note that for proper subset K of A, we have $\mathrm{supp}_{\tau_{odd}}(K) = \mathrm{supp}_{\tau_{even}}(K)$, while

$$\mathrm{supp}_{\tau^n_{odd}}(A) = \begin{cases} 1 & \text{if } n \text{ is odd} \\ 0 & \text{otherwise,} \end{cases} \text{ and } \mathrm{supp}_{\tau^n_{even}}(A) = \begin{cases} 1 & \text{if } n \text{ is even} \\ 0 & \text{otherwise.} \end{cases}$$

Thus, from the point of view of the supports of any proper subset of the attribute set the tables τ^n_{odd} and τ^n_{even} are indiscernible. However, the support of certain queries can be vastly different on these tables. For example, consider the polynomial $p = a_1 \oplus a_2 \oplus \ldots \oplus a_n$. We have $\mathrm{supp}_{\tau^n_{odd}}(p) = 1$ and $\mathrm{supp}_{\tau^n_{even}}(p) = 0$. So, ignoring the term that corresponds to the support for a single attribute set (note that this is also the attribute set with the smallest possible support) has a huge impact on $\mu_\tau(p)$. Note that the result is consistent with Theorem (4.3.1) which gives the set of attributes A a coefficient 2^{n-1}. We stress however that the negative result above does not rule out practical applicability of approximating the values of μ_τ since the parity function query used above is by no means a typical database query.

4.4.2 Support in tables with missing values

Frequently, real world datasets contain missing values; this makes important to adequately address this issue. Here we present a generalization of the notion of support which takes missing values into account. The idea is related to the *hot deck imputation* of missing values where each missing value is replaced by a value randomly drawn from some distribution.

Suppose that $\tau = (T, A, \rho)$ is a table such that $A = \{a_1, \ldots, a_n\}$ and $\mathrm{Dom}(a_i) = \{0, u, 1\}$ for $1 \leq i \leq n$. The symbol u represents *null values*, that is, values that are missing or undefined. With every attribute $a_i \in A$ we associate a real number $\alpha_i \in [0, 1]$. Intuitively, this number corresponds to the probability of $a_i = 1$, and can be obtained using the non-missing values for the attribute or based on background knowledge.

Let α be a non-negative number, and let $b, c \in \{0, 1\}$. Define

$$\alpha^{\langle b,c\rangle} = \begin{cases} \alpha & \text{if } b = 1 \text{ and } c = 0 \\ 1 - \alpha & \text{if } b = 0 \text{ and } c = 0 \\ 1 & \text{if } c = 1, \end{cases} \text{ and } b^{(c)} = \begin{cases} b & \text{if } c = 1 \\ u & \text{if } c = 0, \end{cases}$$

For a table $\tau = (T, A, \rho)$ let $\mu^u_\tau : \mathsf{pol}(A) \longrightarrow \mathbb{R}$ be defined as follows. For a minterm $a_1^{b_1} \wedge \cdots \wedge a_n^{b_n}$ let

$$\mu^u_\tau(a_1^{b_1} \wedge \cdots \wedge a_n^{b_n}) = \sum_{(c_1,\ldots,c_n)\in\{0,1\}^n} \prod_{i=1}^{n} \alpha_i^{\langle b_i,c_i\rangle} \cdot \frac{|\{t \in \rho \mid t(a_i) = b_i^{(c_i)} \text{ for } 1 \leq i \leq n\}|}{|\rho|}$$

For an arbitrary boolean polynomial p define

$$\mu^u_\tau(p) = \sum_{p_{\vec{d}} \in \mathsf{mint}_p} \mu^u_\tau(p_{\vec{d}})$$

where mint_p is the set of minterm implicants of p.

Theorem 4.4.1 *For every table $\tau = (T, H, \rho)$, μ^u_τ is a measure on $\mathsf{pol}(A)$.*

Proof. Since μ^u_τ is clearly non-negative, it remains to be shown that $\mu^u_\tau(p_1 \vee p_2) = \mu^u_\tau(p_1) + \mu^u_\tau(p_2)$ for every $p_1, p_2 \in \mathsf{pol}(A)$ such that $p_1 \wedge p_2 = \mathbf{0}$. Note that if $p_1 \wedge p_2 = \mathbf{0}$ then $\mathsf{mint}_{p_1} \cap \mathsf{mint}_{p_2} = \emptyset$, and

$$\mu^u_\tau(p_1 \vee p_2) = \sum_{p_{\vec{d}} \in \mathsf{mint}_{p_1}} \mu^u_\tau(p_{\vec{d}}) + \sum_{p_{\vec{d}} \in \mathsf{mint}_{p_2}} \mu^u_\tau(p_{\vec{d}}) = \mu^u_\tau(p_1) + \mu^u_\tau(p_2).$$

∎

Example 4.4.2 Let $n = 2$, we have:

$$\begin{aligned}
&\mu^u_\tau(a_1 \oplus a_2) \\
&= \mu^u_\tau(\bar{a_1} \wedge a_2) + \mu^u_\tau(a_1 \wedge \bar{a_2}) \\
&= \mathrm{supp}_\tau(a_1 = 0 \wedge a_2 = 1) + (1 - \alpha_1)\mathrm{supp}_\tau(a_1 = u \wedge a_2 = 1) \\
&\quad + \alpha_2 \mathrm{supp}_\tau(a_1 = 0 \wedge a_2 = u) + (1 - \alpha_1)\alpha_2 \mathrm{supp}_\tau(a_1 = a_2 = u) \\
&\quad + \mathrm{supp}_\tau(a_1 = 1 \wedge a_2 = 0) + \alpha_1 \mathrm{supp}_\tau(a_1 = u \wedge a_2 = 0) \\
&\quad + (1 - \alpha_2)\mathrm{supp}_\tau(a_1 = 1 \wedge a_2 = u) + \alpha_1(1 - \alpha_2)\mathrm{supp}_\tau(a_1 = a_2 = u).
\end{aligned}$$

□

The benefit of using arbitrary measures instead of probabilities or supports in previous sections is that results on inclusion-exclusion principle automatically apply to μ^u_τ. Also, the fact that μ^u_τ is a measure makes the proof of the following theorem straightforward.

