An Analysis on Various Measurements for Association Pattern Identification in a Multiple Database

Dr. Shailendra Singh Sikarwar[1] Mahesh Bansal[2]

[1]Assistant Professor, P. G. V. College, Gwalior

[2]Assistant Professor, P. G. V. College, Gwalior

Abstract – Data mining is an area of data analysis that has arisen in response to new data analysis challenges, such as those posed by massive data sets or non-traditional types of data. Association analysis, which seeks to find patterns that describe the relationships of attributes (variables) in a binary data set, is an area of data mining that has created a unique set of data analysis tools and concepts that have been widely employed in business and science. The objective measures used to evaluate the interestingness of association patterns are a key aspect of association analysis. Indeed, different objective measures define different association patterns with different properties and applications.

This paper first provides a general discussion of objective measures for assessing the interestingness of association patterns. It then focuses on one of these measures, h-confidence, which is appropriate for binary data sets with skewed distributions. The usefulness of h-confidence and the association pattern that it defines—a hyper clique—is illustrated by an application that involves finding functional modules from protein complex data.

----------♦----------

INTRODUCTION

Many different types of data analysis techniques have been developed in a wide variety of fields, including mathematics, statistics, machine learning, pattern recognition, and signal processing. Data mining is an area of data analysis that has arisen in response to new data analysis challenges, such as those posed by massive data sets or non-traditional types of data. In some cases, data mining solves current data analysis problems by combining existing data analysis techniques with innovative algorithms. In other cases, new data analysis techniques have been developed. For example, association analysis, which seeks to find patterns that describe the relationships of attributes (variables) in a binary data set, is an area of data mining that has created a unique set of data analysis tools and concepts that have been widely employed in both business and science.

Association analysis [AIS93, AS94] analyzes transaction data, such as the data generated when customers purchase items in a store. (The items purchased by a customer are a transaction.) A key task of this analysis is finding frequent itemsets, which are sets of items that frequently occur together in a transaction. For example, baby formula and diapers are items that may often be purchased together.

Association analysis of things (variables) and patterns in a database could assume imperative parts in finding answers for numerous issues. In the setting of market crate data, one could perform different sorts of association breakdowns of things obtained. Likewise, there might exist new sorts of example handy in tackling diverse issues. We get to know fascinating purchasing patterns of clients by examining a large volume of data. Past take a shot at mining successive itemsets, association rules, and negative association rules may have not addressed all the inquiries of a data excavator, or a chief.

There are various reasons why mining on multiple databases becomes an important issue in the recent time. In the following paragraph, we mention a few reasons that motivate us to work on mining multiple databases.

n Due to a liberal economic policy adopted by many countries across the globe, the number of branches of a multi-national company as well as the number of multinational companies is increasing over time. Moreover, the economies of many countries are growing at a faster rate. As a result the number of multi-branch companies within a country is also increasing. Many multi-branch companies deal with multiple databases, since local transactions are stored locally. Thus, it is necessary to study data mining on multiple databases. Many decision-making problems are based on knowledge distributed across the branch databases.

Most of the previous pieces of data mining work are based on a single database. Thus, it is necessary to study data mining on multiple databases.

A BROADER VIEW OF ASSOCIATION PATTERNS AND THEIR MEASURES

More generally, an itemset pattern or association rule is defined by the measure that is selected to evaluate the strength of the association. Traditionally, support is used to measure the strength of an itemset, while support and confidence are used to measure the strength of an association rule. However, by defining different association measures, it is possible to find different types of association patterns or rules that are appropriate for different types of data and applications. This situation is analogous to that of using different objective functions for measuring the goodness of a set of clusters in order to obtain different types of clusterings for different types of data and applications.

Thus, association analysis is fundamentally concerned with defining new association measures. These measures, together with a threshold, select itemsets or rules that are of interest. What might motivate the creation of a new association measure? Most often, the development of new measures is motivated by the limitations of support and/or confidence or the desirable properties of some new measure.

However, besides providing new capabilities, these new association measures must be cognizant of the practical realties addressed by the current association measures of support and confidence. In particular, two important goals are computational efficiency and distinguishing interesting patterns from spurious ones. As the size and dimensionality of real world databases can be very large, one could easily end up with thousands or even millions of patterns, many of which might not be interesting. It is therefore important to establish the appropriate criteria for evaluating the quality of the derived patterns. There are two criteria often used to prune uninteresting patterns. First, patterns that involve a set of mutually independent items or cover very few transactions are often considered uninteresting.

