Rare Association Rule Mining and Knowledge Discovery:
Technologies for Infrequent and Critical Event Detection

Yun Sing Koh
Auckland University of Technology, New Zealand

Nathan Rountree
University of Otago, New Zealand

Information Science
REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Detailed Table of Contents

Section 1
Beyond the Support-Confidence Framework

The notion of finding rare association rules is like finding precious gems in an open field; it is a daunting task but, if successful, it is very rewarding. Association rule mining systems, such as Apriori, generally employ an exhaustive search algorithm. While these algorithms are in theory capable of finding rare association rules, they become intractable if the minimum level of support is set low enough to find rare rules. Such algorithms are therefore inadequate for finding rare associations, and also suffer from the rare item problem. Research to solve this problem has become more prevalent in recent times. The main goal of rare association rule mining is to discover relationships among sets of items in a transactional database that occur infrequently. This chapter presents a survey on the current trends and approaches in the area of rare association rule mining.

Association rule mining among frequent items has been extensively studied in data mining research. However, in recent years, there is an increasing demand for mining infrequent items (such as rare but expensive items). Since exploring interesting relationships among infrequent items has not been discussed much in the literature, in this chapter, the authors propose two simple, practical and effective schemes to mine association rules among rare items. Our algorithms can also be applied to frequent items with bounded length. Experiments are performed on the well-known IBM synthetic database. Their schemes compare favorably to Apriori and FP-growth under the situation being evaluated. In addition, they explore quantitative association rule mining in transactional databases among infrequent items by associating quantities of items: some interesting examples are drawn to illustrate the significance of such mining.

Association rules are an intuitive descriptive paradigm that has been used extensively in different application domains with the purpose to identify the regularities and correlation in a set of observed objects. However, association rules' statistical measures (support and confidence) have been criticized because in some cases they have shown to fail in their primary goal: that is to select the most relevant and significant association rules. In this paper the authors propose a new model that replaces the support measure. The new model, like support, is a tool for the identification of reliable rules and is used also to reduce the traversal of the itemsets' search space. The proposed model adopts new criteria in order to establish the reliability of the information extracted from the database. These criteria are based on Bayes' Theorem and on an estimate of the probability density function of each itemset. According to our criteria, the information that we have obtained from the database on an itemset is reliable if and only if the confidence interval of the estimated probability is low compared with the most likely value of it. The authors will see how this method can be computed in an approximate but satisfactory way, with the same algorithms that are usually adopted to select itemsets on support threshold.

A novel approach is presented for effectively mining weighted fuzzy association rules (ARs). The authors address the issue of invalidation of downward closure property (DCP) in weighted association rule mining where each item is assigned a weight according to its significance with some user defined criteria. Most works on weighted association rule mining do not address the downward closure property while some make assumptions to validate the property. The authors generalize the weighted association rule mining problem with binary and fuzzy attributes with weighted settings. Their methodology follows an Apriori approach but employs T-tree data structure to improve efficiency of counting itemsets. Their approach avoids pre and post processing as opposed to most weighted association rule mining algorithms, thus eliminating the extra steps during rules generation. The chapter presents experimental results on both synthetic and real-data sets and a discussion on evaluating the proposed approach.

Section 2
Dealing with Imbalanced Datasets

In this chapter, the authors propose a novel framework for rare class association rule mining. In each class association rule, the right-hand is a target class while the left-hand may contain one or more attributes. This algorithm is focused on the multiple imbalanced attributes on the left-hand. In the proposed framework, the rules with and without imbalanced attributes are processed in parallel. The rules without imbalanced attributes are mined through a standard algorithm while the rules with imbalanced attributes are mined based on newly defined measurements. Through simple transformation, these measurements can be in a uniform space so that only a few parameters need to be specified by user. In the case study, the proposed algorithm is applied in the social security field. Although some attributes are severely imbalanced, rules with a minority of imbalanced attributes have been mined efficiently.

Rare association rule mining has received a great deal of attention in the past few years. In this paper, we propose a multi methodological approach to the problem of rare association rule mining that integrates three different strands of research in this area. Firstly, the authors make use of statistical techniques such as the Fisher test to determine whether itemsets co-occur by chance or not. Secondly, they use clustering as a pre-processing technique to improve the quality of the rare rules generated. Their third strategy is to weigh itemsets to ensure upward closure, thus checking unbounded growth of the rule base. Their results show that clustering isolates heterogeneous segments from each other, thus promoting the discovery of rules which would otherwise remain undiscovered. Likewise, the use of itemset weighting tends to improve rule quality by promoting the generation of rules with rarer itemsets that would otherwise not be possible with a simple weighting scheme that assigns an equal weight to all possible itemsets. The use of clustering enabled us to study in detail an important sub-class of rare rules, which we term absolute rare rules. Absolute rare rules are those are not just rare to the dataset as a whole but are also rare to the cluster from which they are derived.

The authors consider databases in which each attribute takes values from a partially ordered set (poset). This allows one to model a number of interesting scenarios arising in different applications, including quantitative databases, taxonomies, and databases in which each attribute is an interval representing the duration of a certain event occurring over time. A natural problem that arises in such circumstances is the following: given a database D and a threshold value t, find all collections of "generalizations" of attributes which are "supported" by less than t transactions from D. We call such collections infrequent elements. Due to monotonicity, they can reduce the output size by considering only minimal infrequent elements. The authors study the complexity of finding all minimal infrequent elements for some interesting classes of posets. They show how this problem can be applied to mining association rules in different types of databases, and to finding "sparse regions" or "holes" in quantitative data or in databases recording the time intervals during which a re-occurring event appears over time. Their main focus will be on these applications rather than on the correctness or analysis of the given algorithms.

Section 3
Rare, Anomalous, and Interesting Patterns

The paper presents an approach to mining patterns in numerical data without the need for discretization. The proposed method allows for discovery of arbitrary nonlinear relationships. The approach is based on finding a function of a set of attributes whose values are close to zero in the data. Intuitively such functions correspond to equations describing relationships between the attributes, but they are also able to capture more general classes of patterns. The approach is set in an association rule framework with analogues of itemsets and rules defined for numerical attributes. Furthermore, the user may include background knowledge in the form of a probabilistic model. Patterns which are already correctly predicted by the model will not be considered interesting. Interesting patterns can then be used by the user to update the probabilistic model.

In the context of anomaly detection, the data mining technique of extracting association rules can be used to identify rare rules which represent infrequent situations. A method to detect rare rules is to first infer the normal behavior of objects in the form of quasi-functional dependencies (i.e. functional dependencies that frequently hold), and then analyzing rare violations with respect to them. The quasi-functional dependencies are usually inferred from the current instance of a database. However, in several applications, the database is not static, but new data are added or deleted continuously. Thus, the anomalies have to be updated because they change over time. In this chapter, we propose an incremental algorithm to efficiently maintain up-to-date rules (i.e., functional and quasi-functional dependencies). The impact of the cardinality of the data set and the number of new tuples on the execution time is evaluated through a set of experiments on synthetic and real databases, whose results are here reported.

As common criteria in data mining methods, the frequency-based interestingness measures provide a statistical view of the correlation in the data, such as sequential patterns. However, when the authors consider domain knowledge within the mining process, the unexpected information that contradicts existing knowledge on the data has never less importance than the regularly frequent information. For this purpose, they present the approach USER for mining unexpected sequential rules in sequence databases. They propose a belief-driven formalization of the unexpectedness contained in sequential data, with which we propose 3 forms of unexpected sequences. They further propose the notion of unexpected sequential patterns and implication rules for determining the structures and implications of the unexpectedness. The experimental results on various types of data sets show the usefulness and effectiveness of our approach.

