Finding Non-Coincidental Sporadic Rules Using Apriori-Inverse

Yun Sing Koh, University of Otago, New Zealand
Nathan Rountree, University of Otago, New Zealand
Richard O’Keefe, University of Otago, New Zealand

ABSTRACT

Discovering association rules efficiently is an important data mining problem. We define sporadic rules as those with low support but high confidence; for example, a rare association of two symptoms indicating a rare disease. To find such rules using the well-known Apriori algorithm, minimum support has to be set very low, producing a large number of trivial frequent itemsets. To alleviate this problem, we propose a new method of discovering sporadic rules without having to produce all other rules above the minimum support threshold. The new method, called Apriori-Inverse, 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. Rules above maximum support are considered frequent rules, which are of no interest to us, whereas rules that occur by chance fall below the minimum absolute support value. We define two classes of sporadic rule: perfectly sporadic rules (those that consist only of items falling below maximum support) and imperfectly sporadic rules (those that may contain items over the maximum support threshold). This article is an expanded version of Koh and Rountree (2005).

Keywords: Apriori-Inverse; association rules; maximum support; minimum absolute support value; sporadic rules

INTRODUCTION

Data mining is used for extracting useful information and discovering hidden relationships that exist in large databases (Chen & Liu, 2005). Association rule mining has become one of the most popular data exploration techniques, allowing users to generate unexpected rules from market-basket data. Originally proposed by Agrawal et al. (Agrawal, Imielinski & Swami, 1993; Agrawal & Srikant, 1994), association rule mining discovers all rules in the data that satisfy a user-specified minimum support (minsup) and minimum confidence (minconf). Minsup represents the minimum amount of evidence (that is, num-
number of transactions) we require to consider a rule valid, and minconf specifies how strong the implication of a rule must be to be considered valuable.

The following is a formal statement of association rule mining for transaction databases. Let $I = \{i_1, i_2, \ldots, i_m\}$ be the universe of items and $D$ be a set of transactions, where each transaction $T$ is a set of items such that $T \subseteq I$. An association rule is an implication of the form $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. $X$ is referred to as the antecedent of the rule, and $Y$ as the consequent. The rule $X \rightarrow Y$ holds in the transaction set $D$ with confidence $c\%$ if $c\%$ of transactions in $D$ that contain $X$ also contain $Y$. The rule $X \rightarrow Y$ has support $s\%$ in the transaction set $D$, if $s\%$ of transactions in $D$ contain $XY$ (Agrawal & Srikant, 1994). Throughout this article, we shall use $XY$ to denote an itemset that contains both $X$ and $Y$.

One measure of the predictive strength of a rule $X \rightarrow Y$ is its lift value, calculated as $\text{confidence}(X \rightarrow Y) / \text{support}(Y)$. Lift indicates the degree to which $Y$ is more likely to be present when $X$ is present; if lift is less than 1.0, $Y$ is less likely to be present with $X$ than $Y$’s baseline frequency in $D$. The task of generating association rules is that of generating all rules that meet minimum support and minimum confidence and perhaps meet further requirements such as having lift greater than 1.0.

The Apriori algorithm and its variations are used widely as association rule mining methods. However, several authors have pointed out that the Apriori algorithm, by definition, hinders us from finding rules with low support and high confidence (Cohen et al., 2001; Liu, Hsu, & Ma, 1999a; Yun et al., 2003). Apriori generates frequent itemsets (i.e., those that will produce rules with support higher than minsup) by joining the frequent itemsets of the previous pass and pruning those subsets that have a support lower than minsup. Hence, in order to generate rules that have low support, minsup must be set very low, drastically increasing the running time of the algorithm. This is known as the rare item problem. It means that, using the Apriori algorithm, we are unlikely to generate rules that may indicate events of potentially dramatic consequence. For example, we might miss out on rules that indicate the symptoms of a rare but fatal disease due to the frequency of incidence not reaching the minsup threshold. Some previous solutions to this problem are reviewed in the next section.

The aim of our research is to develop a technique to mine low support but high confidence rules effectively. Basically, this method is applicable to find itemsets that do not happen frequently, but when they do occur, they almost always occur together. We call such rules sporadic, because they represent rare cases that are scattered sporadically through the database but with high confidence of occurring together. In order to find sporadic rules with Apriori, we have to set a very low minsup threshold, drastically increasing the algorithm’s running time. In this article, we expand the Apriori-Inverse approach in Koh and Rountree (2005); we propose an algorithm to capture rules using a maximum support threshold and minimum absolute support values. First, we define the notion of a perfectly sporadic rule, where the itemset forming the rule consists only of items that are all below the maximum support threshold. We also establish a way of determining whether a sporadic rule is less likely than chance and should be ignored. To enable us to find imperfectly sporadic rules, we allow maximum support to be in-
Related Content

Knowledge-Based Induction of Clinical Prediction Rules
[www.irma-international.org/chapter/knowledge-based-induction-clinical-prediction/7541/](http://www.irma-international.org/chapter/knowledge-based-induction-clinical-prediction/7541/)

Big Data Literacy: A New Dimension of Digital Divide, Barriers in Learning via Exploring “Big Data”
[www.irma-international.org/chapter/big-data-literacy/150266/](http://www.irma-international.org/chapter/big-data-literacy/150266/)

SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence
[www.irma-international.org/article/slod-bi/130664/](http://www.irma-international.org/article/slod-bi/130664/)

Discovering Hidden Concepts in Predictive Models for Texts’ Polarization
[www.irma-international.org/article/discovering-hidden-concepts-in-predictive-models-for-texts-polarization/130665/](http://www.irma-international.org/article/discovering-hidden-concepts-in-predictive-models-for-texts-polarization/130665/)

Ecological Data Exploration
[www.irma-international.org/chapter/ecological-data-exploration/216092/](http://www.irma-international.org/chapter/ecological-data-exploration/216092/)