DIVERSITY-BASED INTERESTINGNESS MEASURES FOR ASSOCIATION RULE MINING

Huebner, Richard A.
Norwich University
rhuebner@norwich.edu

ABSTRACT

Association rule interestingness measures are used to help select and rank association rule patterns. Diversity-based measures have been used to determine the relative interestingness of summaries. However, little work has been done that investigates diversity measures with association rule mining. Besides support, confidence, and lift, there are other interestingness measures, which include generality (also known as coverage), reliability, peculiarity, novelty, surprisingness, utility, and applicability. This paper investigates the application of diversity-based measures to association rule mining.

INTRODUCTION

Interestingness measures are necessary to help select and rank association rule patterns. Each interestingness measure produces different results, and experts have different opinions of what constitutes a good rule (Lenca, Meyer, Vaillant, & Lallich, 2008). The interestingness of discovered association rules is an important and active area within data mining research (Geng & Hamilton, 2006). The primary problem is the selection of interestingness measures for a given application domain. However, there is no formal agreement on a definition for what makes rules interesting. Association rule algorithms produce thousands of rules, many of which are redundant (Li & Zhang, 2003; McGarry, 2005). In order to filter the rules, the user generally supplies a minimum threshold for support and confidence. Support and confidence are basic measures of association rule *interestingness*. Additionally, these are the most common measures of interest. However, generating rules that meet minimum thresholds for support and confidence may not be interesting. This is because rules are often produced that are already known by a user who is familiar with the application domain.

The challenge in association rule mining (ARM) essentially becomes one of determining which rules are the most interesting. With so many interestingness rules to choose from, it is difficult to determine which one to use for a given domain. This problem is exacerbated by the fact that different interestingness measures produce different results for the same data set, thus making it difficult for the user to interpret the measures (McGarry, 2005).

The purpose of this paper is to review a few of the interestingness measures based on diversity. Diversity of a data set is defined as when comparing two data sets, the one with more diverse rules is more interesting. Diversity will be used to compare two data sets to determine which data set contains rules that are more interesting. Even though diversity is a criterion for measuring summaries, little work has been done that focuses on the diversity of association rules. Measures can be used with either summaries, association rules, or classification rules. This paper focuses exclusively on association rule interestingness measures which are based on diversity.

ASSOCIATION RULE MINING

Association rule mining is a category of data mining tasks that correlate a set of items with other sets of items in a database. Association rules "aim to extract interesting correlations, frequent patterns, associations or causal structures among sets of items in the transaction databases or other repositories"

(Kotsiantis & Kanellopoulos, 2006, p. 71). Association rule mining is one of the most important data mining techniques used today and is a mature field of research (Ceglar & Roddick, 2006; Xu & Li, 2007).

Association rules were first proposed by Agrawal et al. (Agrawal, Imielinski, & Swami, 1993). The main driver for research on association rules was the analysis of customer market basket transactions. An example of an association rule is as follows. 60% of customers that purchase potato chips also purchase soda in the same transaction. Agrawal et al.'s work established a formal model for association rules and establishes algorithms that find large itemsets, confidence, and support of each rule discovered in the itemset. Association rules have been applied to a wide variety of application areas, which will be covered later in this paper. Association rule algorithms can generate thousands of rules, many of which can be redundant. These redundant rules are essentially useless, so researchers have solved this problem by defining new interestingness measures, incorporating constraints, or by designing templates to mine for restricted rules (Xu & Li, 2007). Also, a primary goal of knowledge discovery in databases is to produce interesting rules that can be interpreted by a user (Lenca et al., 2008).

One research team (Lee & Siau, 2001) outlined the requirements and challenges associated with data mining. First, data mining must be able to handle different types of data. Second, data mining algorithms must be scalable and efficient. Third, data mining must be able to handle noisy and missing data. Fourth, data mining techniques should present results in a way that is easy to understand. Fifth, data mining techniques should support requests at different levels of granularity. That is, data mining can be done at different levels of abstraction. Sixth, data mining algorithms should be flexible enough to deal with data from different sources. Finally, a major concern within data mining today is the threat to privacy and data security. This is because data mining makes it easy to establish profiles of individuals based on data from multiple sources (Lee & Siau, 2001).

