An Efficient Association Rules Algorithms for Medical Test Analysis

Dr. Ahmed Tariq Sadiq
Computer Science Department, University of Technology/Baghdad
Email:drahmaed_tark@yahoo.com
Alaa Sameer Ali
Computer Science Department, University of Technology/Baghdad
Email:.Msc.alaa_sameer@yahoo.com

Received on:14/9/2014 & Accepted on:5/3/2015

ABSTRACT

Data Mining denotes mining knowledge from huge quantity of data. All algorithms of association rules mining include 'first finding frequency of item sets, which accept a minimum support threshold, and then calculates confidence percentage for all k-item sets to construct robust association rules'. The trouble is there are some of algorithms that need more time for compute minimum support, minimum confidence and extraction larger item. In this paper one algorithm is proposed (enhanced reduces items Apriori algorithm) to reduce execution time. The proposed algorithm purpose to introduce algorithm to mine association rules to obtain fast algorithm by reducing execute time. Due to many experiments in (enhanced reduces items Apriori algorithm), this algorithm is very fast compared with (to pk-rules and to pk-non redundant rules) algorithms.

Keywords: Apriori Algorithm, Enhanced reduces Items A priori Algorithm, Experimental Results.

INTRODUCTION

Data mining is a process for ridding during lots of data to find beneficial information that helps in decision making. Preferable data means preferable results; there are bounded ways you can ameliorate your result from data mining, preferable data is one. Preferable data means you can construct more extensive and more precise models. Data mining needs lots of data and that means more sources of data but not all data is beneficial to extracting association rules. An association rule is a rule, which implies certain association connections among a set of objects in database. Extracting association rule is an main data-mining problem. Newly there has been huge research in designing fast algorithm for this mission, since the databases to mine are often very huge (measured in gigabyte and even terabytes) [1].Data mining techniques which are exercised to medical data contain association rule mining for finding frequent types, prediction, classification and clustering [2]. Association rules are one of the promising aspects of data mining as knowledge discovery tool and have been widely explored to date, they allow capturing all possible rules that explain the presence of some attributes according to the presence of other attributes [3]. Traditionally data mining techniques were used in different fields. However, it is presented comparatively late into the Healthcare field. Nevertheless, as on today lot of research is existed in the literature. This has led to the evolution of smart systems and decision support systems in Healthcare field for precise diagnosis of diseases, predicting the strength different diseases, and remote health observation. Particular the data mining techniques are more helpful in predicting heart diseases, lung cancer, and breast cancer and so on. [2]. In this paper tried to produce new way to find association rules by using new technique may aid to reduce execution time and present a new method for

https://doi.org/10.30684/etj.34.4B.11
2412-0758/University of Technology-Iraq, Baghdad, Iraq

extracting association rules. Section 2 describes the Related works. Section3 introduce (Apriori algorithm concept). Section4 introduce the (enhanced reduces items Apriori algorithm concept), Section 5experimental results for the algorithm. Finally, Section 6 concludes the paper.

Related Works

The first manifestation of what is now called ***DATA MINING was*** in 1980s by research driven tools centered on lone mission. Through this time there was not much of a requirement to completely grasp multidimensional layers of data [4]. One dimensional analysis tools could solve these missions as constructing a classifier using decision tree or neural network tool, find cluster in data and data visualization of these tools were adequate for data analysis problem and to explain the results at that time, but with advancement of time using more than one of these tools became very uneasy for data analysis problems to solve users problems because the requires of users growth and users search for tools capable to drive largely by the investigation that the knowledge detection operation needs multiple kinds of data analysis [5].

In This paper presents some algorithms which have been proposed:

❖ In 2006 Ordonez, C., N. Ezquerra and C.A. Santana used constraint association rules in the medical area to minimize the number of detected types and introduced a greedy algorithm to calculate rule covers [6].

❖ In 2008 Razavi, A.R., H. Gill, H.Ahlfeldt and N. Shahsavar used decision tree convention to detect kinds of non-compliance with a post-mastectomy radiotherapy guideline, information which should help guideline writers and aid ameliorate the quality of on cological care [6].

❖ In 2009 Delen, D. used a diversity of data mining techniques, including artificial neural networks to determine factors affecting survivability from prostate cancer [6].

❖ In 2011 Abed, H. and L. Zaoui researched the trouble of mining interesting association rule from huge databases and proposed that background knowledge equipe beneficial area information that may aid to subjectively estimate and explain the extracted rules. A framework was sophisticated for mining interesting rules from the medical area. Background knowledge aided to discriminate visible rules from bogus and to aid in rule translation. The framework was tested by using electronic medical data registries from the medical quality amelioration consortium [6].