Second, redundant patterns are considered uninteresting because they correspond to sub-patterns of other interesting patterns. In both cases, various objective interestingness measures have been proposed to help evaluate the patterns.

CAPTURING ASSOCIATION AMONG ITEMS IN A DATABASE

The analysis of relationships around variables is a key errand being at the heart of numerous data mining issues. For example, association rules find relationships between sets of things in a database of transactions. Such rules express purchasing patterns of clients, e.g., finding how the vicinity of one thing influences the vicinity of an alternate et cetera.

Numerous measures of association have been accounted for in the written works of data mining, machine studying, and detail. They could be sorted into two assemblies. A few measures manage a set of items, or could be summed up to manage a set of articles. Then again, the remaining measures couldn't be summed up. Trust, conviction are cases of the second class of measures. Then again, measures, for example Jaccard could be summed up to find association around a set of things in a database. We ought see later why measures, for example uphold , summed up Jaccard, and all-trust have not been viable in measuring association around a set of things in a database.

Different issues could be tended to utilizing association around a set of things in market bushel data. Case in point, an organization could be intrigued by examining things that are obtained often. Let the things P, Q, and R be bought habitually. A couple of particular issues are expressed underneath including these things.

- Some things (items) could be high benefit making. Commonly, the organization might want to advertise them. There are different ways one could push a thing. A circuitous method for pushing a thing P is to advertise things that are exceptionally connected with it. The suggestion of high association between P and an alternate thing Q is that if Q is obtained by a client then P is prone to be bought by the same client in the meantime. Along these lines, P gets by implication pushed.
- Again, a few things could be low-benefit making. Consequently, it is vital to know how they push

offers of different things. Overall, the organization could quit managing such things.

To take care of the above issues, one could bunch the regular things in a database. In the connection of (i), one could push thing P by implication, by advertising different things in the class holding P. In the connection of (ii), the organization could continue managing R if the class size holding R is sensibly large. Along these lines, a suitable metric for catching association around a set of things could empower us to group visit things in a database. All in all, numerous corporate choices could be taken adequately by consolidating knowledge intrinsic in data. Later, we should show that a measure of association dependent upon a 2 x 2 possibility table may not be viable in bunching a set of things in a database. In this manner, we propose measures of association for catching association around a set of things in a database.

In this section, we exhibit two measures of association around a set of things in a database. The second measure of association is dependent upon a weighting model. We furnish hypothetical establishment of the work. With the end goal of measuring association around a set of things, we express second measure regarding underpins of itemsets. The principle commitments of this part are given as takes after: (1) We propose two measures of association around a set of things in a database, (2) We present the thought of acquainted itemset in a database, (3) We give hypothetical establishment of the work, and (4) We express second measure regarding underpins of itemsets.

GLOBAL EXCEPTIONAL PATTERNS IN MULTIPLE DATABASES

Many multi-branch companies transact from different locations. Many of them collect a huge amount of transactional data continuously through their different branches. Due to a growth-oriented and liberal economic policy adopted by many countries across the globe, the number of such companies as well as the number of branches of such a company is increasing over time. Moreover, most of the pieces of data mining work are based on a single database. Thus, it is important to study data mining on multiple databases. Analysis and synthesis of patterns in multiple databases is an important as well as interesting issue.

Based on the number of data sources, patterns in multiple databases could be classified into three categories. They are local patterns, global patterns, and patterns that are neither local nor global. A pattern based on a branch database is called a local pattern. On the other hand, a global pattern is based on all the databases under consideration. Global patterns are useful for global data analyses and global decision making. There exist other types of patterns in multiple databases. For example, frequent itemset, positive associative rule and clustering of relevant objects. There is no fixed set of attributes to describe these patterns, since there are different types of pattern
in a database. Each type of pattern could be described by a specific set of attributes.

Itemset patterns influence KDD research heavily in following ways: Firstly, many interesting algorithms have been reported on mining itemset patterns in a database , , . Secondly, an itemset could be considered as a basic type of pattern in a transactional database, since many patterns are derived from the itemset patterns in a database. Some examples of derived patterns are positive association rule , negative association rule , conditional pattern in a database and high-frequent association rule , heavy association rule , exceptional association rule in multiple databases.

CONCLUSION

In this paper we discussed objective measures for assessing the interestingness of association patterns—itemsets and rules—for binary transaction data. Traditionally, the measures of support and confidence have been used to evaluate itemsets and association rules, but, as was described, these measures are not appropriate in all situations.

