Volume 5, Issue 12, December 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Hiding Sensitive Itemset in Multi-Dimensional Relation

Dr. K. Kavitha
Assistant Professor, Department of Computer Science,
Mother Teresa Women's University,
Kodaikanal, India

Abstract- Extracting frequent pattern from large date repositories is a major task of data mining. The most demanding strategy is to discover an effective rule generation. Association rule mining is a crucial one for finding the frequent itemsets from high voluminous databases. Many algorithms were developed to find the frequently occurred itemsets. The objective of the research is to find the interesting rules using multidimensional dataset. The proposed method is to generate more number of effective interesting rules that satisfy weight based on min-confidence and lift. This paper presents the performance of existing algorithms with the proposed method.

Keywords: Association Rule, Quantitative method, Discretization, Weight, Support, Confidence

I. INTRODUCTION

Data Mining is the process of extracting useful, but unknown information from high voluminous datasets. Mining of data includes formative process such as data preprocessing (i.e) Noise Removal and irrelevant data, Duplicate removal and redundant data etc. Then the data is converted in to the required format it should be able to access. This transformation is done with the help of various methods and tools[1].The Association Rule discovery and generation is most important part of every data mining task. The Association presents the relationship between the antecedent and resultant. Data mining is a procedure of knowledge discovery, "consists of utilizing data analysis and discovery algorithms that make a particular enumeration of patterns over the data". Association rules are employed to distinguish the relationships among a set of items in the databases. Two important parameters considered in the association rule mining, support and trust.
An association rule is of a relation A=>B, where support is the percent of transactions that contains A and B, whereas confidence is the percent of transaction contains both A and B values. Association rules are called strong rule if they fulfill both the minimal support threshold and the minimum confidence threshold.

II. RELATED WORK

Apriori is the most widely used Association rule mining algorithm. This is also undergoing several operations and many algorithms based on Apriori such as Improved Apriori, Custom built Apriori, Weighted Apriori Etc were in being. The Apriori property shows that every subsets of the frequent itemset must also be a frequent item set [2].
Next, FP – Growth (Frequent Pattern) growth algorithm, which produces considerably more number of frequent Itemsets, require iterative scanning of the database [3]. At the first scanning step, 1- itemset is generated, followed by removal of infrequent itemset. The revised FP- growth algorithm includes a minor change that the use of entropy heuristic and new batch search mechanism is introduced. The revised FP-growth algorithms are revised one uses the entropy heuristic during the generation of FP-tree and replaces the one-by one scanning search mechanism with the batch search mechanism.
MQRG Algorithm introduces a new multidimensional quantitative method to handle the categorical attributes and the numerical attributes in an effective manner. The conversion of information into quantitative method is taken place rapidly. Using the binary patterns, frequent patterns are identified using FPgrowth with minimum confidence threshold limit.. The conversion reveals all the frequent patterns from the database [6].
The rest of the paper is organized as follows. The proposed work is presented in Section 2. Section 3 contains the comparative analysis results obtained. Finally, the conclusion is drawn at section4.

III. MULTI-DIMENSIONAL RULE HIDING ALGORITHM

The proposed method filters an interesting rule based on different formats of representation such as multi-dimensional, relational, quantitative database. Binary database returns to generate the frequent itemsets and the effective associative rule generation. Initially, Transactional Database has to be changed into a binary format and then the discretization method is enforced to the Binary Format Table. Grounded on this method, Binary Frequent Patterns are generated using the binary values 0 and 1. Finally, Most Frequent Itemsets are filtered with the patronage of the threshold limit Weight. The following figure 1 illustrates the overall structure of the proposed method.

