Effectiveness of *ERules* in Generating Non Redundant Rule Sets in Pharmacy Database

Kannika Nirai Vaani. M[1], E. Ramaraj[2]

[1]Training Division, Tech Mahindra Ltd, India

[2]Department of Computer Science and Engineering, Alagappa University, India

Abstract: ***'Erules' [3] is an integrated algorithm that is used to mine any data warehouse to extract useful and reliable rule sets effectively. It is used to generate positive &negative; conjunctive & disjunctive rules with the help of genetic algorithm and modified FP growth & Apriori Algorithms accordingly. It is an integrated algorithm for useful and effective association rule mining to capture even useful rare items; Lift Factor is also used to analyze the strength of derived rules. However redundant rules were one of the major challenges which were not addressed. This paper concisely deals the elimination of rule sets with the appropriate modification with the existing algorithm so that it can generate positive and negative rule sets for the non redundant rules with less cost. Besides a voluminous Pharmacy data set has been taken and the effectiveness /performance of 'Erules' got measured on it.***

Keywords: Association Rule Mining, Disjunctive Rules, Multiple Minimum Support, Lift, Redundant rules.

1. Introduction

Association rule mining is a repetitive process that explores and analyzes large data to find out valid, useful and reliable rules, using computationally efficient techniques. It searches for interesting associations among items in a given dataset. The main advantage of association rule mining is that it has ability to discover hidden associations with in the digital data. Two important constraints of association rule mining are support and confidence. Those constraints are used to measure the interestingness of a rule. Therefore, most of the current association rules mining algorithms use these constraints in generating rules. However, choosing support and confidence threshold values is a real dilemma for association rule mining algorithms. The important factor that makes association rule mining practical and useful is the minimum support. It is used to limit the number of rules generated. However, using only a single 'minsup' implicitly assumes that all items in the data are of the same nature and/or have similar frequencies in the database. This is often not the case in real-life applications. In many applications, some items appear very frequently in the data, while others rarely appear. If the frequencies of items vary a great deal, we will encounter two problems [4]:

1. If 'minsup' is set too high, we will not find those rules that involve infrequent items or rare items in the data.
2. In order to find rules that involves both frequent and rare items, then 'minsup' to be kept very low. However, this may produce too many rules.

So when one common support is fixed as minimum support for all the items, the rules which are not frequent occur but majorly contributing towards profit may get lost without notice.

For example, in a supermarket transaction data, in order to find rules involving those infrequently purchased items such as food processor and cooking pan (they generate more profits per item) very minimum support needs to be set ; but due to this the unwanted and rare items will not be get pruned. Hence fixing multiple minimum support for each items have become significant.

Multiple Minimum Supports

In many data mining applications, some items appear very frequently in the data, while others rarely appear. If minsup is set too high, those rules that involve rare items will not be found. To find rules that involve both frequent and rare items, minsup has to be set very low. This may cause combinatorial explosion because those frequent items will be associated with one another in all possible ways. The disadvantage of support is the rare item problem. Items that occur very infrequently in the data set are pruned although they would still produce interesting and potentially valuable rules. The rare item problem is important for transaction data which usually have a very uneven distribution of support for the individual items. In addition, most of the traditional association rules mining algorithms consider all subsets of frequent itemsets as antecedent of a rule [4]. Therefore, when resultant frequent item sets is large, these algorithms produce large number of rules. However, many of these rules have identical meaning or are redundant. In fact, the number of redundant rules is much larger than the previously expected. In most of the cases, number of redundant rules is significantly larger than that of essential rules [4].

1.1 'Erules': A Brief Overview

E-Rules [1] is an algorithm to derive positive & negative and conjunctive & disjunctive rules with the help of genetic algorithm and modified FP growth & Apriori Algorithms accordingly. It is an integrated algorithm for useful and effective association rule mining to capture even useful rare items; Lift Factor is also used to analyze the strength of derived rules. Salient features of Erules in each phase such as,

- Disjunctive rule generation: Algorithm was developed to support disjunctive rules as well and incorporated with 'Erules' [4].

- To reduce the time taken to generate frequent item set, modified FP growth algorithm was used in the place of Apriori algorithm [1].
- Positive and Negative rule sets generation: The significance of negative rule set generation was dealt [2].

In all the phases, Genetic Algorithm is used in generating rule set [4]. To decide about the reliability of the rules an additional factor “Lift” also considered [4]. The lift ratio is the confidence of the rule divided by the confidence assuming independence of consequent from antecedent. A lift ratio greater than 1.0 suggests that there is some usefulness to the rule. The larger the lift ratio, the greater is the strength of the association.

