Topics in Association Rules

by

Mateen Shaikh

A thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Doctor of Philosophy
in
Statistics

Guelph, Ontario, Canada

© Mateen Shaikh, June, 2013
ABSTRACT

Topics in Association Rules

Mateen Shaikh
University of Guelph, 2013
Advisor:
Professor Paul D. McNicholas

Association rules are a useful concept in data mining with the goal of summarizing the strong patterns that exist in data. We have identified several issues in mining association rules and addressed them in three main areas. The first area we explore is standardized interestingness measures. Different interestingness measures exist on different ranges, and interpreting them can be subtly problematic. We standardize several interestingness measures and show how these are useful to consider in association rule mining in three examples. A second area we address is incomplete transactions. By applying statistical methods in new ways to association rules, we provide a more comprehensive means of analyzing incomplete transactions. We also describe how to find families of distributions for interestingness measure values when transactions are incomplete. Finally, we address the common result of mining: a plethora of association rules. Unlike methods which attempt to reduce the number of resulting rules, we harness this large quantity to find a higher-level set of patterns.
Acknowledgements

I have been fortunate to benefit many sources of support during my studies. I first wish to thank Professor Paul McNicholas for his guidance, support, and friendship over the years. He not only influenced me as my advisor but set an example for all his students.

I would also like to thank my advisory committee, Doctor Luiza Antonie and Professor Daniel Ashlock, and my examining committee, Professors Zeny Feng, Ryan Browne, Daniel Ashlock from the University of Guelph and Professor Dennis Lin from The Pennsylvania State University.

This work was also supported by the NSERC Discovery Grant by Natural Sciences and Engineering Research Council of Canada and from the and the Ontario Ministry of Research and Innovation’s Early Researcher Award.

Finally, I would like to thank friends, family, and especially the McNicholas research group for making this a memorable experience.
Publications

The following articles based on the work in this thesis have been submitted or in preparation.


• Shaikh, M., McNicholas, P. D., ‘Mining Missingness Patterns with Association Rules’. In preparation.

Contents

1 Introduction 1
  1.1 Significance 2
  1.2 Thesis Outline 3

2 Literature Review 6
  2.1 Association Rules 6
  2.2 Apriori Algorithm 8
  2.3 Interestingness Measures 9
    2.3.1 Lift 11
    2.3.2 Cosine Similarity 14
    2.3.3 Yule’s Q 14
    2.3.4 Gini Index 15
  2.4 Order Comparisons 15
  2.5 Missing Data 16
  2.6 Distributions to Test Independence 18
    2.6.1 Hypergeometric Distribution 19
    2.6.2 $\chi^2$ distribution 19
  2.7 Generalized Association Rules 21
  2.8 Bayesian Rose Tree 24

3 Standardized Interestingness Measures 28
  3.1 Using Interestingness Measures 28
    3.1.1 New Standardizations 29
    3.1.2 Experimental Results and Evaluation 31
  3.2 Traffic Accidents 32
  3.3 Reuters 38
  3.4 Random Transactions 42
  3.5 Summary 44
## Contents

### 4 Incomplete Transactions

4.1 Chapter Outline ............................................. 46
4.2 Methodology .................................................. 47
  4.2.1 A Discrete Test for Independence ......................... 49
4.3 Hepatitis data ................................................. 50
4.4 Pima data ..................................................... 53
4.5 Interestingness Measures with Missing Data .................... 57
  4.5.1 Simulating values ........................................... 62
4.6 Summary .................................................... 65

### 5 Hierarchies

5.1 Chapter Outline ............................................. 67
5.2 Methodology .................................................. 69
  5.2.1 Balance .................................................. 69
  5.2.2 Scoring .................................................. 71
5.3 Examples ..................................................... 75
  5.3.1 Gas Bar ................................................... 76
  5.3.2 Legionnaires’ Microarray Data ......................... 82
5.4 Summary .................................................... 88

### 6 Conclusion

6.1 Standardizations ............................................. 89
6.2 Incomplete Transactions ..................................... 91
6.3 Association Rule Hierarchies ................................ 93

### Bibliography

### A Standardizations

A.1 Derivations .................................................. 102
  A.1.1 Cosine Similarity ......................................... 103
  A.1.2 Yule’s Q .................................................. 104
  A.1.3 Gini ..................................................... 107
A.2 Scatterplot Matrices ......................................... 110
A.3 Tau-b by Deciles ............................................. 113

### B Hierarchical

B.1 Dirichlet Parameter Estimation ................................ 114
B.2 Dendrograms .................................................. 116

---

**vi**
List of Tables

2.1 Properties of several interestingness measures. .......................... 11
3.1 Rules for the Belgian Accident Data with the consequent \( \{35\} \) sorted by support and confidence. .................................................. 37
3.2 Rule orders for the Belgian accident data. ................................. 37
3.3 Rules for the Reuters Data with the consequent \( \{\text{gold}\} \) sorted by support and confidence. .................................................. 41
3.4 Rule orders for the Reuters data. ............................................. 41
3.5 Rule orders for randomly generated transactions. ....................... 44
4.1 Summary of the hepatitis data. ................................................. 51
4.2 Discretization method of ordinal and continuous variables. .......... 54
4.3 Table of how data manifest with missingness. ............................. 58
5.1 Frequency distribution of transactions from the gasbar simulation. Two beverages and four snacks are available. .................... 76
5.2 Scores for the gasbar hierarchies. Optimal values of \( \delta \) are shown in parenthesis. 79
5.3 Frequency distribution of transactions from the gasbar simulation. Two beverages and four snacks are available. .................... 80
5.4 second gas bar example. ...................................................... 81
5.5 Scores for the trees in Figure 5.4. .......................................... 85
List of Figures

2.1 The bounds on lift when $P(A) = P(B)$.

2.2 A comparison of a binary and non-binary tree.

3.1 Cosine similarity values for randomly rules. Vertical lines represent the range of possible cosine similarity values for the rule with the horizontal dash indicating the achieved value of the cosine similarity. The circle indicates the value of the standardized cosine similarity, the relative position of the achieved value in the range of possible values.

3.2 Plots of interestingness measures against their standardized counterparts for the Belgian accident data. Dark regions indicate a higher density of rules.

3.3 Plots of interestingness measures against their standardized counterparts for the Reuters data. Dark regions indicate a higher density of rules.

3.4 Plots of interestingness measures against their standardized counterparts for the random-transactions data. Dark regions indicate a higher density of rules.

4.1 Relationship between $\chi^2$ approximation and exact $p$-values with the hepatitis data.

4.2 Comparison of rescaled $p$-values and lift on rescaled axes.

4.3 Sample distribution of lift when data are MCAR.

4.4 Sample distribution of lift when items are dependent and data are MCAR.

4.5 Sample distribution of lift when items are dependent and data are missing MNAR.

5.1 Trees corresponding to gas bar data in Table 5.1.

5.2 Simulation with a designed subtree hierarchy.

5.3 Conditional statements from Paper I based on the binarized microarray data. When at least four of the seven are true, the strain is labelled clinical. When no more than three of the seven statements are true, then the strain is labelled environmental.
5.4 Sample trees suggested from hierarchical clustering. Heights have been adjusted to visually distinguish where splits occur, therefore heights are non-informative in the figure. Paper I markers are marked near leaf labels. . . . . 86

A.1 Scatterplot matrices of interestingness measures for the Belgian accident data. 110
A.2 Scatterplot matrices of interestingness measures for the Reuters data. . . . . 111
A.3 Scatterplot matrices of interestingness measures for random transactions. . 112
A.4 Tau-b by deciles for each data set. . . . . . . . . . . . . . . . . . . . 113

B.1 The original hierarchy produced using the $hclust$ algorithm leading to the trees in Figure 5.1 and scores in Table 5.2. . . . . . . . . . . . . . . . . . . . . . 116
B.2 The original hierarchy produced using the $hclust$ algorithm leading to the tree in Figure 5.2 and scores in Table 5.2. . . . . . . . . . . . . . . . . . . . . . 117
B.3 The original hierarchy produced using the $hclust$ algorithm. Nodes above, below, or both above and below 3.5 were collapsed to create the trees shown in Figure 5.4. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 118
Chapter 1

Introduction

Discovering association rules is an effective data mining technique. As an exploration tool, association rules highlight the large quantities of patterns that exist in data. In modern literature, association rules are exemplified (Agrawal et al., 1993) by market basket analysis: trends in purchases. The concept is easy to understand as is its motivation. Identifying all the purchases made during a period of time can highlight interesting patterns like what items are purchased together, what items are only purchased along with others, etc..

Market basket analysis is not the only use of association rules, however. For example, Yairi et al. (2001) detected faults in spacecraft using routine records and Metwally et al. (2005) have employed association rules in fraud detection. One type of association rules called fuzzy association rules have specifically been studied a great deal by Luo and Bridges (2000); Bridges and Vaughn (2000); Tajbakhsh et al. (2009) and others. Text mining has become a particularly fruitful application for association rules and its variants. In one application of generalized association rules, the data itself was the corpus of biomedical
literature (Berardi et al., 2005). In another application, Liu et al. (2012) mined a corpus to construct a hierarchy of terms. As an example, the authors applied their method to insurance queries. Gong et al. (2012) mined text to discover association rules between known autism-related genes, called seeds, and other genes, called candidates, to assess autism risks. Clearly, association rules have extended far beyond the market basket application.

1.1 Significance

This thesis provides three major contributions to association rule literature. The first major contribution is to standardized interestingness rules. We consider the affect of standardizing interestingness measures in a variety of contexts to see what results are true in general and which are specific to certain interestingness measures. We will show that several of the standardized interestingness measures are more correlated to each other than to the original interestingness measures whence they were derived. This intrinsic commonality among standardized interestingness measures is remarkable considering their raw counterparts are so different from each other. The next major contribution addresses incomplete transactions twofold. We first suggest testing whether the data are missing independently from the data as a fast, but necessary, analysis when data are missing. We then provide a method of mining association rules despite the absence of data through random variables. The final major contribution is a summarization technique among association rules. Because mining methods can output so many rules, researchers often prefer to prune rules to the most interesting, relevant, or informative ones. Instead of discarding many rules after mining, we propose and effectively demonstrate a method that uses the abundant rules to create a hierarchy relating items to one another beyond the relationships that traditional association
rules can describe. We show that this method uncovers underlying structures and patterns that association rules alone cannot. All of the contributions presented in this thesis are compatible with many existing association rule methods and work with other approaches and techniques. Using our proposed methods not only improve data analysis but, in some cases, facilitate other existing methods.

1.2 Thesis Outline

This thesis is divided into several main chapters. A comprehensive literature review and the three main areas of contribution are preceded by this introduction and followed by a conclusion and list of references. There are also relevant appendices that are not necessary to follow the material or the impact of the new work presented here, but contain further details for a more comprehensive understanding. Like our contributions, the information in these appendices are original contributions.

Chapter 2 is dedicated to providing the background material of this thesis. The chapter provides precise definitions of concepts including association rules, quantities such as interestingness measures, and other important methods such as mining algorithms. It summarizes many existing contributions in the literature as well as outlines much of the framework for our new approaches.

In Chapter 3, we discuss the usefulness of standardized interestingness measures. Drawing from several interestingness measures discussed in Section 2.3 and a previous standardization technique, we propose new standardized interestingness measures. In Sections 3.2, 3.3, and 3.4 we analyze three data sets with our new standardized interestingness measures. In
the two real data sets, we discover that many rules have interestingness measure values indicating near-independence, but by standardizing we realize they are actually as far away from independence as the rule allows. Because of this, we observe that the ranking of rules varies significantly when using the standardized interestingness measure rather than the original raw interestingness measure. Supplemental material for this work can be found in Appendix A.

We then discuss the realistic association rule problem of incomplete transactions in Chapter 4. Section 4.1 describes missing information in a statistical paradigm. We discuss previous attempts at handling missing data both in a statistical context and an association rule context. We propose employing an existing statistical method to incomplete transactions, a new concept in association rules, and analyze the results through two data sets in Sections 4.4 and 4.3. These examples illustrate how our suggested method of discovering a relationship between missing and observed data is more accurate than other methods. For the remainder of the chapter we consider how to analyze data in light of incomplete transactions. We create a statistical model in Section 4.5 amidst the missing information. Through three simulations in Section 4.5.1, we show how our model works very well at finding traditional quantities of association rules up to an approximate distribution. In particular, we show that this model provides far better impressions of the true underlying values of interestingness measures than an estimate through the strictly non-missing data.

The third major contribution of the thesis is presented in Chapter 5, which addresses the plethora of rules typically encountered with association rules. We define the concept of balance in Section 5.2.1 and use it in a new approach inspired by previous statistical techniques in Section 5.2.2. We then illustrate how well these new methods perform in two types of examples. The first type of example is shown in Section 5.3.1, where the underlying structures
are known. We show how our method will identify the underlying structure using association rules. In the second type of example, we take a real data set where no underlying structure is known and analyze it in Section 5.3.2. This chapter also contains additional material, which can be found in Appendix B.

The thesis concludes in Chapter 6 where we summarize our contributions and discuss what work remains open, followed by a comprehensive list of references and the appendices: Appendix A for Standardizations (Chapter 3) and Appendix B for Hierarchies (Chapter 5).
Chapter 2

Literature Review

2.1 Association Rules

We consider association rules as they were proposed by Agrawal et al. (1993). The notion is defined in the context of the itemset, $I$, which is comprised of individual items. A transaction is a binary vector indicating the presence or absence each item in a particular instance. An association rule is a statement of the form $A \Rightarrow B$, where $A \subseteq I, B \subseteq I$ such that $A \cap B = \emptyset$; $A$ is called the antecedent and $B$ is called the consequent.

One of the most fundamental quantities discussed in the context of association rules is support. The support of $X \subseteq I$, denoted $P(X)$, is the proportion of transactions that contain every item in $X$. Since the number of transactions in a database is discrete and finite, so too is the range of any support. That is, in a database with $m$ transactions, the
CHAPTER 2. LITERATURE REVIEW

support of set of items, say \( X \subseteq \mathcal{I} \), is

\[
P(X) \in \left\{ 0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m-1}{m}, 1 \right\}.
\]

In the context of the rule \( A \Rightarrow B \), \( P(A) \) denotes the support of the antecedent, \( P(B) \) denotes the support of the consequent, and \( P(A, B) \) denotes the support of the rule. Another quantity often used when discussing association rules is confidence. It is the conditional probability \( P(B|A) \) and is the proportion of transactions that contain every item in \( A \) that also contains every item in \( B \). This is formally defined in Equation (2.1).

\[
P(B|A) = \frac{P(A, B)}{P(A)} \tag{2.1}
\]

A scenario which arises when discussing these quantities is independence. In \( A \Rightarrow B \), the antecedent and the consequent are independent when \( P(A, B) = P(A)P(B) \), and dependent otherwise. A rule that is independent is uninteresting since \( P(A, B) \) provides no more information than \( P(A) \) and \( P(B) \). In this light, there is no special relationship between between the antecedent and consequent. Because independence is such a useful case to consider, some interestingness measures quantify a departure from independence. Examples of this include lift (Brin et al., 1997, discussed further in Section 2.3.1), and leverage (Piatetsky-Shapiro, 1991).

There are instances when items not occurring in transactions are useful. \( P(\overline{X}) \) denotes the proportion of transactions that do not contain every item in \( X \), for \( X \subset I \). In the context of the rule \( A \Rightarrow B \), for example, one may be interested in the proportion of transactions
CHAPTER 2. LITERATURE REVIEW

that contain every item in the antecedent but do not contain any item in the consequent. This would be denoted $P(A, \overline{B})$. Note that $P(A, B) + P(\overline{A}, B) + P(A, \overline{B}) + P(\overline{A}, \overline{B}) = 1$.

Rules where part of the focus is the absence of some items are called negative association rules. Though we do not explicitly consider such rules, our calculations are more concise and understandable using the notation for negative association rules.

2.2 Apriori Algorithm

For every data set, rules are mined using the Apriori algorithm (Agrawal and Srikant, 1994). This method uses the Apriori principle, also known as downward closure: if a set of items is deemed to be frequent relative to a threshold, then every subset of this set will also be (at least as) frequent relative to the same threshold. Conversely, if a set of items is judged to be infrequent relative to a threshold, then every superset of this set will also be infrequent relative to the threshold. There are many alternatives to this algorithm such as SETM (Houtsma and Swami, 1995); EClat (Zaki, 2000), which divides the itemset into equivalence classes; frequency pattern trees (Han et al., 2000), which avoid many passes over the data; and variations to the Apriori algorithm itself such as AprioriTID, Apriori-Hybrid (Agrawal and Srikant, 1994), and Direct Hashing and Pruning (Park et al., 1995).

Our methodology focuses on the analysis and affect of standardization on rules which have already been mined. Therefore, as a means of obtaining rules whose interestingness measures we will standardize, the established and well-understood Apriori algorithm will be employed. The interestingness measures for each data set are shown as a pair of scatterplot matrices in Appendix A.2; the first scatterplot matrix shows the raw interestingness measures and the second the standardized interestingness measures.
CHAPTER 2. LITERATURE REVIEW

2.3 Interestingness Measures

Many different interestingness measures have been introduced to analyze association rules as some interestingness measures are more suitable for some purposes than others. Previous work (Piatetsky-Shapiro, 1991) outlined three important properties every interestingness measure should have. The first property, (Prop 1) is that the interestingness measure should achieve the value of 0 at independence. That is, when $P(A, B) = P(A)P(B)$. Another constant known in advance, such as one in the case of lift, satisfies the same purpose. The second desirable property (Prop 2) is that the value of the interestingness measure should be larger when $P(A, B)$ is larger while $P(A)$ and $P(B)$ remain the same. The third property (Prop 3) is related to the second: the interestingness measure should be smaller when either $P(A)$ or $P(B)$ increases but $P(A, B)$ is unchanged; or when $P(B)$ increases but $P(A)$ and $P(A, B)$ are unchanged.

In addition to these three properties, five other properties distinguishing interestingness measures have been discussed (Tan et al., 2002), and considered in (Geng and Hamilton, 2006). These properties are not necessarily desirable or undesirable but are important features which relate or contrast interestingness measures from one another. One property is symmetry (Prop 4): the interestingness measure remains the same when the antecedent and consequent are interchanged. Another property is antisymmetry under negation (Prop 5). That is, the value of the interestingness measure takes on the negative of the original value when the antecedent is replaced with the negation of the antecedent in the rule, or the consequent is replaced with the negation of the consequent in the rule. An additional property (Prop 6) relates to the previous: the interestingness measure achieves the same value when the antecedent is replaced with the negation of the antecedent and the consequent is replaced...
with the negation of the consequent. Prop 5 trivially implies Prop 6 but some interestingness measures, such as the odds ratio and the Gini index, satisfy Prop 6 but not Prop 5. Another suggested property (Prop 7) is that the interestingness measure should achieve the same value under row-wise or column-wise scaling (or both). The final property is Null invariance (Prop 8): additional transactions where neither the antecedent nor the consequent appear do not change the value of the interestingness measure.

