IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Confabulation - Inspired Association Rule Mining for Rare and Frequent Itemsets

K. Jayakumar[1], Mrs.N. Tamilselvi[2]

M.Sc., B.Ed., M.Phil (Scholar), Department of Computer Science, Dr.NGP ASC., Coimbatore, Tamilnadu[1]

M.Sc., M.Phil., Assistant Prof., Department of Computer Science, Dr.NGP ASC., Coimbatore, Tamilnadu[2]

Abstract: Data mining techniques are methods for obtaining useful knowledge from these large databases. One of the main tasks of data mining is association rule mining (ARM), which is used to find interesting rules from large amounts of data. A new confabulation-inspired association rule mining (CARM) algorithm is proposed using an interestingness measure inspired by cogency. Cogency is only computed based on pairwise item conditional probability, so the proposed algorithm mines association rules by only one pass through the file. The proposed algorithm is also more efficient for dealing with infrequent items due to its cogency-inspired approach. The problem of associative classification is used here for evaluating the proposed algorithm. This paper evaluates CARM over data sets. Experiments show that the proposed algorithm is consistently faster due to its one time file access and consumes less memory space than the Conditional Frequent Patterns growth algorithm. In addition, statistical analysis reveals the superiority of the approach for classifying minority classes in unbalanced data sets using dynamic transaction datasets.

Key words: Data mining, Association rule mining, confabulation, cogency, frequent patterns item.

I. INTRODUCTION

Association rule mining is one of the most challenging areas of data mining which was introduced in Agrawal et al., (1993) [1] to discover the associations or co-occurrence among the different attributes of the dataset. Several algorithms like Apriori(Agrawal et. al., 1993), SETM (Houtsma and Swami, 1993), AprioriTID (Agrawal and Srikant, 1994), DIC (Brin et al., 1997), partition algorithm (Savasere et al.,1995), Pincer search (Lin and Kedem, 1998), FP-tree (Han et al., 2000) [1]etc. have been developed to meet the requirements of this problem. These algorithms work basically in two phases: frequent itemset generation and rule generation. Since the first phase is the most time consuming, all of the above mentioned algorithms mainly focus on the first phase. A set of attributes is termed as frequent set if the occurrence of the set within the dataset is more than a user specified threshold called minimum support. After discovering the frequent itemsets, in the second phase rules are generated with the help of another user parameter called minimum confidence.

The aim of association rule mining is to detect interesting associations between items in a database[2]. It was initially proposed in the context of market basket analysis in transaction databases, and has been extended to solve many other problems such as the classification problem. Association rules for the purpose of classification are often referred to as predictive association rules. Usually, predictive association rules are based on relational databases and the consequences of rules are a pre-specified column, called the class attribute.

This paper addresses the problem of finding interesting predictive association rules in datasets with unbalanced class distributions[3]. They proposed two new interestingness measure s for the optimal association rule algorithm developed earlier and use the m to find all interesting association rules in a health dataset containing classes which are very small compared to the population.

However, based on their study, it has been observed that most of these techniques suffer from the following disadvantages[4]:

- A two phase association mining often can be found to be time and resource consuming in case of larger incremental datasets.
- Due to conversion of the real-life data into market-basket domain, information loss occurs.
- Single objective function (i.e. based on only frequency of occurrence) based rules generation often can be found to be non-interesting.

To address these issues, a single phase incremental association mining technique has been reported in this paper, which can extract reduced set of interesting rules from real-life datasets without transforming it into the market basket domain[5]. The proposed technique can be found to be significant in view of the following points:

- During extraction of the rules, it evaluates the rules based not only on the support count, but also on the measures comprehensibility and interestingness.
- It does not require to transforming the dataset into market basket domain.
- It avoids the frequent item set generation phase; rather it generates the rules directly.

II. RELATED WORK

Associative classification is a rule-based approach to classify data relying on association rule mining by discovering associations between a set of features and a class label. Support and confidence are the de-facto

"interestingness measures" used for discovering relevant association rules[6]. The support confidence framework has also been used in most, if not all, associative classifiers. Although support and confidence are appropriate measures for building a strong model in many cases, they are still not the ideal measures and other measures could be better suited[7][8].

