Positive and Negative Association Rule Analysis in Health Care Database

[1]E.Ramaraj [2]N.Venkatesan
[1]Director and Head, Computer Science and Engineering, Alagappa University, Karaikudi, Tamilnadu.

[2]Asst.Prof and Head, Dept of IT, Bharathiyar College of Engg and Tech, Karaikal, Pondichery

Abstract
This paper focuses a new algorithm called BitArrayNegativePos that mines both positive and negative rules from the real time surveyed medical database. Association rules are defined as implication of the form A➔ B where A and B are frequent itemsets in a transaction database. This new algorithm extends this definition to include association rules of forms A ➔ ^B, ^A ➔ B and ^A ➔ ^B, which indicate negative associations between itemsets is called negative rules. Negative rules are generated from infrequent itemsets. Rules of the form A➔ B are called positive rules Negative rules are very useful in association analysis although they are hidden and different from positive rules. BitArrayÉclat algorithm extracts only positive rules. BitArrayNegativePos algorithm is able to find all valid rules in a support-confidence framework Experimental results show the efficiency of our new algorithm.

Keywords:
Association rules, negative rules, BitArrayEclat, BitArrayNegativePos

1. Introduction

Data Mining is used to extract knowledge automatically from large data sets. Association Rules, Classifications, clustering are major areas of interest in data mining. Among these Association Rules has been a very active research area. Association Rule Mining mines useful information huge amount of data by generating rules. The process of mining association rules consists of two steps.

1. Finding the frequent itemset in the database using Support.
2. Constructing the association rule from the frequent itemset with specified confidence.

Frequent itemset finding is the most expensive of the two steps, since the number of item sets grows exponentially with the number of items. A large number of efficient algorithms to mine frequent itemsets has been developed over the years

For example, 1000 items define 2^1000 possible combinations of item sets which results in a large number of rules to explore. The minimum support constraint is used to limit the number of item sets that need to be considered. So there are two things in Association Rules Mining; 1.POSITIVE Rule Mining and 2. NEGATIVE Rule Mining.

However, there is negative rules i.e. in the non frequent itemset sometimes it produces the maximum confidence level, for example a negative rule such as *{^ high* income} ➔ *{^* expensive electronics*}* is also useful because it expresses that people who are not rich generally do not buy expensive electronics. Eclat principle mines only positive association rules with the help of depth first traversal of a prefix tree concept. This paper defines new approaches for Eclat principle which produces both positive and negative rules from the medical datasets.

The structure of the paper is as follows.
Section 2 defines basic Eclat Algorithm, Negative Association Rules and Medical Database Description. Section 3 gives new algorithm for Positive and Negative rules. Experimental results are shown in section 4. Section 5 deals with performance analysis. The study is concluded in the section 6 along with a brief on future work.

2. Basic Algorithms

2.1 Eclat Algorithm

In Eclat algorithm [13][14] implementation the set of transactions as a (sparse) bit matrix and intersects rows to determine the support of item sets. The search space of Eclat algorithm is based on depth first traversal of a prefix tree [15][16].

Éclat principle:-
A convenient way to represent the transactions for the Eclat Algorithm is a bit matrix, in which each row corresponds to an item, each column to a transaction.. A bit is set in this matrix if the item corresponding to the row is contained in the transaction corresponding to the column, otherwise it is cleared. Eclat searches a prefix tree. The transition of a node to its first child consists in constructing a new bit matrix by intersecting the first row

Manuscript received October 5, 2008
Manuscript revised October 20, 2008

with all following rows. For the second child, the second row is intersected with all following rows and so on.

The item corresponding to the row is intersected with the following rows to form the common prefix of the item sets, processed in the corresponding child node. Of course, rows corresponding to infrequent item sets should be discarded from the constructed matrix, which can be done most conveniently if it stores with each row the corresponding item identifier rather than relying on an implicit coding of this item identifier in the row index. For a sparse representation the column indices for the set bits should be sorted ascending for efficient processing. Then the intersection procedure is similar to the merge step of merge sort. In this case counting the set bits is straightforward.

Subset of frequent item set is frequent.
Based on principle, calculate support and confidence level to generate Association rules.

2.2 Negative Association Rule Algorithms

The negative association rule differs from its positive counterpart not only in the mining procedure but also in form. To focus interesting negative rule by incorporating domain knowledge of the data sets.

