IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 1, March, 2013
ISSN: 2320 - 8791
www.ijreat.org

Mine Frequent Patterns using Valid Measures

T.Immanuel[1] and B.Dwarakanath[2]

[1] Student, Department of Information Technology, Hindustan Institute of Technology and Science, Chennai, Tamil Nadu, India.

[2] Assistant Professor, Department of Information Technology, Hindustan Institute of Technology and Science, Chennai, Tamil Nadu, India.

Abstract

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases whose main task is to extract frequent itemsets and to generate association rules for these frequent itemsets. It is intended to identify strong rules discovered in databases using different measures of interestingness. Many techniques for association rule mining require a suitable metric to capture the dependencies among variables in a dataset. Support and Confidence are the interestingness measures used for discovering relevant association rules. Although support and confidence are the appropriate measures for building a strong model in many cases, they are still not the ideal measures. Hence we present an overview of various measures which describes how one should examine in order to select the right measure for a given application domain. A comparative study is made focusing on the objective measures for association rules. The results show that the interestingness measures can significantly reduce the number of rules according to accuracy of the specific domain.

Keywords: *Association Rule Mining, Interestingness Measures, Frequent Itemsets.*

1. Introduction

Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases, such as association rules. Association rules are important in data mining particularly in analyzing and predicting the customer behavior. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Mining frequent patterns usually suffers from the problem of huge output. i.e., the huge number of frequent patterns produced .The number of frequent patterns grows exponentially with respect to the user specified threshold value. Such huge output may include many redundant patterns, which waste resources for later analysis, and makes it difficult for users to utilize data. So that in order to mine interesting patterns, we are going to use interestingness measures.

Frequent pattern mining has been a focused theme in data mining research for over a decade. Abundant literature has been dedicated to this research and tremendous progress has been made, ranging from efficient and scalable algorithms for frequent itemset mining in transaction databases to numerous research frontiers, such as sequential pattern mining, structured pattern mining, correlation mining, associative classification, and frequent pattern-based clustering, as well as their broad applications.

This paper is organized as follows: Section 2 discusses some existing schemes relevant to those proposed in this paper. Section 3 discusses Association Rule Mining. Section 4 describes the Interestingness Measures Section 5 discusses the Experimental Results Section 6 draws conclusions.

2. Related Work

Frequent pattern mining was first proposed by Agrawal et al. (1993) for market basket analysis in the form of association rule mining. It analyses customer buying habits by finding associations between the different items that customers place in their "shopping baskets".

For instance, if customers are buying milk, how likely are they going to also buy cereal (and what kind of cereal) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and arrange their shelf space.
Since the first proposal of this new data mining task and its associated efficient mining algorithms, there have been hundreds of follow-up research publications, on various kinds of extensions and applications, ranging from scalable

data mining methodologies, to handling a wide diversity of data types, various extended mining tasks, and a variety of new applications. With over a decade of Substantial and fruitful research, it is time to perform an overview of this flourishing field and examine what more to be done in order to turn this technology a cornerstone approach in data mining applications.

J. Han et al. [1] developed the research on how frequent pattern mining gets its execution and what are the efficient and scalable methods for mining frequent patterns. They also explains the methodologies which involve the type of algorithm to be used .

P. Lenca et al.[5] developed a two-step solution to the problem of the recommendation of one or more user-adapted interestingness measures. First, a description of interestingness measures, based on meaningful classical properties, is given. Second, a multicriteria decision aid process is applied to this analysis and illustrates the benefit that a user, who is not a data mining expert, can achieve with such methods.

R Vijaya Prakash et al[4] proposed two approaches which measure association rule to help and evaluate their interestingness. They use the measures of diversity and peculiarity to identify those rules that are potentially useful.

Kwang-Il Ahn [3] proposed the algorithm that consists of three processes, which include mining associations among items, nearest neighbor assignments, and updating assignments. The algorithm was tested on synthetic databases. The results show very effective product assignment in terms of the potential for cross-selling to drive maximum sales in retail.

L. Geng et al[2]explains the measure of interestingness for the pattern to be found and the role of measures in the field of data mining. He describes the measures for association and classification rules, Selection Strategies for Probability-Based Objective Measures and the elimination of the uninteresting patterns.He describes a data model which explains how the information relevant to the utility is organized in the dataset.

3. ASSOCIATION RULE MINING

Association rule mining is a category of data mining tasks that correlate a set of items with other sets of items in a database. Association rules "aim to extract interesting correlations, frequent patterns, associations or causal structures among sets of items in the transaction databases or other repositories". Association rules were first proposed by Agrawal et al. (Agrawal, Imielinski, & Swami, 1993). The main driver for research on association rules was the analysis of customer market basket transactions. An example of an association rule is as follows. 60% of customers that purchase potato chips also purchase soda in the same transaction. Agrawal et al.'s work established a formal model for association rules and establishes algorithms that find large itemsets, confidence, and support of each rule discovered in the itemset. Association rule algorithms can generate thousands of rules, many of which can be redundant. These redundant rules are essentially useless, so researchers have solved this problem by defining new interestingness measures, incorporating constraints, or by designing templates to mine for restricted rules. Also, a primary goal of knowledge discovery in databases is to produce interesting rules that can be interpreted by a user (Lenca et al)

One research team (Lee & Siau) outlined the requirements and challenges associated with data mining. First, data mining must be able to handle different types of data. Second, data mining algorithms must be scalable and efficient. Third, data mining must be able to handle noisy and missing data. Fourth, data mining techniques should present results in a way that is easy to understand. Fifth, data mining techniques should support requests at different levels of granularity. That is, data mining can be done at different levels of abstraction. Sixth, data mining algorithms should be flexible enough to deal with data from different sources. Finally, a major concern within data mining today is the threat to privacy and data security. This is because data mining makes it easy to establish profiles of individuals based on data from multiple sources.