A summary of 21 interestingness measures can be found in Tan et al. (2002) and discussed in the specific context of association rules in Geng and Hamilton (2006), which outlines methods of choosing suitable interestingness measures. Part of this analysis of interestingness measures includes identifying which interestingness measures satisfied the three properties from Piatetsky-Shapiro (1991) as well as the five additional properties discussed in Tan et al. (2002). The analysis includes a table (Table 6) which shows whether or not each of the 21 interestingness measures satisfy these eight properties. In an attempt to group these interestingness measures, the analysis also contains a figure (Figure 3) identifying levels of similarity, by correlation, between interestingness measures. It is from this figure where all values of support are considered that that we identify at least three separate groupings of interestingness measures. Interestingness measures which appear in the same group often do not satisfy the same subset of the eight properties but are well correlated by having correlation values of least 0.85 with any other interestingness measure in the same group. Because of this similarity we choose one representative from each of the three groups to analyze. Table 2.1, shows a summary of the three interestingness measures in the same fashion as Table 6 from Tan et al. (2002).
### 2.3.1 Lift

The lift of a rule, originally introduced as the *interest* of a rule (Brin et al., 1997), is denoted here as $\mathcal{L}(A \Rightarrow B)$ and is defined as

$$\mathcal{L}(A \Rightarrow B) = \frac{P(A, B)}{P(A)P(B)}.$$ (2.2)

Lift takes on non-negative values and equals one when the antecedent and consequent are statistically independent. An example of when a measure such as lift can be problematic emerges when we consider an instance where the support of the antecedent of a rule, $A \Rightarrow B$, is 0.5 and the overall lift value for this rule is 1.95. Knowing that the support of the antecedent $P(A) = 0.5$, the value of lift for this rule cannot exceed 2. Suppose, also, that for another rule, $C \Rightarrow D$, the lift of the rule is 1.95 but the support of the antecedent, $P(C)$, is 0.1. In this second rule, the lift value was not bounded above by 2 but by 10. Even though these two rules have the same lift value of 1.95, the interpretation of the two lift values should not be the same because one rule, $A \Rightarrow B$, cannot achieve a value above 2 while the other rule, $C \Rightarrow D$, can achieve a value as high as 10. Such discrepancies occur when using ‘raw’ interestingness measures; standardizing lift addresses this issue McNicholas et al. (2008). In this example, standardized lift values for rules $A \Rightarrow B$ and $C \Rightarrow D$ are 0.975 and 0.195, respectively, if all other factors are equal.
In another example to show just how bounded lift is, consider the values of lift where \( P(A) = P(B) \). The bounds of lift are shown in Figure 2.1. We can see that for larger supports, the value of lift is bounded near one. Without any other consideration, the small differences in the values of lift between rules may be overlooked but considering the narrow range of possible values these differences may be more significant. Though we only highlight this issue in the case when \( P(A) = P(B) \), such problems exist in many cases.

The standardized lift of a rule, \( \mathcal{L}^*(A \Rightarrow B) \), is obtained by finding the suitable upper bound, \( \upsilon \), and lower bound, \( \lambda \), in the following definition:

\[
\mathcal{L}^*(A \Rightarrow B) = \frac{\mathcal{L}(A \Rightarrow B) - \lambda}{\upsilon - \lambda}.
\]

The values of \( \upsilon \) and \( \lambda \) are taken from McNicholas et al. (2008) and shown below. These values are functions of the minimum support and confidence thresholds, denoted in this thesis by \( \sigma \) and \( \kappa \), respectively. These thresholds are assumed to be at least \( 1/n \), where \( n \) is the number of transactions in the data. We believe that this is a reasonable assumption as it corresponds to only considering rules supported by at least one transaction. This method of standardizing an interestingness measure is used for the next three interestingness measures.

\[
\upsilon = \frac{1}{\max\{P(A), P(B)\}}
\]

\[
\lambda = \max \left\{ \frac{P(A) + P(B) - 1}{P(A)P(B)}, \frac{4\sigma}{(1+\sigma)^2}, \frac{\sigma}{P(A)P(B)}, \frac{\kappa}{P(B)} \right\}
\]

One remark about standardized interestingness measures is that the form for a particular
Figure 2.1: The bounds on lift when $P(A) = P(B)$. 
CHAPTER 2. LITERATURE REVIEW

case of interestingness measure may actually correspond to some probability. For example, in the rule $A \Rightarrow B$ if $\lambda = \frac{P(A)+P(B)-1}{P(A)P(B)}$ and $\upsilon = \frac{1}{P(A)}$, then $\mathcal{L}^* = P(\overline{B} \mid \overline{A})$. Another case leads to another conditional probability. However, this will not be true in general. That is, for any interestingness measure, the standardized interestingness measure for a particular rule may not correspond to any meaningful probability.

2.3.2 Cosine Similarity

The definition of the cosine similarity resembles the definition of the lift of a rule and is defined as

$$C(A \Rightarrow B) = \frac{P(A, B)}{\sqrt{P(A)P(B)}}. \tag{2.3}$$

The cosine similarity lies on the closed unit interval, and no constant value is achieved when the antecedent and consequent are statistically independent of one another. When the rule is independent, the cosine similarity attains the value $\sqrt{P(A)P(B)}$.

2.3.3 Yule’s Q

Some interestingness measures take supports of negations into account. One such interestingness measure is Yule’s Q (Yule, 1903), which is based on the odds ratio. Unlike the odds ratio, however, Yule’s Q remains defined in all cases where $P(A) > 0$ and $P(B) > 0$. The range of Yule’s Q is between -1 and 1, inclusively, and is defined as

$$Q(A \Rightarrow B) = \frac{P(A, B)P(\overline{A}, \overline{B}) - P(\overline{A}, B)P(A, \overline{B})}{P(A, B)P(\overline{A}, \overline{B}) + P(\overline{A}, B)P(A, \overline{B})}. \tag{2.4}$$
CHAPTER 2. LITERATURE REVIEW

This interestingness measure obtains a value of zero when the antecedent and consequent are statistically independent. Positive values indicate that the antecedent and consequent are more likely to occur together than if they were independent and negative values indicate that the antecedent and consequent are less likely to occur together than if they were independent. Like lift and the cosine similarity, Yule’s Q is symmetric.

2.3.4 Gini Index

The Gini index is the only measure discussed here that is not symmetric. That is, the Gini index for the rule \( A \Rightarrow B \) may not achieve the same value as the Gini index for the rule \( B \Rightarrow A \). Used in a variety of contexts, the Gini index quantifies impurity and is defined as

\[
G(A \Rightarrow B) = P(A)[P(B \mid A)^2 + P(\overline{B} \mid A)^2] + P(\overline{A})[P(B \mid \overline{A})^2 + P(\overline{B} \mid \overline{A})^2] - P(B)^2 - P(\overline{B})^2.
\]  

(2.5)

This interestingness measure achieves a value of zero when the antecedent and consequent are statistically independent, and achieves a value greater than zero when the antecedent and consequent are dependent, up to a value of 0.5.

2.4 Order Comparisons

A relative measure of rules is through ranking by interestingness measure. Because we will be standardizing interestingness measures in the next chapter, we will analyze how the relative order of rules change before and after an interestingness measure is standardized.
CHAPTER 2. LITERATURE REVIEW

One method of comparing rule orderings is by employing Goodman and Kruskal’s gamma (Goodman and Kruskal, 1954). In this context, it is calculated by determining the difference in the proportions of concordant pairs and discordant pairs while ignoring ties. Kendall’s tau-b is a similar method of measuring the rankings that considers ties to be discordant, making it a more conservative (Agresti, 2002; Blaikie, 2003) description of relative ordering. Because of this, we will use Kendall’s tau-b in the next chapter. If the order of rules by raw interestingness measure is the same as the order of rules by the corresponding standardized interestingness measure, a value of 1 results. If the order of one is the opposite of the other, a value of -1 results. A value of zero results when the two rule orderings have no apparent relationship, such as the case when one of the rule orders have been (uniformly) randomly ordered. In addition to calculating the tau-b value for all rules in each data set, we calculated tau-b by decile. These are shown in plots found in Figure A.4 of Appendix A.3.

2.5 Missing Data

Missing data is a term used to describe a variety of concepts includin latent variables, censoring, or unrecorded data. In this thesis, we will only use the term to refer to data that have not been recorded but is recordable. We will also use the terms incomplete or absent to refer to this same concept.

An existing method to avoid the issues of incomplete transactions with association rules is by partitioning the data set based on which of the data are available (Ragel and Crémilleux, 1998). This leads to different subsets of data being used at different parts of the mining operation, depending on which items are present in each transaction. Though this method
uses the database more fully, it still does not address how missing transactions relate to the
database. This partitioning method is not to be confused with another horizontal partitioning
method (Kantarcioglu and Clifton, 2004) which addresses privacy preservation.

When faced with incomplete data, some investigators simply neglect missing data and focus
strictly on the observed data. This can result in loss of information and in severe cases,
introduce a bias that misleads researchers and affects their conclusions. In a more sophisti-
cated manner of handling incomplete data, missingness can be formalized (Rubin, 1976) for
inferential purposes by considering the manner in which data are missing. The possible dis-
tinction of the underlying statistical parameters in missing data and observed data becomes
a very important problem as the data may or may not be “ignorable". When values are
missing completely at random (MCAR), inference may be drawn directly as the underlying
statistical parameters are the same in both cases. This is one instance when the missingness
itself is deemed “ignorable". The other situation when the missing data are “ignorable” oc-
curs when data are missing but their absence is related to the observed data. Because of this
relationship to observed data, modifications can be made to account for this relationship and
the missingness can again be ignored. These missing data are missing at random (MAR)
and can be handled reasonably well. A scenario which is quite difficult is when missingness
is not ignorable, happens when when data are missing but not at random (MNAR). Data
are missing and cannot be entirely related to the observed data, but must also be related
to unobserved data. This means the missingness cannot be ignored, and often means the
missingness must be included in the model. Without knowing the value of the missing data,
it is difficult to tell whether the missingness is related to the missing data. This challenge
makes the case where data are MNAR difficult to identify, let alone account for.

A test (Little, 1988) has been introduced to identify whether ignorable missing data are
MCAR or MAR. The corresponding test statistic is shown to asymptotically follow an $F$ distribution under certain assumptions. A drawback to this method is that the properties are asymptotic and for continuous data, which is not always appropriate. To tackle discrete cases, the work by Takai and Kano (2008) proposed a test for independence between two binary responses where some of the data are missing. The method uses the expectation-maximization algorithm (Dempster et al., 1977) to estimate the MLEs under nonlinear constraints, which is a challenging problem. Overcoming this challenge comes at a computational cost and still faces the issues that come when asymptotic properties are less applicable.

2.6 Distributions to Test Independence

The test for independence among small data was (a suggestion by Fisher published in Yates, 1934) an improvement to the $\chi^2$ test, which only worked well for tables with large enough counts that the counts themselves could be reasonably considered to be normally distributed. For these large counts, Fisher’s exact test provides similar results to $\chi^2$ test, but the $\chi^2$ is computationally less expensive, giving it an advantage in such cases. As we will show later, the $\chi^2$ test does reasonably well for very small $p$-values under some situations but not others. Two issues arise from a case with smaller counts; one practical and one theoretical. The theoretical issue is that supports range over a countable, and hence discrete, domain since supports are rational. Fisher’s exact test is suitable for discrete domains and the difference from continuous domains is even more pronounced for smaller counts.
CHAPTER 2. LITERATURE REVIEW

2.6.1 Hypergeometric Distribution

The hypergeometric distribution is a discrete mass function which provides the probability how the size of subsets from each of two fixed partitions sum together. More partitions may be defined as easily but it is sufficient to consider the two-partition case. Let there two sets of elements with cardinality $N$, $K < N$ of which possess a common label. In a simple random sample of size $n$ from the set, the number of those elements which possess the label is denoted by the random variable $H$. The probability that $H = k$ is given by

$$P(H = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where $\binom{n}{k}$ denotes the binomial coefficient for the number of ways to choose $k$ elements from a set of $n$.

2.6.2 $\chi^2$ distribution

Given $k$ independent standard normal random variables, $X_i, i = 1, 2, \ldots, k$, The $\chi^2$ random variable is constructed by taking the sum of the squares of the normal random variables. The $\chi^2$ has $k$ degrees of freedom, denoted $\chi^2_k$ as $k$ independent random variables were used to construct it. The $\chi^2$ distribution has been used a great deal to test for a deviation from some pattern. In the case of contingency tables, transformations of functions of the observed and expected counts (under some hypothesis) can be taken to be approximately standard normal, allowing for the a test statistic following the $\chi^2$ distribution.

The function of observed and expected counts is taken from a normal approximation to
the binomial approximation to the hypergeometric distribution. For observed and expected counts, denoted $Y_{\text{obs}}$ and $Y_{\text{exp}}$, respectively,

$$\frac{Y_{\text{obs}} - Y_{\text{exp}}}{\sqrt{Y_{\text{exp}}}} \overset{\text{approx}}{\sim} N(0, 1).$$

This approximation requires that the probability or proportion parameter of the binomial be moderate (close to 0.5) to reduce the effect of skew and for the the parameter representing the number of trials to be large so that the disparity between the continuous normal distribution and discrete binomial is relatively small. In such a case, the normal approximation to the binomial distribution is given by

$$\text{Bin}(n, p) \overset{\text{approx}}{\sim} N(np, np(1-p)).$$

The $z$-score for the number of successes of the binomial is given by

$$\frac{Y_{\text{obs}} - np}{\sqrt{np(1-p)}}$$

and the $z$-score for the number of failures of the binomial is given by

$$\frac{(n - Y_{\text{obs}}) - n(1-p)}{\sqrt{np(1-p)}}$$

making

$$\chi^2 \text{ test-statistic} = \left( \frac{Y_{\text{obs}} - np}{\sqrt{np(1-p)}} \right)^2 = \left( \frac{(n - Y_{\text{obs}}) - n(1-p)}{\sqrt{np(1-p)}} \right)^2 = \frac{(Y_{\text{obs}} - np)^2}{np(1-p)}.$$

There is one degree of freedom as it arises from the $2 \times 2$ contingency table with fixed
marginals and is the sum of only one squared standard normal random variable, making the distribution $\chi^2_1$. For conciseness, we will denote this simply as $\chi^2$ as we will not discuss a $\chi^2$ distribution with any other degrees of freedom. Therefore, the $\chi^2$ test statistic relies on the normal distribution approximating features of the binomial distribution, which is in place of the hypergeometric distribution.

### 2.7 Generalized Association Rules

When the size of the itemset is very large, the resulting rules are masses of many patterns which may be difficult to understand. Several attempts have been made to reduce the number of rules, observing that many of the rules are remarkably similar. One attempt has been to find the rule which Bayardo et al. (2000) describes as providing the greatest improvement over other similar rules. Another attempt to prune from many similar rules is to find a representative subset of rules to present using a set cover (see Toivonen et al., 1995; Kryszkiewicz, 1998, for examples). These approaches are focused on reducing the number of rules, so they do not address why so many of rules of a certain type were found to begin with.

One concept of higher-level rules was considered to construct generalized association rules (Srikant and Agrawal, 1997). What differs about generalized association rules from traditional association rules is the form that a rule can take. Traditional association rules only consider the intersection of transactions while generalized association rules also allow for unions. Consider the items \{disposable forks, disposable spoons, disposable knives, paper plates, plastic cups\}. Supports, and thus many interestingness measures and association rule
learning methods, are limited to conjunctions of the transactions containing these items. The rule shown in (2.6) is, in part, focused on the transactions where both plates and cups appear in transactions, not just the transactions with plates just the transactions with cups.

\[
\{\text{disposable forks, spoons, knives}\} \Rightarrow \{\text{paper plates, plastic cups}\} \quad (2.6)
\]

A generalized association rule will employ a known taxonomy to find the more general rule,

\[
\{\text{some disposable cutlery}\} \Rightarrow \{\text{some disposable dishes}\}, \quad (2.7)
\]

which will consider transactions where at least one item of disposable cutlery is present and at least one item of a disposable dish is present. Here, we carefully use the word some to be equivalent to the phrase at least one item of, purposely excluding the possibility of no such items. The other advantage to this greater expressiveness is that mining methods which typically prune sets of items with low supports may contribute to an association rule through the taxonomy. In this instance the rule \{plastic forks\} \Rightarrow \{paper plates\} would need to be considered interesting in the resultant rules before the rule in Equation (2.6) would be considered interesting, even though Equation (2.6) is closer to the arguably more important and meaningful generalized association rule of \{some disposable cutlery\} \Rightarrow \{some disposable dishes\}. Even if the individual rules such as rule (2.7) did seem interesting, in every combination, they can and arguably should be summarized by this higher-level rule, reducing the number of rules generated while improving the understanding of patterns in the database.

The use of a taxonomy in association rule mining has also for allowed a more efficient method of mining negative association rules. That is, rules that include the explicit absence of items.
An example why this is an important consideration is items which are inherently binary, but the numerical coding is arbitrary with some variables such as gender. It would be useful to consider rules with both males and with females, but only the gender coded as 1 would appear in rules under the traditional paradigm. Negative association rules are known to be more difficult to find because of the prohibitively greater computational time that comes from searching the larger space of rules, most of which are uninteresting. Savasere et al. (1998) proposed employing a taxonomy to reduce the search space and focus on interesting negative association rules. More recently, the GART algorithm (Domingues and Rezende, 2011) provides a useful means of explicitly generalizes existing association rules using a taxonomy. To achieve a similar goal but with less user input by using existing domain knowledge through an ontology and what are called objective interesting measures, Mansingh et al. (2011) proposed a ‘hybrid’ algorithm to guide association rule mining. Song et al. (2007) exploited existing ontologies and mined association rules using the apriori algorithm and applied neuro-linguistic programming to construct, SemanQE, a hybrid method of semantic query expansion. A common theme to all these useful methods is that they require a hierarchy beforehand.

Undoubtedly, the use of a taxonomy is beneficial to mining association rules. This becomes a problem, however, when no taxonomy is known or a meaningful one cannot be found. In a novel method of constructing a taxonomy, albeit for other purposes, Liu et al. (2012) proposed a greedy algorithm to mine a corpus. The algorithm employs the Bayesian Rose Tree (BRT) of Blundell et al. (2010), a generalization of Bayesian Hierarchical Clustering (BHC) (Heller and Ghahramani, 2005). The BHC algorithm is a means of agglomeratively clustering by successively pairwise grouping. The BRT is an extension that merges in a way which introduces the possibility of non-binary trees. These algorithms are discussed in the
next section.