In [1][2][3][15], the traditional definition of itemset is maintained (so $X, Y \in I$), and to each *positive* rule $X \rightarrow Y$ correspond three negative ones, $X \rightarrow \neg Y$, $\neg X \rightarrow Y$ and $\neg X \rightarrow \neg Y$. A transaction t supports $X \rightarrow \neg Y$ if $X \subset t$ and $Y \not\subset t$. Hence, the meaning of a rule like $\{i1\} \rightarrow \neg\{i2, i3\}$ is that "the appearance of $i1$ in a transaction t induces that $i2$ and $i3$ are unlikely to appear simultaneously in t"; hence a record containing $i1$ and $i2$, but not $i3$, supports this rule. It can be verified [15][17] that $supp(X \rightarrow \neg Y) = supp(X\neg Y) = supp(X) - supp(XY)$ for $X, Y \subset I$, and similarly support and confidence of the other kinds of negative Association Rules can be straightforwardly deduced from the corresponding positive itemset supports.

Wu et al [7] presented an Apriori-based framework for mining both positive and negative Association Rules. Another Apriori-based algorithm was given by Antonie and Za¨ıane [3] for the purpose of simultaneously generating positive Association Rules and (a subclass of) negative Association Rules. Chris Cornelis et al [3] proposed a new algorithm S-PNAR for mining both positive and negative rules.

2.3. Medical Database

Experimental data in many domains serves as a basis for predicting useful trends. Association rules are generated from one such medical database [23] with real time surveyed records of 10000 patients. The database which we have chosen depicts the complications occurring in diabetes and/or hypertension (increased Blood pressure). All the patients were in the age group of 25 to 70 years, and the sex ratio was almost equal (M:F of **1.1 : 1**). All of them had either diabetes or hypertension or both for duration of 10 years and more. They were sub-categorized based on the extent of control of these diseases namely diabetes and hypertension. Several complications like kidney disease, heart disease and stroke were evaluated in this group of patients.

Database is the storage which holds data. In the above medical database itemsets are stored in the database corresponding to their transaction which is used for future reference. SQL server is used as the database server. Database server holds the data in string format, string processing requires more time and it is difficult too.

Conversion of the string into corresponding integer value assigned to the itemsets, which reduces the execution time complexity and space complexity of the process.

The data type is used for the storage of transactional data is structure. It has an *int* type data to store the count of the itemset. Also it has a string type data to store the name of the itemset. It has a float type data to store the support count of itemset. Doubly linked structure is defined to use process memory occupation efficient one.

3. Bit Array Procedure

BitArrayNegativePos is a new algorithm that mines both positive and negative rules from the transaction file which array is transformed as sparse bit matrix for processing; this is done from numerically transferred input file.

ALGORITHM 1: BIT_ARRAY_ECLAT

This algorithm shows usage of sparse matrix to mine association rules with bit array data structure. This produces only positive association rules. Array is stored as bits so that memory occupation to do this process is low.

```
1) Initialize bit[n][m]
      //n-> number of itemsets m-> no.of
        transactions
2) for i=0;i<n;i++     do begin
3) for j=I;j<m;j++     do begin
4) if bit[i][m]!=0
5) tcount++
```

```
        //total count to find no. of items
          present in transaction
6) if ((bit[i][m]&&bit[j][m])!=0)
7) count++
8) end
9) support= tcount/m
10) confident = count/tcount
11) end
```

TO FIND FREQUENT ITEM SET

The frequent itemset is defined as the item whose support is greater than or equal to the threshold value. In the database the every items are unique in their integer value. The frequent itemsets are finding using the bit matrix structure. The following are numeric transactions are

1 4 2 7 8 10 0
2 5 1 0
2 5 6 3 1 0
2 1 3 6 0
4 1 2 5 0
2 1 9 10 3 4 0
1 2 8 10 0
1 2 8 5 0
2 4 9 8 7 0
2 1 5 7 8 10 0