Association rule mining has been a highly active research field over the past decade. Extraction of frequency-related patterns has been applied to several domains. However, the way association rules are defined has limited our ability to obtain all the patterns of interest. In this chapter, the authors present an alternative approach that allows us to obtain new kinds of association rules that represent deviations from common behaviors. These new rules are called anomalous rules. To obtain such rules requires that we extract all the most frequent patterns together with certain extension patterns that may occur very infrequently. An approach that relies on anomalous rules has possible application in the areas of counter-terrorism, fraud detection, pharmaceutical data analysis and network intrusion detection. They provide an adaption of measures of interest to our anomalous rule sets, and we propose an algorithm that can extract anomalous rules as well. Their experiments with benchmark and real-life datasets suggest that the set of anomalous rules is smaller than the set of association rules. Their work also provides evidence that our proposed approach can discover hidden patterns with good reliability.

Strong symmetric association rules are defined as follows. Strong means that the association rule has a strong support and a strong confidence, well above the minimum thresholds. Symmetric means that X→Y and Y→X are both association rules. Common objective interestingness measures such as lift, correlation, conviction or Chi-square tend to rate this kind of rule poorly. By contrast, cosine is high for such rules. However, depending on the application domain, these rules may be interesting regarding criteria such as unexpectedness or actionability. In this chapter, the authors investigate why the above-mentioned measures, except cosine, rate strong symmetric association rules poorly, and show that the underlying data might take a quite special shape. This kind of rule can be qualified as rare, as they would be pruned by many objective interestingness measures. Then they present lift and cosine in depth, giving their intuitive meaning, their definition and typical values. Because lift has its roots in probability and cosine in geometry, these two interestingness measures give different information on the rules they rate. Furthermore they are fairly easy to interpret by domain experts, who are not necessarily data mining experts. The authors round off our investigation with a discussion on contrast rules and show that strong symmetric association rules give a hint to mine further rare rules, rare in the sense of a low support but a high confidence. Finally they present case studies from the field of education and discuss challenges.

Section 4
Critical Event Detection and Applications

In this chapter, the authors discuss the characteristics of data collected by the New Zealand Centre for Adverse Drug Reaction Monitoring (CARM) over a five-year period. They begin by noting the ways in which adverse reaction data are similar to market basket data, and the ways in which they are different. They go on to develop a model for estimating the amount of missing data in the dataset, and another to decide whether a drug is rare simply because it was only available for a short time. They also discuss the notion of "rarity" with respect to drugs, and with respect to reactions. Although the discussion is confined to the CARM data, the models and techniques presented here are useful to anyone who is about to embark on an association mining project, or who needs to interpret association rules in the context of a particular database.

Association rule mining produces a large number of rules but many of them are usually redundant ones. When a data set contains infrequent items, we need to set the minimum support criterion very low;

otherwise, these items will not be discovered. The downside is that it leads to even more redundancy. To deal with this dilemma, some proposed more efficient, and perhaps more complicated, rule generation methods. The others suggested using simple rule generation methods and rather focused on the post-pruning of the rules. This chapter follows the latter approach. The classic Apriori is employed for the rule generation. Their goal is to gain as much insight as possible about the domain. Therefore, the discovered rules are filtered by their semantics and structures. An individual rule is classified by its own semantic, or by how clear its domain description is. It can be labelled as one of the following: strongly meaningless, weakly meaningless, partially meaningful, and meaningful. In addition, multiple rules are compared. Rules with repetitive patterns are removed, while those conveying the most complete information are retained. They demonstrate an application of our techniques to a real case study, an analysis of traffic accidents in Nakorn Pathom, Thailand.

In this chapter, the authors explore Area under Curve (AUC) as an error-metric suitable for imbalanced data, as well as survey methods of optimizing this metric directly. We also address the issue of cut-point thresholds for practical decision-making. The techniques will be illustrated by a study that examines predictive rule development and validation procedures for establishing risk levels for violent felony crimes committed when criminal offenders are released from prison in the USA. The "violent felony" category was selected as the key outcome since these crimes are a major public safety concern, have a low base-rate (around 7%), and represent the most extreme forms of violence. They compare the performance of different algorithms on the dataset and validate using survival analysis whether the risk scores produced by these techniques are computing reasonable estimates of the true risk.

Credit scoring is a tool commonly employed by lenders in credit risk management. However credit scoring methods are prone to error. Failures from credit scoring result in granting loans to high risk customers, thus significantly increasing the incidence of overdue payments, or in the worst case, customers defaulting on the loan altogether. In this research the authors use a machine learning approach to improve the identification of such customers. However, identifying such bad customers is not a trivial task as they form the minority of customers and standard machine learning algorithms have difficulty in learning accurate models on such imbalanced datasets. They propose a novel approach based on a data segmentation strategy that progressively partitions the original data set into segments where bad customers form the majority. These segments, known as Majority Bad Payment Segments (MBPS) are then used to train

Foreword

For more than a decade, researches on association rule mining have attracted a huge interest from the data mining communities. Many advances in association rule mining have been proposed in recent years, including more efficient algorithms to process association rules, new data structures to speed up processing, new compression techniques to overcome the memory limitation problem, and so on. Many issues surrounding association rules have been discussed, including security, privacy, and incomplete and inaccurate data. Association rules have also been applied in various domains, including mobile mining, social networking, graph mining, etc.

However, most of the existing research on association rules has been focusing on establishing common patterns and rules; these are patterns and rules based on the majority, some of which may be either obvious or irrelevant. Unfortunately, not enough attentions have been given to mining rare association rules; these are outlier rules and patterns.

Rare association rules are critically important as in many cases they represent outstanding patterns, which cannot be easily discovered by traditional association mining algorithms. This book presents an interesting collection of recent advances in rare association rule mining. This book is certainly an invaluable resource to data mining researchers, especially to those who have strong interest in association rules.

I am pleased to be able to recommend this timely reference source to readers, be they researchers looking for future directions to pursue research in data mining, or practitioners interested in applying data mining concepts in practical situations.

David Taniar
Monash University, Australia
January 2009

***David Taniar** holds Bachelor, Master, and PhD degrees - all in Computer Science, with a particular specialty in Databases. His current research areas are in mobile databases, parallel databases, web databases, GIS, and data mining. He publishes extensively every year, including his recent co-authored book: High Performance Parallel Database Processing and Grid Databases (John Wiley & Sons, 2008). His list of publications can be viewed at the DBLP server (http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/t/Taniar:David.html). He is a founding editor-in-chief of a number of international journals, including Intl J of Data Warehousing and Mining, Intl J of Business Intelligence and Data Mining, Mobile Information Systems, Journal of Mobile Multimedia, Intl J of Web Information Systems, and Intl J of Web and Grid Services. He is currently an Associate Professor at the Faculty of Information Technology, Monash University, Australia. He can be contacted at David.Taniar@infotech.monash.edu.au.*

Preface

This is the third volume of the Advances in Data Warehousing and Mining (ADWM) book series. ADWM publishes books in the areas of data warehousing and mining. This special volume, *Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection*, presents cutting edge research in this newly emerging area. Techniques for rare association mining are quite different from that of traditional rule mining and this book fills an essential gap in this area.

The primary objective of this book is to give readers in-depth knowledge on the current issues in rare association rule mining and critical event detection. The book is designed to cover a comprehensive range of topics related to rare association rule mining and critical event detection: mining techniques, imbalanced datasets, interest metrics, and real-world application domains. We hope this book will highlight the need for growth and research in the area of rare association rule mining and critical event detection. This volume consists of sixteen chapters in four sections.

The first section, Beyond the Support-Confidence Framework, provides an introduction to the area of rare association rule mining, and looks at some of the current proposed techniques which have moved away from the traditional support and confidence measures. This section contains four chapters.

Chapter 1, "*Rare Association Rule Mining: Overview*", by Yun Sing Koh, Auckland University of Technology, New Zealand, and Nathan Rountree, University of Otago, New Zealand, introduces the problem faced in the area of rare association rule mining and the current trends in this area. They provide an extensive literature review on the currently available techniques when dealing with rare itemsets.