Theorem 4.4.3 *For every table $\tau = (T, A, \rho)$ such that $A = \{a_1, \ldots, a_n\}$ and $\mathrm{Dom}(a_i) = \{0, u, 1\}$ for $1 \leq i \leq n$, and every collection of sets of attributes $\mathcal{A} = \{a_{I_1}, \ldots, a_{I_m} \mid I_j \subseteq \{1, \ldots, n\}\}$ there is a probability distribution P over A such that for every $a_{I_r} \in \mathcal{A}$, $P\{\bigwedge_{j \in I_r}(a_j = 1)\} = \mu^u_\tau(\bigwedge_{j \in I_r}(a_j = 1))/|\rho|$.*

Proof. We prove the theorem by showing that $\mu^u_\tau/|\rho|$ is a probability distribution. Since μ^u_τ is a measure, it suffices to show that $\mu^u_\tau(\mathbf{1}) = |\rho|$. For any $a_i \in A$ we have:

$$\begin{aligned}
\mu^u_\tau(\mathbf{1}) &= \mu^u_\tau(a_i \vee \bar{a}_i) = \mu^u_\tau(a_i) + \mu^u_\tau(\bar{a}_i) \\
&= \text{supp}_\tau(a_i = 1) + \alpha_i \text{supp}_\tau(a_i = u) \\
&\quad + \text{supp}_\tau(a_i = 0) + (1 - \alpha_i)\text{supp}_\tau(a_i = u) \\
&= \text{supp}_\tau(a_i = 1) + \text{supp}_\tau(a_i = 0) + \text{supp}_\tau(a_i = u) = |\rho|.
\end{aligned}$$

∎

The importance of the above theorem is that if we use some datamining algorithm (e.g. Apriori) to find μ^u_τ for a collection of sets of attributes, then their values of μ^u_τ are probabilistically consistent. Other approaches to mining frequent itemsets in the presence of missing values can be found in [RC98, NC01]. However, both these approaches can produce probabilistically inconsistent results. Specifically, the technique used in [RC98] is to count the support of an itemset only on the portion of the table where it is valid. For example, consider the table $\tau = (T, a_1a_2, \rho)$, given by

T

a_1	a_2
1	1
1	u
0	u
0	u

Using the method from [RC98] the support of attribute a_2 is counted only in the first row, giving $\text{supp}_\tau(a_2) = 100\%$. Similarly $\text{supp}_\tau(a_1) = 50\%$, and $\text{supp}_\tau(a_1a_2) = 100\%$, but this means $\text{supp}_\tau(a_1a_2) > \text{supp}_\tau(a_1)$, which is impossible. In the method proposed in [NC01] the probability for each attribute is estimated from the part

of the data where the attribute is defined. When computing how much support does a row with a missing value contribute for an itemset, this probabilities are summed for each attribute (see [NC01] for details). In the table above this will give $\text{supp}_\tau(a_1) = 50\%$, $\text{supp}_\tau(a_2) = 100\%$, and $\text{supp}_\tau(a_1a_2) = [(0.5 \cdot 1 + 0.5 \cdot 1) + (0.5 \cdot 1 + 0.5 \cdot 1) + 2(0.5 \cdot 0 + 0.5 \cdot 1)]/4 = 75\%$, and $\text{supp}_\tau(a_1a_2) > \text{supp}_\tau(a_1)$. Using our μ^u_τ measure with $\alpha_2 = 1$ gives consistent values of $\text{supp}_\tau(a_1) = 50\%$, $\text{supp}_\tau(a_2) = 100\%$, and $\text{supp}_\tau(a_1a_2) = 50\%$.

4.5 Approximating Supports of Itemsets Using Bonferroni-type Inequalities

In this section we use Bonferroni inequalities to estimate supports of missing itemsets. In their original form the inequalities require that we know supports of all itemsets up to a given size. We address the problem by using the inequalities recursively to estimate supports of missing itemsets. The advantage of Bonferroni inequalities is that we can choose an arbitrary limit on the size of the marginals involved, thus allowing for trading off accuracy for speed. Our experiments revealed that it is possible to obtain good bounds even if only marginals of small size are used.

Example 4.5.1 Consider a binary table τ whose heading is $A = abc$ and assume that the distribution of the values of the tuples in this table is given by:

a	0	0	0	0	1	1	1	1
b	0	0	1	1	0	0	1	1
c	0	1	0	1	0	1	0	1
Frequency	0	0	0.1	0.25	0.1	0.25	0.05	0.25

A run of the Apriori algorithm ([AMS96]) on a dataset conforming to that distribution, with the minimum support of 0.35 will yield the following itemsets:

Itemset	a	b	c	ac	bc
Support	0.65	0.65	0.75	0.5	0.5

To estimate the unknown support of the itemset abc we can use Bonferroni inequalities of the form:

$$\begin{aligned} \text{supp}_\tau(abc) &\geq 1 - \text{supp}_\tau(\bar{a}) - \text{supp}_\tau(\bar{b}) - \text{supp}_\tau(\bar{c}), & (4.5) \\ \text{supp}_\tau(abc) &\leq 1 - \text{supp}_\tau(\bar{a}) - \text{supp}_\tau(\bar{b}) - \text{supp}_\tau(\bar{c}) & (4.6) \\ &\quad + \text{supp}_\tau(\bar{a}\bar{b}) + \text{supp}_\tau(\bar{a}\bar{c}) + \text{supp}_\tau(\bar{b}\bar{c}). \end{aligned}$$

Note that since the support of ab is below the minimum support its value is not returned by the Apriori algorithm and this creates a problem for this estimation. All the itemset supports, except for $\text{supp}_\tau(\bar{a}\bar{b})$, in the previous expression can be determined from known itemset supports using inclusion-exclusion principle. For example, we have

$$\text{supp}_\tau(\bar{a}\bar{c}) = 1 - \text{supp}_\tau(a) - \text{supp}_\tau(c) + \text{supp}_\tau(ac) = 0.1.$$

Since all needed probabilities are known exactly, the lower bound (4.5) is easy to compute giving

$$\text{supp}_\tau(abc) \geq 1 - 0.35 - 0.35 - 0.25 = 0.05.$$

To compute the upper bound we proceed as follows.