Apriori Algorithm Concept

Apriori equips an iterative approach known as a level wise search, where k-item sets are used to discover (k+1)-item sets. First, the collection of frequent 1-itemsets is found by scanning the database to accumulate the count for each element, and collecting those elements that satisfy minimum support. The resulting collection is denoted L1. Next, L1 is used to find L2, the set of frequent 2- item sets, which is used to find L3, and so on, until no more frequent k-item sets can be found. The finding of each Lk needs one full scan of the database. To enhance the capability of the level-wise creation of frequent item sets, an important characteristic called the Apriori characteristic, presented is used to decrease the search space [7].

Apriori characteristic is based on:

If an item set I does not satisfy the minimum support threshold, min sup, then I is not frequent that is, P(I)<minsup. If an element A is inserted to the item set I, then the resulting item set cannot happen more frequently than I. Therefore, IUA is not frequent either. In public, to find Lk two step operation consisting of joins and prune actions:[8]

1- the join step:

(A) To find Lk , a collection of candidate k-item sets is created by joining Lk-1 with itself. This collection of candidates is denoted Ck. For the (k-1)-item set the elements are isolated such that li[1] < li[2] < . . . < li[k_1].

(B) The join , *Lk*-1 on *Lk*-1,is accomplished, where members of *Lk*-1 are combinable if their first (*k*-2) elements are in common. That is, members *l*1 and *l*2 of *Lk*-1 are combined if (*l*1[1] = *l*2[1]) ^ (*l*1[2] = *l*2[2]) ^. . .^ (*l*1[*k*-2] = *l*2[*k*-2]) ^(*l*1[*k*-1] <*l*2[*k*-1]).

(C) The condition *l*1[*k*-1] <*l*2[*k*-1] simply includes that no duplicates are created. The resulting item set formed by combining *l*1 and *l*2 is *l*1[1], *l*1[2], . . . , *l*1[*k*-2], *l*1[*k*-1], *l*2[*k*-1].

2- the prune step:

(A) *Ck* is a superset of *Lk*, that is, its members may or may not be frequent, but all of the frequent *k*-item sets are included in *Ck*.

(B) A scan of database to define the count of each candidate in Ck would result in the definition of Lk .

Consequently by applying Apriori algorithm, and following the join and prune steps we lastly gain the frequent item sets.

```
Algorithm: Apriori Find frequent itemset using an iterative level-wise approach based on
candidate generation [9].
Input: Database D, of transaction; minimum support; Threshold, min-sup in D.
Output: Association rules.
Begin
L1 = find_frequent_1-itemsets(D);
For (k = 2; Lk-1 ≠Φ; k++) {
 Ck = Apriori _gen(Lk-1, min _sup);
For each transaction t∈ D {//scan D counts
Ct = subset(Ck, t); //get the subsets of t that are candidates
For each candidate c ∈ Ct
  C. count++;
  Lk  ={c ∈ Ck  |c. count ≥ min _sup}
return L = Uk Lk ;
End.
```

Figure (1-a) Apriori Algorithm

```
Procedure has_infrequent_subset (c: candidate k_itemsets);
 // Use prior knowledge
Begin
For each (k-1)_subset s of c
IF s ∉Lk-1 then return TRUE;
Return FALSE;
End.
```

Figure (1-b) Apriori Algorithm Procedure has_infrequent_subset

```
 Procedure Apriori _ gen(Lk-1:frequent(k-1)-itemsets; min _sup:
minimum support threshold)
Begin
 For each itemset l 1 ∈Lk-1
For each itemset l2 ∈ Lk-1
IF (l1[1] = l2[1]) ∧ (l1[2] = l2[2]) ∧…∧(l1[k-2]  =  l2[k-2]) ∧(l1[k-1] <
l2[k-1] ) then  c = l1 join  l2; // join step: generate candidates
IF has _ infrequent _ subset(c,Lk-1) then
      Delete c; // prune step: remove unfruitful candidate
Else add c to Ck ;
Return Ck
End.
```

Figure (1-c) Apriori Algorithm Procedure Apriori_gen

Enhanced Reduces Items Apriori Algorithm (ERIA)

The ERIA algorithm aims to produce an algorithm to mine association rules with aid of set theory and relationship concept to reduce execution time of mining association rules by extraction of a computation of items which contain the large rule with large confidence before computation of the support and the confidence for all rule. An association rule implies that there exists a relationship between all items which form it.