Chapter 2, "*Association Rule and Quantitative Association Rule Mining among Infrequent Items*", by Ling Zhou and Stephen Yau, University of Illinois at Chicago, proposes two new methods to mine infrequent items and find rare association rules. Their approach is versatile and can also be applied to frequent items with bounded length. In addition they explore quantitative association rule mining among infrequent items by associating quantities of items: some interesting examples are drawn to illustrate the significance of such mining.

Chapter 3, "*Replacing Support in Association Rule Mining*", by Rosa Meo and Dino Ienco, Università di Torino, Italy, proposes a new model which adopts criteria based on Bayes' Theorem and on an estimate of the probability density function of each itemset to establish the reliability of the information extracted from the database.

Chapter 4, "*Effective Mining of Weighted Fuzzy Association Rules*", by Maybin Muyeba, Manchester Metropolitan University, UK, M. Sulaiman Khan, Liverpool Hope University, UK, and Frans Coenen, University of Liverpool, UK, presents a novel approach for effectively mining weighted fuzzy association rules. They generalize the weighted association rule mining problem with binary and fuzzy attributes with weighted settings.

The second section, Dealing with Imbalanced Datasets, looks at algorithms and mining frameworks for dealing with datasets where there is uneven representation of various database objects. Imbalanced data is a key issue in rare association rule mining, because: a) it is a *necessary* condition of rare itemsets, and b) it affects the power and accuracy of the statistical models used to perform data mining. This section consists of three chapters, where we look at rare class association rule mining, sub-class association rule mining, and mining minimal infrequent elements.

Chapter 5, "*Rare Class Association Rule Mining with Multiple Imbalanced Attributes*", by Huaifeng Zhang, Yanchang Zhao, Longbing Cao, Chengqi Zhang, University of Technology, Sydney, Australia, and Hans Bohlscheid, Projects Section, Business Integrity Programs Branch, Centrelink, Australia, proposes a framework for rare class association rule mining. In their approach, the rules without imbalanced attributes are mined through a standard algorithm while the rules with imbalanced attributes are mined based on newly defined measurements. In this chapter, they present a compelling case study applied in the social security field.

Chapter 6, "*A Multi-Methodological Approach to Rare Association Rule Mining*" by Yun Sing Koh, Auckland University of Technology, New Zealand, and Russel Pears, Auckland University of Technology, New Zealand, proposes a synthesis of material from three different methodologies to tackle the problem of rare association rule mining: itemset weighting, clustering, and statistical significance testing. They focus on the importance of sub-class rare rules or absolute rare rules. Absolute rare rules are those are not just rare to the dataset as a whole but are also rare to the cluster from which they are derived.

Chapter 7, *"Finding Minimal Infrequent Elements in Multi-Dimensional Data Defined over Partially Ordered Sets and its Applications"*, by Khaled M. Elbassioni, Max-Planck-Institut fur Informatik, Germany, studies the complexity of finding all minimal infrequent elements for some interesting classes of partially ordered set (poset). He looks at a general framework used to mine associations from different types of databases. The rules obtained under this framework are generally stronger than the ones obtained from techniques that use binarization.

In Section 3, Rare, Anomalous, and Interesting Patterns, we look at some of the techniques used to find interesting and unexpected patterns in the area of association rules. Section three consists of five chapters, discussing issues related to discovering interesting patterns in numerical data with background knowledge, discovering quasi-functional dependencies, mining unexpected patterns, and extracting anomalous rules.

Chapter 8, *"Discovering Interesting Patterns in Numerical Data with Background Knowledge"*, by Szymon Jaroszewicz, National Institute of Telecommunications, Warsaw, Poland, presents an approach to mining patterns in numerical data without the need for discretization. The proposed method allows for discovery of arbitrary nonlinear relationships where the user may include background knowledge in the form of a probabilistic model. The patterns that have been previously predicted by the model will not be considered interesting. Interesting patterns can then be used by the user to update the probabilistic model.

Chapter 9, *"Mining Rare Association Rules by Discovering Quasi-functional Dependencies: An Incremental Approach"*, by Giulia Bruno and Paolo Garza, Politecnico di Torino Corso Duca degli Abruzzi, Italy, and Elisa Quintarell, Politecnico di Milano Piazza Leonardo da Vinci, Italy, propose a method of detecting rare rules by first inferring the normal behaviour of objects in the form of quasi-functional dependencies (i.e. functional dependencies that frequently hold), and then analysing rare violations with respect to them. They propose an incremental algorithm to efficiently maintain up-to-date rules.

Chapter 10, *"Mining Unexpected Sequential Patterns and Implication Rules"* by Dong (Haoyuan) Li, LGI2P, École des Mines d'Alès, France, Anne Laurent and Pascal Poncelet, LIRMM, Université Montpellier II, France, presents an approach called USER for mining unexpected sequential rules in sequence databases. They propose a belief-driven formalization of the unexpectedness contained in sequential data.

Chapter 11, *"Mining Hidden Association Rules from Real-Life Data"* by Marco-Antonio Balderas Cepeda, Universidad de Granada, Spain, provides an adaptation of measures of interest to our anomalous rule sets, and proposed an algorithm that can extract anomalous rules as well. Their approach discovered hidden patterns with good reliability.

Chapter 12, *"Strong Symmetric Association Rules and Interestingness Measures"* by Agathe Merceron, University of Applied Sciences TFH Berlin, Germany, proposes a method to find strong symmetric association rules. This approach is slightly different from the conventional rare association rule mining. This kind of rule can be qualified as rare, as they would be pruned by many objective interestingness measures.

In Section 4, Critical Event Detection and Applications, we look at some of the applications of rare association rule mining and critical event detection. In this section, we provide two chapters which specifically look at the usage of association rule mining in different domains. The last two chapters look at a different data mining approach, namely classification techniques, for critical event detection. The areas of application discussed include adverse drug reaction monitoring, analysis of traffic accident, risk levels for violent felony crimes, and financial credit monitoring.

Chapter 13, *"He Wasn't There Again Today"*, by Richard O'Keefe and Nathan Rountree, University of Otago, New Zealand, discusses the characteristics of data collected by the New Zealand Centre for Adverse Drug Reaction Monitoring (CARM) over a five-year period. They discuss the notion of "rarity" with respect to drugs, and with respect to reactions.

Chapter 14, *"Filtering Association Rules by Their Semantics and Structures"* by Rangsipan Marukatat, Mahidol University, Thailand, introduces the filtering of association rules by their patterns and degrees of semantic redundancy. They applied their techniques to a real case study, an analysis of traffic accidents in Nakorn Pathom, Thailand.

Chapter 15, *Creating Risk-Scores in very Imbalanced Datasets: Predicting Extremely Violent Crime among Criminal Offenders Following Release from Prison* by Markus Breitenbach, William Dieterich, Tim Brennan, Northpointe Institute for Public Management, USA, and Adrian Fan, University of Colorado at Boulder, USA, explores the Area under Curve (AUC) as an error-metric suitable for imbalanced data, as well as survey methods of optimizing this metric directly. They conducted a study that examines predictive rule development and validation procedures for establishing risk levels for violent felony crimes committed when criminal offenders are released from prison in the USA.

Chapter 16, *"Boosting Prediction Accuracy of Bad Payments in Financial Credit Applications"*, by Russel Pears and Raymond Oetama, Auckland University of Technology, New Zealand, use a machine learning approach to improve the identification of such customers. They proposed a credit scoring approach to predict bad payments for credit risk management.

We hope that this book will provide readers some specific challenge that motivates the development and enhancement of rare association rule mining and critical event detection area. We also hope that this book will serve as an introductory material to the researchers and practitioners interested in this emerging area of research.