Since $\text{supp}_\tau(\bar{a}\bar{b})$ is not known, we apply Bonferroni inequalities recursively to get an upper bound for it. We have

$$\text{supp}_\tau(\bar{a}\bar{b}) = 1 - \text{supp}_\tau(a) - \text{supp}_\tau(b) + \text{supp}_\tau(ab),$$

and, since ab is not frequent, we know that its support is less than the 0.35 minimum support, giving

$$\mathrm{supp}_\tau(\bar{a}\bar{b}) < 1 - \mathrm{supp}_\tau(a) - \mathrm{supp}_\tau(b) + \mathsf{minsupp} = 0.05.$$

Substituting into (4.7) we get

$$\begin{aligned}\mathrm{supp}_\tau(abc) &< 1 - \mathrm{supp}_\tau(\bar{a}) - \mathrm{supp}_\tau(\bar{b}) - \mathrm{supp}_\tau(\bar{c}) \\ &\quad +0.05 + \mathrm{supp}_\tau(\bar{a}\bar{c}) + \mathrm{supp}_\tau(\bar{b}\bar{c}) \\ &= 1 - 0.35 - 0.35 - 0.25 + 0.05 + 0.1 + 0.1 = 0.3.\end{aligned}$$

Note that both bounds are not trivial since the lower bound is greater than 0, and the upper bound is less than the minimum support. □

4.5.1 A Recursive Procedure for Computing Bonferroni Bounds from Frequent Itemsets

Since the Apriori algorithm only discovers supports of itemsets (as opposed to other types of queries), we need to express all inequalities in terms of supports of itemsets.

Theorem 4.5.2 *Let $q_1, \ldots, q_m$ be m queries in $\mathsf{pol}(A)$. The following inequalities hold for any $t \in \mathbb{N}$:*

$$\begin{aligned}&\sum_{k=0}^{2t+1}(-1)^k \sum_{r<i_1<\ldots<i_k\leq m} \mathrm{supp}_\tau(q_1 \wedge \ldots \wedge q_r \wedge q_{i_1} \wedge \ldots \wedge q_{i_k}) \\ &\qquad \leq \mathrm{supp}_\tau(q_1 \wedge \ldots \wedge q_r \wedge \bar{q}_{r+1} \wedge \ldots \wedge \bar{q}_m) \leq \\ &\qquad \sum_{k=0}^{2t}(-1)^k \sum_{r<i_1<\ldots<i_k\leq m} \mathrm{supp}_\tau(q_1 \wedge \ldots \wedge q_r \wedge q_{i_1} \wedge \ldots \wedge q_{i_k}).\end{aligned}$$

Proof. By Rényi's Theorem [Ren58] it suffices to prove the claim for $q_i \in \{\mathbf{1}, \mathbf{0}\}$ for all $1 \leq i \leq m$. When $q_i = \mathbf{0}$ for some $1 \leq i \leq r$, then both sides of the inequalities reduce to 0 and the result is immediate. For the case $q_i = \mathbf{1}$ for all $1 \leq i \leq r$ we have $\mathrm{supp}_\tau(q_1 \wedge \ldots \wedge q_r \bar{q}_{r+1} \wedge \ldots \wedge \bar{q}_m) = \mathrm{supp}_\tau(\bar{q}_{r+1} \ldots \bar{q}_m)$, and for all k and for all $r < i_1 < \ldots < i_k \leq m$, $\mathrm{supp}_\tau(q_1 \wedge \ldots \wedge q_r \wedge q_{i_1} \wedge \ldots \wedge q_{i_k}) = \mathrm{supp}_\tau(q_{i_1} \wedge \ldots \wedge q_{i_k})$. The result now follows from Bonferroni inequalities. ∎

Corollary 4.5.3 *Let $a_1 \wedge a_2 \wedge \cdots \wedge a_r \wedge \bar{a}_{r+1} \wedge \bar{a}_{r+2} \wedge \cdots \wedge \bar{a}_m$ be a minterm. The following inequalities hold for any natural number t:*

$$\sum_{k=0}^{2t+1}(-1)^k \sum_{r<i_1<\ldots<i_k\leq m} \mathrm{supp}_\tau(a_1 \wedge \cdots \wedge a_r \wedge a_{i_1} \wedge \cdots \wedge a_{i_k})$$
$$\leq \mathrm{supp}_\tau(a_1 \wedge \cdots \wedge a_r \wedge \bar{a}_{r+1} \wedge \cdots \wedge \bar{a}_m) \leq$$
$$\sum_{k=0}^{2t}(-1)^k \sum_{r<i_1<\ldots<i_k\leq m} \mathrm{supp}_\tau(a_1 \wedge \cdots \wedge a_r \wedge a_{i_1} \wedge \cdots \wedge a_{i_k})$$

Proof. This statement follows immediately from Theorem 4.5.2. ∎

Below we present results which form the basis of our algorithm for approximative computations of supports of itemsets. The binomial symbol $\binom{n}{k}$ will allow negative values of n, in which case its value is defined by the usual formula

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!}.$$

Lemma 4.5.4 *For $m, k, h, s \in \mathbb{N}$ we have:*

$$\sum_{k=0}^{s}(-1)^{s-k}\binom{m-k-1}{s-k}\binom{h}{k} = \binom{h-m+s}{s}.$$

Proof. We begin by showing that for every $a, b, c, d \in \mathbb{N}$ we have

$$\sum_{k=0}^{a}(-1)^k\binom{a-k}{b}\binom{c}{k-d} = (-1)^{a+b}\binom{c-b-1}{a-b-d}. \tag{4.7}$$

The proof is by induction on c. The basis step, $c = 0$, follows after elementary algebraic transformations. Suppose that the equality holds for numbers less than c. We have:

$$\begin{aligned}
&\sum_{k=0}^{a}(-1)^k\binom{a-k}{b}\binom{c}{k-d} \\
&= \sum_{k=0}^{a}(-1)^k\binom{a-k}{b}\binom{c-1}{k-d} + \sum_{k=0}^{a}(-1)^k\binom{a-k}{b}\binom{c-1}{k-d-1} \\
&= (-1)^{a+b}\binom{c-b-2}{a-b-d} + (-1)^{a+b}\binom{c-b-2}{a-b-d-1} \\
&\quad \text{(by the inductive hypothesis)} \\
&= (-1)^{a+b}\binom{c-b-1}{a-b-d}.
\end{aligned}$$

By using the complimentary combinations and Lemma 4.5.4 we can write:

$$\begin{aligned}
\sum_{k=0}^{s}(-1)^{s-k}\binom{m-k-1}{s-k}\binom{h}{k} &= \sum_{k=0}^{s}(-1)^{s+k}\binom{m-k-1}{m-s-1}\binom{h}{k} = \\
(-1)^s\cdot\sum_{k=0}^{s}(-1)^k\binom{m-k-1}{m-s-1}\binom{h}{k} &= (-1)^s\cdot(-1)^{2m-2-s}\binom{h-m+s}{s} \\
&= \binom{h-m+s}{s}.
\end{aligned}$$