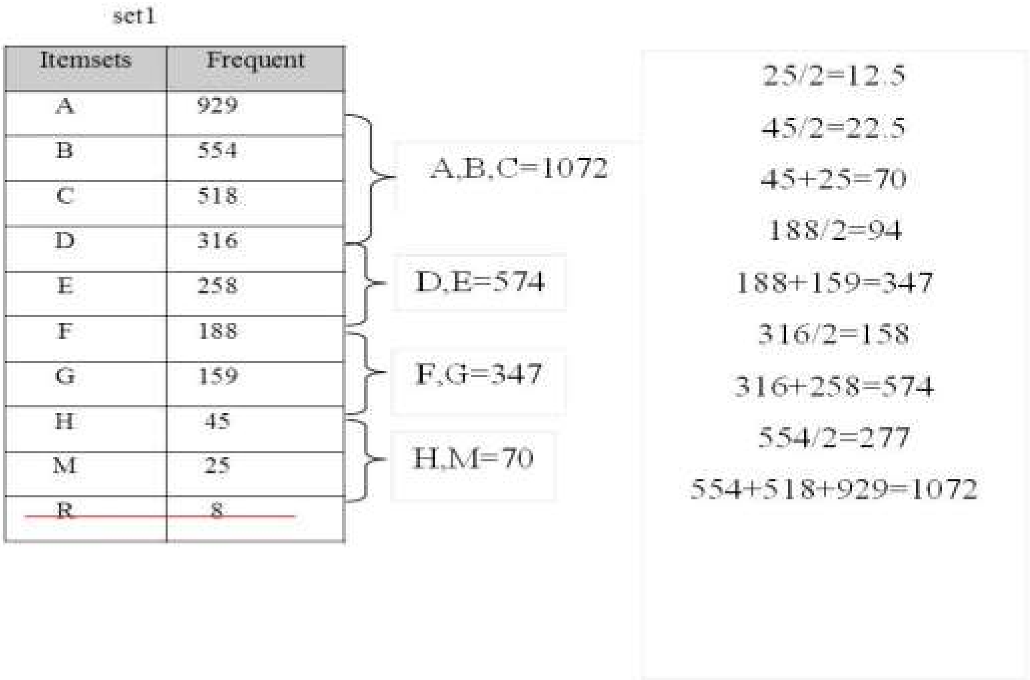
This proposal will introduce new algorithm for extraction of a collection of items which contain the large rule with large confidence. First the ERIA algorithm takes, (minimum support and minimum confidence equal to 0) and input collection of items with aggregate of each item must larger than minimum support and minimum confidence and then order these items descending, after that applies to the set from down to top, taking each two items from down if small item is equal to or large than half large item then gather these two items else omit the small item and then continue this operation applying to all items in the (set1) until at the end two items remain. These two items are gather without action if small item is equal to or large than half large item but if eventually one item remains this item is placed with nearest two items which are gather without gather with them and these two conditions are implemented only on (set1) after that (set2) results and then the same processes are applied to (set2) but without ordering descending for the (set2) and then the (set3) that is resulted from (set2).This is the final result set of items which contain the large rule with large confidence and then we are doing scan only to the final set for compute

Support (X ⇒Y) = frequent (X) / overall number of records in database and Confidence (X ⇒Y) = frequent (X ∪ Y) / frequent (X) for knowledge large item best than doing scan to all items for knowledge large item , The main procedure of ERIA .

```
ERIA Algorithm( R:=0,F:=0,Res1:=0)
Input:S=∑Ai where Ai is item,T=Ai.
Output: set of items contain large rule have large confidence.
1.      Begin
2.      For i=sizeof (s) to 0 decrement by 2
3.      Begin
4.        If (i=2) then goto step 7
5.          Else if (i=1) then goto step 15
6.                Else if (s(i-1)/2)<= s(i) then
7.                     R(F)=s(i-1)+s(i)
8.      L(F)=T(i-1) U T(i)
9.                     F=F+1
10.                 Else Remove s(i),T(i) from (s) and (T)
11.                 End if
12.      End if
13.             End if
14.      End for
15.     F=F-1
16.     L(F)=L(F) U T(i)
17.     For j=0 to F-1 increment by 2
18.     Begin
19.         If (j=1) then goto step 28
20.            Else if (R(j+1)/2)<=R(j) then
21.     Res1(c) = R(j+1) + R(j)
22.                   Res2(c) = L(j+1) U L(j)
23.                   C=C+1
24.     Else Remove R(j), L(j) from (R) and (L)
25.     End if
26.            End if
27.     End for
28.     C=C-1
29.     Res2(C) =Res2(C) U L(J)
30.     Print the result (Res2)
31.                   Res2(c) = L(j+1) U L(j)
32.                   C=C+1
33.     Else Remove R(j), L(j) from (R) and (L)
34.     End if
35.            End if
36.     End for
37.     C=C-1
38.     Res2(C) =Res2(C) U L(J)
39.     Print the result (Res2)
```