General issues related to data mining include the identification of missing information, dealing with noise or missing values, and operating with very large databases (VLDBs). Additionally, data mining is normally used to access data contained in a data warehouse, which contain high degrees of dimensionality, thus making data mining more complex (Marakas, 2003). In order to produce accurate data mining results, it is important that the underlying data is complete. Without complete data, accurate rules cannot be produced. The field of privacy-preserving data mining (PPDM) investigates the issues pertaining to mining association rules when there are missing values in the database or data warehouse. Several privacy-preserving association rule algorithms have also been proposed to address this issue (Chen & Weng, 2008; Zhan, Matwin, & Chang, 2007). However, there are still many open issues related to privacy-preserving association rule mining (PPARM).

INTERESTINGNESS MEASURES

Two important measures within association rule mining are *support* and *confidence*. *Support* for an association rule $X \Rightarrow Y$ is the percentage of transactions in the database that contain $X \cup Y$. *Confidence* for an association rule (sometimes denoted as strength, or α) $X \Rightarrow Y$ is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain X. (Dunham, 2003). In other words, support describes how often the rule would appear in the database, while confidence measures the strength of the rule. A user establishes minimum support (minsup) and minimum confidence (minconf). Rules are then generated based on those criteria. Users can select minsup and minconf parameters before or after rule generation. An example follows. Given a database of supermarket transaction data, a rule might be generated that infers milk → eggs, with support = 40% and confidence = 75%. This means that milk → eggs occurred in 40% of the transactions in the database. It also means that 75% of the time that milk occurs, so do eggs. The antecedent for this rule is milk, while the consequent is eggs. Larger values of confidence and smaller values of support are normally selected when determining which association rules to keep. A third measure of interestingness is *lift*. Lift is a measure of the probability of finding the consequent in any random basket. In other words, lift "measures how well the associative rule performs by comparing its performance to the "null" rule" (Marakas, 2003, p. 342).

What makes a rule *interesting*? One way to define interestingness is that a rule must be valid, new and comprehensive (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Commonly used measures of interestingness include support and confidence as described above. Many other interestingness measures for association rules have been established, but there is no formal agreement on how interestingness should be defined. Some interestingness measures include conciseness, generality (also known as coverage), reliability, peculiarity, diversity, novelty, surprisingness, utility, and applicability (Geng & Hamilton, 2006). Interestingness measures can be classified into two categories: objective and subjective. Objective measures are based on statistics, while subjective measures are based on an understanding of the user's domain knowledge. For example, objective measures include generality and reliability, conciseness, peculiarity, diversity and surprisingness. The subjective interestingness measures include novelty, utility, and applicability. These measures assist in validating association rule results (Tamir & Singer, 2006). Interestingness measures based on diversity have received little attention in the literature (Geng & Hamilton, 2006), thus the need for further research in this particular area.

The generality (or coverage) of a pattern is determined by how comprehensive the pattern is and the fraction of the number of records that match the pattern. General patterns include frequent itemsets, which is also the most frequently studied type of patterns in association rule mining (Geng & Hamilton, 2006). The reliability of a pattern is determined by the percentage of cases found in the itemset. In other words, the rule might be interesting if a high percentage of cases contain the rule. One study applied the reliability measure to evaluate clinical datasets on hepatitis (Ohsaki, Kitaguchi, Okamoto, Yokoi, & Yamaguchi, 2004).

Concise association rules are those that contain few patterns. A pattern is concise only if it contains few attribute-value pairs (Geng & Hamilton, 2006). One method for generating concise association rules was established by Xu and Li (2007). The reason for establishing a method for generating concise association rules is because of the low quality of mined rules. The produced rules are low quality if there are too many redundancies and a large extracted rule set (Xu & Li, 2007). Therefore, Xu and Li's perspective of the problem is how to generate non-redundant and concise association rules, thus producing higher quality association rules.

Tamir and Singer (2006) established an interestingness measure called confidence gain. This measure is unique in that it extracts rules more closely aligned to human free associations rather than associations between itemsets. The confidence gain measure produced better rules than other measures and can be used for personalization of web sites, query expansion and improving classification performance over small itemsets (Tamir & Singer, 2006). In addition to investigating interestingness measures, some research has been done on constraint-based association rule mining. Constraint-based association rule mining typically uses post-processing, pattern filtering, or dataset filtering as a way to filter or constrain the number of patterns produced.