Though the rule generation task is relatively straightforward, there is an important issue lie which is generation of redundant rules. Primary task of this paper is to deal with identifying the redundant rules and eliminate the same, so that the final rule set would be only the non redundant and useful rules.

2. Rule Redundancy

The frequent item sets based association rule mining framework produces large a number of rules, because it considers all subsets of frequent item sets as antecedent of a rule. Therefore, the total number of rules grows as the number of frequent item sets increases. Number of redundant rules is larger than the previously suspected and often reaches in such extend that sometimes it is significantly larger than number of essential rules. The following rules (R1, R2, R3, R4 and R5) can be observed for understanding the duplicate rules in various scenarios.

R1: If rule X=>YZ is redundant when the rules such as XY=>Z, XZ=>Y, *X=>Y,* and *X=>Z* are satisfy the minimum support and confidence. This is because the support and confidence values *X=>YZ* are less than the support and confidence values for the rules *XY=>Z, XZ=>Y, X=>Y,* and *X=>Z* [6].

R2: Check for combination of rules [3]

A rule r in R is said to be redundant if and only if a rule or a set of rules S where S in R, possess the same intrinsic meaning of r. For example, consider a rule set R has three rules such as milk=>tea, sugar=>tea, and milk, sugar=>tea. If we know the first two rules i.e. milk=>tea and sugar=>tea, then the third rule milk, sugar=>tea becomes redundant, because it is a simple combination of the first two rules and as a result it does not convey any extra information especially when the first two rules are present.

R3: Interchange the antecedent and consequence [3]

Swapping the antecedent item set with consequence item set of a rule will not give us any extra information or knowledge.

R4: Redundant Rules with Fixed Consequence Rules [3]

Let us apply this theorem to a rule set R that has three rules such as {AB=>X, AB=>Y and AB=>XY}. Consider the rule AB=>XY has s% support and c% confidence. Then, the rules such as AB=>X and AB=>Y will also have at least s% support and c% confidence because X=>XY and Y=>XY. Since AB=>X and AB=>Y dominate AB=>XY both in support and confidence, for this reason AB=>XY is redundant.

R5: Redundant Rules with Fixed Antecedent Rules [3]

Let us apply this theorem to a rule set R that has three rules such as {XY=>Z, X=>Z and Y=>Z}. Suppose rule XY=>Z has s% support and c% confidence. If n (i.e. number of items in the antecedent) number of rules such as X=>Z and Y=>Z also satisfy s and c then, the rule XY=>Z is redundant because it does not convey any extra information if rule X=>Z and Y=>Z are present. All the above rules cannot be considered for dealing redundant rules in one domain. In this paper only Rules 3, 4 and 5 are taken for consideration and ‘Erules’ have been modified accordingly.

3. Pharmacy Database

Drug Management is an important process in pharmacy field. One important aspect of Pharmacy is having a pre idea of frequently needed medicines and their combinations. Extracting these knowledge will give the pharmacist general awareness of medicines that will help him to keep the stock in balance as per the need. In Pharmacy Database, it is important to analyze such large data because they may contain new knowledge. It extracts patterns that appear more frequently than a user-specified minimum support. ‘Erules’ is applied to find out the frequent item set and generates the useful rule sets. During the generation of the rules the duplicate rules are removed as per the discussion above with respect to Table 1.

The below table gives the details of few transactions happened in a pharmacy for the past 6 months. This data is available in Data warehouse of the Pharmacy. The pharmacy management wants to analyze and understand the buying trend in medicines by different customers. ‘Erules’ is applied on the below dataset to extract the non redundant rules. Using those strategic level managers of the pharmacy will decide about the future business plan in terms of procuring the frequently purchased medicines to improve their business.

Table 1: Pharmacy dataset for a Pharmacy:

Transaction	*Drugs Purchased*
T1	Paracetomol, Diclofenac
T2	Aceclofenac
T3	Diclofenac
T4	Ibuprofen,Diclofenac paracetomol
T5	Metaprolol , Hydrochlorthiazide
T6	Etodolac ,Paracetmol
T7	Lornoxicam, Serratiopeptides+
T8	Aceclofenac ,
T9	Aceclofenac , paracetmol ,
T10	Metaprolol succinate
T11	Cetrizine , Phenylephrine Hcl ,
T12	Losartan potassium,Ibuprofen ,
T13	Aceclofenac ,Drotaverine
T14	Amlodipine besilate
T15	Ambroxol
T16	Levocetirizene
T17	Phenylephrine Hcl
T18	Lornoxicam, Serratiopeptides,
…	…..