In this section, we introduce two measures of association around things in an itemset in a database. An existing measure may not be successful in catching association around things in an itemset of size more amazing than 2. Numerous research issues could come down to catching association around things in an itemset.

We have presented the thought of cooperative itemset in a database. We have furnished numerous handy lemmas and samples to make establishment of proposed measures solid and clear. Utilizing monotone property of a measure of association, we have indicated that A2 measures association around things in an itemset more precisely than A With the end goal of processing A2, we express it as far as backings of itemsets. The measure of association A2 is adequate in catching factual association around things in a database.

The customary help certainty schema for mining association rules is dependent upon a binary database. It has constrained use in association analysis of things, since a genuine transaction may hold a thing various times.

The accepted backing certainty structure is dependent upon the recurrence of an itemset in a binary database. In a Timt sort database, there are two sorts of recurrence of an itemset viz., transaction recurrence, and database recurrence. Because of these explanations, we get the accompanying classifications of association rules in a Timt sort database: (i) Association rules impelled by transaction recurrence of an itemset, (ii) Association rules actuated by database recurrence of an itemset, and (iii) Association rules affected by both transaction recurrence and database recurrence of an itemset. We have presented a structure for mining every classification of association rules. The proposed skeletons are viable for contemplating association around things in genuine market wicker bin data.

Universal backing certainty schema has not been adequate in finding association rules in true market wicker container data. A thing in a database could be acquired various times in a transaction. Hence, there are two sorts of recurrence of an itemset in a database: the amount of transactions in the database holding the itemset, and the amount of events of the itemset in the database. Hence, one could study association rules regarding these sorts of recurrence of an itemset. We have proposed structures for three separate classes of association rules in a database. We accept that such skeleton might help mulling over association between a couple of itemsets in genuine market wicker container data.

REFERENCES

- Adhikari, P. R. Rao, "Study of select Items in multiple databases by grouping", *Proceedings of 3rd Indian International Conference on Artificial Intelligence,* 2007, pp. 1699 - 1718.
- Adhikari, P. R. Rao, "Synthesizing heavy association rules from different real data sources", *Pattern Recognition Letters* 29(1), 2008, pp. 59-71.
- C. C. Aggarwal and P. S. Yu, Mining associations with the collective strength ap- proach, IEEE Transactions on Knowledge and Data Engineering 13 (2001), no. 6, 863–873.
- D. Cheung, V. Ng, A. Fu, Y. Fu, "Efficient mining of association rules in distributed databases", *IEEE Transactions on Knowledge and Data Engineering* 8(6), 1996, pp. 911 - 922.
- Edward R. Omiecinski, Alternative interest measures for mining associations in databases, IEEE TKDE 15 (2003), no. 1, 57–69.
- G. K. Palshikar, M. S. Kale, M. M. Apte, "Association rule mining using heavy itemsets", *Proceedings of Eleventh International Conf on Management of Data,* 2005, pp. 148 - 155.
- H. Liu, H. Lu, J. Yao, "Toward multi-database mining: Identifying relevant databases", *IEEE Transactions on Knowledge and Data Engineering* 13(4), 2001, pp. 541 - 553.
- Jeudy, J. F. Boulicaut, "Using condensed representations for interactive association rule mining", *Proceedings of PKDD,* LNAI 2431, 2002, pp. 225 - 236.
- K. Ali, S. Manganaris, R. Srikant, "Partial classification using association rules", *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining,* 1997, pp. 115-118.
- M.-L. Antonie, O.R. Zaiane, "Mining positive and negative association rules: An approach for confined rules", *Proceedings of PKDD,* 2004, pp. 27 - 38.
- Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava, Selecting the right objective measure for association analysis, Information Systems 29 (2004), no. 4, 293–313.
- R. Agrawal, J. Shafer, "Parallel mining of association rules", *IEEE Transactions on Knowledge and Data Engineering* 8(6), 1999, pp. 962 - 969.
- R. J. Hilderman, H. J. Hamilton, "Knowledge discovery and interestingness measures: A survey", Technical Report CS-99-04, *Department of Computer Science, University of Regina,* 1999.
- Richard J. Bolton, David J. Hand, and Niall M. Adams, Determining hit rate in pattern search, Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery (London, UK), Springer-Verlag, 2002, pp. 36–48.
- Savasere, E. Omiecinski, S. Navathe, "An efficient algorithm for mining association rules in large databases", *Proceedings of the 21st International Conference on Very Large Data Bases,* 1995, pp. 432 - 443.