2.8 Bayesian Rose Tree

The BHC algorithm uses trees to represent data. A tree is comprised of nodes with a specific structure. A node is either a representation of a single observation making the node a leaf node, or it is the parent of other nodes, thereby representing a collection of observations via recursion. The method of merging is what introduces potentially non-binary merges. Two nodes merge in one of three ways. One method to merge is by introducing a new node to become the parent of the two nodes. Using only this method results in the special case of the algorithm that is the BHC algorithm. Another method is to make one of the nodes the parent of the other. The third possible method of merging is for a node to adopt all the children of the node and then delete the barren node. The pair to merge and merging method are selected at every iteration by considering all possible pairs and merges and choosing based on the highest resulting likelihood ratio.

The likelihood is chosen in using a density, say $f$, a function of a node as a collection of observations. The likelihood ratio of two nodes being considered to merge, say $A$ and $B$, into a resulting merged node, say $C$, is given in (2.8).

$$\frac{f(C)}{f(A)f(B)}$$

The algorithm is a greedy search, and so only constructs one tree and has no sense of the quality of the hierarchy relative to other hierarchies. In part, because this is not a searching algorithm, the resulting tree is unlikely a tree which maximizes the likelihood, which would
be ideal. The resulting tree, however, performs reasonably well in practice.

This method defines the density of a tree as a mixture model (Wolfe, 1963). More specifically, it is a mixture of partitions of the observations. The density $f$ as a mixture model of the tree $T$ is given in Equation (2.9).

$$f(T) = \sum_{t \in T} \phi_t f(t)$$  \hspace{1cm} (2.9)

Here, $\phi_t$ denote mixing proportions such that $\sum_{t \in T} \phi_t = 1$, and $t \in T$ represents a subtree of $T$. Generally this form is intractable and an equivalent recursive method of finding the likelihood is given in Equation (2.10).

$$f(T) = \pi_T f(T) + (1 - \pi_T) \prod_{t \in \text{child}(T)} \pi_t f(t)$$  \hspace{1cm} (2.10)

The mixing proportion $\pi_T$ is the probability that a collection of observations under node $T$ are immediately leaf children rather than partitioned further. The term $\text{child}(T)$ denotes the immediate children, possibly parents themselves to other nodes, of the tree $T$. Letting $|\text{child}(T)|$ denote the number of children of the tree $T$, the choice $\pi_t = \delta^{|\text{child}(t)|-1}$ for any (but constant throughout the tree) choice $\delta \in (0,1)$ ensures that $f$ is indeed a density as it can be shown that mixing proportions will sum to 1. The choice of $\delta$ can be found to maximize $f(T)$ through an optimization algorithm. Our choice of $\delta$ is equivalent to $1 - \gamma$, where $\gamma$ is defined in Blundell et al. (2010). Larger values of $\delta$ correspond to optimal trees with more partitions while smaller values of $\delta$ correspond to optimal trees with fewer partitions.

The BRT is considered an improvement over the traditional binary tree of BHC for a number of reasons. First, we define the concept of a binary cascade. We will refer to a binary cascade...
as a binary (sub)tree where at least one child of every parent is a leaf. A binary cascade of $n$ leaves necessarily has $n - 1$ levels. Every non-leaf node is the parent of one leaf, and the parent of a binary cascade, except the last parent which has two leaves. We will contrast this to a two-level tree. A two-level tree with $n$ leaves has two levels: a single node which is the parent of $n$ leaf nodes. An example of a binary cascade is shown in Figure 2.2a while an example of a two-level tree is shown in Figure 2.2b.

![Figure 2.2: A comparison of a binary and non-binary tree.](image)

The BRT is a generalization of a binary tree so all possible binary trees exist as a subset of all possible BRTs. Another reason that we prefer non-binary trees over binary trees is that the set of non-binary trees is more parsimonious as it has less internal nodes on the same number of leaves. We will consider that if two sibling nodes are switched with each other, the tree before and the tree after the switch are considered the same (or in the same equivalence class if more rigorously defined). If any two nodes which do not share the same parent are switched, the resulting tree is not the same (nor in the same equivalence class). Therefore, when a binary cascade occurs on $n$-leaves, a permutation of any pair of nodes leaves (except for the final two sibling leaves) results in a different tree. The $n$ leaves can be permuted to produce $(n - 1)!$ distinct binary cascades but there is only one subtree structure where all $n$ leaves are siblings. In fact, binary cascades are not the only way a binary tree
can form and on \( n \) leaves, there would be up to

\[
\frac{(2n - 2)!}{n!(n-1)!}
\]  \hspace{1cm} (2.11)

possible distinct binary trees, each with at least \( \lceil \log_2(n) \rceil \) levels. This is equal to the \( n^{th} \) Catalan number\(^1\). Having to choose among all these binary trees would be incredibly difficult. Worse yet, many of these trees may score very similar to each other as so many subtrees would be in common, especially if they are all different binary-tree approximations to the same two-level tree. These, among other concerns motivating the BRT are why we prefer some non-binary structure over binary cascades. See Blundell et al. (2010) for further details.

\[^1\text{The sequence now known as Catalan numbers was discussed for combinatorial purposes in letters between Euler and Goldbach in 1751 (Euler, 1843), and published in two articles in Segner (1759) and Euler (1759). The sequence was derived independently to estimate ratios by Minggatu, most likely decades earlier. Minggatu’s book was finished after his death by his student, Chen Jixen, in 1774 and published in Jixin (1839).}\]
Chapter 3

Standardized Interestingness Measures

3.1 Using Interestingness Measures

This chapter is dedicated to the derivation, analysis, and impact of standardized interestingness measures. Given the analysis of several interestingness measures (see Piatetsky-Shapiro, 1991; Freitas, 1999; Tan et al., 2002; Geng and Hamilton, 2006, for discussions), we standardize the interestingness measures discussed in Section 2.3. Specifically, using many ideas of standardized lift in Section 2.3.1, we provide the standardized forms for the cosine similarity, Yule’s Q, and the Gini index (these three interestingness measures are introduced in Sections 2.3.2, 2.3.3, and 2.3.4, respectively). These standardized forms are provided in the next section, Section 3.1.1.

The remainder of the chapter is devoted to the analysis of three data sets. In Section 3.2, we analyze how association rules from the Belgian Accident data are impacted with standardized
CHAPTER 3. STANDARDIZED INTERESTINGNESS MEASURES

interestingness measures. We then consider a similar analysis on the famous Reuters-21578 data set in Section 3.3. Following this, we consider a set of random transactions in Section 3.4 to understand how standardized interestingness measures fare against rules where no true patterns exist. In all three analyses, we use Kendall’s tau-b, discussed in Section 2.4, to compare how the relative ranking of rules have changed. Finally, the work presented in the chapter is summarized in Section 3.5 with an in-depth discussion in Section 6.1. Supplementary material can be found in Appendix A.

3.1.1 New Standardizations

Using the same standardization technique as lift, we standardize the cosine similarity, Yule’s Q, and the Gini index. The standardized cosine similarity of a rule, \( C^*(A \Rightarrow B) \), is defined by

\[
C^*(A \Rightarrow B) = \frac{C(A \Rightarrow B) - \lambda}{\nu - \lambda},
\]

where

\[
\nu = \min \left\{ \sqrt{\frac{P(A)}{P(B)}}, \sqrt{\frac{P(B)}{P(A)}} \right\}
\]

and

\[
\lambda = \max \left\{ \frac{2\sigma}{1 + \sigma}, \frac{\sigma}{\sqrt{P(A)P(B)}}, \frac{P(A) + P(B) - 1}{\sqrt{P(A)P(B)}}, \sqrt{\kappa \frac{\sigma}{P(B)}}, \kappa \sqrt{\frac{P(A)}{P(B)}} \right\}.
\]

Derivations for \( \nu \) and \( \lambda \) are found in Appendix A.1.1.
CHAPTER 3. STANDARDIZED INTERESTINGNESS MEASURES

The standardized Yule’s Q of a rule, $Q^*(A \Rightarrow B)$, is defined by

$$Q^*(A \Rightarrow B) = \frac{Q(A \Rightarrow B) - \lambda}{\upsilon - \lambda},$$

where $\upsilon = 1$ and

$$\lambda = \max \left\{ -1, \frac{\sigma - P(A)P(B)}{\sigma + P(A)P(B) - 2\sigma(P(A) + P(B) - \sigma)}, \frac{\kappa - P(B)}{\kappa + P(B) - 2\kappa(P(A) + P(B) - \kappa P(A))} \right\}.$$

Derivations for $\upsilon$ and $\lambda$ are found in Appendix A.1.2.

The standardized Gini index of a rule, $G^*(A \Rightarrow B)$, is defined by

$$G^*(A \Rightarrow B) = \frac{G(A \Rightarrow B) - \lambda}{\upsilon - \lambda},$$

where

$$\upsilon = \begin{cases} 
2\frac{(\min\{P(A),P(B)\} - P(A)P(B))^2}{P(A)(1-P(A))} & \text{if } P(A, B) \geq P(A)P(B) \\
2\frac{(\max\{\sigma,\kappa P(A),P(A) + P(B) - 1\} - P(A)P(B))^2}{P(A)(1-P(A))} & \text{if } P(A, B) < P(A)P(B)
\end{cases}$$

and

30
CHAPTER 3. STANDARDIZED INTERESTINGNESS MEASURES

\[ \lambda = \begin{cases} 
2 \left( \frac{\max\{\sigma, \kappa P(A), P(A)+P(B)-1, P(A)P(B)-P(A)P(B)\}}{P(A)(1-P(A))} \right)^2 & \text{if } P(A, B) \geq P(A)P(B) \\
0 & \text{if } P(A, B) < P(A)P(B) 
\end{cases} \]

Derivations for \( \upsilon \) and \( \lambda \) are found in Appendix A.1.3.

3.1.2 Experimental Results and Evaluation

We explore the effects of standardizing interestingness measures with three different data sets. In the first data set, the values obtained for lift, cosine similarity, Yule’s Q, and Gini index, is clearly reordered when standardized, making previously unremarkable rules more noteworthy by reporting the value of an interestingness measure relative to its attainable range. The second data set that we analyze is comparably more sparse. We observe that standardizing interestingness measures on these data has some of the same effects as the first data set but also shows other effects. We consider a third data set that is completely simulated random data, with independent items. In particular, these data highlight how standardized interestingness measures, taking into account the low probabilities of co-occurrences, do not suffer from suggesting that rules in random transactions are interesting.

There are two important remarks regarding support and confidence thresholds in the following examples. The first is that we employ the same support and confidence thresholds throughout each example. It would be inappropriate to compare one rule’s standardized interestingness measure against another rule’s standardized interestingness measure when different thresholds were used. Using different thresholds would mean that interestingness measures have been standardized by different standardizing functions, which cannot be so
CHAPTER 3. STANDARDIZED INTERESTINGNESS MEASURES

easily compared. The second remark is that the thresholds chosen in our analyses may not be typical for specific applications as our goal is to explore the effects of standardizing interestingness measures on a wide variety of values. Specific applications often use higher support thresholds and impose higher confidence thresholds. Such thresholds typically result in fewer rules, most often those indicating a positive correlation between the antecedent and consequent. To appropriately explore the effect of standardizations, we intentionally keep these rules in our analysis when they result.

3.2 Traffic Accidents

Information from over 340 000 traffic accidents from Belgium’s National Institute of Statistics Geurts et al. (2003) is analyzed by considering up to 572 factors that may have been involved. These factors include the type of road, intersection, features of the road, conditions, time, type of vehicle, direction, severity of accident, age of road user, obstacles, and other relevant criteria recorded by a police officer. As Table 3.1 will indicate, information identifying the actual factors have not been made available. Rather, factors are simply enumerated.

We set minimum support and confidence thresholds to 0.3 and 0.4, respectively, for these data. These thresholds were chosen because when considering entire rules that were no more than five items in length, approximately 90 000 rules were produced; we deemed this adequate for analysis.

To illustrate how just one interestingness measure is standardized, Figure 3.1 shows ten randomly selected rules sorted by $C^*$. Each vertical bar represents the range of values of $C$ that were attainable by the rule based on the bounds discussed in Section 2.3.2. The
Figure 3.1: Cosine similarity values for randomly rules. Vertical lines represent the range of possible cosine similarity values for the rule with the horizontal dash indicating the achieved value of the cosine similarity. The circle indicates the value of the standardized cosine similarity, the relative position of the achieved value in the range of possible values.
horizontal tick within the bar shows where the actual value of $C$ lies. The value of $C^*$ is the relative location of $C$ in the bar shown with circles. These values are overlain because the range of $C$ and $C^*$ are the same but the range of the raw and standardized interestingness measures are not coincident, in general. The rules are sorted by ascending values of $C^*$ and show that standardizing the interestingness measure is not simply a consistent transformation of the raw interestingness measure.

When comparing the standardized values against raw values, a variety of behaviours are observed (see Figure 3.2). Standardizing $L$ results in two very noticeable trends in these rules. The values of $L$ and $L^*$ are generally positively correlated when the value of $L$ is not near one. Rules that have values of $L$ near one obtain many values of $L^*$ in the unit interval with a concentration near one. Even though the cosine similarity is defined in a similar form as lift, no constant value of $C$ indicates independence and so the trend observed in the plot with lift is not identifiable. Generally, high values of $C$ identify positively correlated items and low values of $C$ identify negatively correlated items.

The plot of Yule’s $Q$ reveals several features. There is a high concentration of values with $Q$ near zero indicating the same independent rules indicated $L$ near one. There are also no rules with low values of $Q$ while achieving high values of $Q^*$. This is not particularly surprising because the upperbound of $Q$ is constant. The plot of the Gini index is the similar to the others. The majority of rules have values of $G$ very near zero. No values of $G$ are high that obtain a low value of $G^*$. In contrast to the other standardized interestingness measures, $G^*$ maps rules indicating independence to a single value, namely zero, as $G$ would already be very close to zero.

In Figure 3.2, orange points highlight rules that obtained values of $L^*$ more than 0.9 and
Figure 3.2: Plots of interestingness measures against their standardized counterparts for the Belgian accident data. Dark regions indicate a higher density of rules.
CHAPTER 3. STANDARDIZED INTERESTINGNESS MEASURES

blue points highlight rules that obtained values of $C^*$ greater than 0.95. We choose these values to observe how high values of a standardized interestingness measure correspond to other interestingness measures, including the other standardized interestingness measures. In this figure as well as similar figures in following sections, the raw interestingness measure is on the horizontal axis and the standardized on the vertical axis. In this example, lift and the cosine similarity are specifically selected together because we know that the similar form of $L$ and $C$ results in similar raw interestingness measures. Not surprisingly, the relationship between high values of $L^*$ and high values of $C^*$ is strong. These points fall with relatively high values of $Q$, $Q^*$, and even $G^*$ but not $G$. The correspondence of all four standardized interestingness measures speaks to a common type of measurement. This contrasts with the raw interestingness measures: mostly moderate values of $L$, a wide range of values of $C$, mostly high values of $q$, and mostly low values of $G$. These are reflected more generally in Appendix A.2.

Table 3.1 shows rules with the consequent $\{35\}$. These rules show similar values for an interestingness measure as well as similar values for a standardized interestingness measure. There is, generally, greater disparity in the standardized interestingness measure than in the raw interestingness measure. For example, all the rules obtained values of $L$ of approximately 1.95, but $L^*$ ranged from 0.32 to 0.39 with the lowest values of $L^*$ occurring when item $\{18\}$ is in the antecedent. The standardized cosine similarity and standardized Yule’s $Q$ exhibits the same behaviour and to a lesser extent, we observe this with $Q^*$. Table 3.2 contains Kendall’s tau-b for these data. Standardizing lift reorders the rankings of rules the most, while standardizing Yule’s $Q$ reorders rules the least. The cosine similarity and the Gini index maintain a more moderate relationship when each is standardized. These trends are shown by deciles in Figure A.4a. The values of tau-b are fairly close to zero, suggesting a
CHAPTER 3. STANDARDIZED INTERESTINGNESS MEASURES

<table>
<thead>
<tr>
<th>antecedent</th>
<th>supp</th>
<th>conf</th>
<th>raw lift</th>
<th>s. lift</th>
<th>raw cos</th>
<th>s. cos</th>
<th>raw Q</th>
<th>s. Q</th>
<th>raw Gini</th>
<th>s. Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>{12,18,23}</td>
<td>0.203</td>
<td>0.411</td>
<td>1.951</td>
<td>0.317</td>
<td>0.630</td>
<td>0.317</td>
<td>0.959</td>
<td>0.334</td>
<td>0.079</td>
<td>0.868</td>
</tr>
<tr>
<td>{17,18,23}</td>
<td>0.203</td>
<td>0.411</td>
<td>1.951</td>
<td>0.325</td>
<td>0.630</td>
<td>0.325</td>
<td>0.960</td>
<td>0.342</td>
<td>0.079</td>
<td>0.870</td>
</tr>
<tr>
<td>{18,23}</td>
<td>0.203</td>
<td>0.411</td>
<td>1.951</td>
<td>0.327</td>
<td>0.630</td>
<td>0.327</td>
<td>0.960</td>
<td>0.343</td>
<td>0.079</td>
<td>0.870</td>
</tr>
<tr>
<td>{12,17,23}</td>
<td>0.204</td>
<td>0.411</td>
<td>1.952</td>
<td>0.368</td>
<td>0.631</td>
<td>0.368</td>
<td>0.962</td>
<td>0.386</td>
<td>0.079</td>
<td>0.878</td>
</tr>
<tr>
<td>{12,23}</td>
<td>0.204</td>
<td>0.411</td>
<td>1.952</td>
<td>0.370</td>
<td>0.631</td>
<td>0.370</td>
<td>0.962</td>
<td>0.388</td>
<td>0.079</td>
<td>0.878</td>
</tr>
<tr>
<td>{17,23}</td>
<td>0.204</td>
<td>0.411</td>
<td>1.952</td>
<td>0.385</td>
<td>0.631</td>
<td>0.385</td>
<td>0.963</td>
<td>0.403</td>
<td>0.079</td>
<td>0.881</td>
</tr>
<tr>
<td>{23}</td>
<td>0.204</td>
<td>0.411</td>
<td>1.952</td>
<td>0.387</td>
<td>0.631</td>
<td>0.387</td>
<td>0.963</td>
<td>0.405</td>
<td>0.079</td>
<td>0.881</td>
</tr>
</tbody>
</table>

Table 3.1: Rules for the Belgian Accident Data with the consequent \{35\} sorted by support and confidence.

<table>
<thead>
<tr>
<th>Interestingness Measure</th>
<th>tau-b</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mathcal{L}) and (\mathcal{L})(^*)</td>
<td>-0.033</td>
</tr>
<tr>
<td>(\mathcal{C}) and (\mathcal{C})(^*)</td>
<td>0.380</td>
</tr>
<tr>
<td>(\mathcal{Q}) and (\mathcal{Q})(^*)</td>
<td>0.657</td>
</tr>
<tr>
<td>(\mathcal{G}) and (\mathcal{G})(^*)</td>
<td>0.250</td>
</tr>
</tbody>
</table>

Table 3.2: Rule orders for the Belgian accident data.

significant reordering of rules for all interestingness measures and all deciles, except for high deciles of Yule’s Q, which obtain higher values of tau-b. The maintained ordering is most evident in higher deciles of Yule’s Q, which may be due to the interestingness measure being the only of the four discussed which always has a constant for an upperbound.