This transaction is converted into bit matrix as in the table 3.1

ALGORITHM 2: BIT_ARRAY_NEGATIVEPOS

This algorithm mines both positive and negative association rules in one database scan.

```
1) Initialize bit[n][m]
     //n-> number of itemsets m-> no.of
       transactions
2) Initialize nbit[n1][m]              //nbit[][] -
> non frequent itemset
3) for i=0;i<n;i++     do begin
4) for j=I;j<m;j++     do begin
5) if bit[i][m]!=0
6) tcount++
        //total count to find no. of items
        present in transaction
7) if ((bit[i][m]&&bit[j][m])!=0)
8) count++
9) end
10) support= tcount/m
11) confident = count/tcount
12) end
13) for i=0;i<n;i++   do begin
14) for j=I;j<m;j++   do begin
15) if nbit[i][m]!=0
16) tcount++
        //total count to find no. of items
        present in transaction
17) if ((nbit[i][m]&&nbit[j][m])!=0) ||
        ((nbit[i][m]&&bit[j][m])!=0)
18) count++
19) end
20) support= tcount/m
21) confident = count/tcount
22) end
```

In the above bit matrix itemset number represent the item is present in the transaction. '0' Represent the item is not present in the transaction.

If the minimum support is 3 then item 1,2,4,5,7,8,10 are taken as frequent item and 3,6,9 are taken as non frequent item set.

The table 3.1 has the frequent itemsets and also non frequent itemsets for further comparison:

1. frequent itmesets are compared with only frequent item not with non frequent items.

   The frequent itemsets is present more than the threshold level but non frequent itemset is present only few times so if we take the confidence of frequent with non frequent ie very less value when compared with minimum confidence level .

   eg:- let

   2 is frequent and 9 is non frequent

   2 is present in 10 times and 9 is present only 2 times

   Hence the confidence of 2→ 9 is 20 % only

   So there is no need to compare frequent with non frequent items

2. Non frequent items – they have to compared with frequent and non frequent items sets

a. Compared with non frequent – in that there may be two items that one is present in every transaction that another one is present on that, hence the confidence of that two may be 100% but they are omitted because they are not frequent itemasets and hence the negative rule mining is rule generated

Eg :-

3 is non frequent and 6 is non frequent
3 is present in 2 times and 6 is present only 2 times

Hence the confidence of 3→6 is 100 % only

1	1	1	1	1	1	1	1	0	1
2	2	2	2	2	2	2	2	2	2
0	0	3	3	0	0	0	0	0	0
4	0	0	0	4	4	0	0	4	0
0	5	5	0	5	0	0	5	0	5
0	0	6	6	0	0	0	0	0	0
7	0	0	0	0	0	0	0	7	7
8	0	0	0	0	8	8	8	8	8
0	0	0	0	9	0	0	0	9	0
10	0	0	0	0	10	1	0	0	10

Table 3.1: Sparse matrix of data set

b. Compared with frequent – in that there may be two items that one is present in every transaction that another one is present on that, hence the confidence of that two may be more than the threshold value but they are omitted because they are not frequent itemasets and hence the negative rule mining is rule generated

Eg :-

9 is non frequent and 2 is non frequent
9 is present in 2 times and 2 is present only 10 times

Hence the confidence of 9→2 is 100 % and the rule is reverse also.

4. Experimental Results

For the experiments, Intel Pentium IV dual core processor, Windows XP with 256 MB RAM is used. These two algorithms are experimented with real time surveyed database which is generated by us. To test the efficiency of the new algorithm, data sets are experimented with previous S-PNAR algorithm. Diagrams are represented as the comparison of various support level and execution time which are discussed in the following section The results for these data sets are compared.

5. Performance Analyses

Rule is generated; confidence and support level for frequent and non-frequent itemsets are calculated. Negative rule mining is generated in a separate bit matrix that hold the non-frequent itemsets bit matrix..

Four different data sets are implemented to our algorithms. Figure 4.1 to 4.4 show the data sets 2000, 3000, 5000 and 10000 records respectively, Total number of itemsets is only fifty. In this comparison, existing SPNAR algorithm and BitArrayNegativePos algorithm are taken for discussion. Between these two results, BitArrayNegativePos has less execution time. Figure 4.5 shows the result of Association Rules generation of BitArrayEclat and BitArrayNegativePos for the above mentioned same data sets. Number of useful rules generated is higher in BitArrayNegativePos algorithm.