∎

Note that if $h = m$, the previous lemma implies

$$\sum_{k=0}^{s}(-1)^{s-k}\binom{m-k-1}{s-k}\binom{m}{k} = 1.$$

Our method of obtaining bounds is based on the following theorem

Theorem 4.5.5 *Let $\tau = (T, A, \rho)$ be a table and let $a_1, \ldots, a_m$ be attributes in A.*

The following inequalities hold for any natural number t:

$$\text{supp}_\tau(a_1 \wedge a_2 \wedge \ldots \wedge a_m) \leq \sum_{k=0}^{2t}(-1)^k\binom{m-k-1}{2t-k}S_k \tag{4.8}$$

$$\text{supp}_\tau(a_1 \wedge a_2 \wedge \ldots \wedge a_m) \geq \sum_{k=0}^{2t+1}(-1)^{k+1}\binom{m-k-1}{2t+1-k}S_k \tag{4.9}$$

where

$$S_k = \sum_{1\leq i_1<\ldots<i_k\leq m} \text{supp}_\tau(a_{i_1} \wedge \ldots \wedge a_{i_k}),$$

and $S_0 = 1$.

Proof. We use the method of indicators previously discussed.

Let ν_m be a random variable equal to the number of events $A_1, \ldots, A_m$ that actually occur. By Lemma 4.5.4 we have:

$$\sum_{k=0}^{s}(-1)^{s-k}\binom{m-k-1}{s-k}\binom{\nu_m}{k} = \binom{\nu_m - m + s}{s}$$

$$= \begin{cases} 1 & \text{if } \nu_m = m \\ 0 & \text{if } \nu_m < m \text{ and } \nu_m \geq m-s \\ \binom{\nu_m - m + s}{s} & \text{if } \nu_m < m - s. \end{cases}$$

By taking expectations of the above equation we get

$$\sum_{k=0}^{s}(-1)^{s-k}\binom{m-k-1}{s-k}S_k = \text{supp}_\tau(\nu_m = m)$$
$$+ \sum\left\{\binom{\nu_m(\omega) - m + s}{s}\text{supp}_\tau(\omega) : \omega \in \Omega, \nu_m(\omega) < m - s\right\},$$

where Ω denotes the space of elementary events. Note that when $\nu_m < m - s$ the sign of $\binom{\nu_m - m + s}{s}$ is identical to that of $(-1)^s$. Replacing s by $2t$ or $2t+1$ yields the result. ∎

4.6 The Estimation Algorithm

The main problem in using Bonferroni-type inequalities on collections of frequent itemsets is that some of the probabilities in the S_k sums are not known. We solved this problem by estimating the missing probabilities using Theorem 4.5.5.

Given in Figure 4.1 is an algorithm that computes bounds on support of an itemset based on a collection of itemsets with known supports. The algorithm consists of two functions L and U returning the lower and upper bounds respectively. Additional functions for the algorithm are given in Fugure 4.2.

Of course upper and lower bounds for itemsets are cached during computations to avoid repeated evaluations for the same itemset. The parameter r controls the maximum size of marginals (itemsets) used in the estimation.

The use of minsupp in step 5 of function U requires some comment. Including the value of minsupp in the minimum is possible only if we can determine that the estimated itemset I is not frequent. This can be done for example if $\mathcal{F}$ contains all frequent itemsets, or when $\mathcal{F}$ contains all frequent itemsets up to a given size k, and $|I| \leq k$. If we don't know whether I is frequent or not, we have to drop minsupp from the minimum.

4.6.1 Experimental results

In this section we present experimental evaluation of the bounds. Our algorithm works best on dense datasets, which are more difficult to mine for frequent itemsets than sparse ones. However, the algorithm was tested on both dense and sparse data (artificial market basket data was used). The rest of the chapter is focused on experiments performed on dense databases.

Input: Itemset I, natural number r, collection $\mathcal{F}$ of itemsets, supports of itemsets in $\mathcal{F}$

Output: Bounds $L(I), U(I)$ on the support of I

Function $L(I, \mathcal{F}, r)$:

1. If $I \in \mathcal{F}$
2. return $\text{supp}_\tau(I)$
3. else
4. return $\max_{-1 \leq 2t+1 \leq r} \sum_{k=0}^{2t+1} S^L \left((-1)^{k+1} \binom{m-k-1}{2t+1-k}, k, I, \mathcal{F} \right)$

Function $U(I, \mathcal{F}, r)$:

1. If $I \in \mathcal{F}$
2. return $\text{supp}_\tau(I)$
3. else
4. $U \longrightarrow \min_{0 \leq 2t \leq r} \sum_{k=0}^{2t} S^U \left((-1)^k \binom{m-k-1}{2t-k}, I, k, \mathcal{F} \right)$
5. $U \longrightarrow \min\{U, \mathsf{minsupp}, \min_{J \subset I} U(J)\}$
6. return U

Figure 4.1: Itemset support estimation algorithm.

Function S^L(real coefficient c, itemset $I = a_1 \dots a_m$, $\mathcal{F}$, integer k)

1. If $k = 0$ return c
2. If $c \geq 0$
3. return $c \cdot \sum_{i_1 < \dots < i_k \leq m} L(a_{i_1} a_{i_2} \dots a_{i_k}, \mathcal{F}, k-1)$
4. else
5. return $c \cdot \sum_{i_1 < \dots < i_k \leq m} U(a_{i_1} a_{i_2} \dots a_{i_k}, \mathcal{F}, k-1)$

Function S^U(real coefficient c, itemset $I = a_1 \dots a_m$, $\mathcal{F}$, integer k)

1. If $k = 0$ return c
2. If $c \geq 0$
3. return $c \cdot \sum_{i_1 < \dots < i_k \leq m} U(a_{i_1} a_{i_2} \dots a_{i_k}, \mathcal{F}, k-1)$
4. else
5. return $c \cdot \sum_{i_1 < \dots < i_k \leq m} L(a_{i_1} a_{i_2} \dots a_{i_k}, \mathcal{F}, k-1)$

Figure 4.2: Itemset support estimation algorithm, additional functions.