Through this example, we have shown that standardized interestingness measures can alter the interpretation of many rules. The relative order of rules often changes when an interestingness measure is standardized. In the cases of interestingness measures which can measure a departure from independence, such as lift and Yule’s Q, the values of many of the interestingness measures were as high or low as the rule allows, even though the actual value itself indicated independence or near independence. In the next section, we change the thresholds in the apriori algorithm and obtain a different number of rules but observe many of the same results.
3.3 Reuters

The Reuters-21578 Distribution 1.0 Lewis (Lewis) data are well-known data in text categorization. They are a collection of documents appearing on the Reuters newswire in 1987, made available in 1990, and reformatted in 1996. The reformatted version of the data are analyzed here. Words in documents are also reduced to word stems using the Rstem package (Lang, 2010) for R (R Development Core Team, 2012) and stop words are removed. There were over 1100 key word stems and 17,000 documents containing at least one of these key word stems, with an average of 38 unique key word stems.

Support and confidence thresholds, both set to 0.005, provide over 500,000 rules with the Apriori algorithm. Figure 3.3 shows the plots of raw interestingness measures against their standardized counterparts. The highest values of lift are higher than the previous data, indicating greater dependence. Similar to the previous data, most rules have values of $L$ near one and these rules have values of $L^*$ spread throughout the unit interval with concentrations near zero and one. Most rules obtain positive values of $Q$ with these data. When standardized, the range of values of $Q^*$ between zero and one is well represented. One concentration occurs at the maximum value of $Q$ near one, which has a large proportion of values of $Q^*$ near its maximum of one. The plot of $G^*$ against $G$ shows that most values of $G$ are close to zero and correspond to a variety of $G^*$ values, not just those near zero. However, most of the rules with $G$ near zero also have values of $G^*$ near zero.

Rules represented in Figure 3.3 are highlighted in orange when $G$ is above 0.05 and in blue when $G^*$ is above 0.8. High values of $G$ correspond to high values of $C$ and $Q$ while rules with high values of $G^*$ correspond to high values of $L, L^*, C$ and $C^*$ and very high values of $Q$ and hence $Q^*$. We also observe a few rules which exhibit the opposite behaviour among the neg-
Figure 3.3: Plots of interestingness measures against their standardized counterparts for the Reuters data. Dark regions indicate a higher density of rules.
CHAPTER 3. STANDARDIZED INTERESTINGNESS MEASURES

Atively correlated rules. This most interestingly manifests in the Yule’s Q subplot, showing the few negatively correlated rules with values of $Q$ below zero. Similar to the previous data, these data indicate high values of $G$ related to highly dependent rules. In this second data set, we again observe that values of raw interestingness measures indicating independence have standardized counterparts spread throughout the unit interval rather than being centered around a single value. This shows that while the value of raw interestingness measures of several rules may achieve values indicating independence (Prop 1 from Section 2.3), the rules may have different standardized interestingness measures, which highlights the importance of considering the standardized interestingness measure.

Table 3.4 shows values of tau-b for the Reuters-21578 data. Unlike the previous data, lift and the cosine similarity are reordered substantially. Figure A.4c shows how most rules have been reordered across deciles and interestingness measures. An example of rules is illustrated in Table 3.3 with rules containing the consequent \{golden\}. Unlike the previous data, we observe greater differences in raw interestingness measures and even greater differences in the standardized interestingness measures. Once again, we see that the standardized interestingness measures share some similarities, including a generally monotonic relationship among each other (see Figure A.2 in Appendix A.2). We also observe the near-boundary case for the standardized Gini index. The raw values for the Gini index are all very, very close to zero but the standardized interestingness measures clearly differ from zero in two rules while remaining essentially zero in another rule. Despite being so close to the raw interestingness measure’s global lower bound, we see that the standardized interestingness measure for the rule \{reuter\} $\Rightarrow$ \{gold\} is relatively high; this shows that this indication of nearly independent rules may not be so close to independence, after all. A somewhat similar phenomenon is observed in the previous data in Table 3.1, where $Q^*$ values were moderate.
CHAPTER 3. STANDARDIZED INTERESTINGNESS MEASURES

<table>
<thead>
<tr>
<th>antecedent</th>
<th>supp</th>
<th>conf</th>
<th>raw lift</th>
<th>s. lift</th>
<th>raw cos</th>
<th>s. cos</th>
<th>raw Q</th>
<th>s. Q</th>
<th>raw Gini</th>
<th>s. Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>{march}</td>
<td>0.011</td>
<td>0.659</td>
<td>1.001</td>
<td>0.198</td>
<td>0.198</td>
<td>0.198</td>
<td>0.195</td>
<td>3 ×10−8</td>
<td>7×10−11</td>
<td></td>
</tr>
<tr>
<td>{said}</td>
<td>0.016</td>
<td>0.895</td>
<td>1.183</td>
<td>0.754</td>
<td>0.754</td>
<td>0.754</td>
<td>0.754</td>
<td>0.623</td>
<td>0.00062</td>
<td>0.325</td>
</tr>
<tr>
<td>{reuter}</td>
<td>0.017</td>
<td>0.983</td>
<td>1.074</td>
<td>0.960</td>
<td>0.960</td>
<td>0.960</td>
<td>0.960</td>
<td>0.691</td>
<td>0.828</td>
<td>0.325</td>
</tr>
</tbody>
</table>

Table 3.3: Rules for the Reuters Data with the consequent {gold} sorted by support and confidence.

<table>
<thead>
<tr>
<th>Interestingness Measure</th>
<th>tau-b</th>
</tr>
</thead>
<tbody>
<tr>
<td>L and L*</td>
<td>-0.264</td>
</tr>
<tr>
<td>C and C*</td>
<td>0.023</td>
</tr>
<tr>
<td>Q and Q*</td>
<td>0.353</td>
</tr>
<tr>
<td>G and G*</td>
<td>0.234</td>
</tr>
</tbody>
</table>

Table 3.4: Rule orders for the Reuters data.

even though Q values were near the global maximum of one.

Through these we again observe similar features as the last data when standardizing interestingness measures. In particular, we see that values of interestingness measures which indicate that the antecedent and consequent of a rule are nearly independent were, in many instances, as far away from independence as possible given the attainable range of the interestingness measure. We again saw how the relative order of rules change if standardized rather than raw interestingness measures are used to rank the interestingness of rules. In the next example, we consider rules which are designed to be independent and observe that some of the standardized interestingness measures do not behave the same as the nearly-independent rules in this the previous data sets.
3.4 Random Transactions

Two real data sets have been investigated where it was reasonable to assume some associations exist. We now consider a simulated data set with completely random transactions of independent items. This investigation is necessary to understand how randomness, such as noise, affects standardizing interestingness measures. We expect that many of the raw interestingness measures that fulfil Prop 1 of Section 2.3 will achieve values very near the constant implying independence. For this reason we expect no pattern to the values of $C$, and by extension, $C^*$. As in special cases of previous examples, we expect that the standardized counterparts will have no pattern for these rules.

We generate 100 000 random transactions of up to 10 000 items using `random.transactions` from the `arules` package in R. The probability of items appearing in transactions was left at the default value of 0.01. We then mine the rules with support and confidence thresholds each set to 0.0001 to obtain over 630 000 rules.

Figure 3.4 shows standardized interestingness measures are plotted against their raw counterparts with these random transactions. The scale of each of the plots is important to note: some only show a small range near zero. The ‘shape’ of the plots are similar for the interestingness measures. In three cases, the transformation is nearly linear and the maximum standardized values of $L^*$ and $C^*$ are small. The relationship between $Q$ and $Q^*$ is also linear but a larger range is observed, in part because of the constant upperbound, but many of the rules are very close to zero. Contrary to this trend, $G$ is very near zero but $G^*$ polarizes to either zero or one. We investigated this further and the values that tended to one are the entire set of negatively correlated rules. All the positively correlated rules tended to zero. Though we are observing a dichotomy, whether positively correlated or negatively correlated,
Figure 3.4: Plots of interestingness measures against their standardized counterparts for the random-transactions data. Dark regions indicate a higher density of rules.
CHAPTER 3. STANDARDIZED INTERESTINGNESS MEASURES

<table>
<thead>
<tr>
<th>Interestingness Measure</th>
<th>tau-b</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{L} ) and ( \mathcal{L}^* )</td>
<td>0.678</td>
</tr>
<tr>
<td>( \mathcal{C} ) and ( \mathcal{C}^* )</td>
<td>0.777</td>
</tr>
<tr>
<td>( Q ) and ( Q^* )</td>
<td>0.809</td>
</tr>
<tr>
<td>( G ) and ( G^* )</td>
<td>0.284</td>
</tr>
</tbody>
</table>

Table 3.5: Rule orders for randomly generated transactions.

we see independent transactions tending to a particular standardized value, depending on the case. With previous data, we highlighted high values of an interestingness measure. With these data, this is not necessary as the transformations appear to be generally monotonic and most rules achieve standardized values very near zero.

Values of tau-b are shown in Table 3.5. Though one may have expected that with random transactions, rule orders would also appear random, this was not the case. The standardized rule orders are surprisingly much more similar to the original interestingness measure rankings compared to previous data. The Gini index was reordered the most, most likely because of how negatively correlated rules score higher than independent rules. Also interesting with these random transactions, Figure A.4c shows a generally monotonically increasing relationship between the decile and value of tau-b for all interestingness measures.

3.5 Summary

In this chapter we introduced the standardized cosine similarity, standardized Yule’s Q, and standardized Gini Index. We compared these three standardized interestingness measures along with standardized lift against each other and their raw counterparts to see how standardized interestingness impact analysis. The impacts were analyzed primarily in two data
sets: the Belgian Accidents data set and the Reuters-21578 data set. We also analyzed how standardized interestingness measures fare when there are no true underlying transactions, as in the case of the random-transactions data set.

In the three cases where the interestingness measure is monotonic with $P(A,B)$, we saw that the standardized interestingness measures generally quantify similar features of a rule. In most scenarios, we discovered that the relationship between the raw and standardized counterpart is fairly monotonic, except when rules are, or nearly are independent. In that case, the standardized interestingness measure has no predictable response, sometimes tending to one extreme or the other. Therefore, we find that the value of an interestingness measure is often as extreme as possible given the attainable range imposed by the rule. In the case of random transactions we saw that except for Yule’s Q, all of the standardized interestingness measures tend to zero. Yule’s Q remained almost entirely unchanged except being rescaled to lie on the unit interval.

The behaviour of standardized interestingness measures for values that seemed to be so clearly independent or nearly independent was somewhat curious. We explore how actually close to independence these rules are in the next chapter, with hypothesis tests for independence. Though the context of the hypothesis testing is for missing data, the methodology is just as applicable for testing for the independence discussed in this chapter.
Chapter 4

Incomplete Transactions

4.1 Chapter Outline

Incomplete transactions are problematic in analysis. Mining algorithms often cannot handle missing data so incomplete transactions must be addressed properly. The simplest approach is to discard the incomplete transactions, but this loses the information in those transactions. Further, the absence of data can be a source of information, not simply an obstacle to overcome. In this chapter, we address this problem for a scenario that has been lacking a proper methodology.

This chapter is an exploration into methods to mine association rules with incomplete transactions. In Section 4.2 we specifically discuss a method which handles missingness. Section 4.2.1 describes Fisher’s exact test and its application for association rules. We then begin analyzing data in Section 4.3. We discover rules regarding missingness and illustrate
the difference between Fisher's exact test and the $\chi^2$ test. In Section 4.4 we analyze the Pima diabetes data and discover how the missing information provides interesting rules. We also use the data set to compare the relationship with Fisher’s exact test for independence with an interestingness measure which measures departure from independence. We consider how supports and interestingness measures behave when data are missing in Section 4.5. This leads to a framework analyzing rules when data are missing. We illustrate this with simulated rules in Section 4.5.1 and show how our proposed is used. Finally, we provide a synopsis of our contributions in Section 4.6 with detailed in Section 6.2 of Chapter 6, including future work.

4.2 Methodology

To investigate missingness patterns, we create indicator variables that identify whether an item is missing in a transaction. We consider binary data to be a special case of categorical data. Categorical data can simply be binarized, with an additional item indicating missingness. Thus, we consider a missing observation as another level of categorical data. A categorical variable with $k$ possible values is replaced by $k + 1$ variables: one variable for each level plus one more to indicate if the information is missing. The apriori algorithm discards rules where two items from the same category are present in a set of items as its support must be zero. This occurs in the first iteration of finding frequent sets of items. After this iteration, no redundancies from multiple items from a category are even considered in the algorithm, making this a simple and fairly efficient modification for the apriori algorithm and its variants.
CHAPTER 4. INCOMPLETE TRANSACTIONS

Ordinal variables with only a few distinct values observed in the data can be treated categorically but will benefit from grouped intervals. Grouping helps to keep individual item supports high enough to remain interesting, or at least avoid being pruned. Ordinal variables with more distinct values may undergo a process of discretization by creating appropriate intervals and treating the intervals categorically. A continuous variable can be discretized in several ways. Various discretization methods are compared in Dougherty et al. (1995) but have some disadvantages Vannucci and Colla (2004) that can be overcome by using a self-organizing map (SOM) (Kohonen, 1990). However, we only have a few continuous-type variables and discretize them according to functional or previously used intervals. Further details are explained with each specific data set.

Once the data are in the proper form, we mine rules using the apriori algorithm. We choose minimum support and confidence thresholds small enough to provide a fair number of rules but not so small that the algorithm outputs a plethora of rules. Note that the original information from complete transactions is intact, and so mining algorithms will discover rules both with and without missingness indicators. This means rules that would have traditionally been mined will be a subset of the rules discovered, provided that the absence of data did not deflate supports and confidences below their respective minimum thresholds.

Association rules which consider missingness fills a surprisingly high proportion of the space of all rules. Mining association rules from binary data with $n$ singleton items, we know that there are exactly $3^n - 2 \cdot 2^n + 1$ possible rules. If we were to mine rules while also considering missingness in each possible item in the binary data, we would actually have $5^n - 2 \cdot 2^n + 1$ rules to consider employing in the framework we propose in Section 4.2. Considering only 10 items, over 90% of possible rules include an item regarding the missingness; considering more than 25 items, this proportion is more than 99.9%. Clearly, by considering the missingness...
of items, the majority of rules considered will include at least one missingness indicator and this merits consideration when mining. With many of these rules which include a missingness indicator, it becomes necessary to identify which missingness indicators are independent and which are dependent on the rest of the data. Such tests for independence are discussed in the next section.

### 4.2.1 A Discrete Test for Independence

Testing if the observed data are independent of the missing data falls under the scenario where data are MCAR (see Section 2.5). The test for independence among discrete data was (a suggestion by Fisher published in Yates, 1934, and) an improvement to the $\chi^2$ test, which only worked well for tables with large enough counts that the counts themselves could be reasonably considered to be normally distributed. For large counts, Fisher’s exact test approaches the $\chi^2$ test, but the $\chi^2$ is still computationally less expensive, making it desirable in those cases. As we will show later, the $\chi^2$ test does reasonably well for very small $p$-values under some situations but not others. Two issues arise from a case with smaller counts: one practical and one theoretical. Note that we use small counts to refer either the scenario when the total size of the sample is small, or when the proportion is small even if the sample size is large. The theoretical issue is that since supports are rational, supports range over a countable, and therefore discrete, domain. Fisher’s exact test is suitable for discrete domains and the difference from continuous domains is more pronounced for smaller counts even if the total number remains high.

We now provide Fisher’s exact two-tailed test for independence between the antecedent and consequent. Using Fisher’s exact test has recently been suggested for association rules as
an alternative to the $\chi^2$ but not in the context of missingness. See Ableson and Glasgow (2003) for details on other applications. If $P(A, B) < P(A)P(B)$, and $m$ transactions when $0 \neq P(A, B) \neq 1$,

$$p\text{-value} = \frac{1}{\binom{m}{mP(B)}} \sum_{i=0}^{mP(A,B)} \sum_{i=m(1-P(A,B))+1}^{m} \binom{mP(A)}{i} \binom{m(1-P(A))}{m(P(B) - i)}$$

where $\binom{n}{k}$ denotes the binomial coefficient. In the case where $P(A, B) > P(A)P(B)$ the indices change to reflect the nature of the tails of the hypergeometric distribution. We choose to sum across potential consequents given an antecedent as this seems most consistent with the definition of the rule. This means that the hypothesis test not symmetric in general. If a symmetric version were required, we would take a symmetric function of the $p$-value from the rule $A \Rightarrow B$ and the $p$-value from the rule $B \Rightarrow A$, such as the supremum of the two $p$-values.

Using Fisher’s exact test, we analyze the rules which are mined from two data sets. In the first data set, discussed in the next section we will compare this test against the $\chi^2$ test and in the following section we will compare this test against an interestingness measure.

### 4.3 Hepatitis data

The hepatitis data set is a well-known example used in classification. Hepatitis is an inflammation of the liver and can be caused from a variety of sources. The cause of hepatitis is not given in the data set and it remains unclear whether the patients developed hepatitis in similar or different ways. Various observations from 155 people, including the class attribute
whether the individual lived or died, are recorded, but many of the other attributes are missing from the data set. The proportions of missing attributes are show in Table 4.1.

<table>
<thead>
<tr>
<th>attribute</th>
<th>proportion missing (frequency missing)</th>
<th>nature</th>
<th>binarized levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>class</td>
<td>0.000 (0)</td>
<td>discrete</td>
<td>died,lived</td>
</tr>
<tr>
<td>age</td>
<td>0.000 (0)</td>
<td>continuous</td>
<td>≤ 47, &gt; 47</td>
</tr>
<tr>
<td>sex</td>
<td>0.000 (0)</td>
<td>discrete</td>
<td>male,female</td>
</tr>
<tr>
<td>antivirals</td>
<td>0.000 (0)</td>
<td>discrete</td>
<td>yes,no</td>
</tr>
<tr>
<td>histology</td>
<td>0.000 (0)</td>
<td>discrete</td>
<td>yes,no</td>
</tr>
<tr>
<td>steroid</td>
<td>0.006 (1)</td>
<td>discrete</td>
<td>yes,no,missing</td>
</tr>
<tr>
<td>fatigue</td>
<td>0.006 (1)</td>
<td>discrete</td>
<td>yes,no,missing</td>
</tr>
<tr>
<td>malaise</td>
<td>0.006 (1)</td>
<td>discrete</td>
<td>yes,no,missing</td>
</tr>
<tr>
<td>anorexia</td>
<td>0.006 (1)</td>
<td>discrete</td>
<td>yes,no,missing</td>
</tr>
<tr>
<td>sgot</td>
<td>0.026 (4)</td>
<td>continuous</td>
<td>≤ 50, &gt; 50,missing</td>
</tr>
<tr>
<td>spleen-palpable</td>
<td>0.032 (5)</td>
<td>discrete</td>
<td>yes,no,missing</td>
</tr>
<tr>
<td>spiders</td>
<td>0.032 (5)</td>
<td>discrete</td>
<td>yes,no,missing</td>
</tr>
<tr>
<td>ascites</td>
<td>0.032 (5)</td>
<td>discrete</td>
<td>yes,no,missing</td>
</tr>
<tr>
<td>varices</td>
<td>0.032 (5)</td>
<td>discrete</td>
<td>yes,no,missing</td>
</tr>
<tr>
<td>bilirubin</td>
<td>0.039 (6)</td>
<td>continuous</td>
<td>≤ 0.3, ∈ (0.3, 1.9], &gt; 1.9,missing</td>
</tr>
<tr>
<td>liver.big</td>
<td>0.065 (10)</td>
<td>discrete</td>
<td>yes,no,missing</td>
</tr>
<tr>
<td>liver-firm</td>
<td>0.071 (11)</td>
<td>discrete</td>
<td>yes,no,missing</td>
</tr>
<tr>
<td>albumin</td>
<td>0.103 (16)</td>
<td>continuous</td>
<td>≤ 3.4, ∈ (3.4, 5.4], &gt; 5.4,missing</td>
</tr>
<tr>
<td>alk-phosphate</td>
<td>0.187 (29)</td>
<td>continuous</td>
<td>≤ 40, ∈ (40, 120], &gt; 120,missing</td>
</tr>
<tr>
<td>protime</td>
<td>0.432 (67)</td>
<td>continuous</td>
<td>present, missing</td>
</tr>
</tbody>
</table>

Table 4.1: Summary of the hepatitis data.