Yun Sing Koh and Nathan Rountree
January 2009

Chapter 1
Rare Association Rule Mining: An Overview

Yun Sing Koh
Auckland University of Technology, New Zealand

Nathan Rountree
University of Otago, New Zealand

ABSTRACT

The notion of finding rare association rules is like finding precious gems in an open field; it is a daunting task but, if successful, it is very rewarding. Association rule mining systems, such as Apriori, generally employ an exhaustive search algorithm. While these algorithms are in theory capable of finding rare association rules, they become intractable if the minimum level of support is set low enough to find rare rules. Such algorithms are therefore inadequate for finding rare associations, and also suffer from the rare item problem. Research to solve this problem has become more prevalent in recent times. The main goal of rare association rule mining is to discover relationships among sets of items in a transactional database that occur infrequently. This chapter presents a survey on the current trends and approaches in the area of rare association rule mining.

INTRODUCTION

The most popular pattern discovery method in data mining is association rule mining. Association rule mining was introduced by Agrawal, Imielinski, and Swami (1993). It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in transaction databases or other data repositories. The relationships are not based on inherent properties of the data themselves but rather based on the co-occurrence of the items within the database. The associations between items are commonly expressed in the form of *association rules*.

The original motivation for seeking association rules came from the need to analyse supermarket transaction data to examine customer behaviour in terms of the purchased products. This is known as market basket analysis. Market basket analysis begins by finding all *frequent itemsets*; that is, combinations of items that appear together in at least *m* transactions in the database, where *m* is specified by

DOI: 10.4018/978-1-60566-754-6.ch001

the analyst in advance. This user-specified parameter is called minimum support (minsup). Then association rules are derived in the form of X → Y where XY is a frequent itemset. Strong association rules are derived from frequent itemsets and constrained by another user-specified parameter: minimum confidence (minconf). Confidence is the percentage of transactions containing X that also contain Y. For example, suppose in a database 27% of all transactions contain both bread and milk, and 30% of all transactions contain bread. An association rule mining system might therefore derive the rule *bread → milk* with 27% support and 90% confidence. (Confidence can be treated as the conditional probability of a transaction containing bread also containing milk.) In classical association rule mining systems, the user must set minsup to 27% or lower, and minconf to 90% or lower for this rule to have been produced. For instance, if minconf had been set to 35%, the {*bread, milk*} itemset would never have been spotted by the system – and a rule of high confidence would have been missed.

Currently most association mining algorithms are dedicated to frequent itemset mining. These algorithms are defined in such a way that they only find rules with high support and high confidence. Most of these approaches adopt an Apriori-like approach (Agrawal & Srikant, 1994). A much less explored area in association mining is *infrequent* itemset mining. Intuitively, we can define rare itemsets as those that appear together in very few transactions, or some very small percentage of the transactions in the database. However, the key motivation of infrequent itemset mining is that, although two items may appear in very few transactions, it may be that when they do appear, they typically appear together. Therefore, it may be possible to form an association rule that has very low support, but very high confidence. For example, suppose that {espresso machine} appears in only 1% of a department store's transactions, and that {coffee grinder} appears in about 1.2%. Both items could be said to be rare. Furthermore, suppose that {*coffee grinder, espresso machine*} appears in 0.8% of the store's transactions: even more rare. But with this information, we can derive the rule *espresso machine → coffee grinder* with a confidence of 80%.

A characteristic of frequent itemset mining is that it relies on there being a meaningful minimum support level that is sufficiently *high* to reduce the number of frequent itemsets to a manageable level. However, in some data mining applications relatively infrequent associations are likely to be of great interest as they relate to rare but crucial cases. Examples of mining infrequent itemsets include identifying relatively rare diseases, predicting telecommunication equipment failure, and finding associations between infrequently purchased (e.g. expensive or high-profit) retail items. Indeed, infrequent itemsets warrant special attention because they are more difficult to find using traditional data mining techniques.

This chapter introduces the current approaches to rare association rule mining. The chapter is divided into several sections that include an introduction to association rule mining, the rare item problem, current trends and approaches, and discussion on the future development of this area. Finally, we provide a section, Further Information, summarizing the key papers in the area of rare association rule mining.

Association Rule Mining

The following is a formal statement of association rule mining for a transaction database. Let I = {i_1, i_2, ..., i_m} be the universe of items. A set X ⊆ I of items is called an itemset or pattern. In particular, an itemset containing *k* items is called a *k*-itemset. Every transaction contains a unique transaction ID tid. A transaction t = (tid,X) is a tuple where X is an itemset. A transaction t = (tid, X) is said to contain itemset Y if Y ⊆ X. The main function of a unique transaction ID is to allow instances of an itemset to occur more than once in a database. Let {t_1, t_2, ..., t_n} be the set of all possible transactions

T. A transaction database D is a set of transactions, such that D ⊆ T. In effect, D is really a multiset of itemsets. An association rule is an implication of the form X → Y, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$. Let sup(X) be the number of transactions containing all the items in itemset X. The rule X → Y has support of *s* in the transaction set D, if *s* = sup(XY), where XY refers to an itemset that contains all of the items in X and all of the items of Y. Alternatively, sup(X) can be expressed in relative terms, as the percentage of transactions that contain X in the dataset. The rule X → Y holds in the transaction set D with confidence *c* where *c* = conf(X→Y), the *confidence* of the rule X→Y. This may be expressed as sup(XY)/sup(X). Note that in the calculation of confidence, it does not matter whether support is absolute or relative.

The Apriori algorithm is the "classical" method of finding frequent k-itemsets, which may then be formed into rules. Its purpose is to avoid counting the support of every possible itemset derivable from I (since there are 2^m possible itemsets to be checked if there are m items). Apriori exploits the property of *downward closure*, which is that if any k-itemset is frequent, all of its subsets must be frequent too. Assume that the items in a k-itemset are always stored in lexicographic order. Apriori proceeds as follows:

1. Calculate the support of all 1-itemsets (this information is often readily available from the system anyway) and prune any that fall under minimum support. These are the frequent 1-itemsets.

Loop

2. Form candidate k-itemsets by taking each pair *p, q* of itemsets in the (k-1)-itemsets where *all but* the last item match. Form each new k-itemset by adding the last item of *q* onto the items of *p*.
3. Prune the candidate k-itemsets by eliminating any itemset that contains a subset not in the (k-1)-frequent itemsets
4. Count the supports in the database of the remaining candidate k-itemsets and eliminate any that fall below minsup. The result is the frequent k-itemsets.

Until Step 2 fails to produce any candidates (which will also occur if Steps 3 or 4 resulted in an empty set of candidates).

It is clear that finding association rules with low support but high confidence using Apriori-like methods would face difficulties. To find these rules the minimum support threshold would need to be set quite low to enable rare items to be let in. The Apriori heuristic is only able to reduce the size of candidate itemsets if the minimum support is set reasonably high. However, in situations with abundant frequent patterns, long patterns, or a low minimum support threshold, an Apriori-like algorithm may still suffer from the following nontrivial cost: most of the items would be allowed to participate in itemset generation. This will have an effect on the scalability of the Apriori algorithm. It is costly to handle a huge number of candidate items. It is time consuming to repeatedly scan the database and check the support of each of the candidate itemsets generated. The complexity of the computation will increase exponentially with regard to the number of itemsets generated.

Let us consider the work that the Apriori algorithm does in terms of the specified minimum support and the number and length of frequent itemsets in the database. Apriori may be able to cut out a lot of candidates; however, it is still costly to handle a huge number of candidate itemsets in large transaction databases. For example, consider the case where there are 1 million items but only 1% (1000 items) are frequent 1-items. Apriori still has to generate more than 5×10^5 candidate 2-itemsets, and evaluate and store the support for the generation of candidate 3-itemsets. It is expensive to repeatedly scan the database and check a large set of frequent itemsets by pattern matching, especially if the length of the itemset is long. Apriori does a level-by-level candidate generation and test. If it has a frequent itemset X = {i^1 ... i^k}, Apriori has to scan the database k times. For example, if k is 100 we would have to scan the database 100 times for that particular itemset. Apriori encounters difficulty in mining long patterns. To find a frequent itemset X = {i^1 . . . i^{100}} it has to generate and test $2^{100} - 1$ candidate itemsets.