Association rule algorithms are designed to efficiently find large itemsets. Large itemsets are those that have a number of occurrences above some minimum threshold (Dunham, 2003). The reason we are more interested in large itemsets is that many of the produced association rules may not be interesting. Apriori is an important algorithm in association rule mining (Dunham, 2003; El-Hajj & Zaiane, 2003). The apriori algorithm was first established by Agrawal and Srikant (1994). It is the major technique used to detect large itemsets within a database of transactions. It also forms the basis of many association rule algorithms (El-Hajj & Zaiane, 2003).

Hilderman and Hamilton established three primary principles that a good interestingness measure should satisfy (Hilderman & Hamilton, 2001):

- The *minimum value principle*, which states that a uniform distribution is the most uninteresting.
- The *maximum value principle*, which states the most uneven distribution is the most interesting.
- The *skewness principle*, which states that the interestingness measure for the most uneven distribution will decrease when then number of classes of tuples increases.
- The *permutation invariance principle*, which states that interestingness for diversity is unrelated to the order of the class and it is only determined by the distribution of counts.
- The *transfer principle*, which states that interestingness increases when a positive transfer is made from the count of one tuple to another whose count is greater.

DIVERSITY-BASED INTERESTINGNESS MEASURES

According to Geng and Hamilton, a "pattern is diverse if its elements differ significantly from each other, while a set of patterns is diverse if the patterns in the set differ significantly from each other" (Geng & Hamilton, 2006, p. 3). Summaries can be measured using diversity-based interestingness measures. While there has been research on using diversity to measure summaries, there is little, if any, research that focuses on measuring the interestingness of association or classification rules (Geng & Hamilton, 2006). Therefore, this study suggests that diversity can be used to measure association rule interestingness.

Diversity us generally determined by two factors: 1) the proportional distribution of classes in the population, and 2) the number of classes. Consider two different sets of rules mined from a dataset. We can consider the rules that are more diverse to be more interesting. Additionally, we can consider a set of rules less interesting if the rules are less diverse. Another way of thinking about this is by considering that too many similar rules will convey less knowledge to a user.

BRIEF METHOD

DATA SET	CHARACTERISTICS
Breast Cancer data set	• Multivariate data • Contains categorical and integer data • Number of instances(rows): 266 • Number of attributes: 9 + the class attribute Source: UCI Machine Learning Library
Adult data set	• Multivariate data • Contains categorical and integer data • Number of instances (rows): 48842 • Number of attributes: 14 • There are some missing values Source: UCI Machine Learning Library

The open-source data mining toolkit, Orange, was used to conduct the study. Orange uses the Python programming language, which allows the programmer to extend or adapt modules for specific experiments. The adult data set was analyzed and compared using the basic interestingness measures: support, confidence, and lift. Minimum support was set at .25. Characteristics of the data set can be found in Table 1. The adult data set is a benchmark data set that is frequently used in the machine learning and data mining community. It is available online at the University of California at Irvine in their machine learning library. Their online library consists of hundreds of synthetic and real-world data sets collected over the past 20 years. The top 20 rules were produced and sorted by confidence. Data sets are described in the tables below. Once the rules were produced, diversity measures were applied to the data sets to determine relative interestingness. Two diversity measures were used: variance and Shannon. Shannon's measure is based on information theoretic function and entropy. It measures the relative information content present in a dataset. A data mining practitioner should have a deep understanding of the data and application domain before analyzing and interpreting the data output. Otherwise, results will be meaningless to the untrained eye. Therefore, interpreting data mining output requires expertise in statistics, machine learning, and data mining techniques.