Discrete attributes were binarized with an additional indicator included for missingness. Continuous attributes were binarized by levels which separate what is normal and healthy from those that are not. The attribute for prothrombic time was not discretized as the units are unclear. Prothrombic time is often recorded in seconds, or given as a ratio of the observed time relative to an ideal time for a healthy individual in the patient’s circumstances. The observed values do not correspond well to the data unless almost all individuals are clotting very quickly or very slowly. Because of this ambiguity we only binarize the attribute into
CHAPTER 4. INCOMPLETE TRANSACTIONS

whether the attribute is missing or recorded in the transaction. There are no expected patterns except that many rules are expected to include missing prothrombic time because of its high prevalence.

The sets of rules with confidence precisely 1 are useful to analyze because they identify when some items appear only as a subset of a larger set of items. This often happens with structured non-response and can actually be an effective method of screening. This may be considered uninteresting in a sense because it should be expected, or known in advance. However, we have no such knowledge with these data so such discoveries are interesting to us. Liver tests for size and firmness are always missing together, and always occurs with those indicating antivirals. We also find that the attributes spleen-palpable, spiders, varices, and ascites are always missing together, and again as a subset of transactions where antivirals is present. This is unusual as the presence of spiders is a simple visual assessment. Only a subset of one gender (males) and not the other gender have missing observations for phosphate alkalinity. Individuals with low (below 40) phosphate are always missing a prothrombic time measurement. This may not be very revealing given how often the prothrombic time is missing.

As an example why using Fisher’s exact test is more beneficial than the \( \chi^2 \) test, observe the plots in Figure 4.1. Rules are mined and the tests for independence are conducted, with the \( p \)-values plotted. In the first plot we see that the approximation is actually fairly good globally, and deviation between the two diminish for smaller \( p \)-value but when highlighting smaller values, when it is arguably very important to have a good approximation, we can see a problem with the \( \chi^2 \) test. A line indicates where points would appear if the two tests completely agreed. Unfortunately, in most instances the true \( p \)-value is lower than the approximated \( p \)-value from the \( \chi^2 \) test and would significantly effect any inferences.
with these small values. Under the null hypothesis that the antecedent and consequent are independent, the hypothesis would not be rejected many times when there is evidence against it. This is a problem with the granularity of discrete data having larger relative deviations from the normal with small proportions. Fortunately, Fisher’s exact test is computationally inexpensive for small counts and can always be used in these cases, avoiding the error of the approximation.

### 4.4 Pima data

The data set about diabetes among the Pima people is a well-known classification data set. The original analysis (Smith et al., 1988) constructed a function to predict type II diabetes among the Pima people known, to have a high prevalence of the disease. Among the collected data for 768 women at least 21 years old are nine variables including the indicator if the
CHAPTER 4. INCOMPLETE TRANSACTIONS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Support</th>
<th>Discretization</th>
</tr>
</thead>
<tbody>
<tr>
<td>number times pregnant</td>
<td>0.000 (0)</td>
<td>(0), (1,2), (3,4,5,6), (7,8,...)</td>
</tr>
<tr>
<td>glucose concentration in mmol/L</td>
<td>0.007 (5)</td>
<td>missing, &lt; 140mmol/L, ≥ 140</td>
</tr>
<tr>
<td>diastolic bp in mmHg</td>
<td>0.046 (35)</td>
<td>missing, &lt;60, [60,80), ≥80</td>
</tr>
<tr>
<td>triceps skin fold mm</td>
<td>0.296 (227)</td>
<td>missing, skin≤25, (25,32], &gt;32</td>
</tr>
<tr>
<td>insulin level in µU/mL</td>
<td>0.487 (374)</td>
<td>insulin is missing, &lt;100, ≥100</td>
</tr>
<tr>
<td>bmi</td>
<td>0.014 (11)</td>
<td>missing, &lt;18.5, [18.5,25], &gt;25</td>
</tr>
</tbody>
</table>

Table 4.2: Discretization method of ordinal and continuous variables.

individual has type II diabetes by WHO guidelines. The other covariates consist of six continuous variables measured (or functions of variables measured such as the BMI), an ordinal variable indicating the number of times the woman was pregnant, and the paper’s contribution of a diabetes predictive pedigree function: a continuous variable on the unit interval. In our analysis we discard the pedigree function and discretize the other variables based on relevant WHO standards and when those are unavailable, the discretization in the original analysis. Glucose level up to 140mmol/L is considered normal and diastolic blood pressure between 60 and 80 mmHg is considered normal so we discretize according to this. These discretizations differ from the ones used in the original analysis. The other continuous variables are discretized to coincide with the original analysis.

For the variables where it was relevant, we also included an indicator of missingness. Missingness was determined based on the observation (Breault, 2001) that there are recorded zeroes where a zero is practically impossible, such as a blood pressure. Missingness is somewhat sparse, so we impose a low support threshold of just two transactions. The confidence threshold was set to 0.5. This generally focuses on rules where the antecedent and consequent co-occur. We only consider singleton consequents for interpretability and focus on rules with missing indicators. There are no identifiable forms of structured missingness that we are aware of but expect one trend again due to prevalence. Insulin was missing so often
CHAPTER 4. INCOMPLETE TRANSACTIONS

in the database (almost half of observations) that the original analysis completely discarded the variable.

We discovered 3254 rules with antecedents comprised of one to six items. We discovered several rules that indicated a strong departure from independence. We record values of support, confidence, lift, and standardized lift. In some instances, standardized lift required using limits to evaluate where the support of the antecedent equalled the minimum support, an issue arising from a low threshold. In the case that minimum support is \( \sigma \) where some supports can exactly equal \( \sigma \), standardized lift may be undefined. However, we find the value by considering a minimum support threshold of \( \sigma + \epsilon \) for some small \( \epsilon > 0 \) in (4.1).

\[
\mathcal{L}^* = \frac{\mathcal{L} - \lambda}{v - \lambda} = \lim_{\epsilon \to 0} \frac{P(A, B)}{P(A)P(B)} - \frac{\sigma + \epsilon}{(P(A)P(B)} - \frac{\sigma + \epsilon}{(P(A)P(B)}
\]

since \( \lambda = \frac{\sigma + \epsilon}{(P(A)P(B)}, v = \frac{P(A)}{P(A)P(B)} \)

\[
= \lim_{\epsilon \to 0} \frac{P(A, B) - \sigma + \epsilon}{P(A) - \sigma + \epsilon}
\]

\[
= \lim_{\epsilon \to 0} \frac{\sigma - \sigma + \epsilon}{\sigma - \sigma + \epsilon} = \lim_{\epsilon \to 0} 1 = 1
\]

(4.1)

The rule \{more than 7 pregnancies, glucose below 140, low blood pressure} \( \Rightarrow \) \{missing skin measurement\} achieved a lift value of 2.25. In traditional mining methods this would likely be considered interesting since it seems the antecedent and the consequent are far from independent but the \( p \)-value testing that the antecedent and consequent were independent was 0.21 using Fisher’s exact test. This indicates that there was more than a 20% chance of observing such a rule if the antecedent and the consequent are actually independent. This discrepancy comes from the inflation of lift because of a low antecedent support. Standardized lift is not so influence and obtained a value of 0 in this instance.
As we alluded with the rule above, we can analyze the relationship between $p$-values and lift. In Figure 4.2, as we may expect, values of lift near one have higher $p$-values. Surprisingly, for values of lift as high as 3 we obtain $p$-values which are fairly moderate under the null hypothesis that the antecedent and consequent are independent of one another. Although lift is a fast method of measuring departure from independence, we can see that using it to strictly judge how close we are to independence has a drawback as well. This inflation of lift making it seem fairly dependent is due to smaller supports and smaller counts, highlighting the importance of using Fisher’s exact test.

![Test for independence vs lift](image)

Figure 4.2: Comparison of rescaled $p$-values and lift on rescaled axes.

Many of the rules we discovered related to the missing skin measurements and missing insulin values since they were the most prevalent missing attributes. Interestingly, one rule we discovered, \{more than 7 pregnancies\} $\Rightarrow$ \{missing insulin measurement\}, had a lift value of 1.24 which may not seem particularly interesting but the $p$-value for the hypothesis that the antecedent and consequent are unrelated is below .001.
The rule \{missing glucose concentration, bp normal\} ⇒ \{high skin triceps measurement\} had a value of lift 2.723. Without the missing glucose concentration, the value of lift is 0.9149899 meaning it was actually less likely to have a high skin triceps measurement. The rule with present, rather than missing glucose had a lift of 0.90, with a p-value of .08973 when testing for independence with a two-tailed alternative.

In this example we have shown the usefulness of Fisher’s exact test over an interestingness measure to find evidence of dependence. In the previous example, we showed how Fisher’s exact test is more suitable than the $\chi^2$ test in many cases. In both instances, we were using the test to identify how missing data relate, if at all, to other data. In the next section, we will focus on how to analyze association rules in light of missing data.

### 4.5 Interestingness Measures with Missing Data

We now consider analysis where the absence of information is not our focus, but the possible patterns in light of missing information. Distributions for the data may be found under some mild assumptions. We will assume that transactions are independent, so assuming they are following Bernoulli distributions is reasonable. The model that we present in this section is suitable for data that are MAR, and therefore also the case when MCAR as a special case.

In traditional association rule mining, every item is recorded as either 1 or 0. For two items, there are four distinct possible transactions. When a missing value is observed, we observe up to three possible values including missingness but know the missing values represent up to four values. Thus, an item in a transaction may either be observed to be 1, observed to be 0, missing but truly 1, or missing but truly 0. It is worthwhile emphasizing that we do
not consider this to be four possible responses but two responses, each of which can manifest two different ways. For two items, there are 16 unique ways that a transaction truly exists, but only a maximum of 9 ways to distinguish them, as shown in Table 4.3. This observation is the basis for our method.

\[
\begin{array}{c|c|c|c|c}
A & B = 1 & B = 0 & \bar{A} & B = 1 \\
\hline
1 & x_{11} & x_{10} & \bar{x}_{11} & \\
0 & x_{01} & x_{00} & \bar{x}_{01} & \\
1 & \bar{x}_{1} & \bar{x}_0 & \bar{x}_{10} & \\
0 & \bar{x}_{10} & \bar{x}_{0} & \bar{x}_{1} & \\
\end{array}
\]

Table 4.3: Table of how data manifest with missingness.

The most basic quantity to first consider is support. Let \( x_1 \) and \( x_0 \) denote the number of transactions where the item has been observed to be 1 and 0, respectively. Let \( \bar{x}_1 \) denote the number of transactions where the item is missing. We use this notation because we will be decomposing what is observed to the underlying 0 and 1: \( x_+ = x_1 + x_0 \) where \( x_1 \) and \( x_0 \) represent the number of transactions that are missing but the item is truly 1 and 0, respectively. This means \( P(X = 1) = \frac{x_1 + \bar{x}_1}{m} \) where \( x_1 \in \{0, 1, 2, \ldots, x_+\} \), and so

\[
\frac{x_1}{m} \leq P(X = 1) = \frac{x_1 + \bar{x}_1}{m} \leq \frac{x_1 + \bar{x}_1}{m} = \frac{m - x_0}{m}.
\]

Assuming that transactions are independent of one another, we may consider that

\[
\bar{X}_1 \sim \text{Bin}(x_+ = m - x_+, p)
\]

for some unknown probability \( p \). The value of \( p \) may be estimated through various means.
such as classifiers. One clear choice, which corresponds to independence, is \( p = x_1/x_+ \), the proportion of observed values which are 1. The binomial distribution has previously been used with association rules to construct intervals and calculate exact \( p \)-values in Weiß (2008). The key difference is our method exploits the genuine uncertainty coming from missing values while the other method assumes that databases are some kind of simple random samples of larger populations and randomness comes from sampling.

We extend this case of a single item to more items to construct a joint distribution. Let a double-subscript be used where the first subscript refers to the antecedent and the second subscript refers to the consequent. For example \( x_{+1} \) denotes the number of transactions where the first item is missing and the second item is present and observed to be 1.

We have that the joint support is

\[
\frac{x_{11}}{m} \leq P(X_{11}) = \frac{x_{11} + X_{1\bar{1}} + X_{\bar{1}1} + X_{\bar{1}\bar{1}}}{m} \leq \frac{m + x_{\bar{1}+} + x_{\bar{1}+} + x_{++}}{m}.
\]

To consider more than just the range of the support, we may make a distributional assumption such as the \( X \)-s being distributed binomially. Of course, we now have twelve random variables in total, with several restrictions based on how items are missing in various transactions. The eight variables corresponding to exactly one missing item are related through the restrictions shown in (4.2).

\[
\begin{align*}
X_{1\bar{1}} + X_{\bar{1}0} &= x_{1\bar{1}} \\
X_{\bar{0}1} + X_{0\bar{0}} &= x_{0\bar{1}} \\
X_{1\bar{1}} + X_{\bar{0}1} &= x_{\bar{1}1} \\
X_{\bar{1}0} + X_{\bar{0}0} &= x_{\bar{1}0} \\
\end{align*}
\]

(4.2)
CHAPTER 4. INCOMPLETE TRANSACTIONS

The four random variables when both items are missing are more intimately bound with the restriction shown in (4.3).

\[ X_{\hat{1}\hat{1}} + X_{\hat{0}\hat{0}} + X_{\hat{1}\hat{0}} + X_{\hat{0}\hat{1}} = x_{++} \]  

Each restriction from (4.2) can be used to construct a pair of complementary binomial random variables with appropriate probabilities while the restriction from (4.3) can be used to construct a multinomial distribution. Therefore, we propose the distribution shown in (4.4) to model our missing variables, under the assumption of independent transactions.

\[
\begin{pmatrix}
    X_{\hat{1}\hat{1}} \\
    X_{\hat{0}\hat{0}} \\
    X_{\hat{1}\hat{0}} \\
    X_{\hat{0}\hat{1}}
\end{pmatrix}
\sim \text{Mult}(x_{++}, p)
\]

\[
p = \frac{1}{x_{++}}
\begin{pmatrix}
    x_{11} \\
    x_{10} \\
    x_{01} \\
    x_{00}
\end{pmatrix}
\]  

for MAR

\[
X_{\hat{1}\hat{1}} \sim \text{Bin}(x_{+1}, q_3)
\]

\[ X_{\hat{0}\hat{1}} = x_{+1} - X_{\hat{1}\hat{1}} \quad q_3 = \frac{x_{1+} + x_{+1}}{x_{++} + x_{+1}} \]  

for MAR

\[
X_{\hat{1}\hat{0}} \sim \text{Bin}(x_{+0}, q_2)
\]

\[ X_{\hat{0}\hat{0}} = x_{+0} - X_{\hat{1}\hat{0}} \quad q_2 = \frac{x_{1+} + x_{+1}}{x_{++} + x_{+1}} \]  

for MAR

\[
X_{\hat{1}\hat{1}} \sim \text{Bin}(x_{1+}, q_1)
\]

\[ X_{\hat{0}\hat{0}} = x_{1+} - X_{\hat{1}\hat{1}} \quad q_1 = \frac{x_{+1} + x_{+1}}{x_{++} + x_{+1}} \]  

for MAR

\[
X_{\hat{0}\hat{1}} \sim \text{Bin}(x_{0+}, q_0)
\]

\[ X_{\hat{0}\hat{0}} = x_{0+} - X_{\hat{0}\hat{0}} \quad q_0 = \frac{x_{+1} + x_{+1}}{x_{++} + x_{+1}} \]  

for MAR

We have included the suggestions for probabilities assuming that the missing data are MAR. The choice of these parameters is actually not simple. Partly foreshadowed by the notation, a great deal of sampling theory may go into calculating these probabilities if the model from (4.4) is chosen. In the case that the missing data are modelled to be MCAR, the complete-transaction estimates alone are sufficient and a special case of probability estimates when data are MAR.
We can use these distributions of supports to find quantities including interestingness measures. Clearly, supports themselves do not require much manipulation because values from binomial and multinomial distributions are well documented. Values such as confidence and lift, though defined somewhat simply, become far more difficult to evaluate. For example, lift would be

\[
\frac{P(X_{11})}{P(X_{+1})P(X_{1+})} = m\frac{x_{11} + X_{1\dagger} + X_{\dagger1} + X_{\dagger\dagger}}{(x_{1+} + x_{1\dagger} + X_{\dagger+} + X_{\dagger\dagger})(x_{+1} + X_{+\dagger} + x_{\dagger+} + X_{\dagger\dagger})}.
\]

The expected values are sensible but calculating them theoretically is slightly trickier. Rather than finding the closed form, we can estimate them numerically. We can very easily simulate several values from (4.4) to understand how the interestingness measure could behave. Simulating values would then become similar to a multiple-imputation method, without the drawback of having to repeatedly mine the database.

If only the expected lift was of interest, we can simply calculate this average. Substituting the random variables by their respective expected value is an appropriate method of doing so if we considered each of the models to be independent of one another. If we introduced dependence, which would be a more realistic approach, the expected value needs to be calculated from first principles. Numerically estimating this quantity would have the advantage of providing not just the expected value of lift but information of the distribution of lift. We could use the approximate distributional information to calculate more than just the expected value of lift, but other relevant quantities such as the probability that lift is at most one.

We will note that if antecedents or consequents are comprised of more than one item, we can still use this method. A set of items evaluates to 0 if at least one item is 0. If no items are 0, the set evaluates to being missing if at least one of the item is missing. Otherwise, the set of
items evaluates to 1. Thus, the methodology above works for antecedents and consequents of any non-zero length.