These drawbacks suggest that it would not be efficient to use Apriori to generate rare rules as rare itemsets have low frequencies in the database by definition. Hence if we use the Apriori algorithm we would need to lower the minimum support threshold close to 0. This would allow most of the items within the dataset to be extended and used in the next iteration. As Apriori will not be able to prune a lot of the candidate itemsets, the repeated scan through the database becomes very expensive.

RARE ITEM PROBLEM

Traditional association rule mining algorithms, such as Apriori, focus on mining association rules in large databases with a single minimum support (minsup) threshold. Since a single threshold is used for the whole database, it assumes that all items in the database are of the same nature and/or have similar frequencies. As such, Apriori works best when all items have approximately the same frequency in the data. Apriori exploits the downward closure property that states that if an itemset is frequent so are all its subsets. As such, it is not possible to use Apriori with multiple user-defined minsups without modification to the algorithm. Consider the case where the user-defined minsup of {A,B,C} is 2% and the user-defined minsup of {A,B}, {A,C}, and {B,C} is 5%. It is possible for {A,B,C} to be frequent with respect to its minsup but none of {A,B}, {A,C}, {B,C} to be frequent with respect to their minsup. Suppose {A,B}, {A,C}, {B,C} have support of 4%, and {A,B,C} has support of 3%. In this case itemset {A,B,C} should be considered frequent because the user has specified a minsup of 2% for it, and it is above that. However, Apriori will not generate it, because {AB} and {AC} fall below their user-specified minsup of 5%. In reality, some items may be very frequent while others may rarely appear. Hence minsup should not be fixed because deviation and exceptions generally have a much lower support than general trends. Note that support requirements vary as the support of items contained in an itemset varies. Given that the existing Apriori algorithm assumes a uniform support, rare itemsets can be hard to find. Rare items are by definition in very few transactions and will be pruned because they do not meet the minsup threshold. In data mining, rare itemsets may be obscured by common cases. Weiss (2004) calls this relative rarity. This means that items may not be rare in the absolute sense but are rare relative to other items. This is especially a problem when data mining algorithms rely on greedy search heuristics that examine one item at a time. Since rare cases may depend on the conjunction of many conditions, analysing any single condition alone may not be interesting (Weiss, 2004).

As a specific example of the problem, consider the association mining problem where we want to determine if there is an association between buying a food processor and buying a cooking pan (Liu et al., 1999a). The problem is that both items are rarely purchased in a supermarket. Thus, even if the two items are almost always purchased together, this association may not be found, because the 1-itemsets are pruned out before they can be used to generate 2-itemsets. Modifying the minsup threshold to take into account the importance of the items is one way to ensure that rare items remain in consideration. To find this association minsup must be set low. However setting this threshold low would cause a combinatorial explosion in the overall number of itemsets generated. Frequently occurring items will be associated with one another in an enormous number of ways simply because the items are so common that they cannot help but appear together. This is known as the rare item problem (Liu et al., 1999a). It means that, using the Apriori algorithm, we are unlikely to generate rules that may indicate rare events of potentially dramatic consequence. For example, we might prune out rules that indicate the symptoms of a rare but fatal disease due to the frequency of occurrence not reaching the minsup threshold. As rare rule mining is still an area that has not been well explored, there is some groundwork that needs to be established. A real dataset will contain noise, possibly at levels of low support. Normally, noise has low support. In Apriori, setting a high minimum support threshold would cut the noise out. Inherently we are looking for rules with low support that could make them indistinguishable from coincidences (that is, situations where items fall together no more often than they would normally by chance). Although Apriori is the most commonly used association mining technique, it is far from efficient when we try to find low support rules. Using Apriori, we would still need to wade through thousands of itemsets (often having high support) to find the rare itemsets that are of interest to us.

Although rare rule mining has many potential possibilities, like frequent pattern mining, there could be a large number of rules generated from a database. We would need to find ways to generate only the potentially useful rare rules.

Current Trends and Approaches

Classic association mining techniques, such as Apriori, rely on uniform minimum support. These algorithms either miss the rare but interesting rules or suffer from congestion in itemset generation caused by low support. Driven by such shortcomings, some research has been carried out in developing new rule discovery algorithms to mine rare rules. Currently there are several different approaches to deal with the shortcoming of using support threshold and the rare item problem. In this section we take a look at the mainstream research effort in this particular area. There approaches to mining rare rules include using a variable support threshold, mining without support threshold, constraint-based mining, and structure-based mining. Here we take a look general idea behind these approaches.

There have been several approaches taken to ensure that rare items are considered during itemset generation. One of the approaches is association rule mining with variable support threshold. In this approach, each itemset may have a different support threshold. The support threshold for each itemset is dynamically lowered to allow some rare items to be included in the rule generation. Some of the research using this approach includes Multiple Supports Apriori (MSApriori) (Liu et al., 1999a), Relative Support Apriori (RSAA) (Wang et al., 2003), Weighted Association Rules (WARM) (Tao et al., 2003), Adaptive Apriori (Wang et al., 2003), LPMiner (Seno & Karypis, 2001), and NB model (Hashler, 2006). These approaches try to vary the support constraint in some fashion to allow some rare items to be included in frequent itemset generation. These approaches are exhaustive in their generation of rules, and so spend time looking for rules with high support and high confidence. If the varied minimum support value is set close to zero, they will take a similar amount of time to that taken by Apriori to generate low-support rules in amongst the high-support rules. These methods generate all rules that have high confidence and high support. To include truly rare items, the minsup threshold must be set very low, which consequently generates an enormous set of rules consisting of both frequent and infrequent items.

A fixed minimum support threshold is not effective for datasets with a skewed distribution because they tend to generate many trivial patterns or miss potential low-support patterns. Hence another approach uses association rule mining without support threshold, but it usually introduces another constraint to solve the rare item problem. We discussed some of the approaches that use a variable support threshold to include some rare items in rule generation. But to ensure each rare item is considered, the minimum support threshold must still be pushed low, resulting in a combinatorial explosion in the number of rules generated. To overcome this problem, some researchers have proposed to remove the support-based threshold entirely. Instead they use another constraint such as similarity or confidence-based pruning. Techniques in this area includes Min-Hashing and its variations (Cohen et al., 2001), Confidence-Based Pruning (Wang et al., 2001), and H-Confidence (Xiong et al., 2003). Similar to the techniques in the previous approach, these algorithms suffer from the same drawback of generating all the frequent rules as well as the rare rules. In both of these approaches we need post-pruning methods to filter out the frequent rules or the trivial rules produced.

Using a variable support threshold or no support threshold would generate frequent rules as well as rare rules. There are some approaches that try to generate only rare rules. For example, providing a list of those items that may or may not take part in a rule and then modifying the mining process to take advantage of that information. One of the restrictions that may be imposed is called consequent constraint-based rule mining. In this approach, an item constraint is used which requires mined rules to satisfy a given constraint. Techniques that use this approach include Dense-Miner (Bayardo et al., 2000), DS (Direction Setting) rules (Liu, Hsu & Ma, 1999b), EP (Emerging Pattern) (Li et al., 1999), and Fixed-Consequent ARM (Association Rule Mining) (Rahal et al., 2004). These algorithms are only useful when we have prior knowledge that a particular consequent is of interest. Since rare items occur infrequently by definition, they may go undetected by prior processes that seek to identify what itemsets *should* be participating in consequents. This makes it unsuitable for generating rare item rules efficiently because we want to generate rules without needing prior knowledge of which consequents ought to be interesting.

Another way to encourage low-support items to take part in candidate rule generation is by imposing structure constraints. Techniques in this approach usually use an extra boundary to only allow the generation of rare rules. Techniques in this approach includes Apriori-Inverse (Koh & Rountree, 2005), MIISR (Mining Interesting Imperfectly Sporadic Rules) (Koh et al., 2006), and Apriori-Rare (Szathmary et al, 2007). These approaches are reliant on the fixed upper-boundaries. Setting the correct boundaries is still an open research question in rare association rule mining.