As dense databases we used the `mushroom` database from the UCI Machine Learning Archive [BM98], and a census data of elderly people from the University of Massachusetts at Boston Gerontology Center available at

`http://www.cs.umb.edu/~sj/datasets/census.arff.gz`.

Since both datasets involve multivalued attributes, we replaced each attribute (including binary ones) with a number of Boolean attributes, one for each possible value of the original attribute.

Before we present a detailed experimental study of the quality of bounds, we present the results of applying the bounds to a practical task. Suppose that we did not have enough time or computational resources to run the Apriori (or similar) algorithm completely, and we decided to stop the algorithm after finding frequent itemsets of size less than or equal to 2. We then use lower bounds to find frequent itemsets of size greater than 2. The experimental results for `mushroom` and `census` databases are shown in Figures 4.3 and 4.4 respectively.

The figures show, for various values of minimum support, the true number of frequent itemsets of sizes 3 and 4, the number of itemsets that we discovered to be frequent by using our bounds, and the ratio of the two numbers.

For large values of minimum support we are more likely to classify an itemset correctly than for smaller ones. The data shows that for itemsets with largest support the chances of actually being determined to be frequent without consulting the data can be as high as 80%.

We now present an experimental analysis of the bounds obtained. In what follows, by *trivial bounds* for the support of an itemset I we mean 0 for the lower bound, and for the upper bound: the minimum of the upper bounds of the supports of all proper subsets of I and of the minimum support. As in the example above

Itemset size	Min. support	18%	25%	30%	37%	43%	49%	55%	61%	73%
3	Frequent	1761	893	498	308	152	70	45	23	13
	Est. Freq.	345	244	179	127	86	54	34	19	10
	ratio (%)	19.6%	27.3%	35.9%	41.2%	56.6%	77.1%	75.6%	82.6%	76.9%
4	Frequent	4379	1769	795	368	147	48	29	16	6
	Est. Freq.	298	202	131	85	53	31	18	10	2
	ratio (%)	6.8%	11.4%	16.5%	23.1%	36.1%	64.6%	62.1%	62.5%	33.3%

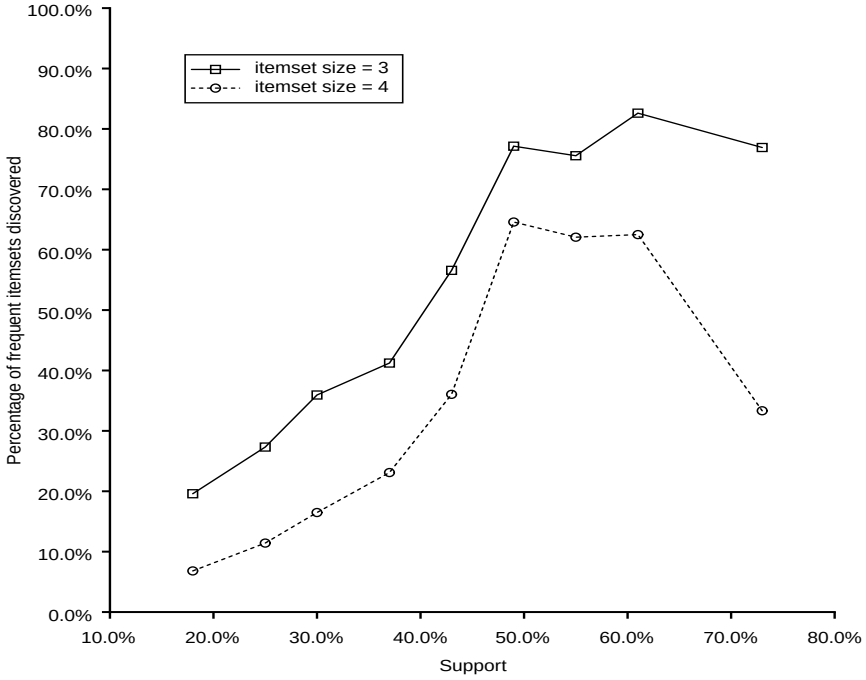


Figure 4.3: Discovered vs. total frequent itemsets for the `mushroom` dataset

Itemset size	Min. support	1%	2%	3%	5%	10%	15%	30%	50%
3	Frequent	1701	1377	1145	879	503	312	112	40
	Est. Freq.	154	149	146	137	108	90	47	21
	ratio (%)	9.1%	10.8%	12.8%	15.6%	21.4%	28.9%	42.0%	52.6%
4	Frequent	5050	3560	2728	1901	852	485	105	20
	Est. Freq.	103	98	94	85	64	48	18	3
	ratio (%)	2.0%	2.8%	3.5%	4.5%	7.5%	9.9%	17.1%	15.0%

Figure 4.4: Ratios of discovered to total frequent itemsets for the `census` data

itemset size	3	4	5	6
average interval width	0.0482797	0.0313103	0.0228579	0.0196316
average upper bound	0.0568679	0.0319395	0.0228771	0.0196316
average lower bound	0.00858817	0.000629199	1.925e-05	0
itemsets with nontrivial bounds	7.04%	0.59%	0.04%	0.00%
itemsets with nontrivial lower	4.06%	0.39%	0.02%	–
average lower improvement	0.211321	0.161151	0.0962518	–
itemsets with nontrivial upper	6.43%	0.47%	0.03%	–
average upper improvement	0.0225656	0.00983444	0.00262454	–
time [ms/itemset]	0.2	0.3	1	7

(a) 1.8% minimum support, all itemsets

itemset size	3	4	5	6
average interval width	0.102848	0.105024	0.106997	0.110767
average upper bound	0.127438	0.109572	0.107491	0.110767
average lower bound	0.0245896	0.00454846	0.00049354	0
itemsets with nontrivial bounds	20.17%	4.25%	0.58%	0.02%
itemsets with nontrivial lower	11.64%	2.82%	0.46%	–
average lower improvement	0.211321	0.161151	0.106164	–
itemsets with nontrivial upper	18.41%	3.43%	0.40%	0.02%
average upper improvement	0.0225656	0.00983444	0.00333985	0.00338427

(b) 1.8% minimum support, frequent itemsets only

itemset size	3	4	5	6
average interval width	0.171608	0.205194	0.222602	0.231362
average upper bound	0.235004	0.223174	0.225491	0.231362
average lower bound	0.0633963	0.0179804	0.00288882	0
itemsets with nontrivial bounds	48.55%	16.79%	3.40%	0.14%
itemsets with nontrivial lower	30.00%	11.16%	2.72%	–
average lower improvement	0.211321	0.161151	0.106164	–
itemsets with nontrivial upper	44.00%	13.56%	2.33%	0.14%
average upper improvement	0.0238776	0.00983444	0.00333985	0.00338427

(c) 9% minimum support, frequent itemsets only

Table 4.1: Results for the `census` dataset

here too we mine frequent itemsets with at most two items, and compute bounds for larger ones.