Rare association rule mining has received a great deal of attention in the recent past. In this research they use transaction clustering as a pre-processing mechanism to generate rare association rules. The basic concept underlying transaction clustering stems from the concept of large items as defined by traditional association rule mining algorithms. They make use of an approach proposed by Koh & Pears (2008) to cluster transactions prior to mining for association rules [9] [10] [11]. They show that pre-processing the dataset by clustering will enable each cluster to express their own associations without interference or contamination from other sub groupings that have different patterns of relationships. Their results show that the rare rules produced by each cluster are more informative than rules found from direct association rule mining on the un partitioned dataset[12 [13].

Frequent patterns are an important class of regularities that exist in a transaction database. Certain frequent patterns with low minimum support (minsup) value can provide useful information in many real-world applications. However, extraction of these frequent patterns with single minsup based frequent pattern mining algorithms such as Apriori and FP-growth leads to "rare item problem[1][16][17] [18]." That is, at high minsup value, the frequent patterns with low minsup are missed, and at low minsup value, the number of frequent patterns explodes. In the literature, "multiple minsup frameworks" was proposed to discover frequent patterns. Furthermore, frequent pattern mining techniques such as Multiple Support Apriori and Conditional Frequent Pattern-growth (CFP-growth) algorithms have been proposed. As the frequent patterns mined with this framework do not satisfy downward closure property, the algorithms follow different types of pruning techniques to reduce the search space.
In this paper, they proposed an efficient[14] [15] CFP-growth algorithm by proposing new pruning techniques. Experimental results show that the proposed pruning techniques are effective [19][20].

III. PROPOSED METHODOLOGY

A new confabulation-inspired association rule mining (CARM) algorithm is proposed using an interestingness measure inspired by cogency. Cogency is only computed based on pair wise item conditional probability, so the proposed algorithm mines association rules by only one pass through the file. The proposed algorithm is also more efficient for dealing with infrequent items due to its cogency-inspired approach. The problem of associative classification is used here for evaluating the proposed algorithm. They evaluate CARM over both synthetic and real benchmark data sets obtained from the UC Irvine machine learning repository.

Experiments design that the proposed algorithm is consistently faster due to its one time file access and consumes less memory space than the Conditional Frequent Patterns growth algorithm. In addition, statistical analysis reveals the superiority of the approach for classifying minority classes in unbalanced data sets.

Confabulation Theory

Confabulation theory offers a comprehensive detailed explanation of the mechanism of thought (i.e., "cognition": vision, hearing, reasoning, language, planning, origination of movement and thought processes, etc.) in humans and other vertebrates (and possibly in invertebrates, such as octopi and bees). For expositional simplicity, only the human case is considered here.

Confabulation (verb: confabulate) is a memory disturbance, defined as the production of fabricated, distorted or misinterpreted memories about oneself or the world, without the conscious intention to deceive. Confabulation is distinguished from lying as there is no intent to deceive and the person is unaware the information is false. Although individuals can present blatantly false information, confabulation can also seem to be coherent, internally consistent, and relatively normal. Individuals who confabulate present incorrect memories ranging from "subtle alterations to bizarre fabrications", and are generally very confident about their recollections, despite contradictory evidence.

Rare Item Mining

The discovery of new and interesting patterns in large datasets, known as data mining, draws more and more interest as the quantities of available data are exploding. Data mining techniques may be applied to different domains and fields such as computer science, health sector, insurances, homeland security, banking and finance, etc. In this project they are interested by the discovery of a specific category of patterns, known as rare and non-present patterns. They present a novel approach towards the discovery of non-present patterns using rare item-set mining.

CARM Algorithm

In CARM, only one-item consequent association rules are generated, where there can be multiple antecedent items. The proposed CARM approach using a cogency inspired measure for generating rules. Cogency inspiration can lead us to more intuitive rules. Moreover, cogency-related computations only need pair-wise item co-occurrences, hence, They can find rules only by one file scan. Rule mining is performed in two main phases: knowledge acquisition and structure construction and rule generation by confabulation and cogency measure. In this algorithm, only one item con-sequent association rules are generated, which means that the consequents of these rules only contain one item.