Currently there are various techniques in the area of rare association rule mining. Nonetheless there is still room for expansion. The capability of current techniques is limited to particular types of rare rules. It is a difficult task to determine and generate all useful rare rules. This process is often bounded by the nature the dataset. Rare rules often consist of a combination of frequent items that separately have high support, but together have low support. Thus we can not rely on normal frequent mining techniques to detect rare rules. The low support of the itemsets also makes it difficult for us to tell apart rare rules from noise.

Discussion: Where is this Heading?

Mining rare association rule mining goes beyond techniques and approaches which generate the rules. Rare association rules require different pre-processing and post-pruning techniques as compared to frequent rule mining. Despite being in the same area, the properties of the rules are substantially different. Current pre-processing and post-pruning techniques which cater for frequent rule mining are designed to suit the characteristics of frequent rules. The development in this area of rare association rule mining has room for expansion in several different significant directions.

Rare Itemset Detection and Noise Detection. The first direction is to find a theoretically-sound way to find rare itemsets. While showing promise, current rare association rule mining (RARM) techniques use arbitrary thresholds for finding rare itemsets. While the current techniques are sound, many do not consider noise detection in the technique. One of the crucial factors in finding rare itemsets, is being able to differentiate valid itemsets from noise.

Rare Rule Generation. The second direction addresses the different types of rare rules that can be found. It has been commonly observed, especially in medical domains, that certain items might occur frequently on their own but rarely as a group (itemset). For instance, two common allergens combined can produce a rare allergic reaction. When such a situation arises, there are usually a few rare rules that one could have mined. Even recent developments only allow us to generate a subset of these rules. We acknowledge the fact that not all types of rare rules are interesting. However there still lacks a generic framework to produce all useful rare rules. One problem with rare rule mining is the possibility of generating too many rules many which are not useful. Real-world datasets contain noise. This part of the nature of rare rules means they are susceptible to being drowned out in the noise; or, maybe worse, that we incorrectly treat noise-rules as valid rules.

Post-pruning Metrics. The third direction focuses on developing post-pruning methods, i.e. interest measures, to examine rare rules. Existing interest measures are inaccurate when dealing with low support rules (i.e. rare rules). Given that there has not yet been a substantial amount of work carried out in this area of rare association rule mining, there is currently no method that can be used to rank or prune these rules. A complementary research line is devoted to mining a concise set of frequent association rules. Most interest measures, such as the Cosine, Jaccard, and Confidence measures, are biased towards high support rules. The current proposed techniques are designed for frequent association rule mining and are not suited for rare rule mining.

CONCLUSION

Rare rule mining is a fairly new area in association rule mining research and has gained some attention in the past few years. Rare association rule mining can be viewed as an extension in the area of association rule mining. However the properties of rare rules are inherently different to their counterpart, frequent rules, and warrants further research. Currently there is still no ideal solution that allows us to find all possible interesting rare association rules, and there is much room for expansion in this area.

FURTHER INFORMATION

Multiple Supports Apriori (MSApriori)

Liu et al. (1999a) deal with the rare item problem by using multiple minimum support thresholds. They note that some individual items can have such low support that they cannot contribute to rules generated by Apriori, even though they may participate in rules that have very high confidence. They overcome this problem with a technique whereby each item in the database can have its own minimum item support (MIS). By providing a different MIS for different items, a higher minimum support can be set for rules that involve frequent items and lower minimum support for rules that involve less frequent items. The minimum support of an itemset is the lowest MIS among those items in the itemset. For example, let MIS(i) denote the MIS value of item i. The minimum support of a rule R is the lowest MIS value of items in the rule. A rule, R: AB → C satisfies its minimum support if the rule has an actual support greater or equal to: min(MIS(A), MIS(B), MIS(C)). However consider four items in a dataset, A, B, C, and D with MIS(A) = 10%, MIS(B) = 20%, MIS(C) = 5%, and MIS(D) = 4%. If we find that {A,B} has 9% support at the second iteration, then it does not satisfy min(MIS(A), MIS(B)) and is discarded.

Then potentially interesting itemsets {A,B,C} and {A,B,D} will not be generated in the next iteration. By sorting the items in ascending order of their MIS values, the minimum support of the itemset never decreases as the length of an itemset grows, making the support a more general support constraint. In general, it means that a frequent itemset is only extended with an item having a higher (or equal) MIS value. The MIS for each data item i is generated by first specifying LS (the lowest allowable minimum support), and a value β, $0 \leq \beta \leq 1.0$. MIS(i) is then set according to the following formula:

MIS(i) = max(β.sup(i), LS)

The advantage of the MSApriori algorithm is that it has the capability of finding some rare-itemset rules. However, the actual criterion of discovery is determined by the user's value of β rather than the frequency of each data item.

Relative Support Apriori (RSAA)

Determining the optimal value for β could be tedious especially in a database with many items where manual assignment is not feasible. Thus Yun, Ha, Hwang and Ryu (2003) proposed the RSAA algorithm to generate rules in which significant rare itemsets take part, without any set number specified by the user. This technique uses relative support: for any dataset, and with the support of item i represented as sup(i), relative support (RSup) is defined as:

$$\mathrm{RSup}(i_1, i_2, \ldots, i_k) = \frac{\mathrm{sup}(i_1, i_2, \ldots, i_k)}{\min(\mathrm{sup}(i_1), \mathrm{sup}(i_2), \ldots, \mathrm{sup}(i_k))}$$

Thus this algorithm increases the support threshold for items that have low frequency and decreases the support threshold for items that have high frequency. Using a non-uniform minimum support threshold leads to the problem of choosing a suitable minimum support threshold for a particular itemset. Each item within the itemset may have a different minimum support threshold. MSApriori and RSAA sort the items within the itemset in non-decreasing order of support. Here the support of a particular itemset never increases and the minimum support threshold never decreases as the itemset grows.

Adaptive Apriori

Wang, He and Han (2003) proposed Adaptive Apriori which has a variable minimum support threshold. Adaptive Apriori introduces the notion of support constraints (SC) as a way to specify general constraints on minimum support. In particular, they associate a support constraint with each of the itemsets. They consider support constraints of the form $SC_i(B_1, \ldots, B_s) \geq \theta_i$, where $s \geq 0$. Each B_j, called a bin, is a set of items that need not be distinguished by the specification of minimum support. θ_i is a minimum support in the range of [0 . . . 1], or a function that produces minimum support. If more than one constraint is applicable to an itemset, the constraint specifying the lowest minimum support is chosen. For example, given $SC_1(B_1, B_3) \geq 0.2$, $SC_2(B_3) \geq 0.4$, $SC_3(B_2) \geq 0.5$, and $SC_0() \geq 0.9$, if we have an itemset containing $\{B_1, B_2, B_3\}$ the minimum support used is 0.2. However, if the itemset only contains $\{B_2, B_3\}$ then the minimum support is 0.4. The key idea of this approach is to push the support constraint following the dependency chain of itemsets in the itemset generation. For example, we want to generate itemset $\{B_0B_1B_2\}$, which uses SC_3, which is 0.5. $\{B_0B_1B_2\}$ is generated by using $\{B_0B_1\}$ with SC_0 and $\{B_1B_2\}$ with SC_3. This requires the minsup, which is 0.5 from $\{B_0B_1B_2\}$, to be pushed down to $\{B_0B_1\}$, and then pushed down to $\{B_0\}$ and $\{B_1\}$. The pushed minimum support is 0.5, which is lower than the specified minsup for $\{B_0B_1\}$, $\{B_0\}$, or $\{B_1\}$, which is 0.9. The pushed minimum support of each itemset is forced to be equal to the support value corresponding to the longest itemset.