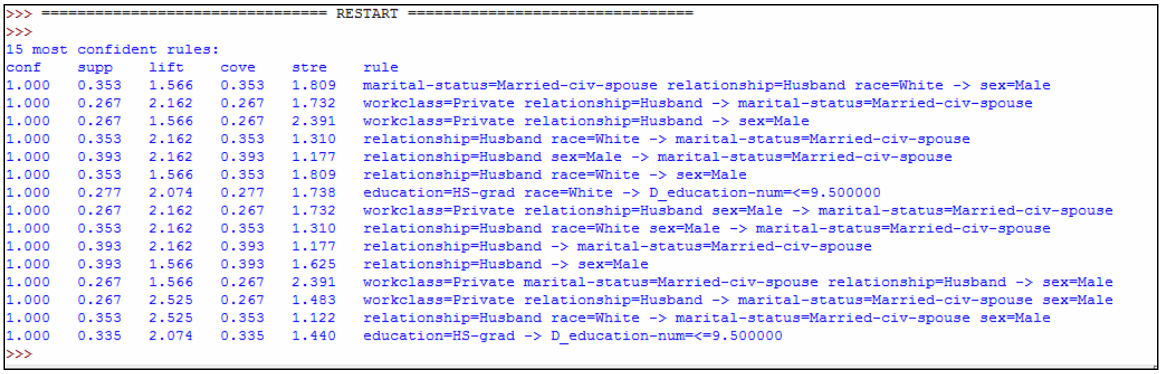
The equations for variance and the Shannon (Shannon, 1948) measure are shown in the table below. These are only two measures based on diversity. There are fourteen other measures based on diversity that are available. Most measures are designed to be used with a specific application domain. We chose variance and Shannon because of their wide-spread use.

Variance	Claude Shannon (Entropy)
$\dfrac{\sum_{i=1}^{m}(p_i - \bar{q})^2}{m-1}$	$-\sum_{i=1}^{m} p_i \log_2 p_i$

In the above equation for variance, p_i is the probability for class i, and $\bar{q}$ is the average probability for all classes. For each dataset, only the top 15 rules will be displayed, sorted by confidence. This is done because too many rules would be produced, which would not be useful to the data mining user. We are essentially only concerned with the most interesting rules that the algorithm produces. It is also important to note that there are several different measures for entropy, which is a mathematical measure of information loss. Typically, we try to minimize this loss, so lower values would indicate less information loss.

PRELIMINARY RESULTS

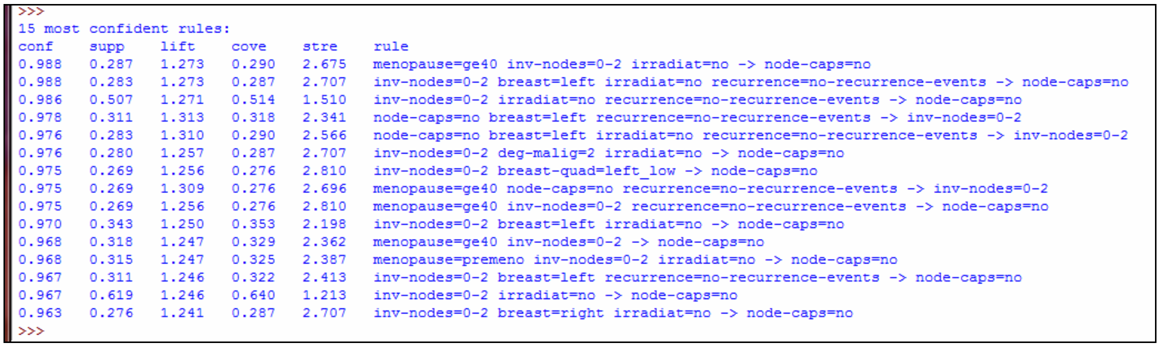
The primary goal of the adult data set is to determine the type of adult that makes greater than $50,000. In this case, we are using association rules to help make predictions. The adult data set showed 15 rules ranked in order by confidence. Support ranges from .267 to .393. A value of 26.7% means that the algorithm found 26% of the transactions contained the left hand side of the rule. For example, in examining the second rule produced from the adult data set, we find 26.7% (.267) of the records (transactions) contained both workclass=private and relationship=husband.



```
>>> =============================== RESTART ===============================
>>>
15 most confident rules:
conf    supp    lift    cove    stre    rule
1.000   0.353   1.566   0.353   1.809   marital-status=Married-civ-spouse relationship=Husband race=White -> sex=Male
1.000   0.267   2.162   0.267   1.732   workclass=Private relationship=Husband -> marital-status=Married-civ-spouse
1.000   0.267   1.566   0.267   2.391   workclass=Private relationship=Husband -> sex=Male
1.000   0.353   2.162   0.353   1.310   relationship=Husband race=White -> marital-status=Married-civ-spouse
1.000   0.393   2.162   0.393   1.177   relationship=Husband sex=Male -> marital-status=Married-civ-spouse
1.000   0.353   1.566   0.353   1.809   relationship=Husband race=White -> sex=Male
1.000   0.277   2.074   0.277   1.738   education=HS-grad race=White -> D_education-num=<=9.500000
1.000   0.267   2.162   0.267   1.732   workclass=Private relationship=Husband sex=Male -> marital-status=Married-civ-spouse
1.000   0.353   2.162   0.353   1.310   relationship=Husband race=White sex=Male -> marital-status=Married-civ-spouse
1.000   0.393   2.162   0.393   1.177   relationship=Husband -> marital-status=Married-civ-spouse
1.000   0.393   1.566   0.393   1.625   relationship=Husband -> sex=Male
1.000   0.267   1.566   0.267   2.391   workclass=Private marital-status=Married-civ-spouse relationship=Husband -> sex=Male
1.000   0.267   2.525   0.267   1.483   workclass=Private relationship=Husband -> marital-status=Married-civ-spouse sex=Male
1.000   0.353   2.525   0.353   1.122   relationship=Husband race=White -> marital-status=Married-civ-spouse sex=Male
1.000   0.335   2.074   0.335   1.440   education=HS-grad -> D_education-num=<=9.500000
>>>
```