Below is the pseudo code for the CARM algorithm:

DOI 10.17148/IJARCCE.2015.412156

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Algorithm 1: CARM

```
1-  i=1
2-  while not EOF
3-  read transaction t_i
4-  for each x ∈ t_i
5-    for each y ∈ t_i
6-      L xy = L xy + 1
7-    end for
8-  end for
9-  i=i+1
10- end while
11- end
```

Algorithm 2: D-CARAM

```
1- S_1 = Fr
2- find S_2,the set of all frequent 2-itemsets
3- find all rules from 2-itemsets(according their support and
   confidence)
4- k=2
5- while S_k ≠ ∅
     5.1. S_{k+1} = {}
     5.2. for each X ∈ S_k do
          i. find the set Y(X) such that
               Y(X) = {y | y ∈ S_1 & y ∩ X = ∅ &
                cogency X → y > mincog_1^|X|}
          ii. for each y ∈ Y(X)
               a. Z = X + y
               b. S_{k+1} = S_{k+1} + Z
               c. if Cogency X → y > mincog_2^|X|
                  Add rule X → y to the association rule list
               d. end for
          iii. end for
     5.3. k=k+1;
     5.4. end while
```

IV. RESULTS AND DISCUSSIONS

ILLUSTRATION OF PROPOSED SYSTEM

Stock item and transaction items are followed:

Table 1: Transaction ITEMS

TID	Items	TID	Items
T1001	J,A,B	T1006	B,C,E
T1002	I,H,G	T1007	A,B,D
T1003	D,E,F	T1008	H,J,I
T1004	I,J,A	T1009	E,F,G
T1005	F,G,H	T1010	F,I,C

Items	Class	Items	Class
apples	**A**	mangoes	**F**
bananas	**B**	oranges	**G**
cherries	**C**	pineapples	**H**
grapes	**D**	plums	**I**
lemons	**E**	strawberries	**J**

Step 1: The total transaction items are **30 items**. So **cuT** is set 30 items.

Step 2: **cuT** is 30 items so 10 items for each schema for **cs1**, **cs2** and **c3** store. Transaction is partition of **cuT** limit and store the **cs1, cs2** and **cs3**.

Table 2:Confabulation 1

cs2	fs1	temp	aso	count
J	A	J	A	3
A	B	E	B	3
F	C		F	2
G			G	2
H			H	2
B	**fs2**		C	1
C				
E	F			
A	G			
B	H			

Step 3: Fuzzy items is set the frequent items, the limit **cuF** is **3**. Fuzzy items are partition of **cuF** limit and store the **fs1**and **fs2** schema.

Table 3: Confabulation 2

cs1	cs1	cs2	cs2	cs3	cs3
J	G	J	B	D	F
A	D	A	C	H	G
B	E	F	E	J	F
I	F	G	A	I	I
H	I	H	B	E	C

Step4: Finding the frequent itemset between **fs** and **cs**.

Table 4: Confabulation 3

cs3	fs1	temp	aso	count
D	A	D	A	3
H	B	J	B	3
J	C	I	F	4
I		E	G	3
E		I	H	3
F			C	2
G				
F	**fs2**			
I				
C	F			
	G			
	H			

cs1	fs1	temp	aso	count
J	A	J	A	1
A	B	I	B	1
B	C	D	F	1
I		E	G	1
H		I	H	1
G				
D				
E	**fs2**			
F				
I	F			
	G			
	H			

First check the **cs1** compare the **fs1** finding the frequent items and frequent count. Second check the **cs1**compare the **fs2** finding the frequent items and frequent count. The non frequent items are temporary store the list and future checks the frequent items. Similar checks **cs2 & fs1**, **cs2 & fs2**, and **cs3 & fs1, cs3 & fs2.**

DOI 10.17148/IJARCCE.2015.412156

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Step 5 & 6: Temporary items is store the **ds** and **ds** limit **cuD** is 3. The temp item partitions of **cuD** set and stores the **ds**.