Table 4.1 (a) contains the results for the `census` dataset with minimum support of 1.8%.

The parameter r in the estimation algorithm was chosen for each itemset I to be $|I| - 1$ for maximum accuracy. This causes an increase in estimation time for larger itemsets. Later in the section we present results showing that limiting the value of r can give very fast estimates with a very small impact on the quality of

the bounds. All experiments were run on a 100MHz Pentium machine with 64MB of memory.

The bounds obtained are fairly accurate. The width of the interval between the lower and upper bounds varied from 0.048 to 0.019 for itemsets of size 3. Note that the estimates become more and more accurate for larger itemsets. The reason is that the bulk of large itemsets will have subsets whose support is very small, thus giving better average trivial bounds. Nontrivial upper bounds occur slightly more frequently than nontrivial lower bounds; however, lower bounds give on average much better improvement over the trivial bounds (this is due to the fact that our trivial upper bounds are quite sophisticated, while the trivial lower bound is just assumed to be 0).

The percentage of itemsets having nontrivial bounds is quite small. However those itemsets who have high support (and thus are the most interesting) are more likely to get interesting nontrivial bounds. This can be seen in Tables 4.1(b) and 4.1(c), where up to 48% of itemsets have nontrivial bounds proving the usefulness of Theorem 4.5.5. Note that in this case the interval width increases with the size of the itemsets. This is due to the fact that for high supports we don't have large number of itemsets with low supports that would create trivial upper bounds.

The conclusions were analogous for the `mushroom` database.

Table 4.2 shows how the choice of the argument r in the estimation algorithm influences the computation speed and the quality of the bounds. The results when r is set to the highest possible value (size of the estimated itemset minus one) is given in Table 4.1(a).

The results show that limiting the value of r to 2 or 3 gives a large speedup at a negligible decrease in accuracy. This is the approach we recommend. Also note

Census Data with 1.8% Minimum Support

$r = 2$				
itemset size	3	4	5	6
average interval width	0.0482797	0.0315442	0.022993	0.0196671
average upper bound	0.0568679	0.0321734	0.0230122	0.0196671
average lower bound	0.00858817	0.000629199	1.925e-05	0
itemsets with nontrivial bounds	7%	1%	0.10%	0%
time [ms/itemset]	0.18	0.24	0.34	0.46
$r = 3$				
itemset size	3	4	5	6
average interval width	0.0482797	0.0313103	0.0228666	0.0196328
average upper bound	0.0568679	0.0319395	0.0228859	0.0196328
average lower bound	0.00858817	0.000629199	1.925e-05	0
itemsets with nontrivial bounds	7%	0.50%	0%	0%
time [ms/itemset]	0.18	0.3	0.53	0.92

Table 4.2: Influence of the order of inequalities on the bounds

Census Data with 1.8% Minimum Support

itemset size	3	4	5	6	7
avg interval width	0.040498	0.081989	0.0668155	0.0392651	0.0180174
average upper bound	0.171319	0.120666	0.0685168	0.0392925	0.0180174
average lower bound	0.130821	0.0386768	0.00170127	2.73405e-05	0
time [ms/itemset]	0.24	0.46	0.96	2.54	5.12

Table 4.3: Estimates for itemsets with negations

that the proportion of itemsets with nontrivial bounds is higher for lower values of r. The same experiments repeated for frequent itemsets only yielded analogous results, so we omitted the data here.

Our last experimental result concerns estimating support of conjunctions allowing negated items using Corollary 4.5.3. Table 4.3 shows the results for the `census` ataset, with supports of all frequent 1- and 2-itemsets known (1.8% minimum support). In each of the itemsets exactly two of the items were negated. Again the inequalities gave fairly tight bounds.

CHAPTER 5

CONCLUSIONS

This work presented various applications of information theoretical and combinatorial methods in data mining. The most important contributions are summarized below.

5.1 Generalized Entropy Distances with Applications to Decision Tree Construction

In Section 2.2 an axiomatization has been introduced for a family of entropies including both Shannon entropy and the Gini index as special cases. This entropies were then applied to decision tree construction.

One of the most important criteria for splitting attribute selection in decision tree construction is Shannon entropy gain or the Shannon entropy gain ratio. In [LGR93] it has been shown that the expression $H(Y|X) + H(X|Y)$, where H is the Shannon Entropy, is a distance between attribute sets, and that using this distance as a splitting criterion during decision tree construction often leads to much smaller trees with almost no loss in accuracy.

The results of [LGR93] have been extended in Section 2.6 to generalized entropies axiomatized in Section 2.2

It is shown experimentally in Section 2.8 that generalized entropies are useful

as a splitting criterion for building decision trees as well. Generalized entropies produced in many cases even smaller decision trees without significant loss in accuracy.

Future research will include evaluations of various other measures as splitting criteria. We have obtained some promising results with expressions related to Goodman-Kruskal association index well known in statistical literature.

5.2 Association rule pruning and interestingness

One of major problems in association rule mining is huge number of rules produced, creating a secondary data mining problem. Indeed, even a toy `contact-lenses` database produces hundreds of association rules, most of them between independent attributes. There are two methods of dealing with this problem: sorting rules based some interestingness measure and rule pruning.

Chapter 3 contains contributions to both methods. Section 3.4 presents a new interestingness measure generalizing three well-known measures: chi-squared, entropy gain and Gini gain.

Section 3.9 contains a method of pruning association rules using Maximum Entropy Principle. Usefulness of both methods is shown experimentally. It is worth mentioning that the Maximum Entropy pruning gives high reduction in the number of rules while retaining most of the interesting ones.

Future research will concentrate on further improvements to rule pruning and incorporating background knowledge in the rule selection and pruning process.

5.3 Bonferroni inequalities

A series of seminal papers by H. Mannila et al. [MT96, Man01, PMS01] introduced the idea of using supports of itemsets discovered with a data mining algorithm to obtain the size of arbitrary database queries.

In Chapter 4 this work a solution to the problem is presented using a variation of the so called Bonferroni inequalities [GS96].