Example of the ERIA Algorithm:-

1.Arrange these items in descending order, and then apply the set from down to top, take each two items from down if small item is equal to or large than half large item then sum these two items else delete the small item.



set1

Itemsets	Frequent
A	929
B	554
C	518
D	316
E	258
F	188
G	159
H	45
M	25
~~R~~	~~8~~

A,B,C=1072

D,E=574

F,G=347

H,M=70

25/2=12.5

45/2=22.5

45+25=70

188/2=94

188+159=347

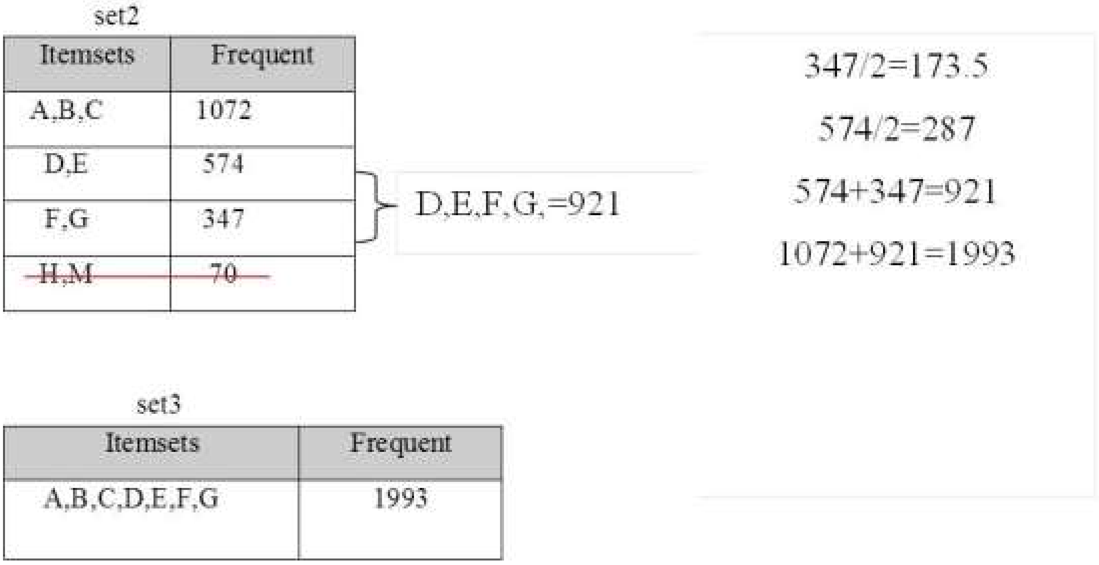
316/2=158

316+258=574

554/2=277

554+518+929=1072

2. The same operations that are used for calculating the (set 1) is done but without arranging in descending order for the (set 2) and then the (set 3).



set2

Itemsets	Frequent
A,B,C	1072
D,E	574
F,G	347
~~H,M~~	~~70~~

D,E,F,G,=921

347/2=173.5

574/2=287

574+347=921

1072+921=1993

set3

Itemsets	Frequent
A,B,C,D,E,F,G	1993

The Advantages and Disadvantages of the ERIA Algorithm:-

★ **The Advantages:-**

1. Reducing the time and potential.
2. Obtaining the group containing large rule without need to compute the confidence and the support for all rules.
3. Delete some of the items not useful from start so there is no need to scan all items.