4. The Overview of Proposed Work

The proposed idea can be understood by the below diagram. Erules [2] was supporting to derive positive &negative disjunctive as well as conjunctive association rules. But eliminations of redundant rules were not addressed. The effectiveness of Erules can be estimated by applying it on a large dataset like pharmacy database as in Table 1. It will generate the frequent item sets and disjunctive and conjunctive rules [1] for the same.

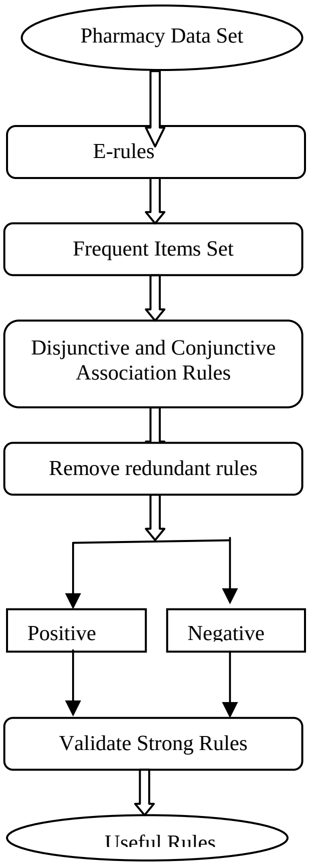

Figure 1: Framework of proposed work

In this paper the major work is spent on identifying and removing the redundant rules before they generate the positive and negative rules.

Pseudo code for Eliminating Redundant Rules

'Erules' algortithm is modified accordingly as follows,

Step 1: Create function whose parameters are Dataset, list_of_antecedents-A, list_of_consequent –C
List of A and C can be generated using the following conditions using Genetic Algorithm.
Min (A) =Cnt (FIS)-[Cnt (FIS)-1]; Max (A) =Cnt (FIS)-[(Cnt (FIS)-(Cnt (FIS)-1))]
[A- Antecedent; C-Consequent; Cnt- number of frequent item set]

Step 2: Find out allowed antecedents (A) and Consequent(C). Here A and C contains list of FIS.

Step 3:to remove redundant rules as per R4[3],
For all rules r € R

```
'r=U{A,C}
        'n=length(A)
        If(n>1)
         For all (n-1) – subsets e €A
If ($r_i$ =U{e,C})
                e.i++
        end for
if (i==n)
$R_1$=R-r
End if
End for
```

R1 returns the collection of same consequents but different antecedents.

Step 4:to remove redundant rules as per R5[3],
For all rules r € R

```
'r=U{A,C}
        'n=length(A)
        If(n>1)
         For all (n-1) – subsets e €A
If ($r_i$ =U{A,e })
                e.i++
        end for
if (i==n)
$R_2$=R-r
End if
End for
```

R2 returns the collection of same antecedents but different consequents.

Step 4: Generation of Positive and Negative rules from R1 and R2 [1]
The output of this algorithm is an non redundant and combination of positive and negative rules.

5. Result and Discussion

From the below data available in Table II and Figure 2, it is inferred very clearly that the time taken to extract the useful rules from a non redundant rule set is much faster than from redundant rule set. However there is no difference in time taken to generate frequent item set as there is no modification done. But there is a obvious difference in time taken to generate the rules from non redundant and from redundant rule set. The modified 'Erules' reduces the time taken to almost 3 times when compared with redundant rule set generate. The comparison can be realized as below,

Table 2: Time taken to generate useful rules by E-Rules using Modified Erules

Algorithm	*Time Taken to generate frequent itemset (in Sec)*	*Time Taken to generate non redundant useful rules (in minutes)*
Erules with redundant Rules	0.008	8.225
Erules with non redundant rules	0.008	3.201

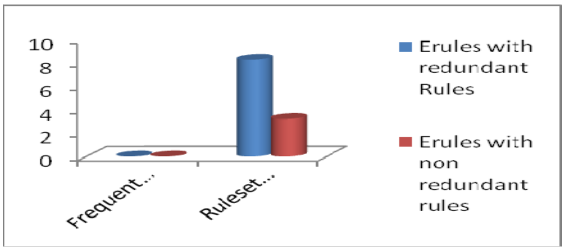

Figure 2: Comparison between the Erules [2] and the modified Erules Algorithms

The following information can be understood for the rules generated.