Table 5: Confabulation 4

fs1	fs2
A	F
B	G
C	H

In this step first check **ds1** and **ds2** finding the rare frequent items and frequent count. Next check the **ds3** finding the frequent items and frequent count. The minimum frequent support items are storing the **aso** schema.

Finally numbers of frequent items and frequent count are entering the transaction items. Compute the association rule mining for the minimum support frequent items and update the original database future select and finding fuzzy and frequent items easily.

The above **Table 5** shows the Number of transaction Items [N] and Average frequency item set for the transactions in %. **Figure 5** describes the percentage of average frequency item sets for the number of transactions [N].

Experimental results in this section show that the proposed research outperforms the state-of-the-art algorithms almost in all cases on medicine transaction data sets.

This thesis is used to eliminate time complexity rate while finding high utility item sets in a transaction database. In this proposed paper, tree construction process using two strategies, namely CARM (Confabulation Association Mining Rule) and D-CARM (Dynamic Confabulation Association Mining Rule)). It also used to reduce number scans to the database.

Transaction No	Data item set (N)	CARM [%]	D-CARM [%]
1	100	55.33	61.33
2	200	62.30	64.33
3	300	67.33	69.22
4	400	72.11	72.45
5	500	76.57	78.01
6	600	80.08	81.09
7	700	81.44	81.98
8	800	82.55	83.08
9	900	83.55	84.03
10	1000	84.67	85.04

Table 5- Performances for CARM and D-CARM Algorithm

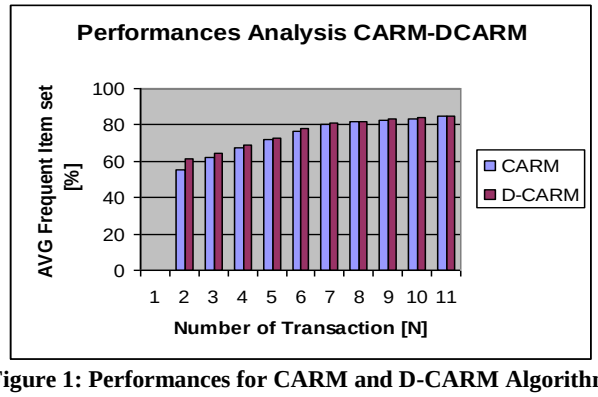

Figure 1: Performances for CARM and D-CARM Algorithm

V. CONCLUSION FOR FURTHER ENHANCEMENTS

A novel confabulation-inspired association rule mining (CARM) algorithm is proposed using an interestingness measure inspired by cogency. Cogency is only computed based on pair wise item conditional probability, so the proposed algorithm mines association rules by only one pass through the file. The proposed algorithm is also more efficient for dealing with infrequent items due to its cogency-inspired approach. The problem of associative classification is used here for evaluating the proposed algorithm.

The new system become useful if the below enhancements are made in future.

- ✓ In future work, the method can be applied to real data sets. In addition, the CTMSP-Mine can be applied to other applications, such as GPS navigations, with the aim to enhance precision for predicting user behaviors.
- ✓ If the application is developed as web based application, then it can be used from anywhere.

The new system is designed such that those enhancements can be integrated with current modules easily with less integration work.

REFERENCES

[1] Agrawal R. and Srikant R., "Fast Algorithms for Mining Association Rules," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB), pp. 487-499, 1994.

[2] Cai, C.H. Fu A.W.C., Cheng C.H., and Kwong W.W., "Mining Association Rules with Weighted Items," Proc. Int'l Database Eng.and Applications Symp. (IDEAS '98), pp. 68-77, 1998.

[3] Chen M.-S., Park J.-S., and Yu P.S., "Efficient Data Mining for Path Traversal Patterns," IEEE Trans. Knowledge and Data Eng., vol. 10, no. 2, pp. 209-221, Mar. 1998

[4] Creighton C. and Hanash S., "Mining Gene Expression Databases for Association Rules," Bioinformatics, vol. 19, no. 1, pp. 79-86, 2003.