★ **The Disadvantages:-**

1. In this algorithm some items are not useful interference in final set but they are no more than three items.

2. If the comparison between sums of items is low, the small item will not be deleted.

Experimental Results

Medical data mining represents an main part in detecting diseases from ancient data of different areas. Data mining techniques have turn into a general research tool for medical researchers to determine and invest types and relations between a big number of variables, and make them able to portend the result of a disease using the ancient datasets.In this paper all algorithms are run on medical blood tests database which contains 10 items (blood sugar , cholesterol , triglyceride , creatinine , blood urea , potassium , sodium , calcium , alk-phosphatase , ferric) and 3120 patients were taken from (medical city hospital). In this work (top-k association rules , top-k non redundant rules , enhanced top-k association rules , enhanced top-k non redundant rules and enhanced reduces items Apriori algorithm) are run on the datasets while different the number of the parameter k from 1.000 to 10.000 with minconf=0 to evaluate its impact on the execution time of the algorithm . All experiments were performed on a computer with a core i5 processor running Windows 7 and 2 GB of free RAM. This paper is helpful to the doctors for auguring the relation between blood diseases and this operation is used in hospital . Then Table (1) to (4) show results of rules and show in other tables comparisons in time between all algorithms. And then show in Table (5) show comparisons in time between all algorithms.

Conf: 90%→80%

Table (1):Results of Rules have confidence between (90-80)

No	Rules	Confidence
1	K(low)→Na(low)	84.87%

Conf: 80%→70%

Table (2):Results of Rules have confidence between (80-70)

No	Rules	Confidence
1	Cr(high)→Bur(high)	79.34%
2	Bur(high)→Cr(high)	74.19%
3	Chol(high)→Tri(high)	72.09%
4	Cr(low)→Bur(low)	70.83%

Conf: 70%→60%

Table (3): Results of Rules have confidence between (70-60)

No	Rules	Confidence
1	K(high)→Bur(high)	67.3%
2	K(high)→Cr(high)	66.04%
3	Chol(high)→Glu(high)	65.89%
4	Tri(high)→Glu(high)	65.51%
5	Ca(low)→Na(low)	64.63%

Conf: 60%→50%

Table (4):Results of Rules have confidence between (60-50)

No	Rules	Confidence
1	Na(low)→K(low)	59.43%
2	Tri(high)→Chol(high)	58.86%
3	Ca(low)→K(low)	53.38%
4	Alk. phosphate(low)→Cr(low)	50.36%

Table(5): Comparisons in time between algorithms

Data Sets	Execution Time (sec)				
	K=1000	K=3000	K=5000	K=7000	K=10.000
TopK-Rules Algorithms	3	4	5	6	7
TopK-Non Redundant Rules Algorithms	3	8	22	37	78
ERIA Algorithm	1	3	8	13	30

CONCLUSION

Base on the select of parameters, association rule mining algorithms can create an very big number of rules which drive algorithms to trouble from long execution time and large memory consuming, To manipulate this matter, we proposed (ERIA algorithm), an algorithm to discover the set of items contain The large rule have large confidence, where k is collection by the user. Experimental results show that ERIA algorithm has excellent accomplish, and that it is an advantageous instead to ancient association rule mining algorithms when the user obtain some known which rule is have large confidence.

REFERENCES

[1] Executive Briefing SPSS " Field tested data mining" U.S.A 1999.

[2] Mohammed Abdul Khaleel, Sateesh Kumar Pradham, G.N. Dash, "A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases",International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 8, August 2013.

[3] Emad Kadhiem Jabbar and Waheed Abd Al-Kadhiem Salman, " Proposed Parallel Association Rules Algorithm ", Eng & Tech. Journal. Vol. 32, Part(B), No. 1,2014.

[4] Piatetsky Shapiro,Gregor,"Data mining industry coming of age",IEEE intelligent system 2000.

[5] Charu C. Agrawal and philip S. Ya, "Mining large itemesets for association rules", Bulletin of the IEEE computer society Technical Committe on Data Engineering, 21(1) March 1998.

[6] Bakheet Al dosari, Ghada Al modaifer, Alaa eldin Hafez and Hassan Mathkour,"Constrained Association Rules For MedicalData",Journal of Applied Sciences,2012.

[7] Girja Shankar and Latita Bargadiya, " A New Improved Apriori Algorithm For Association Rules Mining ", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 2 Issue 6, June - 2013.

[8] M.Naga Sushma, Md.Rabiya Begum, K.Pavani Swarupa and D.Sruthi, " Mining Of Frequent,Closed ,Maximal Frequent Itemsets and Association Rules ", Department Of Computer Science And Engineering Gokaraju Rangaraju Institute Of Engineering &Technology(Affiliated To J.N.T.University, Hyderabad) Bachupally, Kukatpally, Hyderabad-50007.

[9] Jiawei Han, Micheline Kanmber, "data mining concepts and techniques", Academic press, USA, 2001.