Figure 1: Results of association mining (apriori algorithm) from the Adult dataset

Variance is computed by taking the variance of the top 15 produced association rules. The adult data set produced a variance of .002643 and a Shannon value of .50401.



```
>>>
15 most confident rules:
conf    supp    lift    cove    stre    rule
0.988   0.287   1.273   0.290   2.675   menopause=ge40 inv-nodes=0-2 irradiat=no -> node-caps=no
0.988   0.283   1.273   0.287   2.707   inv-nodes=0-2 breast=left irradiat=no recurrence=no-recurrence-events -> node-caps=no
0.986   0.507   1.271   0.514   1.510   inv-nodes=0-2 irradiat=no recurrence=no-recurrence-events -> node-caps=no
0.978   0.311   1.313   0.318   2.341   node-caps=no breast=left recurrence=no-recurrence-events -> inv-nodes=0-2
0.976   0.283   1.310   0.290   2.566   node-caps=no breast=left irradiat=no recurrence=no-recurrence-events -> inv-nodes=0-2
0.976   0.280   1.257   0.287   2.707   inv-nodes=0-2 deg-malig=2 irradiat=no -> node-caps=no
0.975   0.269   1.256   0.276   2.810   inv-nodes=0-2 breast-quad=left_low -> node-caps=no
0.975   0.269   1.309   0.276   2.696   menopause=ge40 node-caps=no recurrence=no-recurrence-events -> inv-nodes=0-2
0.975   0.269   1.256   0.276   2.810   menopause=ge40 inv-nodes=0-2 recurrence=no-recurrence-events -> node-caps=no
0.970   0.343   1.250   0.353   2.198   inv-nodes=0-2 breast=left irradiat=no -> node-caps=no
0.968   0.318   1.247   0.329   2.362   menopause=ge40 inv-nodes=0-2 -> node-caps=no
0.968   0.315   1.247   0.325   2.387   menopause=premeno inv-nodes=0-2 irradiat=no -> node-caps=no
0.967   0.311   1.246   0.322   2.413   inv-nodes=0-2 breast=left recurrence=no-recurrence-events -> node-caps=no
0.967   0.619   1.246   0.640   1.213   inv-nodes=0-2 irradiat=no -> node-caps=no
0.963   0.276   1.241   0.287   2.707   inv-nodes=0-2 breast=right irradiat=no -> node-caps=no
>>>
```

Figure 2: Results of association mining (apriori algorithm) from the Breast Cancer dataset.

The primary goal in the breast cancer dataset is to determine the characteristics of women that might be more susceptible to breast cancer. The adult data set produced a variance of .009925. Shannon's measure produced a value of .50171.