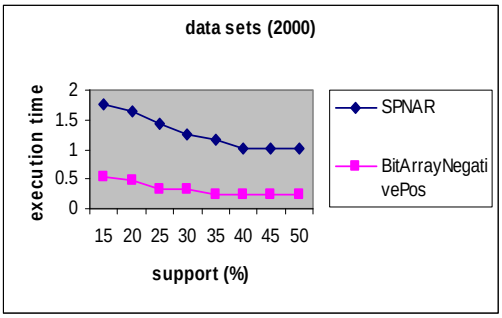

Figure 4.1 comparison of BitArrayNegativePos with SPNAR for data set 2000

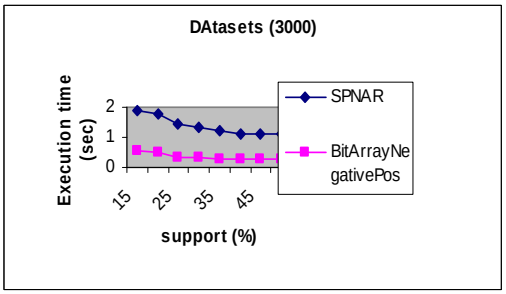

Figure 4.2 comparison of BitArrayNegativePos with SPNAR for data set 3000

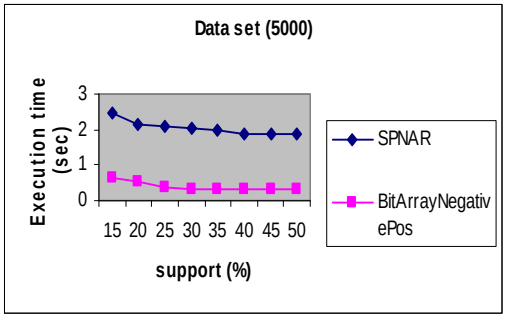

Figure 4.3 comparison of BitArrayNegativePos with SPNAR for data set 5000

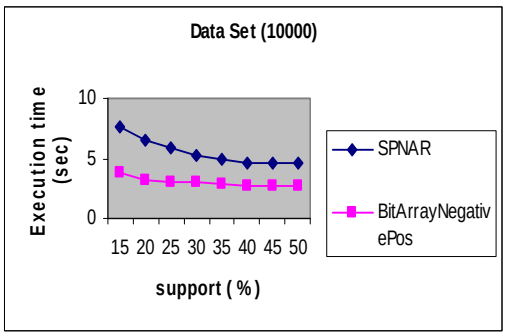

Figure 4.4 comparison of BitArrayNegativePos with SPNAR for data set 10000

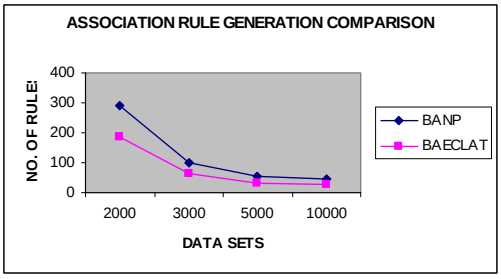

Figure 4.5 comparison of BitArrayNegativePos with BitArrayEclat

The above data sets are created and executed on our own development. In figure 4.5, twenty per cent support level for the various data sets is given in our comparison. From these diagrams, BitArrayNegativePos is low execution time with processing more association rules.

The following are the interesting information observed from from this new Positive and Negative Association Rule mining approach.

- Patients with stroke developing as a complication of diabetes or hypertension have every chance of having other complications namely kidney disease or heart disease.
- Patients with poorly controlled diabetes have the highest chance of developing one or all end organ complications.
- Patients with poorly controlled Hypertension have a certain possibility of developing heart disease and stroke.
- Patients with poorly controlled hypertension have greater chance of developing heart disease when compared to any other complications.

6. Conclusion and Future Work

This paper has proposed a new algorithm - BitArrayNegativePos algorithm - to mine interesting relationship among data sets which yields positive and negative association rules from the medical datasets. The advantages of BitArrayNegativePos over existing algorithms include (1) Faster execution time compared with previous algorithms (2) Using negative rule mining mines more interesting rules for our further interestingness (3) Representation of Sparse matrix as bit reduces the process memory occupation which yields low execution time. (4) Negative rules extracts hidden knowledgeable useful information from this medical database. Interesting rules are extracted from this new algorithm. Hence, BitArrayNegativePos seems to be more efficient than other prior algorithms. Extension of this work to the very large data sets and for mining knowledge rules poses many interesting issues for future investigation.