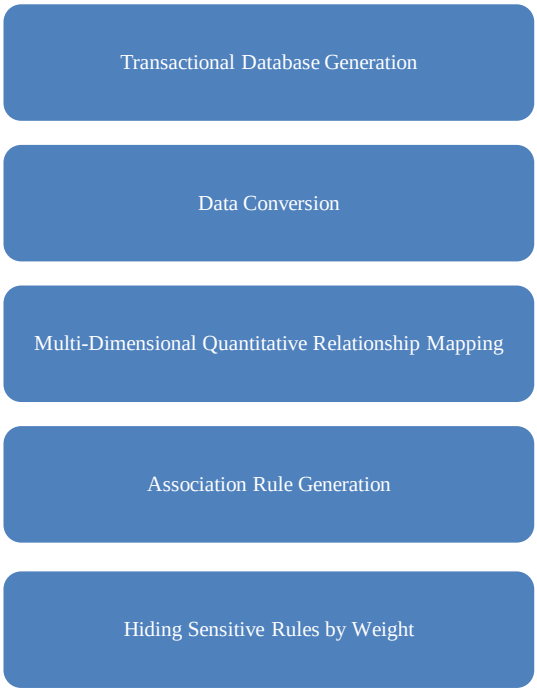

Figure-1 Overall Structure

Algorithm

Input: Dataset composed of N-Tuples, minsup, minconfidence and number of attributes.
Output: multidimensional association rules R
Step 1: Select particular set of attributes for the specified database.
Step 2: Let C_t be a set of constraints defined on these attributes
Step 3: Generate concept hierarchy for attributes
Step 4: Compute data conversion based on quantitative and binary values
Step 5: Associates Multi-Dimensional Relationship and find the relationship between multiple relations.
Step 6: Generate the rules from data extraction
Step 7: Filter the rules which satisfy the threshold limit attribute weight [(MinConfidence + Lift)/2].

A relational database consists of a table in which rows a represents a tuple which columns represents an attribute of the databases. Boolean, numerical and character attributes are shown in the database. For simplification, the proposed method is enforced in the bank loan database having six attributes. It consists of 8 records as shown in table 1.

Table-1 Student Database

T id	Sex	Age	Annual_Income	Category	Property_ Worth
1	Male	50	4.5 Lakhs	Govt.employed	10 Lakhs
2	Female	35	10 Lakhs	Self employed	5 Lakhs
3	Male	52	4.3 Lakhs	Self employed	8 Lakhs
4	Male	38	15 Lakhs	Govt employed	5 Lakhs
5	Female	57	3 Lakhs	Self employed	10 Lakhs
6	Female	40	4 Lakhs	Govt employed	4 Lakhs
7	Male	55	8 Lakhs	Govt employed	15 Lakhs
8	Male	56	2.5 Lakhs	Govt employed	3.5 Lakhs

The database includes data about the customer who applied for a loan, such as client ID, sex, age, annual income etc. Among those attributes, properties with annual income and age are regarded as numerical attribute where category sex are categorical attribute. The attributes are chosen according to the constraints. These attributes are useful to extract association rules.

Database contains large number of class attributes[9]. Before mining the transactional database, It maps all type of categories while converts numerical values and converts the relational data table into binary table form as shown in table 2.

Table-2 Binary Table Conversion

T id	Sex	Age	Annual_Income	Category	Property_ Worth
1	**1**	**0**	**0**	**1**	**1**

2	**0**	**1**	**1**	**0**	**1**
3	**1**	**0**	**0**	**0**	**1**
4	**1**	**1**	**1**	**1**	**0**
5	**0**	**0**	**0**	**0**	**1**
6	**0**	**1**	**0**	**1**	**0**
7	**1**	**0**	**1**	**1**	**1**
8	**1**	**0**	**0**	**1**	**0**

The proposed method is applied in table 1 to convert the data storage format and then pre-processing strategy is conducted. For example, the attribute is categorized as government employed and self-employed category for while substitute the values input and output respectively. Similarly, all the categorical attribute are classified into input and output respectively. For numerical attributes, data discretization techniques can be used to reduce the number of values by dividing the range into itemsets[6]. From table 1 attribute age has values from 30 to 60. In such situations all the values cannot be observed during rule generating. Therefore, the values are divided into two ranges such as one above the average and other below the average values.

In order to identify the in divided frequent items the proposed method assign the coordinate as the combination of attribute column number and the binary value presented in each position of the binary table as shown in table 3.