The model presented in this section is flexible enough to account for data missing in a variety of ways. In the next section, we will employ this model in three simulations. The first two simulations will be instances where data are MCAR, and the data will be MNAR in the third simulation.

### 4.5.1 Simulating values

Data can be missing in a variety of ways. To explore this variety, we create several data sets with data selectively removed to illustrate how flexible our method is. We use 1000 simulated sets of values distributed as described in (4.4) for our estimates. We will refer to complete-transaction estimate of lift, and by that we mean to estimate the value of lift found using only the $x_{ij}$ values, so the complete-transaction estimate of lift is given by

$$\frac{x_{11}}{x_{1+}x_{+1}}.$$ 

The first rule we analyze has antecedent support of 0.28 and consequent support of 0.70. The two are simulated independently. Of the 50 transactions, we independently remove 10 from the antecedent and 10 from the consequent, making the missing data MCAR. The 90% credible interval centered around the median of lift using our distribution is (0.88,1.32). This median is based on the sample of simulated values. To avoid ambiguity, we will call this median the simulated median and will serve as the point estimate for the value of lift. The true value of lift is 1.04 and the estimated value of lift from only complete transactions is
CHAPTER 4. INCOMPLETE TRANSACTIONS

1.16. The simulated median lift was 1.12, closer than the complete-transaction estimate but the complete-transaction estimate is also close. This is not surprising since the missing are MCAR. A summary of the data and estimates for lift are shown in Figure 4.3.

<table>
<thead>
<tr>
<th></th>
<th>( B \parallel 1 )</th>
<th>( B \parallel 0 )</th>
<th>( \tilde{B} \parallel 1 )</th>
<th>( \tilde{B} \parallel 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A = 1 )</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>( A = 0 )</td>
<td>18</td>
<td>8</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>( A = \tilde{1} )</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( A = \tilde{0} )</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 4.3: Sample distribution of lift when data are MCAR.

In a second set, we use supports 0.42 and 0.52 but with 0.28 overlapping so that the antecedent and consequent are positively correlated. The antecedent and consequent of the rules is not independent but the distributions are still independent of one another. We remove data randomly within the antecedent regardless of its value, again regardless of its value. By design, a higher proportion of the antecedent is removed but because there is no relationship between the observed (or the missing) data and the actually missingness, data are MCAR.
We find that with this example the true and estimated values were fairly close while the complete-data estimate is poor. The 90% credible interval centered around the simulated median of lift using our distribution is (1.09, 1.42). The true value of lift was 1.28, and the complete-transaction estimate of lift was 1.58, while the simulated median was 1.25, considerably closer than the complete-transaction estimate. A summary of the data and estimates for lift are shown in Figure 4.4.

In this third data set, we intentionally remove data with higher proportion when the antecedent is 1. The remaining observed data alone do not indicate that there is a relationship between the missingness and the antecedent being 1. Therefore, these data are MNAR. However, we will see that finding estimates as though the data are MAR works reasonably well.
The 90% credible interval centered around the median of lift using our distribution is (0.88, 1.34). The true value was 1.02, and the complete-transaction estimate was 1.31 while the simulated median lift value was 1.11. A summary of the data and estimates for lift are shown in Figure 4.5. It should be noted here that this model presented here has the potential for doing poorly for the same reason that a model assuming data are MCAR when they are really MAR can do poorly. The suggested estimates for proportions take advantage of the available information in incomplete transactions as much as possible but by nature of MNAR, that is insufficient. As we discuss in future work in Section 6.2, a more complicated model is required.

4.6 Summary

We have proposed applying a statistical test to find where missing data are related to the items of a data base. Through Fisher’s Exact Test, we can find exact $p$-values which are superior to the probabilities exploiting the $\chi^2$ approximation. The disparity between the
two can be large when $p$-values are not moderate, making it particularly important for smaller $p$-values. We also discovered that interestingness measures that measure a departure from independence like lift should be used cautiously to judge independence. Though a general trend did exist, there were many instances where the a missing item was clearly not independent, or may easily have been independent according to the $p$-value while the interestingness measure indicates otherwise.

Tackling another task related to incomplete transactions, we facilitated a method of inference when an interestingness measure is desired, in light of the missing data and how the data are missing in various transactions. The model presented in 4.4 provide systems of distributions to relate the missing and observed data and to make analysis requiring complete supports. In the cases where data were MCAR and MAR, we provided good estimates of an interestingness measure and even in the case where data were MNAR we still obtained a reasonably good estimate.
Chapter 5

Hierarchies

5.1 Chapter Outline

Association rule mining is a data mining technique which typically results in many rules. It can be a challenge to focus on learning the most interesting rules when so many rules are discovered. A common means of finding the most interesting rules in the data is through the use of an interestingness measure, which quantify an aspect of a rule. Even with the use of interestingness measures, there may still be so many rules that it is beneficial to summarize or generalize the rules to understand the collection of rules.

The method that we will present will superficially appear to be a variant of frequency patterns (FP trees in Han et al., 2000) as both use tree representations. They are, however, significantly different in methodology and application. The most significant difference is that frequency patterns constructs a unique tree to facilitate association rule discovery. Based on
CHAPTER 5. HIERARCHIES

supports, the tree becomes an object possessing all the original information of the database, so it can be mined in place of the database. This makes it more efficient than some other mining techniques which directly use the database. Conversely, the method that we propose is not meant to mine rules but requires rules to already be mined to construct a non-unique hierarchy, representing the higher-level patterns of rules. In fact, frequency patterns may even be implemented to discover the rules required by our method to find higher-level patterns. In a completely different approach, rather than try to prune the rules which are similar, we exploit the large number of similar rules to discover a higher-level of patterns.

Attempts to construct hierarchies with association rules have been limited. The only method that exists to our knowledge is similar to a generalized association rule (Nanavati et al., 2001) but requires fixing the consequent, which the authors admit is rather limited. This would be beneficial to problems which emphasize one variable such as a response variable. This is further discussed in the Section 5.3.2.

The remainder of the chapter is outlined as follows. In the next section, we will discuss our contributions including the concept of balance in Section 5.2.1. In Section 5.2.2 we consider how to assess different trees of the same type using some old concepts as well a promising new alternative. In Section 5.3.1 we present a simulated example to illustrate our method and assess various scoring methods. In Section 5.3.2 we apply our method to a microarray data on various strains of Legionella bacteria. We summarize this chapter in Section 5.4. A detailed discussion, including areas of future research, is in Section 6.3 of Chapter 6.
CHAPTER 5. HIERARCHIES

5.2 Methodology

Our methodology is organized into two sections. In Section 5.2.1, we introduce new terms for convenience and briefly describe how the concept is used for scoring. Section 5.2.2 is devoted to various methods of assessing the new method we propose.

5.2.1 Balance

We will say that a set of items, say $X$, balances another itemset, say $Y$, if there are interesting rules where $X$ and $Y$ always appear on opposite sides of the rule. That is, when $X$ is a subset of the antecedent and $Y$ is a subset of the consequent in a rule, or $Y$ is a subset of the antecedent and $X$ is a subset of the consequent in a rule. More rigorously, we can define some function which quantifies the degree to which $X$ balances $Y$ such as the number of mined rules where $X$ balances $Y$ or a function of the interestingness measures of each of the rules involving $X$ and $Y$. Let $b(X,Y)$ be this function of $X,Y \subset I$ that quantifies the degree to which $X$ balances $Y$. Here, we will use the sum of the values lift for every rule where $X$ and $Y$ appear on opposite sides of the rule. The more rules there are, and where the value of lift is high, indicating a strong relationship between $X$ and $Y$, the higher $b(X,Y)$ will be. Note that we have defined balance symmetrically here and will only consider this case but some applications may warrant the more general, non-symmetric concept of balance. Another important application would be temporal association rules (see Roddick et al., 2002, for a summary), especially evolution rules.

We will construct a hierarchy by finding disjoint sets of items which tend to balance the same items and group them together. The resulting taxonomy thus summarizes the mined
association rules to create generalizations where possible. This is an important advantage to traditional association rule mining as the common method of mining association rules often focuses on co-occurrences. Consider the case where $A_1$ and $A_2$ separately balance $B$ but $P(A_1A_2)$ is low, perhaps even 0. The relationship between $A_1$ and $A_2$ may only be highlighted when $P(A_1, A_2)$ is sufficiently high. However, finding sets of items which balance common items will highlight groups such as $\{A_1, A_2\}$ regardless if the items within the sets themselves are positively correlated. That is, the rule $\{A_1 \text{ or } A_2\} \Rightarrow \{B\}$ can be found to be interesting. This uses the strengths of ontological ideas (Mansingh et al., 2011) which are both hierarchical and associative. To find the items that would be grouped requires employing a hierarchical clustering algorithm.

Given a list of rules, we use a matrix to represent the degree of balance between the corresponding items from the row and column. Because we defined balance symmetrically, this matrix will also be symmetric. Let $u_i$ denote the $i^{th}$ row of this matrix, where $i = 1, 2, \ldots, |I|$. We then use the rows of this matrix and calculate a distance matrix. If items $X_1$ and $X_2$ have a similar degree of balance for every item, then distance between the two, $\|u_i - u_j\|$, will be very small, whereas if the two have differing degrees of balance for some items, the distance between the two increase. Choosing the Euclidean distance,

$$\text{dist}(X_1, X_2) = \|u_1 - u_2\| = \sqrt{\sum_{Y \in I \setminus \{X_1, X_2\}} \left(e^{-b(X_1, Y)} - e^{-b(X_2, Y)}\right)^2}.$$ (5.1)

The distances are then used by a hierarchical clustering algorithm which provides a hierarchy of the groupings of the data. We use agglomerative hierarchical clustering, which begins with every item as a leaf and successively groups items together pairwise. This is more similar to the approach of BHC and the BRT, than a divisive clustering algorithm (such as
diana in Kaufman and Rousseeuw, 1990), which begins with all items in a single group and successively divides non-singleton clusters. The clustering algorithm ends when only single hierarchy remains and is ready to be scored. Methods of scoring are discussed in the next section.

5.2.2 Scoring

If a taxonomy is unknown, it is important to at least consider other trees that may better represent the rules, especially if the taxonomy will be used for the methods discussed earlier. This can be calculated several ways. One method is to find the rules suggested by the constructed taxonomy, as described in Srikant and Agrawal (1997), ignoring the existing rules. A faster method would be to try to generalize the existing rules, pruning rules where antecedents and consequents begin to match other generalized association rules. There are drawbacks to this. One problem is the possibility that rules which were undiscovered because of low supports but are interesting generalized rules would remain undiscovered. Adjusting the initial mining method through changed thresholds can avoid such situations but at great computational cost. In either scenario, the issue of assessing the mass of rules remains, so we propose a scoring method. This would exploit the BRT’s likelihood function, a mixture model, to obtain a score for the tree that is fair to compare to other such scores from other trees. This would require that a density of a collection be defined. One major difference to the BRT is that we are not going to group transactions (observations), but items (variables). We consider several possible solutions. One is to use the binomial distribution, as the database is binary data. An extension to this is to use a beta-binomial which allows a model to incorporate the uncertainty that may exist with supports but as we see this poses other
challenges. We also propose an alternative method of scoring the tree which does not use a function that is strictly a density, but another function which serves the same purpose and provides promising results.

Equation (2.10) makes finding the likelihood function relatively simple when a marginal density, $f$, is defined. Using the binomial distribution for $f$ is a convenient choice. Given that the number of transactions in the database is set to be fixed, say $n$, and the support of a set of items is $k/n$ for $0 \leq k \leq n$, the log-likelihood may be simplified because we are evaluating it at its expected value, as shown in Equation (5.2).

$$\log \left( \binom{n}{k} \left( \frac{k}{n} \right)^k \left( \frac{n-k}{n} \right)^{n-k} \right) = g(n) - g(k) - g(n-k) \quad \text{where} \quad g(x) = \log(x!) - x \log(x)$$

(5.2)

Rather than using the discrete binomial density, we may assign a prior distribution to the support, namely the beta distribution, conjugate prior distribution for the binomial distribution. The beta-binomial distribution has parameters $\alpha$ and $\beta$, as shown in Equation (5.3), where $B(\cdot)$ refers to the beta function.

$$f(k; \alpha, \beta) = \binom{n}{k} \frac{B(k + \alpha, n - k + \beta)}{B(\alpha, \beta)}$$

(5.3)

Parameter estimates for $\alpha$ and $\beta$ requires a repeated sampling of many proportions, but a database only exhibits one proportion. To rectify this we propose bootstrapping from the database to create several databases to make several samples, simulating a repeated sampling. Estimates $\hat{\alpha}$ and $\hat{\beta}$ do not exist in closed form in all cases, so it must be estimated through numerical means. Minka (2003) provides several methods for estimating parameters.
CHAPTER 5. HIERARCHIES

from the Dirichlet distribution, the generalization of the beta distribution. In our case, we find \( \hat{\alpha} \) and \( \hat{\beta} \) using these methods, described in Appendix B.1. One issue that is likely to arise is that the leaf items of a node may have a disjunctive support of 1. Such nodes should not be modelled with the beta-binomial distribution as there is no evidence in the data indicating that the support is variable let alone a random variable. Rather, this indicates that the support is constant and the union of items is ubiquitous, so the likelihood function will be fixed to 1 for these computational and philosophical reasons. Without this measure, we would require another technique such as perturbing proportions in parameter estimation.

A final method of scoring using the mixture model framework is by employing the interestingness measure used to construct the tree. Let \( \{ x_i \} \) denote the leaves, which are items, of a subtree, \( T \), in the tree, and let \( \bar{r}(x_i) \) denote the mean interestingness measure of rules involving \( x_i \) in the mining step. Let \( d(x_i) \) represent the distance of a node to its leaf, \( x_i \). A parent is one unit away from one of its immediate children, two units away from its children’s children, etc.. Using these, we define the score of a non-leaf node, \( T \), in place of the density of the node, shown in Equation (5.4). Note that the value of \( f \) for a leaf node is defined to be 1.

\[
f(T) = \sum_{i} \frac{1}{d(x_i)} \bar{r}(x_i)
\] (5.4)

Essentially, \( f \) is higher when it possesses leaves involved in strong rules, discounted for the height of the node which will naturally incur more rules the higher it is. The distance \( d \) could also be substituted with the number descendants or the number of leaves of the node as they are all generally proportional to one another. These alternate choices may be of greater importance with special cases of trees, such as trees with many levels like a binary tree or few levels like when nodes have many children. However, it will be sufficient to consider distances for our purposes. Another method of scoring we can consider penalizes rules that
are not entirely contained in the subtree when calculating the average interestingness measure of a node. Using this method is shown in Equation (5.5) while using both proportions and distance penalties simultaneously is shown in Equation (5.6).

\[ f(T) = \sum_{i} \bar{r}_p(x_i) \]  

\[ f(T) = \sum_{i} \frac{1}{d(x_i)} \bar{r}_p(x_i) \]  

Here, \( \bar{r}_p(x_i) \) denotes the average penalized interestingness measure of rules containing the leaf \( x_i \). Only a proportion of each rule’s interestingness measure contributes to the score. The proportion corresponds to the proportion of items in the rule which fall under node \( T \).

In the case that every item of the rule is a leaf of \( T \), the proportion is 1 and so the full value of the interestingness measure contributes. A rule composed partly of items that do not fall under \( T \) will be penalized by a fraction between 0 and 1 to reduce the contribution of the interestingness measure. Thus, higher scores result when subtrees represent entire rules well.

The penalty factor can approach, but will always remain above 0 as there must be some leaf, \( x_i \), under \( T \) to elicit the calculation for \( \bar{r}_p(x_i) \).

There are some considerations to employing these choices of \( f \). First, we assume that the value of the interestingness measure is positive with higher values indicating greater interestingness. Using support, confidence, or lift, for instance, works quite well but an interestingness measure which can attain negative values requires shifting the scale so that \( f \) remains non-negative like a density. Second, the fitness of the node is dictated by the interestingness measure so any advantages or disadvantages of an interestingness measure can also be inherited by \( f \). Despite these drawbacks, however, we have the advantage that the actual value of an interestingness measure rather than just rule supports are used to assess quality.
of a node in a tree unlike the binomial or beta-binomial cases. We will use the lift of a rule to indicate interestingness in our scoring methods. It resembles the likelihood ratio used in the BRT, and like densities, remains non-negative and globally unbounded from above. We expect that for these reasons, it will be a suitable replacement for the likelihood in our analysis.

5.3 Examples

We will now present two examples to serve two distinct purposes. First, we will introduce a simulation of a small gas bar whose transactions primarily consist of fuel purchases but also the purchase of other items. Because we will construct this simulation, we can explore the effects of different choices in our method and see how it affects our results. In a real example, we take genomic data from a study of the pathogenicity of strains of bacteria related to Legionnaires’ disease.

The branch lengths in our diagrams are not informative in the method that we proposed. However, one advantage to using a clustering method which constructs a dendrogram is that small branch lengths between splits that are close together may indicate a potential improvement in a mixture-model scoring by collapsing two splits into one, introducing non-binary splits in the tree. In our first example, we will ignore heights as they will be not be informative in such a small scale. In our second example, however, we will take advantage of the heights provided by dendrograms to suggest where non-binary structures exist and explore those trees.
5.3.1 Gas Bar

In a simulation of a gas bar, we construct 1550 transactions which involve purchases of fuel, beverages, snacks, and chewing gum in various combinations. The frequency distribution of the purchases are shown in Table 5.1. There are two choices of beverage and four choices of snack but no more than one item from each category appears in a transaction. The choice between the two beverages or between the four snacks was determined randomly from the discrete uniform distribution.

<table>
<thead>
<tr>
<th>fuel</th>
<th>gum</th>
<th>a beverage</th>
<th>a snack</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>⬤</td>
<td></td>
<td>⬤</td>
<td></td>
<td>500</td>
</tr>
<tr>
<td>⬤</td>
<td>⬤</td>
<td></td>
<td></td>
<td>350</td>
</tr>
<tr>
<td>⬤</td>
<td></td>
<td>⬤</td>
<td></td>
<td>250</td>
</tr>
<tr>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td></td>
<td>250</td>
</tr>
<tr>
<td>⬤</td>
<td></td>
<td>⬤</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>⬤</td>
<td></td>
<td>⬤</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total 1550</td>
</tr>
</tbody>
</table>

Table 5.1: Frequency distribution of transactions from the gasbar simulation. Two beverages and four snacks are available.