Modifications of Bonferroni inequalities have been developed which allow for estimating bounds of arbitrary database queries based on supports of frequent itemsets. Special cases like estimating support of an itemset with an unknown support, or of an itemset with negated attributes are also considered. Experiments show that useful bounds can be obtained from the inequalities in many significant cases.

Future work will include finding Bonferroni inequalities for other types of queries like for example monotonic functions, and examining approximation methods other than Bonferroni inequalities.

REFERENCES

[Adw97] Ratnaparkhi Adwait. "A Simple Introduction to Maximum Entropy Models for Natural Language Processing." IRCS Report 97–08, University of Pennsylvania, 3401 Walnut Street, Suite 400A, Philadelphia, PA, May 1997.

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. "Mining Association Rules between Sets of Items in Large Databases." In *Proc. ACM SIGMOD Conference on Management of Data*, pp. 207–216, Washington, D.C., 1993.

[AL99] Y. Aumann and Y. Lindell. "A Statistical Theory for Quantitative Association Rules." In *Knowledge Discovery and Data Mining*, pp. 261–270, 1999.

[AMS96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Inkeri Verkamo. "Fast Discovery of Association Rules." In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 307–328. AAAI Press, Menlo Park, 1996.

[And74] A. H. Andersen. "Multidimensional Contigency Tables." *Scand. J. Statist*, **1**:115–127, 1974.

[BA99] R. J. Bayardo and R. Agrawal. "Mining the Most Interesting Rules." In *Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 145–154, August 1999.

[Bad95] J. Badsberg. *An Environment for Graphical Models*. PhD thesis, Aalborg University, 1995.

[BFO98] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall, Boca Raton, 1998.

[BG99] L. Buzzigoli and A. Giusti. "An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals." In *Statistical Data Protection (SDP'98), Eurostat*, pp. 131–147, Luxembourg, 1999.

[BM98] C. L. Blake and C. J. Merz. *UCI Repository of machine learning databases*. University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[BVW00] T. Brijs, K. Vanhoof, and G. Wets. "Reducing redundancy in characteristic rule discovery by using integer programming techniques." *Intelligent Data Analysis Journal*, **4**(3), 2000.

[CA90] K.T. Cheng and V. Agrawal. "An Entropy Measure of the Complexity of Multi-Output Boolean Function." In *Proc. of the 27th Design Automation Conference*, pp. 302–305, 1990.

[CG02] T. Calders and B. Goethals. "Mining All Non-derivable Frequent Itemsets." In *6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2002)*, pp. 74–85, Helsinki, Finland, August 2002.

[Csi89] I. Csiszar. "A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling." *The Annals of Staticstics*, **17**(3):1409–1413, 1989.

[CSS98] V. Cheushev, V. Shmerko, D.A. Simovici, and S. Yanushkevich. "Functional Entropy and Decision Trees." In *Proc. ISMVL'98*, pp. 257–262, 1998.

[Czi72] I. Cziszar. "A Class of Measures of Informativity of Observation Channels." *Periodic Math. Hungarica*, **2**:191–213, 1972.

[D 00] L. Cristofor D. A. Simovici, D. Cristofor. "Generalized Entropy and Projection Clustering of Categorial Data." In *Principles of Data Mining and Knowledge Discovery (PKDD)*, volume 1910 of *Lecture Notes in Artificial Intelligence*, pp. 619–625, Lyon, 2000. Springer-Verlag.

[Dar70] Z. Daróczy. "Generalized Information Functions." *Information and Control*, **16**:36–51, 1970.

[Dev74] P.A. Devijer. "Entropie quadratique et reconnaisance des formes." In *Computer Oriented Learning Processes, Proceedings of the NATO Advanced Study Institute*, pp. 257–278, Château de Bonas, France, 1974.

[DLS80] J. N. Darroch, S. L. Lauritzen, and T. P. Speed. "Markov fields and log-linear interaction models for contingency tables." *Annals of Statistics*, **8**:522–539, 1980.

[Dob01] A. Dobra. "Computing sharp integer bounds for entries in contingency tables given a set of fixed marginals." Technical report, Department of Statistics, Carnegie Mellon University, 2001.

[DR72] J. N. Darroch and D. Ratcliff. "Generalized Iterative Scaling for Log-Linear Models." *Annals of Mathematical Statistics*, **43**:1470–1480, 1972.

[Fad56] D.K. Faddeev. "On the notion of entropy of finite probability distributions." *Usp. Mat. Nauk*, **11**:227–231, 1956. (in Russian).

[FRB02] U. Fayyad, N. Rothleder, and P. Bradley. "Enterprise Customer Data Mining for E-Business." Tutorial at the Second SIAM Conference on Data Mining, April 2002.

[GHK94] A. J. Grove, J. Y. Halpern, and D. Koller. "Random Worlds and Maximum Entropy." *Journal of Artificial Intelligence Research*, **2**:33–88, 1994.

[GS96] J. Galambos and I. Simonelli. *Bonferroni-type Inequalities with Applications*. Springer, 1996.

[HH99] R. Hilderman and H. Hamilton. "Knowledge discovery and interestingness measures: A survey." Technical Report CS 99-04, Department of Computer Science, University of Regina, 1999.

[HR66] P. L. Hammer and S. Rudeanu. *Pseudo-Boolean Methods for Bivalent Programming*, volume 23 of *Lecture Notes in Mathematics*. Springer-Verlag, Cambridge, 1966.

[HW97] C.H. Hwang and A.C.H. Wu. "An Entropy Measure for Power Estimation of Boolean Function." In *Proc. of Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 101–106, 1997.

[IU62] R.S. Ingarden and K. Urbanik. "Information without Probability." *Coll. Math.*, **1**:281–304, 1962.

[JS99] S. Jaroszewicz and D. A. Simovici. "On Axiomatization of Conditional Entropy of Functions between Finite Sets." In *Proc. of the 29th ISMVL*, pp. 24–31, Freiburg, Germany, 1999.

[JS01] S. Jaroszewicz and D. A. Simovici. "A General Measure of Rule Interestingness." In *5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2001)*, pp. 253–265, 2001.

[JS02] S. Jaroszewicz and D. A. Simovici. "Pruning Redundant Association Rules Using Maximum Entropy Principle." In *Advances in Knowledge Discovery and Data Mining, 6th Pacific-Asia Conference, PAKDD'02*, pp. 135–147, Taipei, Taiwan, May 2002.

