Chapter 6
A Multi–Methodological Approach to Rare Association Rule Mining

Yun Sing Koh
Auckland University of Technology, New Zealand

Russel Pears
Auckland University of Technology, New Zealand

ABSTRACT

Rare association rule mining has received a great deal of attention in the past few years. In this chapter, the authors 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 that are not just rare to the dataset as a whole but are also rare to the cluster from which they are derived.

INTRODUCTION

The main goal of association rule mining is to discover relationships among sets of items in a transactional database. Association rule mining was introduced by Agrawal, Imielinski and Swami (1993). It aims to extract interesting correlations, frequent patterns, associations or causal 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 also known as association rules. In the classi-
A Multi-Methodological Approach to Rare Association Rule Mining

cal association rule mining process, all frequent itemsets are found, where an itemset is said to be frequent if it appears at least a given percentage of time in all transactions called minimum support s. Association rules are then generated in the form of $A \rightarrow B$ where $AB$ is a frequent itemset. Strong association rules are derived from frequent itemsets and constrained by minimum confidence c (the percentage of transactions containing A that also contain B).

A much less explored area in association mining is infrequent itemset mining or rare association rule mining. Items that rarely occur are in very few transactions and are normally pruned out in the classical association mining process. One limitation of common association rule mining approaches, i.e. Apriori, are that they rely on there being a meaningful minimum support level that is reasonable (sufficiently strong) 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 rare itemsets include identifying relatively rare diseases, predicting telecommunications equipment failure, and finding associations between infrequently purchased supermarket items. Indeed, infrequent itemsets warrant special attention because they are more difficult to find using traditional data mining techniques.

In this chapter we present a multi-methodological approach to finding rare association rules. Our approach integrates three separate strands of research, each of which contributes in different ways to finding rules of interest which would not be discovered using conventional methods.

The first and most fundamental challenge that needs to be overcome in the rare mining context is to ensure that candidate itemsets do not occur together purely by chance. Various different schemes have been proposed to combat this problem, including the use of the Pearson correlation coefficient to measure the strength of the relationship between co-occurring itemsets (Xiong, Shekhar, Tan, & Kumar, 2004). Our approach to this problem is to use an inverted Fisher test to find itemsets that co-occur together with a frequency greater than chance occurrence. The use of a rigorous statistical method such as the Fisher test was demonstrated in previous research (Koh, Rountree, & O’Keefe, 2006), (Koh, Rountree, & O’Keefe, 2008) to be superior to the use of subjective Pearson correlation coefficients.

We hypothesize that rare rules can manifest in only certain segments of a dataset and not appear across a dataset as a whole. Such rules can only be found by clustering transactions and then applying association mining on each of the clusters found. In previous research (Koh & Pears, 2008), we demonstrated that strong rules that were found in clusters could not be found by mining across the entire un-partitioned dataset due to contamination that occurs between different parts of the dataset.

Current methods in rare association rule mining do not discriminate between itemsets that have widely different levels of support. As long as an itemset meets the lower bound support threshold it is considered as a candidate itemset for rule generation. All such candidates can now be extended with other candidate itemsets to form composite itemsets which are guaranteed to have support lesser than or equal to the minimum of the individual itemsets. However, depending on the support threshold chosen, this could lead to a combinatorial explosion in the number of candidate itemsets, which in turn will lead to degradation in performance. Apart from this the number of rules generated could also grow exponentially, especially for dense datasets. We thus impose a control over the number of candidates by implementing a weighting scheme that associates a weight to each candidate itemset. A lower bound support threshold is initially chosen by the miner. As candidate generation progresses this threshold does not remain constant but is updated by the weight associated with an itemset that is...
Related Content

Clustering Mixed Incomplete Data
Jose Ruiz-Shulcloper, Guillermo Sanchez-Diaz and Mongi A. Abidi (2002). Heuristic and Optimization for Knowledge Discovery (pp. 89-106).
www.igi-global.com/chapter/clustering-mixed-incomplete-data/22151?camid=4v1a

Finding Persistent Strong Rules: Using Classification to Improve Association Mining
www.igi-global.com/chapter/finding-persistent-strong-rules/46892?camid=4v1a

A Multi-Methodological Approach to Rare Association Rule Mining
Yun Sing Koh and Russel Pears (2010). Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection (pp. 76-97).
www.igi-global.com/chapter/multi-methodological-approach-rare-association/36901?camid=4v1a

Incorporating Correlations among Gene Ontology Terms into Predicting Protein Functions
Pingzhao Hu, Hui Jiang and Andrew Emili (2011). Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances (pp. 154-173).
www.igi-global.com/chapter/incorporating-correlations-among-gene-ontology/53885?camid=4v1a