Lornoxicam=> Serratiopeptides OR Paracetamol
Lornoxicam=> Serratiopeptides AND Paracetamol
Lornoxicam OR Serratiopeptides => Paracetamol
Lornoxicam AND Serratiopeptides => Paracetamol
¬(Lornoxicam AND Serratiopeptides) => Paracetamol
Lornoxicam AND Serratiopeptides => ¬Paracetamol
¬ (Lornoxicam AND Serratiopeptides) => ¬Paracetamol

The above are some of the association rules generated after removed redundant rules, and as per the minimum confidence (0.75); the following rules are selected as useful rules.

Lornoxicam=> Serratiopeptides OR Paracetamol
Lornoxicam=> Serratiopeptides AND Paracetamol
Lornoxicam OR Serratiopeptides => Paracetamol
Lornoxicam AND Serratiopeptides => Paracetamol
Few of the selected negative rules whose confident is greater than the predefined confident
Indinavir => ¬triazolam
Indinavir => ¬midazolam
Indinavir => ¬alprazolam

To validate the strength of the rules the lift factor is used. As a result the following rules can be selected as a reliable rules at the end as its confident is greater than 0.75. And the useful rules were finalized by checking lift value (>1) as well.

The following information is very useful for the stakeholders to analyze and improve their business.

- Customers who will buy Lornoxicam and Serratiopeptides will also buy Paracetamol;
- Customers who will buy Lornoxicam or Serratiopeptides will also buy Paracetamol;
- Customers who will buy Lornoxicam will also buy Serratiopeptides or Paracetamol;
- Customers who will buy Lornoxicam will also buy Serratiopeptides and Paracetamol;
- Customers who will buy Indinavir cannot buy triazolam
- Customers who will buy Indinavir cannot buy midazolam
- Customers who will buy Indinavir cannot buy alprazolam

The above negative rules make logic, as per the chemical reaction also. Because the above combinations will increase the risk of serious adverse effects resulting from indinavir increasing the blood levels of these drugs. Hence the generating and validating the association rules with non redundant rules makes always sense especially in the case where large dataset is involved.

6. Conclusion

In this paper the main idea was to reduce the time and cost involved in deriving strong and useful rules from Pharmacy data ware house in order to improve their business trend. Hence 'Erules' have been modified accordingly to support the elimination of redundant rules and the effectiveness of the same has been estimated and compared. And it is realized that it has reduced the time drastically as per the claim.

7. Future Works

Since an integrated complete framework has been proposed to generate positive & negative and conjunctive & disjunctive and non redundant rules, an idea of creating a tool to support these features could be feasible in future.

References

[1] Kannika Nirai Vaani.M, Ramaraj E "E-Rules: An Enhanced Approach to derive disjunctive and useful Rules from Association Rule Mining without candidate item generation",VOL 72,NO 19,(June 2013).

[2] Kannika Nirai Vaani.M, Ramaraj E "An Optimal Approach to derive Disjunctive Positive and Negative Rules from Association Rule Mining using Genetic Algorithm",VOL 72,NO 19,(June 2013).

[3] Mohammed Javeed Zaki, "Scalable Algorithms for Association Mining" IEEE Transactionson Knowledge and Data Engineering, Vol. 12 No.2 pp. 372-390 (2000).

[4] Kannika Nirai Vaani.M, Ramaraj E "An integrated approach to derive effective rules from association rule mining using genetic algorithm", Pattern Recognition, Informatics and Medical Engineering (PRIME), 2013 International Conference, (2013), pp: 90–95.

[5] Jiawei Han, Jian Pei, and Yiwen Yin. "Mining Frequent Patterns without Candidate Generation", Data Mining and Knowledge Discovery (8), (2004), pp: 53-87.

[6] Charu C. Aggarwal and Philip S. Yu, "A new Approach to Online Generation of Association Rules". IEEE TKDE, Vol. 13, No. 4 pages 527- 540.

Author Profile

Kannika Nirai Vaani is presently working as Lead Trainer in Tech Mahidnra Ltd, Bangalore.She has got around 13 years of training experience. She has presented 3 international research papers and 2 national research papers. Her area of expertise is Database, Datamining and Informatica.

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 45 national and international conferences and published more than 35 papers in national and international journals. His research areas include Data mining and Network security.