Table-3 Binary Frequent Patterns

T id	**Sex**	**Age**	**Annual_Income**	**Category**	**Property_ Worth**
1	1,1	2,0	3,0	4,1	5,1
2	1,0	2,1	3,1	4,0	5,1
3	1,1	2,0	3,0	4,0	5,1
4	1,1	2,1	3,1	4,1	5,1
5	1,0	2,0	3,0	4,0	5,1
6	1,0	2,1	3,0	4,1	5,0
7	1,1	2,0	3,1	4,1	5,1
8	1,1	2,0	3,0	4,1	5,0

Frequent item plays an important role in the frequent pattern mining. It is necessary to perform transaction scan to identify the set of frequent items. It extracts the database information from the given set of database based on the given constraint. The effective associative rule mining is done by making relationship between multi-dimensional data which is based on the constraints weight [(lift + confidence)/2]. The threshold limit filters and produces an effective association rule.

IV. EXPERIMENTAL ANALYSIS

The aim of the work is to obtain an associative model that allows input variables and related to loan database and output variables related to the software product and the software process. For experimental study, various types of datasets are used to justify the efficiency of proposed method. The datasets are iris and mushroom. Some characteristics of the above datasets are shown in the following table.

Table-4 Datasets Applied

Data	**Items**	**Transaction_Length**
Mushroom	120	23
Iris	130	43

The proposed method was compared with the classical algorithm, Apriori and MQRG Algorithm using java. The following table shows the run time of the compared algorithm on iris data with different number of records represented by total transactions in % under minsup and minimum confidence and Lift runs faster than apriori and MQRG.

Table-5 RunTime for Mushroom Data

Number of Records	**Apriori in MSec**	**MQRG in MSec**	**MDRH in Msec**
100	188	37	32
200	196	84	80
500	156	81	75
1000	379	152	136
2000	564	212	192

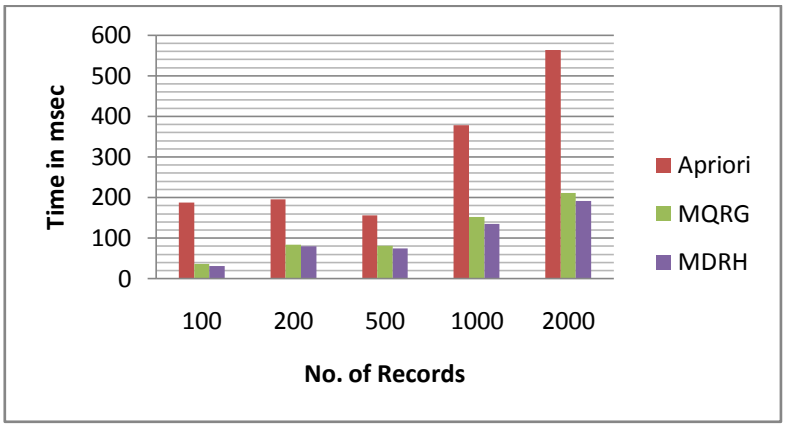

Table-5 and Figure-2 shows the performance comparison of the existing algorithm on Mushroom Dataset with proposed algorithm. For this dataset, MDRH algorithm runs faster than other two algorithms.
Table-6 shows the relative performance of the algorithms on Iris Dataset as shown below.

Table-6 RunTime for IRIS Data

Number of Records	Apriori in MSec	MQRG in MSec	MDRH in Msec
100	140	32	28
200	157	41	35
500	219	59	47
1000	307	146	133
2000	432	231	202