Weighted Association Rules (WARM)

We can determine the minimum support threshold of each itemset by using a weighted support measurement. Each item or itemset is assigned a weight based on its significance. Itemsets that are considered interesting are assigned a larger weight. Weighted association rule mining (WARM) (Tao, Murtagh and Farid, 2003) is based on a weighted support measurement with a weighted downward closure property. They propose two types of weights: item weight and itemset weight. Item weight w(i) is assigned to an item representing its significance, whereas itemset weight w(X) is the mean of the item weight.

$$w(X) = \frac{\sum_{k=1}^{X} w(i_k)}{|X|}$$

The goal of using weighted support is to make use of the weight in the mining process and prioritise the selection of targeted itemsets according to their significance in the dataset, rather than by their frequency alone. The weighted support of an itemset can be defined as the product of the total weight of the itemset (sum of the weights of the items) and the weight of the fraction of transactions that the itemset occurs in. In WARM, itemsets are no longer simply counted as they appear in a transaction. The change in the counting mechanism makes it necessary to adapt the traditional support to a weighted support. An itemset is significant if its support weight is above a pre-defined minimum weighted support threshold. Tao et al. (2003) also proposed a weighted downward closure property as the adjusted support values violate the original downward closure property in Apriori. The rules generated in this approach rely heavily on the weights used. Thus to ensure the results generated are useful, we have to determine a way to assign the item weights effectively.

LPMiner

Previous approaches vary the minimum support constraint by using a particular weighting method using either the frequency or significance of the itemsets. LPMiner (Seno & Karypis, 2001), also varies the minimum support threshold. It uses a pattern-length-decreasing support constraint that tries to reduce support so that we favour smaller itemsets which have higher counts over larger itemsets with lower counts. They propose a support threshold that decreases as a function of itemset length. A frequent itemset that satisfies the length-decreasing support constraint can be frequent even if the subsets of the itemset are infrequent. Hence the downward closure property does not hold. To overcome this problem, they developed a property called smallest valid extension (SVE). In this property, for an infrequent itemset to be considered it must be over a minimum pattern length before it can potentially become frequent. Exploiting this pruning property, they propose LPMiner based on the FP-tree algorithm (Han, Pei & Yin, 2000). This approach favours smaller itemsets; however, longer itemsets could be interesting, even if they are less frequent. In order to find longer itemsets, one would have to lower the support threshold, which would lead to an explosion of the number of short itemsets found.

Min-Hashing and its Variations

Variations on the Min-Hashing technique were introduced by Cohen et al. (2001) to mine significant rules without any constraint on support. Transactions are stored as a 0/1 matrix with as many columns as there are unique items. Rather than searching for pairs of columns that have high support or high confidence, the technique searches for columns that have high similarity, where similarity is defined as the fraction of rows that have a 1 in both columns when they have a 1 in either column. Although this is easy to do by brute force when the matrix fits into main memory, it is time-consuming when the matrix is disc-resident. Their solution is to compute a hashing signature for each column of the matrix in such a way that the probability that two columns have the same signature is proportional to their similarity. After signatures are calculated, candidate pairs are generated, and then finally checked against the original matrix to ensure that they do indeed have strong similarity. It should be noted that the hashing solution will produce many rules that have high support and high confidence, since only a minimum acceptable similarity is specified. It is not clear whether the method will extend to rules that contain more than two or three items, since $\binom{m}{k}$ checks for similarity must be done, where m is the number of unique items in the set of transactions, and k is the number of items that might appear in any one rule. Removing the support requirement entirely is an elegant solution, but it comes at a high cost of space: for n transactions containing an average of k items over m possible items, the matrix will require n × m bits, whereas the primary data structure for Apriori-based algorithms will require $n \times \log_2 m \times k$ bits. Note that itemsets with low similarity may still produce interesting rules.

Confidence-Based Pruning

Another constraint known as confidence-based pruning was proposed by Wang et al. (2001). It finds all rules that satisfy a minimum confidence, but not necessarily a minimum support threshold. They call the rules that satisfy this requirement "confident rules." The problem with mining confident rules is that, unlike support, confidence does not have a downward closure property. Wang et al. (2001) proposed a confidence-based pruning that uses the confidence requirement in rule generation. Given three rules R_1: A → B, R_2: AC → B, and R_3: AD → B, R_2 and R_3 are two specialisations of R_1, having additional items C and D. C and D are exclusive and exhaustive in the sense that exactly one will hold up in each itemset but they will not appear together in the same itemset. The confidence of R_2 and R_3 must be greater than or equal to R_1. We can prune R_1 if neither R_2 nor R_3 is confident. This method has a universal existential upward closure. This states that if a rule of size k occurs above the given minimum confidence threshold, then for every other attribute not in the rule (C and D in the given example), some specialisation of size k+1 using the attribute must also be confident. They exploit this property to generate rules without having to use any support constraints.

H-Confidence

Xiong et al. (2003) try to improve on the previous confidence-based pruning method. They propose the h-confidence measure to mine hyperclique patterns. A hyperclique pattern is a type of association containing objects that are highly affiliated with each other, that is, every pair of objects in a hyperclique pattern is guaranteed to have a cosine similarity (uncentered correlation coefficient) above a certain level. They show that h-confidence has a cross-support property which is useful for eliminating candidate patterns having items with widely different supports. The h-confidence of an itemset $P = \{i_1, i_2, ..., i_m\}$ in a database D denoted by hconf(P, D), is a measure that reflects the overall affinity among items within the itemset.

$$\mathrm{hconf}(P) = \begin{array}{l} \min(\ \mathrm{conf}(\{i_1 \rightarrow i_2, \ldots, i_m\}), \\ \quad \mathrm{conf}(\{i_2 \rightarrow i_1, i_3, \ldots, i_m\}), \\ \quad \ldots \\ \quad \mathrm{conf}(\{i_m \rightarrow i_1, i_2, \ldots, i_{m-1}\})) \end{array}$$

A hyperclique pattern P is a strong-affinity association pattern because the presence of any item x ∈ P in a transaction strongly implies the presence of P\{x} in the same transaction. To that end, the h-confidence measure is designed specifically for capturing such strong affinity relationships. Nevertheless, even when including hyperclique patterns in rule generation, we can also miss interesting patterns. For example, an itemset {A,B,C} that produces low confidence rules A → BC, B → AC, and C → AB, but a high confidence rule AB → C, would never be identified.

Dense-Miner

Bayardo et al. (2000) noted that the candidate frequent itemsets generated are too numerous in dense data, even when using an item constraint. A dense dataset has many frequently occurring items, strong correlations between several items, and many items in each record. Thus Bayardo et al. (2000) use a consequent constraint-based rule mining approach called Dense-Miner. They require mined rules to have a given consequent C specified by the user. They also introduce an additional metric called improvement. The key idea is to extract rules with confidence above a minimum improvement value greater than any of the simplifications of a rule. A simplification of a rule is formed by removing one or more items from its antecedent. Any positive minimum improvement value would prevent unnecessarily complex rules from being generated. A rule is considered overly complex if simplifying its antecedent results in a rule with higher confidence. The improvement of a rule A → C is defined as the minimum difference between its confidence and the confidence of any proper sub-rule with the same consequent.

$$\text{improvement}(A \rightarrow C) = \text{conf}(A \rightarrow C)$$
$$- \max\{\text{conf}(A' \rightarrow C) | A' \subset A\}$$

If the improvement of a rule is greater than 0, then removing any non-empty combination of items from the antecedent will lower the confidence by at least the improvement. Thus every item and every combination of items present in the antecedent of a rule with a large improvement is an important contributor to its predictive ability. In contrast, it is considered undesirable for the improvement of a rule to be negative, as it suggests that the extra elements in the antecedent detract from the rule's predictive power.