General issues related to data mining include the identification of missing information, dealing with noise or missing values, and operating with very large databases (VLDBs). Additionally, data mining is normally used to access data contained in a data warehouse, which contain high degrees of dimensionality, thus making data mining more complex. In order to produce accurate data mining results, it is important that the underlying data is complete. Without complete data, accurate rules cannot be produced.

4. INTERESTINGNESS MEASURES

Two important measures within association rule mining are support and confidence. Support for an association rule is the percentage of transactions in the database that contain

XUY. Confidence for an association rule (sometimes denoted as strength, or α) X⇒Y is the ratio of the number of transactions that contain XUY to the number of transactions that contain X. In other words, support describes how often the rule would appear in the database, while confidence measures the strength of the rule. A user establishes minimum support (minsup) and minimum confidence (minconf). Rules are then generated based on those criteria. Users can select minsup and minconf parameters before or after rule generation.

Measuring the interestingness of discovered patterns is an active and important area of data mining research. These measures are intended for selecting and ranking patterns according to their potential interest to the user. Good measures also allow the time and space costs of the mining process to be reduced.

Association rule algorithms are designed to efficiently find large itemsets. Large itemsets are those that have a number of occurrences above some minimum threshold (Dunham, 2003). The reason we are more interested in large itemsets is that many of the produced association rules may not be interesting. Apriori is an important algorithm in association rule mining. The apriori algorithm was first established by Agrawal and Srikant. It is the major technique used to detect large itemsets within a database of transactions. It also forms the basis of many association rule algorithms .

Hilderman and Hamilton established three primary principles that a good interestingness measure should satisfy : The minimum value principle, which states that a uniform distribution is the most uninteresting. The maximum value principle, which states the most uneven distribution is the most interesting. The skewness principle, which states that the interestingness measure for the most uneven distribution will decrease when then number of classes of tuples increases. The permutation invariance principle, which states that interestingness for diversity is unrelated to the order of the class and it is only determined by the distribution of counts. The transfer principle, which states that interestingness increases when a positive transfer is made from the count of one tuple to another whose count is greater.

5. EXPERIMENTAL RESULTS

The open-source data mining toolkit, Orange, was used to conduct the study. Orange uses the Python programming language, which allows the programmer to extend or adapt modules for specific experiments. The Breast-cancer data set was analyzed and compared using the basic interestingness measures: support, confidence.

Table 1: Measures Used

supp	*conf*	*Leve*	*Stre*	*Lift*	*Cove*	*Info-Gain*
0.703	0.944	0.125	1.042	1.216	0.745	0.085
0.657	0.862	0.66	1.018	1.111	0.762	0.046
0.640	0.839	0.072	0.977	1.127	0.762	0.052
0.619	0.967	0.122	1.213	1.246	0.640	0.096

Table 2:Rules Generated

Rule
inv-nodes=0-2 -> node-caps=no
irradiat=no->node-caps=no
inv-nodes=0-2->irradiat=no
inv-nodes=0-2 irradiat=no->node-caps=no

5. CONCLUSION

Associative classification is a relatively new paradigm for Classification relying on association rule mining and naturally inherits the most commonly used interestingness measures, support and confidence. These are not necessarily the best choice and no systematic study was undertaken to identify the most appropriate measures from the myriad measures already used as filters or rankers for relevant rules in different fields. This study is to answer the question whether other measures are more suited for the different phases of the associative classifier, and an attempt to identify the best measure for each phase. The results clearly indicate that many interestingness measures can indeed provide a better set of classification rules. Another observation is that using the combination of the best measures in pruning and selection phases does not improve the accuracy of the classifier which means that the best selecting measure for an original rule set is not the best for the pruned version of that rule set. This observation shows that there might exist some rule set characteristics that have effect on selecting the best measure.

References

[1] Jiawei Han,Hong Cheng,Dong xin,Xifeng yan "Frequent Pattern Mining:Current status & future directions"Springer,Data Min Knowl Disc,2007.
[2] Liqiang Geng and Howard J.Hamilton "Interestingness Measures for Data Mining: A Survey", ACM Computing Surveys, Vol. 38, No. 3, Article 9, September 2006.
[3] Kwang-Il Ahn "Effective product assignment based on association rule mining in retail"Elsevier,2012.
[4] R Vijaya Prakash, Dr. A. Govardhan, Sarma "Interestingness Measures for Multi-level Association

Rules", Information and Knowledge Management
ISSN 2224-5758 (Paper) ISSN 2224-896X (Online)
Vol 2, No.6, 2012.

[5] Philippe Lenca "On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid",sma preprint pm-pp-06-01-v01.