[Khi57] A. Ia. Khinchin. *Mathematical Foundations of Information Theory*. Dover Publ., New York, 1957.

[KK92] J.N. Kapur and H.K. Kesavan. *Entropy Optimization Principles with Applications*. Academic Press, San Diego, 1992.

[Ler81] I. C. Lerman. *Classification et analyse ordinale des données*. Dunod, Paris, 1981.

[LGR93] A. Lloris-Ruiz, J.F. Gomez-Lopera, and R. Roman-Roldan. "Entropic Minimization of Multiple-Valued Logic Functions." In *Proc. ISMVL'93*, pp. 24–28, 1993.

[LHM99] Bing Liu, Wynne Hsu, and Yiming Ma. "Pruning and Summarizing the Discovered Associations." In Surajit Chaudhuri and David Madigan, editors, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 125–134, N.Y., August 15–18 1999. ACM Press.

[Man91] R. López de Mántaras. "A Distance-Based Attribute Selection Measure for Decision Tree Induction." *Machine Learning*, **6**:81–92, 1991.

[Man01] H. Mannila. "Combining Discrete Algorithms and Probabilistic Approaches in Data Mining." In L. DeRaedt and A. Siebes, editors, *Principles of Data Mining and Knowledge Discovery*, volume 2168 of *Lecture Notes in Artificial Intelligence*, p. 493. Springer-Verlag, Berlin, 2001.

[McE77] R.J. McEliece. *The Theory of Information and Coding. A mathematical Framework for Communication*. Encyclopedia of Mathematics and its Applications. Addisson-Wesley, Reading Massachusetts, 1977.

[MFM98] Y. Morimoto, T. Fukuda, H. Matsuzawa, T. Tokuyama, and K. Yoda. "Algorithms for Mining Association Rules for Binary Segmentations of Huge Categorical Databases." In *Proc. of the 24th Conf. on Very Large Databases*, pp. 380–391, 1998.

[Min89] J. Mingers. "An Empirical Comparison of Selection Measures for Decision Tree Induction." *Machine Learning*, **3**:319–342, 1989.

[Mit97] T.M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[Mor98] S. Morishita. "On Classification and Regression." In *Proc. of the First Int'l Conf. on Discovery Science*, number 1532 in Lecture Notes in Artificial Intelligence, pp. 40–57, 1998.

[MR75] A.M. Mathai and P.N. Rathie. *Basic Concepts in Information Theory and Statistics — Axiomatic Foundations and Applications*. Halsted Press, John Wiley & Sons, 1975.

[MT96] H. Mannila and H. Toivonen. "Multiple Uses of Frequent Sets and Condensed Representations." In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pp. 189–194, Portland, Oregon, 1996.

[NC01] J. R. Nayak and D. J. Cook. "Approximate Association Rule Mining." In *Proceedings of the Florida Artificial Intelligence Research Symposium*, 2001.

[PMS01] D. Pavlov, H. Mannila, and P. Smyth. "Beyond Independence: Probabilistic Models for Query Approximation on Binary Transaction Data." Technical Report ICS TR-01-09, Information and Computer Science Department, UC Irvine, 2001.

[PT00] B. Padmanabhan and A. Tuzhilin. "Small is beautiful: discovering the minimal set of unexpected patterns." In Raghu Ramakrishnan, Sal Stolfo, Roberto Bayardo, and Ismail Parsa, editors, *Proceedinmgs of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-00)*, pp. 54–63, N. Y., August 2000. ACM Press.

[Qui93] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[Rat96] Adwait Ratnaparkhi. "A Maximum Entropy Model for Part-of-Speech Tagging." In Eric Brill and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 133–142. Association for Computational Linguistics, Somerset, New Jersey, 1996.

[RC98] A. Ragel and B. Crémilleux. "Treatment of Missing Values for Association Rules." In Xindong Wu, Ramamohanarao Kotagiri, and Kevin B. Korb, editors, *Proceedings of the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining (PAKDD-98)*, volume 1394 of *LNAI*, pp. 258–270, Berlin, April 15–17 1998. Springer.

[Ren58] A. Rényi. "Quelques Remarques ser les Probabilites des Evenements Dependants." *Journal de Mathematique*, **37**:393–398, 1958.

[Rud74] S. Rudeanu. *Boolean Functions and Equations*. North-Holland, Amsterdam, 1974.

[SBM98] C. Silverstein, S. Brin, and R. Motwani. "Beyond Market Baskets: Generalizing Association Rules to Dependence Rules." *Data Mining and Knowledge Discovery*, **2**:39–68, 1998.

[SJ02] D. A. Simovici and S. Jaroszewicz. "An Axiomatization of Partition Entropy." *IEEE Transactions on Information Theory*, **48**:2138–2142, 2002.

[SK98] E. Suzuki and Y. Kodratoff. "Discovery of Surprising Exception Rules Based on Intensity of Implication." In *Proc of PKDD-98, Nantes, France*, pp. 10–18, 1998.

[SLR99] D. Shah, L. V. S. Lakshmanan, K. Ramamritham, and S. Sudarshan. "Interestingness and Pruning of Mined Patterns." In *1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1999.

[SR93] D.A. Simovici and C. Reischer. "On Functional Entropy." In *Proc. ISMVL'93*, pp. 100–104, 1993.

[ST95] D. A. Simovici and R. L. Tenney. *Relational Database Systems*. Academic Press, New York, 1995.

[Suz97] E. Suzuki. "Autonomous Discovery of Reliable Exception Rules." In David Heckerman, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy, editors, *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, p. 259. AAAI Press, 1997.

[TKR95] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hätönen, and H. Mannila. "Pruning and grouping discovered association rules." In *MLnet Workshop on Statistics, Machine Learning, and Discovery in Databases*, pp. 47–52, Heraklion, Crete, Greece, April 1995.

[TKS02] P.-N. Tan, V. Kumar, and J. Srivastava. "Selecting the Right Interestingness Measure for Association Patterns." In *Proc of the Eighth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD-2002)*, pp. 32–41, 2002.

[Weh96] L. Wehenkel. "On uncertainty Measures Used for Decision Tree Induction." In *Info. Proc. and Manag. of Uncertainty in Knowledge-Based Systems (IPMU'96)*, pp. 413–418, Granada, Spain, July 1996.

[WF00] I. H. Witten and E. Frank. *Data Mining.* Morgan-Kaufmann, San Francisco, 2000.