Because we constructed the transactions, the choice of thresholds is not as important as in other applications. We know there will be no supports less than $50/1550$ so we set this to be our minimum support threshold and confidence threshold. After mining, we analyze the discovered rules to create the appropriate similarity matrix using the sum of lift values for our balance function to create dissimilarity through Equation (5.1) and use complete linkage in an agglomerative clustering algorithm to group all items. Other linkage methods to combine or divide clusters such as average distance produced similar results. The structure is shown in Figure 5.1.
Figure 5.1: Trees corresponding to gas bar data in Table 5.1.
As desired, items which appear to balance the same items in rules were grouped together and so our items comprising beverages are together and items comprising snacks are together. Beverages and gas ‘balanced’ snacks the most, as 350 of the transactions included all these items. One feature that appears is that beverages and gas as grouped closer together than beverages and snacks. As there are more transactions with beverages and fuel than there are transactions with snacks and fuel (350 transactions compared to 250 transactions) beverages and gas together behave similarly and balance snacks.

The method for scoring the binomial log-likelihood and the beta-binomial log-likelihood are through the BRT. The value of $\delta$ which maximizes the likelihood is selected so the maximum log-likelihood of one tree is compared against the maximum log-likelihood of another tree, though they may differ in value of $\delta$. Similarly, we employ Equations (5.4), (5.5), and (5.6) recursively through the mixture model to score the trees exploiting interestingness measures. Rather, we compare the scores of different trees with the same model. Note that it is inappropriate to compare a log-likelihood of a binomial to the log-likelihood of a beta-binomial or even scores using (5.5) to scores using (5.6) so comparisons must be made using the same scoring method.

We adjust the existing binary tree to be non-binary by collapsing the snacks together into a single group rather than maintaining the binary cascades and observe that four scoring methods prefer the reduced trees. The exception preferred the suggested initial binary tree which does a good job of grouping items into a high-level pattern. Of the trees discovered, the more sensible tree is favoured over a random one, both when the tree was binary and when it was not. The results are shown in Table 5.2. Our proposed scoring methods choose the non-random trees over the random ones, and methods from (5.4) and (5.6) choose the true structure while (5.5), which does not use a distance penalty at all, favoured the binary
Using the interestingness measure based scoring methods fare better than using the likelihoods as they only consider support, not the interestingness of the rule, which is what we focus on. From this, it becomes clear that searching through possible tree structures is an important concept as trees which truly describe trees better will result in higher scores and give a sense of how much better some trees are compared to others.

This example can be seen as clustering, using our proposed method of quantifying similarity and then dissimilarity. Though this is clearly useful as is, we wish to show that our method extends beyond that by extending the example to have a designed subtree. We will again have fuel in high proportion, but replace gum with another kind of snack, so that snacks are either one of three kinds of candy, or one of three kinds of chips. We keep several types of beverages as well. The distribution of the items is shown in Table 5.3. We introduce a small number of transactions where both types of snacks are purchased together, but the majority of transactions do not contain both types of snack together.

We find that the tree chosen through a hierarchical clustering is the tree we desire, shown in Figure 5.2. The underlying structure we used to construct transactions has been discovered by our method. This resulting tree could be used by other methods discussed in Section 2.7.
### Table 5.3: Frequency distribution of transactions from the gasbar simulation. Two beverages and four snacks are available.

<table>
<thead>
<tr>
<th>fuel</th>
<th>chips</th>
<th>candy</th>
<th>beverage</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>1000</td>
</tr>
<tr>
<td>●</td>
<td>●</td>
<td></td>
<td>●</td>
<td>160</td>
</tr>
<tr>
<td>●</td>
<td></td>
<td>●</td>
<td>●</td>
<td>160</td>
</tr>
<tr>
<td>●</td>
<td></td>
<td></td>
<td>●</td>
<td>180</td>
</tr>
<tr>
<td>●</td>
<td></td>
<td></td>
<td>●</td>
<td>180</td>
</tr>
<tr>
<td>●</td>
<td></td>
<td></td>
<td>●</td>
<td>60</td>
</tr>
<tr>
<td>●</td>
<td></td>
<td></td>
<td>●</td>
<td>60</td>
</tr>
<tr>
<td>●</td>
<td></td>
<td></td>
<td>●</td>
<td>50</td>
</tr>
<tr>
<td>●</td>
<td></td>
<td></td>
<td>●</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>1890</td>
</tr>
</tbody>
</table>

The best scores generally come from the reduced tree. Using (5.5) resulted in a tie between the initial binary tree and the collapsed tree. A complete summary of results is shown in Table 5.4.

The method by Savasere et al. (1998) which uses taxonomies to find both positive and negative association rules can clearly exploit the tree resulting from our method. We can see that many negative rules, such those including more than one kind of beverage, or more than one kind of snack, can be mined by constructing rules only using sibling leaf items.

We were curious about a variant of the tree shown in Figure 5.2, which illustrates the tree where the fuel item is alone directly under the root so that the root had had 3 children. We labelled this $T'$ in Table 5.4, but is otherwise the same as $T$, which did well. We expected that this tree may not perform as well as $T$ but score nearly as high and much better than the binary tree. Instead, we found that the binary tree outperformed $T'$ in one instance, and in the other two instances the score for $T'$ was closer to the binary tree’s score than the scores for $T$. This was surprising because $T$ and $T'$ are the same except for one branch. We suspect
this came from the near-ubiquity of the fuel item. Its high prevalence means that many rules that do not possess the item achieve similar supports, and hence similar lift values as the rules with the prevalent fuel item. This illustrated a caveat to our method: prevalent items are highly influential in scoring. We address a similar issue in the next example.
5.3.2 Legionnaires’ Microarray Data

Legionnaires’ disease is potentially fatal infection caused by strains of *Legionella* bacteria. There has been a clear rise in the number of cases over recent decades (Joseph et al., 2010) in many travelled parts of the world. Variants of the bacteria responsible have been studied as not all strains of *Legionella* bacteria are pathogens. Yzerman et al. (2010) (Paper I) analyzed 480 DNA markers of 257 different strains collected in the Netherlands and employed genetic programming to predict whether a strain is pathogenic (a clinical strain) with 100% accuracy and 69% specificity using just five markers. The training set was made of 133 strains and the test set made of 213 strains. The output of the genetic programming were seven sets of ‘rules’ which, when at least four statements hold true, indicate a clinical strain, otherwise the strain was labelled environmental. The term ‘rules’ is also used in the context of genetic programming but refers to a different concept than those of association rule mining. To avoid confusion, we will continue to use the term rules in the association rules context and instead use *conditional statements* or *conditions* to refer to those in the genetic programming context.

The conditional statements predicting pathogenicity involve the markers: LePn.011A2-b, LePn.019H4, LePn.010B12, LePn.024B1, and LePn.008D6. The statements are summarized in Figure 5.3, a summary of Table 2 from Paper I.

The genetic algorithm results in statements which are more expressive than those of association rules and that added expressiveness is necessary to obtain such high sensitivity and specificity in determining the pathogenic nature of a strain. This has no bearing on our method as we are not predicting a singular response from the data. Rather, our method is used to discover and to express high-level patterns, a confluence of trait and markers, making
1. If \((\text{LePn.011A2-b} < \text{LePn.019H4})\)
   then \([\neg(\text{LePn.011A2-b} = \text{LePn.010B12})]\)
   else \([\text{LePn.011A2-b} = \text{LePn.024B1}] \land (\text{LePn.008D6} = \text{LePn.010B12})]\)

2. If \((\text{LePn.011A2-b} < \text{LePn.010B12})\)
   then \([\neg(\text{LePn.019H4} < \text{LePn.024B1})]\)
   else \([\text{LePn.008D6} > \text{LePn.011A2-b}] \downarrow (\text{LePn.008D6} < \text{LePn.024B1})]\)

3. If \((\text{LePn.011A2-b} < \text{LePn.010B12})\)
   then \([\neg(\text{LePn.024B1} > \text{LePn.019H4})]\)
   else \([\text{LePn.010B12} < \text{LePn.008D6}] \downarrow (\text{LePn.024B1} > \text{LePn.008D6})]\)

4. If \((\text{LePn.019H4} = \text{LePn.008D6})\)
   then \([\text{LePn.008D6} \leq \text{LePn.019H4}]\)
   else \([\neg(\text{LePn.019H4} < \text{LePn.008D6})] \downarrow (\text{LePn.010B12} \leq \text{LePn.011A2-b})]\)

5. If \((\text{LePn.019H4} = \text{LePn.008D6})\)
   then \([\neg(\text{LePn.024B1} > \text{LePn.019H4})]\)
   else \([\text{LePn.010B12} \leq \text{LePn.011A2-b}] \downarrow (\neg(\text{LePn.024B1} \leq \text{LePn.019H4}))]\)

6. If \((\text{LePn.019H4} = \text{LePn.008D6})\)
   then \([\neg(\text{LePn.024B1} > \text{LePn.019H4})]\)
   else \([\text{LePn.024B1} > \text{LePn.019H4}] \downarrow (\text{LePn.011A2-b} \geq \text{LePn.010B12})]\)

7. If \((\text{LePn.019H4} = \text{LePn.008D6})\)
   then \([\neg(\text{LePn.019H4} < \text{LePn.024B1})]\)
   else \([\neg(\text{LePn.010B12} = \text{LePn.011A2-b})] \downarrow (\text{LePn.019H4} < \text{LePn.024B1})]\)

Figure 5.3: Conditional statements from Paper I based on the binarized microarray data. When at least four of the seven are true, the strain is labelled clinical. When no more than three of the seven statements are true, then the strain is labelled environmental.
ours an inherently different task than predicting pathogenicity. We will, however, include pathogenicity among the DNA markers as it is another trait to analyze.

The binarized data used in the original analysis is also used here. A recorded 1 indicates relatively high expression of the DNA marker in the strain while a recorded 0 indicates relatively low expression of the marker in the strain. For our analysis, we combine the data sets and remove strains with incomplete genomic data leaving 157 strains, 49 of which are pathogenic. To better illustrate our method, we reduce the number of markers to Yzerman’s five indicators, an indicator whether the strain is a pathogen, and 93 other uniformly randomly selected markers for a total of 99 items. As we will see, this is sufficient to illustrate our method’s ability to organize items into a hierarchy based on association rules.

We first mine association rules of lengths two to four. The minimum support threshold is set to be 0.2 and minimum confidence threshold to be 0.95 to obtain over 2.5 million rules. Values of lift range from 0.96 to 4.24 but are heavily right-skewed; the median lift value is 1 and only 3% of rules have a value of lift above 1.5.

We purposely mined all association rules rather than ones which focus on the items indicating pathogenicity. Focusing the mining to include only rules of that form would inflate the occurrence of some items and would make them separate themselves from other items as they would accumulate higher scores. For example, suppose that there was another other item which was a duplicate of the pathogen indicator in every transaction. We would want that item to appear with the pathogen indicator as it is so intimately associated, but if we employed a mining method which forced rules to only mine rules including the indicator and, the indicator would accumulate scores both with and without the duplicate. The indicator
could then be separated from the duplicate item. There is also the additional risk that inflating the number of rules involving the indicator will artificially group it with ubiquitous items making it seem more related to the ubiquitous items than its own duplicate. Finally, as we witness in the last example, the location of a relatively prevalent item significantly influences the score of the entire tree. Exaggerating an item’s presence in rules by focused mining distorts the nature of which rules are mined and can have a significant impact on the scoring. Using unfocused rules allows scores to accumulate more objectively giving a better sense of the relationships and high-level patterns.

One issue we had was deciding which tree to consider. With so many possible permutations of branches, we accepted the trees suggested by the algorithm by taking the hierarchy resulting from the algorithm as the binary tree, and various node groupings based on the heights. The initial tree including heights is shown in Figure B.3 of Section B.2 in Appendix B. We applied a complete-linkage clustering algorithm and modified the resulting hierarchy in several ways to obtain the scores shown in Table 5.5. We did not rearrange any branches, but instead collapsed subtrees partly (Figure 5.4b) or collapsed to make a three-level tree (Figure 5.4c). This third tree is a simple case of a tree which corresponds to groups in conventional clustering.

We found that three-level trees performed worse than binary trees in two cases. The binary tree is optimal when distance penalty is employed, either alone or in conjunction with the
Figure 5.4: Sample trees suggested from hierarchical clustering. Heights have been adjusted to visually distinguish where splits occur, therefore heights are non-informative in the figure. Paper I markers are marked near leaf labels.
proportions penalty. Using only the distance penalty, this score was only slightly higher than
the score obtained in the partly collapsed tree which is significantly more parsimonious. The
same ranking was observed when using Equation (5.5). The partially-collapsed tree was opti-
mal when only proportions were used as a penalty. In the cases where the proportion penalty
was used, the differences between trees were more significant. An interesting note is that the
same score is achieved with the three-level tree using either method where proportions are
involved in scoring; Equations (5.5) and (5.6). This results from the tree only containing two
distinct levels (excluding the root-node level) so distances provided no penalty from the leaf
nodes to their parents. Once again, this case reflects traditional clustering, which clearly
does not correspond to an optimal hierarchy. Before our analysis, we preferred the case
which had fewer binary cascades and ‘fuller’ trees for the aforementioned reasons but this
data set could be a case where the underlying hierarchy we are trying to model is truly more
binary. It is also possible that the true underlying structure is not binary, but more binary
than the partially collapsed tree we suggested, or even a different tree with different branches
that we explored. In the instance that some structures which are more binary are preferred
it can be difficult to search through the possible permutations of binary cascades because
there would be so many and they would be very similar to each other among different trees.
Even without binary cascades, there are still many non-binary trees to search through. Our
method would prefer the less binary trees in many cases but if the constructed tree were
used for another purpose requiring the taxonomy, then desirable properties of the tree could
guide the tree selection process.
5.4 Summary

In this chapter, we have outlined a method of summarizing many association rules. In our first example, we showed how our method of summarizing association rules discovers the underlying structure of the data itself, which were not summarized using association rules alone. In our second example we applied our method to a real data set to illustrate how several hierarchies can be constructed and in doing so understood the need for an efficient search-and-score algorithm among different hierarchies. Using many of the concepts from the BRT, we provided a means of scoring the trees and showed how the scores we presented in this chapter captures features of the tree we believe to be more desirable. The context of the problem would guide the investigation further. These further steps may be any of the existing methods (see Section 2.7) which require a taxonomy but previously had no means of deriving one.
Chapter 6

Conclusion

6.1 Standardizations

As we have shown through several examples, standardizing interestingness measures serves several purposes which are not met by raw interestingness measures alone. We have also investigated the behaviour of standardized interestingness measures in different data sets. In the Belgian accidents data, we observed that high values of one standardized interestingness measure correspond to high values of another standardized interestingness measure. Comparing the relative order of rules of a raw against standardized interestingness measure resulted in rules being reordered to various degrees, depending on the interestingness measures; some interestingness measures reordered rules considerably while others maintained much of the original order. In the Reuters-21578 data, high values of both the Gini index and standardized Gini index related to relatively high values of the standardized interestingness measures and some of the raw interestingness measures. Exceptions coincide with the nature
Gini index scoring high for very negatively correlated data as well as very positively correlated data. The relative order of rules also changed dramatically with these data, such as when standardizing lift and the cosine similarity. We once again observed greater similarities between standardized interestingness measures than between raw interestingness measures. We also highlighted an example when an interestingness measure took on one of the smallest possible values on its range. The value of the interestingness measure indicated that the antecedent and consequent were very close to being independent, but the standardized interestingness measure was quite high indicating that the value obtained was nearly as far from independence as bounds permit. In the last example we observed how standardizing interestingness measures are affected by data designed to have no pattern. Although three interestingness measures indicate some form of relationship, standardizing interestingness measures typically resulted in values near zero. Only in the case of $Q^*$ were values above 0.1 even observed. The order of rules by raw interestingness measures generally similar to the standardized interestingness measure.

In the case of the random-transactions, standardizing interestingness measures shows how data with no patterns does not result in an interestingness measure that is misleadingly interesting whereas the raw interestingness measure alone may have indicated patterns. Standardized interestingness measures often appear to capture the same aspect of a rule by providing similar values; plots in Appendix A.2 show how standardized interestingness measures are more similar to each other than raw interestingness measures. In the cases of $L^*$ and $Q^*$, independent rules do not evaluate to a single value but $G^*$ often maps independent rules to zero like its raw counterpart; this is not surprising as this is the lower bound of the interestingness measure. Random transactions, on the other hand, showed standardized interestingness measures of rules near zero in most cases.
One may further study the effects of standardizing interestingness measures on specific properties by considering two families of interestingness measures: one containing interestingness measures possessing the property and the other containing interestingness measures that do not possess the property. A thorough evaluation of several data sets with known patterns and features with members from each of the two families can address more specific aspects of the impact of standardizing interestingness rules.

### 6.2 Incomplete Transactions

We introduced several important methods in Chapter 4 for handling incomplete transactions when mining association rules. It is clear that asymptotic properties of data are often inappropriate for inferences. When possible, the analysis of the nature of missing data and how it relates to existing data or even other missing attributes should be incorporated in an investigation. We have provided a method, through Fisher’s exact test, of identifying these relationships. We showed how hypothesis tests, which have traditionally been used by exploiting the asymptotic properties, simply do not hold and are not reliable for many kinds of data. We also showed how an interestingness measure which quantifies the degree of departure from independence is also inappropriate for this kind of inference.

Once a relationship between missing data and observed data has been considered, mining rules with non-missing items is facilitated through a new model, which we presented. Under the fairly reasonable assumption that transactions are independent of one another, we constructed a model for the missing data whence we can derive our traditional interestingness measures up to a distribution and consistently provided better estimates for the value of lift of
the rule than the complete-transaction estimate. We provided suggestions for the parameters of the model which correspond to data MAR and our method still performs well even when the assumptions are violated. However, some extensions to our method can be made to suit different situations including relationships between proportions of the underlying missing binary data. The case of data which are MAR is well suited in our model but the probabilities can be adjusted if other information, such as some kind of prior, was known. Other means of calculating the probability parameters can create a more flexible model by making them more than just the constants we suggested. Sampling theory provides many of the tools necessary for this task including blocking, for which our model is well designed, and post-stratification. Data which are MNAR are difficult to account for in various cases but if some mechanism for missing data were known, it would most easily be incorporated in the model through the probability parameters. This added complexity of the model would be necessary to handle data which are MNAR and to create a model which incorporates the missingness.

The model we presented is a collection of five distributions connecting twelve unobserved random variables. These five distributions are marginals of a broader joint distribution. This broader joint distribution is what is required to find the true expected value of variables such as confidence, lift, or other another interestingness measure and may be required for a model when data are MNAR. If the distributions are truly independent of one another they may be multiplied together to make this broader joint distribution. If, however, these five distributions harbour some dependence, another method of combining marginal distributions is required. Even though we have shown our method performs well, a suitable extension to our method would be to construct a joint distribution incorporating this dependence. Copulas are a way to tie marginal distributions together and explicitly states the dependence structure.
6.3 Association Rule Hierarchies

Previous efforts have outlined methods of mining association rules given a hierarchy. Little has been devoted to determining that hierarchy for association rules, however, so that those efforts could be utilized in more situations. We have proposed a suitable method of finding such a hierarchy by addressing the related issue of a large number of association rules. When association rules are mined, the mass of patterns that result is often so large that those rules themselves require some kind of analysis. We have adapted several past contributions and introduced new ideas to construct an ontology that is suitable for providing a higher-level structure to the resulting association rules. We have shown that our new method of organizing association rules leads to useful results and overcomes some of the challenges faced in association rule mining. While other methods discard one rule for another, we have taken advantage of the quantity of similar rules. We use all of the discovered rules to find relationships between items based on the strengths of rules, and by doing so, learn higher-level information than what association rules alone provide. This higher-level pattern thus becomes the hierarchy that other techniques require.