[5] Erwin A., Gopalan R.P., and Achuthan N.R., "Efficient Mining of High Utility Itemsets from Large Data Sets," Proc. 12th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 554-561, 2008.

[6] Georgii E., Richter L., Ru ¨ ckert U., and Kramer S., "Analyzing Microarray Data Using Quantitative Association Rules," Bioinformatics, vol. 21, pp. 123-129, 2005.

[7] Han J., Dong G., and Yin Y., "Efficient Mining of Partial Periodic Patterns in Time Series Database," Proc. Int'l Conf. on Data Eng., pp. 106-115, 1999.

[8] Han J. and Fu Y., "Discovery of Multiple-Level Association Rules from Large Databases," Proc. 21th Int'l Conf. Very Large Data Bases, pp. 420-431, Sept. 1995.

[9] Han J., Pei J., and Yin Y., "Mining Frequent Patterns without Candidate Generation," Proc. ACM-SIGMOD Int'l Conf. Management of Data, pp. 1-12, 2000.

[10] Lee S.C., Paik J., Ok J., Song I., and Kim U.M., "Efficient Mining of User Behaviors by Temporal Mobile Access Patterns," Int'l J. Computer Science Security, vol. 7, no. 2, pp. 285-291, 2007.

[11] Li H.F., Huang H.Y., Chen Y.C., Liu Y.J., and Lee S.Y., "Fast and Memory Efficient Mining of High Utility Itemsets in Data Streams," Proc. IEEE Eighth Int'l Conf. on Data Mining, pp. 881-886, 2008.

[12] Li Y.-C., Yeh J.-S., and Chang C.-C., "Isolated Items Discarding Strategy for Discovering High Utility Itemsets," Data and Knowledge Eng., vol. 64, no. 1, pp. 198-217, Jan. 2008.

[13] Lin C.H., Chiu D.Y., Wu Y.H., and Chen A.L.P., "Mining Frequent Itemsets from Data Streams with a Time-Sensitive Sliding Window," Proc. SIAM Int'l Conf. Data Mining (SDM '05), 2005.

[14] Y. Liu, W. Liao, and A. Choudhary, "A Fast High Utility Itemsets Mining Algorithm," Proc. Utility-Based Data Mining Workshop, 2005.
[15] Martinez R., Pasquier N., and Pasquier C., "GenMiner: Mining nonredundant Association Rules from Integrated Gene Expression Data and Annotations," Bioinformatics, vol. 24, pp. 2643-2644, 2008.
[16] Pei J., Han J., Lu H., Nishio S., Tang S., and Yang D., "H-Mine: Fast and Space-Preserving Frequent Pattern Mining in Large Databases," IIE Trans. Inst. of Industrial Engineers, vol. 39, no. 6, pp. 593-605, June 2007.
[17] Pei, J. Han J., Mortazavi-Asl, H. Pinto H., Chen Q., Moal U., and M.C. Hsu, "Mining Sequential Patterns by Pattern-Growth: The Prefixspan Approach," IEEE Trans. Knowledge and Data Eng., vol.16, no.10, pp. 1424-1440, Oct. 2004.
[18] Pisharath J., Y. Liu, Ozisikyilmaz B., Narayanan R., Liao W.K., Choudhary A., and Memik G. MineBench NU- Version 2.0 Data Set and Technical Report, http://cucis.ece.northwestern.edu/projects/DMS/MineBench.html, 2012
[19] Shie B.-E., Hsiao H.-F., Tseng V., S., and Yu P.S., "Mining High Utility Mobile Sequential Patterns in Mobile Commerce Environments," Proc. 16th Int'l Conf. DAtabase Systems for Advanced Applications (DASFAA '11), vol. 6587/2011, pp. 224-238, 2011
[20] Shie B.-E., Hsiao H.-F., Tseng V., S., and Yu P.S.,, "Online Mining of Temporal Maximal Utility Itemsets from Data Streams," Proc. 25th Ann. ACM Symp. Applied Computing, Mar. 2010

DOI 10.17148/IJARCCE.2015.412156