Reference

[1] Chris Cornelis, Peng Yan, Xing Zhang, Guoqing Chen *Mining Positive and Negative Rules from Large databases* IEEE conference 2006.

[2] Xiaohui Yuan, Bill P. Buckles, Xhaoshan Yuan, Jian Zhang *Mining Negative Association Rules*

[3] Mario Luiza antonic, Osmar R. Zaiane *Mining Positive and Negative Association Rules – an approach for confine rules*

[4] M.L. Antonie and O.R. Za¨ıane, ”Mining Positive and Negative Association Rules: an Approach for Confined Rules”, *Proc. Intl. Conf. on Principles and Practice of Knowledge Discovery in Databases*, 2004, pp 27–38.

[5] A. Savasere, E. Omiecinski and S. Navathe, ”Mining for Strong Negative Associations in a Large Database of Customer Transactions”, *Proc. Intl. Conf. on Data Engineering*, 1998, pp 494–502.

[6] D.R. Thiruvady and G.I. Webb, ”Mining Negative Association Rules Using GRD”, *Proc. Pacific-Asia*

Conf. on Advances in Knowledge Discovery and Data Mining, 2004, pp 161–165.
[7] X. Wu, C. Zhang and S. Zhang, "Efficient Mining of Both Positive and Negative Association Rules", *ACM Trans. on Information Systems*, vol. 22(3), 2004, pp 381–405.
[8] P. Yan, G. Chen, C. Cornelis, M. De Cock and E.E. Kerre, "Mining Positive and Negative Fuzzy Association Rules", LNCS 3213, 2004, pp 270–276.
[9] X. Yuan, B.P. Buckles, Z. Yuan and J. Zhang, "Mining Negative Association Rules", *Proc. Seventh Intl. Symposium on Computers and Communication*, Italy, 2002, pp 623–629.
[10] X. Wu, C. Zhang, and S. Zhang. Mining both positive and negative association rules. In *Proc. of ICML*, 2002.
[11] Tan, P., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proc. of SIGKDD. (2002) 32–41
[12] Lars Schmidt, Thieme *Algorithmic features of Eclat* – 2004
[13] Christian Borgelt *Efficient implementation of Aprirori and Eclat* FIMI 2004
[14] Bart Goethals *Survey on Frequent Patten Mining* 2004
[15] FIMI dataset – http://fimi.cs.helsinki.fi/
[16] Mingju Song and Sanguthevar Rajasekaran *A transaction mapping for frequent itemsets mining* IEEE transactions on Knowledge and Data Engineering 18(4):472-480, April 2006.
[17] Ke Su, Fengsdhan Bai *Mining weighted Association Rules* IEEE transactions on KDE 489-5=495, April 2008
[18] Data Mining – Concepts and Techniques, Jiawei Han, Micheline Kamber
[19] Wemnin Li, Jiawei Han, Jian Pei, *CMAR: Accurate and Efficient classification base on Multiple class Association Rules*
[20] Rajeev Rastogi, Kyuseok Shim *Mining Optimized Association Rules with categorical and numerical attributes* IEEE trans on KDE vol 14 No.1 Jan/Feb '02.
[21] Hui Xiong Michael Steinbach *Enhancing Data Analysis with Noise Removal* IEEE trans on KDE vol 18 NO.3 March '06
[22] Balaji Padmanabhan, Alexandar Tuzhilin *On Characterization and Discovery of Minimal unexpected pattern in Rule Discovery* IEEE trans on KDE vol 18 no. 2 Feb '06
[23] Dr. E. Ramaraj, N. Venkatesan *Discrete Topological Mining Association Rules – An Approach IRIS '06*

Authors

E. Ramaraj is presently working as a Director and Head, Computer Science and Engineering at Alagappa University, Karaikudi. He has 24 years teaching experience and 6 years research experience. He has presented research papers in more than 40 national and international conferences and published more than 30 papers in national and international journals. His research areas include Data mining and Network security.

N.Venkatesan is working as an Assistant Professor & Head, Information Technology Department, Bharathiyar college of Engineering and Technology, Karaikal. Currently pursuing Ph.D in Computer Science at SASTRA University.. He has been member of ISTE. He published 12 papers in National and International conferences. He has authored a book ***Data Mining and Warehousing***.