In our gasbar example we show that our method does indeed construct hierarchies which truly represent the higher-level patterns. We did this by exploiting a scoring method compatible with a mixture model representations of trees. Of the methods we have analyzed, we see that the proposed scoring functions (5.4), (5.5) and (5.6) were more promising than the likelihoods modelling supports. This was in part because we directly used interestingness measures rather than just the supports and focused on constructing distances based on the new concept of balance.

One example came from real genomic data on strains of *Legionella* bacteria. Our analysis
resulted in a tree structure, including a subtree structure, which identifies predictive markers in other analysis as being similar to pathogenicity according to our new method. We saw some evidence that a more binary structure is suitable, and there may be many very similar such structures to choose from. In such a case, the context of an actual problem would suggest what properties of a tree are more desirable than others in a selection process.

Our new method is promising and opens possibilities for further exploration. The first clear extension is the variety of ways to construct a balance function. In different steps, we chose functions which were sensible in our analysis and worked very well but other functions and worthy of exploration. The construction of the tree used a hierarchical clustering method with a purposeful choice of a distance metric. We have shown that our choice in clustering method and distance works well but variants of either may also be suitable. As we have shown, we can also exploit the heights from the hierarchical method we used, as long branches can indicate a clear subtree structure. Though we did not witness such subtrees, we can conceivably encounter trees where subtree structures are less clear. In such situations alternative methods are useful to consider. Ward’s method (Ward, 1963) in particular, may identify where subtrees should collapse or remain intact if such ambiguity arose.

We selected some trees over others using a scoring method in conjunction with heights from dendrograms with good results. However, if many trees are being investigated, a natural extension to our method is to have an appropriate search among many trees. The method we have proposed lends itself well to a genetic search-and-score algorithm because some subtrees improve scores and would supposedly occur in many other high-scoring trees. In a genetic algorithm, these subtrees would likely survive a selection process from one generation to the next. Many search-and-score algorithms exist, and many of them would take advantage of the recursive nature of trees, and the similarities between trees, further supporting our
The mixture model expression of trees leads to several related extensions. The definition of mixing proportions leads Equation (2.10) to be a recursive polynomial function of $\delta$. This may be estimated if its form were simplified by being fixed, rather than calculating the score for different parameter values. A simplified expression could suit some applications. This extension would be reflected in trees having some specified type of structure. Trees would share some common traits, excluding trees missing these traits, and therefore, reduce the space of possible trees. This is would benefit search-and-score algorithms. Finally, incorporating model-based clustering would be interesting to investigate. A variety of mixture models, including those which use ‘soft’ group memberships, could yield further extensions to our new method.
Bibliography


Han, J., J. Pei, and Y. Yin (2000, May). Mining frequent patterns without candidate generation. *SIGMOD Rec.* 29, 1–12. (Cited on pages 8 and 67.)


Lang, D. T. (2010). *Rstem: Interface to Snowball implementation of Porter’s word stemming algorithm*. R package version 0.4-0. (Cited on page 38.)

Lewis, D. D. *Reuters-21578*. (Cited on page 38.)


Tan, P.-N., V. Kumar, and J. Srivastava (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD ’02, New York, NY, USA, pp. 32–41. ACM. (Cited on pages 9, 10, and 28.)


Appendix A

Standardizations

A.1 Derivations

Before deriving any inequalities, denote the following:

- $\sigma$ is the minimum support threshold. When not explicitly given, $\sigma = \frac{1}{n}$.
- $\kappa$ is the minimum confidence threshold. When not explicitly given, $\kappa = \frac{1}{n}$.
- $m = \min \{P(A), P(B)\}$
- $M = \max \{P(A), P(B)\}$
A.1.1 Cosine Similarity

The lower bounds of $C$ are obtained by considering several cases. First, because $P(A, B) \geq \sigma$,

$$ C \geq \frac{\sigma}{\sqrt{P(A)P(B)}}. $$

Because $P(A, B) \leq m$,

$$ C = \frac{P(A, B)}{\sqrt{P(A)P(B)}} = \frac{P(A, B)}{\sqrt{mM}} \leq \frac{m}{\sqrt{mM}} = \sqrt{\frac{m}{M}}. $$

We obtain another by considering that $P(A \cup B) = P(A) + P(B) - P(A, B) \leq 1$, and so $P(A, B) \geq P(A) + P(B) - 1$, making

$$ C \geq \frac{P(A) + P(B) - 1}{\sqrt{P(A)P(B)}}. $$

By considering support threshold $\sigma$, if $P(A, B) \geq \sigma$, the cosine similarity is smallest if $P(A) = P(B) = \frac{1}{2}(1 - \sigma)$ giving

$$ C \geq \frac{2\sigma}{1 + \sigma}. $$

If a confidence threshold, $\kappa$, is employed, then $\frac{P(A, B)}{P(A)} \geq \kappa$ and it follows that

$$ C \geq \frac{P(A, B)}{\sqrt{\kappa \frac{P(A, B)}{P(A)P(B)}}} = \sqrt{\kappa \frac{P(A, B)}{P(B)}}. $$

and since $P(A, B) \geq \sigma$, we have

$$ C \geq \sqrt{\kappa \frac{\sigma}{P(B)}}. $$

103
APPENDIX A. STANDARDIZATIONS

Also, because \( P(A, B) \geq \kappa P(A) \),

\[
C \geq \frac{\kappa P(A)}{\sqrt{P(A)P(B)}} = \kappa \sqrt{\frac{P(A)}{P(B)}}.
\]

Upper bounds of \( C \) come from considering that \( P(A, B) \leq m \) and \( P(A)P(B) = Mm \), giving

\[
C \leq \sqrt{\frac{m}{M}}.
\]

A.1.2 Yule’s Q

Yule’s Q is given by

\[
Q = \frac{P(A, B)P(\overline{A}, \overline{B}) - P(A\overline{B})P(\overline{A}B)}{P(A, B)P(\overline{A}, \overline{B}) + P(A\overline{B})P(\overline{A}B)}.
\]

We will reexpress the numerator and denominator separately. The numerator is

\[
P(A, B)P(\overline{A}, \overline{B}) - P(A\overline{B})P(\overline{A}B)
\]

\[
= P(A, B) \left[ 1 - P(A) - P(B) + P(A, B) \right] - \left[ P(A) - P(A, B) \right] \left[ P(B) - P(A, B) \right]
\]

\[
= P(A, B) - P(A, B)(P(A) + P(B) - P(A, B)) - P(A)P(B) + P(A, B)(P(A) + P(B) - P(A, B))
\]

\[
= P(A, B) - P(A)P(B)
\]

(A.1)
\textbf{APPENDIX A. STANDARDIZATIONS}

\[ P(A, B)P(\overline{A}, \overline{B}) + P(\overline{A}B)P(\overline{A}B) \]
\[ = P(A, B) [1 - P(A) - P(B) + P(A, B)] + [P(A) - P(A, B)] [P(B) - P(A, B)] \]
\[ = P(A, B) - P(A, B)(P(A) + P(B) - P(A, B)) + P(A)P(B) - P(A, B)(P(A) + P(B) - P(A, B)) \]
\[ = P(A, B) + P(A)P(B) - 2P(A, B)(P(A) + P(B) - P(A, B)) \quad (A.2) \]

Thus,

\[ Q = \frac{(A.1)}{(A.2)} = \frac{P(A, B) - P(A)P(B)}{P(A, B) + P(A)P(B) - 2P(A, B)(P(A) + P(B) - P(A, B))} \]

Yule’s Q achieves the maximum value of 1 iff \( P(A, B) = \min\{P(A), P(B)\} \) and achieves the minimum value of -1 only when \( P(A, B) = 0 \) or when \( P(\overline{A}B) = 0 \). It is a strictly monotonically increasing function of \( P(A, B) \) so only these endpoints will be considered.

Given the minimum support threshold \( \sigma \), \( P(A, B) \geq \sigma \) so the minimum value of -1 is not attained. If, however, \( P(A, B) = \sigma \), then Yule’s Q attains the value

\[ \frac{\sigma - P(A)P(B)}{\sigma + P(A)P(B) - 2\sigma(P(A) + P(B) - \sigma)} \quad (A.3) \]

Given the minimum confidence threshold \( \kappa \), because \( P(A, B) = P(B \mid A)P(A) \), we know that \( P(A, B) > \kappa P(A) \) and so Yule’s Q is greater than the value.

105
From (A.3) and (A.4), the lower bound of Yule’s Q is given by

\[
\lambda = \max \left\{ -1, \frac{\sigma - P(A)P(B)}{\sigma + P(A)P(B) - 2\sigma(P(A) + P(B) - \sigma)}, \frac{\kappa - P(B)}{\kappa + P(B) - 2\kappa(P(A) + P(B) - \kappa P(A))} \right\}.
\]
A.1.3 Gini

Derivation of an alternative form of the Gini index.

\[
G = P(A) \left[ P(B \mid A)^2 + P(\overline{B} \mid A)^2 \right] + P(\overline{A}) \left[ P(B \mid \overline{A})^2 + P(\overline{B} \mid \overline{A})^2 \right] - P(B)^2 - P(\overline{B})^2
\]

\[
= P(A) \left[ \frac{P(A, B)^2}{P(A)^2} + \frac{P(A, \overline{B})^2}{P(A)^2} \right]
\]

\[
+ P(\overline{A}) \left[ \frac{P(\overline{A}, B)^2}{P(\overline{A})^2} + \frac{P(\overline{A}, \overline{B})^2}{P(\overline{A})^2} \right] - P(B)^2 - (1 - P(B))^2
\]

\[
= \left[ \frac{P(A, B)^2 + P(A, \overline{B})^2}{P(A)} \right] + \left[ \frac{P(\overline{A}, B)^2 + P(\overline{A}, \overline{B})^2}{P(\overline{A})} \right] - 2P(B)^2 + 2P(B) - 1
\]

\[
= \left[ \frac{(P(A, B) + P(A, \overline{B}))^2 - 2P(A, B)P(A, \overline{B})}{P(A)} \right] - 2P(B)^2 + 2P(B) - 1
\]

\[
= \left[ \frac{(P(A))^2 - 2P(A, B)P(A, \overline{B})}{P(A)} \right] + \left[ \frac{(P(\overline{A}))^2 - 2P(\overline{A}, B)P(\overline{A}, \overline{B})}{P(\overline{A})} \right] - 2P(B)^2 + 2P(B) - 1
\]

\[
= P(A) + P(\overline{A}) - 1 - 2P(B)^2 + 2P(B) - 2 \frac{P(A, B)P(A, \overline{B})}{P(A)} - 2 \frac{P(\overline{A}, B)P(\overline{A}, \overline{B})}{P(\overline{A})}
\]

\[
= -2P(B)^2 + 2P(B) - 2 \frac{P(A, B)[P(A) - P(A, B)]}{P(A)} - 2 \frac{P(\overline{A}, B)[P(\overline{A}) - P(\overline{A}, B)]}{P(\overline{A})}
\]

\[
= -2P(B)^2 + 2P(B) - 2 \left[ \frac{P(A, B)[P(A) - P(A, B)]}{P(A)} - \frac{P(\overline{A}, B)[P(\overline{A}) - P(\overline{A}, B)]}{P(\overline{A})} \right]
\]

\[
= -2P(B)^2 + 2P(B) - 2 \left[ P(A, B) - \frac{P(A, B)^2}{P(A)} + P(\overline{A}, B) - \frac{P(\overline{A}, B)^2}{P(\overline{A})} \right]
\]

\[
= -2P(B)^2 + 2P(B) - 2P(\overline{A}, B) + 2 \left[ \frac{P(A, B)^2}{P(A)} + \frac{P(\overline{A}, B)^2}{P(\overline{A})} \right]
\]

\[
= -2P(B)^2 + 2P(B) - 2 \left[ \frac{P(A, B)^2}{P(A)} + \frac{[P(B) - P(A, B)]^2}{1 - P(A)} \right]
\]
APPENDIX A. STANDARDIZATIONS

\[
-2P(B)^2 + 2 \left[ \frac{P(A,B)^2[1-P(A)] + P(A)[P(B)^2 - 2P(B)P(A,B) + P(A,B)^2]}{P(A)(1-P(A))} \right]
\]

\[
= -2P(B)^2 + \left[ \frac{2}{P(A)(1-P(A))} \times [P(A,B)^2 - P(A)P(A,B)^2 + P(A)P(B)^2] \right]
\]

\[
= \frac{-2P(B)^2P(A)(1-P(A))}{P(A)(1-P(A))} + 2 \left[ \frac{P(A,B)^2 + P(A)P(B)^2 - 2P(A)P(B)P(A,B)}{P(A)(1-P(A))} \right]
\]

\[
= 2 \left[ \frac{-P(B)^2P(A) + P(B)^2P(A)^2 + P(A,B)^2 + P(A)P(B)^2 - 2P(A)P(B)P(A,B)}{P(A)(1-P(A))} \right]
\]

\[
= 2 \left[ \frac{P(B)^2P(A)^2 + P(A,B)^2 - 2P(A)P(B)P(A,B)}{P(A)(1-P(A))} \right]
\]

\[
= 2 \frac{[P(A,B) - P(A)P(B)]^2}{P(A)(1-P(A))} \quad (A.5)
\]

\[
= \frac{2}{1-P(A)} [P(B \mid A) - P(B)][P(A,B) - P(A)P(B)] \quad (A.6)
\]

Using (A.5), we know immediately that the (global) lowerbound of zero is achieved iff \(A\) and \(B\) are independent. When \(A\) and \(B\) are not independent, the Gini index is non-zero.

Using Expression (A.5), we can write the Gini index as the product

\[
G = \frac{2}{P(A)(1-P(A))} \times \left( P(A,B) - P(A)P(B) \right)^2.
\]

The first factor is largest when \(P(A) = \frac{1}{2}\), making it smaller than 8. The second term can be no larger than \(\frac{1}{16}\), precisely when \(A\) and \(B\) are most positively correlated. Let \(z = P(A) = P(B) = P(A,B)\) in this instance.

\[
(P(A,B) - P(A)P(B))^2 \leq (z - z^2)^2 \leq \left( \frac{1}{2} - \left( \frac{1}{2} \right)^2 \right)^2 = \frac{1}{16}
\]
APPENDIX A. STANDARDIZATIONS

From these deductions, we know that

$$G \leq 8 \times \frac{1}{16} = \frac{1}{2} \quad (A.7)$$

We may obtain other inequalities by considering the highest and lowest values of $P(A, B)$ influences the numerator of (A.5). Let $l = \max\{\sigma, P(A) + P(B) - 1, \kappa P(A)\}$. From Section A.1.1, $l \leq P(A, B)$ and $P(A, B) \leq m$ as before.

Because of the quadratic nature of the numerator, we must be slightly more cautious. When $P(A, B) - P(A)P(B) \geq 0,$

$$2 \frac{(l - P(A)P(B))^2}{P(A)(1 - P(A))} \leq G \quad (A.8)$$

When $P(A, B) - P(A)P(B) > 0$, the above inequalities are reversed to obtain

$$2 \frac{(m - P(A)P(B))^2}{P(A)(1 - P(A))} \leq G \quad (A.9)$$

Considering when $P(A, B) < P(A)P(B)$, we can also employ $P(A, B) \geq \kappa P(A)$ to obtain

$$G \geq 2 \frac{(P(A, B) - P(A)P(B)))(\kappa - P(B))}{1 - P(A)}$$

and when $P(A, B) > P(A)P(B)$, we find that

$$G \geq 2 \frac{(P(A, B) - P(A)P(B))(1 - P(B))}{1 - P(A)}.$$
A.2 Scatterplot Matrices

(a) Raw Interestingness Measures

(b) Standardized Interestingness Measures

Figure A.1: Scatterplot matrices of interestingness measures for the Belgian accident data.
Figure A.2: Scatterplot matrices of interestingness measures for the Reuters data.
APPENDIX A. STANDARDIZATIONS

Figure A.3: Scatterplot matrices of interestingness measures for random transactions.
A.3 Tau-b by Deciles

Figure A.4: Tau-b by deciles for each data set.
Appendix B

Hierarchical

B.1 Dirichlet Parameter Estimation

The database only provides one true proportion. Unless the databases are split into several smaller databases, this poses a problem in estimating parameters. We suggest building many alternate databases by bootstrapping from the transactions of the database and proceed with the problem of finding parameter estimates. We can use an existing method for the Dirichlet distribution which finds parameter estimates, which is what required that provided that many proportions are available. The estimates are found iteratively. At iteration \( k + 1 \), the estimate for parameter \( \alpha_i^{(k+1)} \), \( i = 1, 2 \) is

\[
\hat{\alpha}_i^{(k+1)} = \psi^{-1} \left( \psi \left( \sum_i \alpha_i^{k+1} + \log(p_i) \right) \right),
\]
where \( \psi^{-1} \) is the inverse of the digamma function. Minka (2003) also outlines a method for estimating \( \psi^{-1} \).

In our scenario, we are only using the beta-binomial so \( \alpha_1 \) and \( \alpha_2 \) for this method correspond to \( \alpha \) and \( \beta \) in

\[
f(x) = \frac{1}{\mathrm{B}(\alpha, \beta) x^{\alpha-1} (1-x)^{\beta-1}},
\]

the beta distribution. The method iterates until the likelihood of the beta-binomial converges. For convergence we employ a method introduced by Böhning et al. (1994) based on Aitken’s acceleration (Aitken, 1926), which stops iterating when the estimate of where the likelihood is moving toward is sufficiently close to the likelihood at the current iteration. Specifically, the algorithm stops when

\[
l^{(k+1)}_\infty - l^{(k)} < \epsilon
\]

for some small \( \epsilon > 0 \). Here, \( l^{(k)} \) denotes the log-likelihood at iteration \( k \), \( l^{(k+1)}_\infty \) is the estimate of where the likelihood is moving toward and is given by

\[
l^{(k+1)}_\infty = l^{(k)} + \frac{1}{1 - a^{(k)}} \left( l^{(k+1)} - l^{(k+1)} \right)
\]

where

\[
a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}}.
\]
APPENDIX B. HIERARCHICAL

B.2 Dendrograms

Figure B.1: The original hierarchy produced using the \texttt{hclust} algorithm leading to the trees in Figure 5.1 and scores in Table 5.2.
Figure B.2: The original hierarchy produced using the \texttt{hclust} algorithm leading to the tree in Figure 5.2 and scores in Table 5.2.
Figure B.3: The original hierarchy produced using the \texttt{hclust} algorithm. Nodes above, below, or both above and below 3.5 were collapsed to create the trees shown in Figure 5.4.