Emerging Pattern (EP)

The Emerging Pattern (EP) method was proposed by Li et al. (1999). Given a known consequent C, a dataset partitioning approach is used to find top rules, zero-confidence rules, and μ-level confidence rules. The dataset, D, is divided into sub-datasets D_1 and D_2; where D_1 consists of the transactions containing the known consequent and D_2 consists of transactions which do not contain the consequent. All items in C are then removed from the transactions in D_1 and D_2. Using the transformed dataset, EP then finds all itemsets *X* which occur in D_1 but not in D_2. For each *X*, the rule *X* → *T* is a top rule in D with confidence of 100%. On the other hand, for all itemsets, Z, that only occur in D_2, all transactions in D which contain Z must not contain C. Therefore Z → C has a negative association and is a zero-confidence rule. For μ-level confidence rules Y → C the confidences are greater than or equal to 1 − μ. The confidences of μ-level rules must satisfy:

$$\frac{\sup(Y) | D_1 |}{\sup(Y) | D_1 | + \sup(Y) | D_2 |} \geq 1 - \mu$$

Note that $\sup(Y)|D_1|$ is the number of times itemset *YC* appears together in dataset *D* and $\sup(Y)|D_1| + \sup(Y)|D_2|$ is the number of times itemset *Y* appears in dataset *D*. This approach is considered efficient as it only needs one pass through the dataset to partition and transform it. Of course, in this method one must supply C.

Fixed-Consequent Association Rule Mining

Rahal et al. (2004) proposed a slightly different approach. They proposed a method that generates the highest support rules that matched the user's specified minimum without having to specify any support threshold. Fixed-consequent association rule mining generates confident minimal rules using two kinds of trees (Rahal et al., 2004). Given two rules, R_1 and R_2, with confidence values higher than the confidence threshold, where R_1 is A → C and R_2 is AB → C, R_1 is preferred, because the antecedent of R_2 is a superset of the antecedent of R_1. The support of R_1 is necessarily greater than or equal to R_2. R_1 is referred to as a minimal rule ("simplest" in the notation of Bayardo et al. (2000)) and R_2 is referred to as a non-minimal rule (more complex). The algorithm was devised to generate the highest support rules that match the user specified minimum confidence threshold without having the user specify any support threshold.

Apriori-Inverse

Apriori–Inverse (Koh et al., 2005) is a variation of the Apriori algorithm that uses the notion of maximum support instead of minimum support to generate candidate itemsets. Candidate itemsets of interest to us fall below a maximum support value but above a minimum absolute support value. Given a user-specified maximum support threshold, maxsup, and a generated minabssup value, we are interested in a rule X if sup(X) < maxsup and sup(X) > minabssup. Rules above maximum support are considered frequent rules, which are of no interest to us, whereas we consider rules appearing below the minimum absolute support value as coincidence. Rare rules are generated in the same manner as in Apriori rule generation. Apriori-Inverse produces rare rules which do not consider any itemsets above maxsup.

Apriori-Rare

Szathmary et al (2007) presented an approach for rare itemset mining from a dataset that splits the problem into two tasks. The first task, the traversal of the frequent zone in the space, is addressed by two different algorithms, a naıve one, Apriori-Rare, which relies on Apriori and hence enumerates all frequent itemsets; and MRG-Exp, which limits the considerations to frequent generators only. They consider computation of the rare itemsets that approaches them starting from the lattice bottom, from the frequent zone. They defined a positive and the negative border of the frequent itemsets, and a negative lower border and the positive lower border of the rare itemsets, respectively. An itemset is a maximal frequent itemset (MFI) if it is frequent but all its proper supersets are rare. An itemset is a minimal rare itemset (mRI) if it is rare but all its proper subsets are frequent.

REFERENCES

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In P. Buneman & S. Jajodia (Eds.), *Proceedings of the 1993 ACM SIGMOD international conference on management of data* (pp. 207–216). New York, NY: ACM Press.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Databases* (pp. 487–499). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Bayardo, R. J. (1998). Efficiently mining long patterns from databases. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, SIGMOD '98* (pp. 85–93). New York, NY: ACM Press.

Bayardo, R. J., & Agrawal, R. (1999). Mining the most interesting rules. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '99.,* (pp. 145–154). New York, NY: ACM Press.

Bayardo, R. J., Agrawal, R., & Gunopulos, D. (2000). Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery, 4*(2/3), 217–240. doi:10.1023/A:1009895914772

Cohen, E., Datar, M., Fujiwara, S., Gionis, A., Indyk, P., & Motwani, R. (2001). Finding interesting association rules without support pruning. *IEEE Transactions on Knowledge and Data Engineering, 13*, 64–78. doi:10.1109/69.908981

Hahsler, M. (2006, September). A model-based frequency constraint for mining associations from transaction data. *Data Mining and Knowledge Discovery, 13*(2), 137–166. doi:10.1007/s10618-005-0026-2

Koh, Y. S., & Rountree, N. (2005). Finding sporadic rules using Apriori-Inverse. In *Advances in Knowledge Discovery and Data Mining, 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining 2005* (pp. 97–106). Berlin / Heidelberg: Springer.

Koh, Y. S., Rountree, N., & O'Keefe, R. A. (2008, January). Mining interesting imperfectly sporadic rules. *Knowledge and Information Systems, 14*(2), 179–196. doi:10.1007/s10115-007-0074-6

Li, J., Zhang, X., Dong, G., Ramamohanarao, K., & Sun, Q. (1999). Efficient mining of high confidence association rules without support threshold. In *Proceedings of the 3rd European Conference on Principle and Practice of Knowledge Discovery in Databases, PKDD '99* (pp. 406 – 411).

Liu, B., Hsu, W., & Ma, Y. (1999a). Mining association rules with multiple minimum supports. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 337–341). New York, NY: ACM Press.

Liu, B., Hsu, W., & Ma, Y. (1999b). Pruning and summarizing the discovered associations. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 125–134). New York, NY: ACM Press.

Rahal, I., Ren, D., Wu, W., & Perrizo, W. (2004). Mining confident minimal rules with fixed-consequents. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence, ICTAI'04* (pp. 6–13). Washington, DC: IEEE Computer Society.

Seno, M., & Karypis, G. (2001). LPMINER: An algorithm for finding frequent itemsets using length-decreasing support constraint. In N. Cercone, T. Y. Lin, & X. Wu (Eds), In *Proceedings of the 2001 IEEE International Conference on Data Mining ICDM* (pp. 505–512). Washington, DC: IEEE Computer Society.

Szathmary, L., Napoli, A., & Valtchev, P. (2007). Towards Rare Itemset Mining. In *Proceedings of the 19th IEEE international Conference on Tools with Artificial intelligence - Vol.1 (ICTAI 2007) - Volume 01 (October 29 - 31, 2007). ICTAI. (pp. 305-312).* Washington, DC: IEEE Computer Society

Tao, F., Murtagh, F., & Farid, M. (2003). Weighted association rule mining using weighted support and significance framework. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03* (pp. 661–666). New York, NY: ACM Press.

Wang, K., He, Y., & Cheung, D. W. (2001). Mining confident rules without support requirement. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, (pp. 89–96). New York, NY: ACM Press.

Wang, K., He, Y., & Han, J. (2003). Pushing support constraints into association rules mining. *IEEE Transactions on Knowledge and Data Engineering, 15*(3), 642–658. doi:10.1109/TKDE.2003.1198396

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *SIGKDD Exploration Newsletter, 6*(1), 7–19. doi:10.1145/1007730.1007734

Xiong, H., Tan, P.-N., & Kumar, V. (2003). Mining strong affinity association patterns in data sets with skewed support distribution. In *Proceedings of the Third IEEE International Conference on Data Mining* (pp. 387 – 394). Washington, DC: IEEE Computer Society.

Yun, H., Ha, D., Hwang, B., & Ryu, K. H. (2003). Mining association rules on significant rare data using relative support. *Journal of Systems and Software, 67*(3), 181–191. doi:10.1016/S0164-1212(02)00128-0