	Adult data set	Breast cancer data set
Support range	0.126	0.35
Variance	0.002643	0.009925
Shannon's Entropy	0.50401	0.50171

DISCUSSION

According to these preliminary results, we would decide that the breast cancer data set would be considered more interesting based on the variance. A higher variance indicates that the rules are more diverse. The major challenge in deciding which data set is more diverse is that different measures produce different values. As we have discussed earlier in this paper, diversity as an interestingness measure has received little attention in the literature. It is simple to take an interestingness measure that is usually used for summaries and apply it elsewhere (to association rules). Variance is a simple measure that can be used to compare two data sets and the rule diversity of each data set.

Analysis of the breast cancer data provided an entropy measure of .50171. Notice that the entropy for the breast cancer data set is lower than entropy for the adult data set. This is expected since lower entropy measures show less information loss. Using entropy as the measure of interest for these two data sets support our previous conclusion that the breast cancer data set has more diverse rules. Interpretation of all three measures tells us that the breast cancer data set has more diverse rules. This is encouraging since these measures tell us different things about the rules.

The results of this paper have several implications. First, data mining users and decision-makers must realize that there are additional interestingness measures besides support, confidence, and lift. These measures can be meaningless unless one has a firm grasp on the dataset itself. Additionally, the measures only tell us the probability of finding frequent itemsets and how frequently those itemsets occur in the data set. It is far more interesting to compare several data sets to determine which data set will produce more interesting rules.

One possible weakness of this study is that variance was computed against only the top 15 rules. Future research may examine the application of the interestingness measure to more rules, say the top 50 rules.

The reason why only the top 15 rules were used is because too many rules can be overwhelming to the data mining user. Conversely, using additional rules may give us a greater understanding of the diversity of several data sets. In the future, we can also apply additional diversity measures to association rules or classification rules.

Management should have an interest in learning how to glean as much information from data mining tasks as possible. This paper shows one way to do so. In this case, association rules are relatively easy to understand. However, the results from a single data set may not provide us with useful information. We should be comparing the results from multiple data sets to determine which data set will provide the most interesting results. This is, of course, highly based on the context of the data set and management's understanding of the data.

REFERENCES

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record, 22*(2), 207-216.

Ceglar, A., & Roddick, J. F. (2006). Association mining. *ACM Computing Surveys, 38*(2), Article 5.

Chen, Y.-L., & Weng, C.-H. (2008). Mining association rules from imprecise ordinal data. *Fuzzy Sets and Systems, 159*, 460-474.

Dunham, M. (2003). *Data mining: Introductory and advanced topics*. Upper Saddle River, NJ: Prentice Hall.

El-Hajj, M., & Zaiane, O. R. (2003). Inverted matrix: Efficient discovery of frequent items in large datasets in the context of interactive mining. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 109-118.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 1-24). Boston, MA: MIT Press.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys, 38*(3), Article 5.

Hilderman, R. J., & Hamilton, H. J. (2001). *Knowledge Discovery and Measures of Interest*. Boston, MA: Kluwer Academic.

Kotsiantis, S., & Kanellopoulos, R. (2006). Association Rules Mining: A Recent Overview. *GESTS International Transactions on Computer Science and Engineering, 32*(1), 71-82.

Lee, S. J., & Siau, K. (2001). A review of data mining techniques. *Industrial Management & Data Systems, 101*(1), 41-46.

Lenca, P., Meyer, P., Vaillant, B., & Lallich, S. (2008). On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operations Research, 184*, 610-626.

Li, J., & Zhang, Y. (2003). Direct Interesting Rule Generation. *Proceedings of the Third IEEE International Conference on Data Mining (ICDM '03)*, 155-167.

Marakas, G. (2003). *Decision Support Systems* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review, 20*(1), 39-61.

Ohsaki, M., Kitaguchi, S., Okamoto, K., Yokoi, H., & Yamaguchi, T. (2004). Evaluation of rule interestingness measures with a clinical dataset on hepatitis. *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 362-373.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379-423.

Tamir, R., & Singer, Y. (2006). On a confidence gain measure for association rule discovery and scoring. *The VLDB Journal, 15*(1), 40-52.

Xu, Y., & Li, Y. (2007). Generating concise association rules. *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, 781-790.

Zhan, J., Matwin, S., & Chang, L. (2007). Privacy-preserving collaborative association rule mining. *Journal of Network and Computer Applications, 30*, 1216-1227.