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
the analyst in advance. This user-specified parameter is called minimum support (minsup). Then association rules are derived in the form of $X \rightarrow Y$ 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 $bread \rightarrow milk$ 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 \{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 $espresso machine \rightarrow coffee grinder$ 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 $X \subseteq I$ 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 $Y \subseteq X$. 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.
Related Content

Protocol Identification of Encrypted Network Streams
Matthew Gebski, Alex Penev and Raymond K. Wong (2010). Complex Data Warehousing and Knowledge Discovery for Advanced Retrieval Development: Innovative Methods and Applications  (pp. 328-341).
www.igi-global.com/chapter/protocol-identification-encrypted-network-streams/39599?camid=4v1a

Heuristic Knowledge Discovery for Archaeological Data Using Genetic Algorithms and Rough Sets
Alina Lazar (2002). Heuristic and Optimization for Knowledge Discovery (pp. 263-278).
www.igi-global.com/chapter/heuristic-knowledge-discovery-archaeological-data/22159?camid=4v1a

Mining XML Schemas to Extract Conceptual Knowledge
Ivan Bedini, Benjamin Nguyen, Christopher Matheus, Peter F. Patel-Schneider and Aidan Boran (2012). Semi-Automatic Ontology Development: Processes and Resources  (pp. 79-105).
www.igi-global.com/chapter/mining-xml-schemas-extract-conceptual/63899?camid=4v1a

Drivers of Innovation
www.igi-global.com/chapter/drivers-of-innovation/117852?camid=4v1a