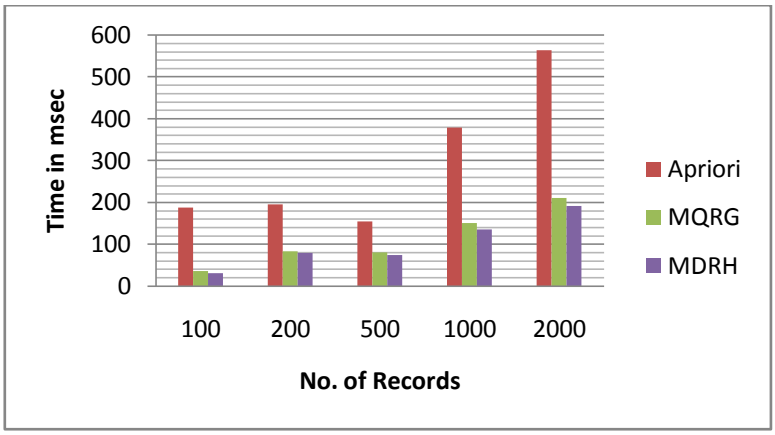

Figure-3 illustrates the comparative study of the existing algorithms(Apriori and MQRG) and proposed(MDRH) work for the Iris Dataset.

V. CONCLUSION

In this paper, a new method is proposed to generate efficient frequent rules. The main feature of the method is to reduce the time complexity. The quantitative and binary representation helps many users to find the frequent items for the transactional database. Association rule mining deployed from binary frequent patterns database which helps to filter the Sensitive itemsets effectively with the help of Constraints Weight. Finally, it is proven that the proposed MDRH algorithm serves better than existing methods.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proc.1993 ACM-SIGMOD Int. Conf. Management of Data, Washington, D.C., May 1993, pp 207–216.
[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In VLDBY94, pp. 487-499.
[3] C.Borgelt. “Efficient Implementations of Apriori and Eclat”. In Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations, CEUR Workshop Proceedings 90, Aachen, Germany 200.
[4] J.S .Park, M.S.Chen and P.S.Yu. An effective hash based algorithm for mining association rules. In SIGMOD1995, pp 175-186.
[5] J. Han, J. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation (PDF), (Slides), Proc. 2000 ACM-SIGMOD Int. May 2000.

[6] E.Ramaraj and R.Sridevi A general Survey on multidimensional and Quantitative Association Rule mining Algorithms. International Journal of Engineering Research and Applications(IJERA) 2013.

[7] A.B.M.Rezbaul Islam, Tae-Sun Chung An Improved Frequent Pattern Tree Based Association Rule Mining Techniques Department of Computer Engineering Ajou University Suwon, Republic of Korea.

[8] E. Ramaraj and N. Venkatesan, ― Bit Stream Mask Search Algorithm in Frequent Itemset Mining,‖ European Journal of Scientific Reasearch,‖ Vol. 27 No.2 (2009),

[9] G. Grahne, J. Zhu, Fast algorithms for frequent itemset mining using FP-Trees, IEEE Transactions on Knowledge and Data Engineering 17 (10) (2005) 1347–1362.

[10] J. Han, H. Cheng, D. Xin, X. Yan, Frequent pattern mining: current status and future directions, Data Mining and Knowledge Discovery (2007). 10th Anniversary Issue.

[11] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 2000, pp. 1–12.

[12] T.-P. Hong, C.-W. Lin, Y.-L. Wu, Incrementally fast updated frequent pattern trees, Expert Systems with Applications 34 (4) (2008) 2424–2435.

[13] H. Huang, X. Wu, R. Relue, Association analysis with one scan of databases, in: Proceedings of the IEEE International Conference on Data Mining, 2002,pp. 629–632.

[14] G. Grahne, and J. Zhu, "Efficiently using prefix-trees in mining frequent itemsets", IEEE ICDM Workshop on Frequent Itemset Mining Implementations, 2003.

[15] C. Lucchese, S. Orlando, and R. Perego, "DCI_CLOSED: a Fast and Memory Efficient Algorithm to Mine Frequent Closed Itemsets", IEEE Transaction On Knowledge And Data Engineering, Vol. 18, No. 1, PP. 21-35, 2006.

[16] D.Lin, Z. M. Kedem, Pincer-Search: A New Algorithm for Discovering the Maximum Frequent Itemset, EDBT Conference Proceedings, 1998, pages 105-110.

