ASSOCIATION MODELS FOR PREDICTION WITH APRIORI CONCEPT

Smitha.T[1], V.Sundaram[2]

[1]PhD-Research Scholar, Karpagam University, Coimbatore
Asst. Prof., Department of Computer Application, SNGIST, N. Paravoor, Kerala, India
[2]Director-MCA, Karpagam College of Engineering, Coimbatore, India

ABSTRACT

Data mining techniques have led over various methods to gain knowledge from vast amount of data. So different research tools and techniques like classification algorithm, decision tree, association rules etc are available for bulk amount of data. Association rules are mainly used in mining transaction data to find interesting relationship between attribute values and also it is a main topic of data mining There is a a great challenge in candidate generation for large data with low support threshold. Through this paper we are making a study to show how association rules will be effective with the dense data and low support threshold. The data set which we have used in this paper is real time data of certain area and we are applying the data set in association rules to predict the chance of disease hit in that area using A Priori Algorithm. In this paper three different sets of rules are generated with the dataset and applied the apriori algorithm with it. With the algorithm, found the relation between the parameters in the database.

KEYWORDS: *APriori algorithm, Association rules, Data mining, item based partitioning, multi Dimensional analysis.*

I. INTRODUCTION

Association rules discover correlation among data items in a transactional data base. It involves the discovery of rules that satisfy defined threshold from tabular database. Here the rule how often it occurs in the data base which is known as its frequency is important. Association rule mining is the process of finding frequent set with minimum support and confidence. The first phase is support counting phase where we have to find the frequent set generation. Effective partitioning may help for this process

We also have to create a border set to avoid frequent updating of real time data. In real life applications, the number of frequent sets are large in number and as the result, the number of association rules are also very large.. We are selecting only the rules which we have interested for disease prediction in this context. The discovery of frequent item sets with item constraints is also very much important.

There are many data mining algorithms such A priori Algorithm, Partition algorithm, Pincer-Search Algorith, Dynamic Itemset Counting Algorithm [2], FP-Tree Growth etc are used for finding the discovery of frequent sets. are related with association rules.Here we are applying A Priori algorithm to the dataset to find the frequent sets and with the help of the algorithm we are predicting the chances of disease hit in the particular area.

1.1 ASSOCIATION RULE DEFINITION

The basic definition of association rule states that Let A={l1,l2,l3,........ln} be a set of items and T is the transactional database where t is the set of items of each transaction, then t is the subset of A.

A transaction t is said to support an item li if li is present in t, t is said to be support a subset of items X€A has a support s in T, denoted by s(X)t , if s% of transaction in T support X.[4]

The key feature of association rule algorithm is that each of the methods assume that the underlying database size is enormous and they require minimum passes over the database and the data must run thousands of transactions per second. So to make efficient computing the problem of mining association rules must be decomposed into sub problems.

Association rule mining searches for interesting relationships among items in a given set. Here the main rule interestingness are rule support and confidence which reflect the usefulness and certainty of discovered rules. Association rule mining algorithm is a two step process where we should have to find all the frequent item sets and generate strong association rules from the frequent item sets.[9]If a rule concerns association between the presence or absence of items, it is a Boolean association rule. If a rule describes association between quantitative items or attributes, then it is known as quantitative association rules. Here the quantitative values for items are partitioned into intervals. The algorithm can be formed based on dimensions, based on level of abstractions involved in the rule set and also based on various extensions to association mining such as correlation analysis. [27]

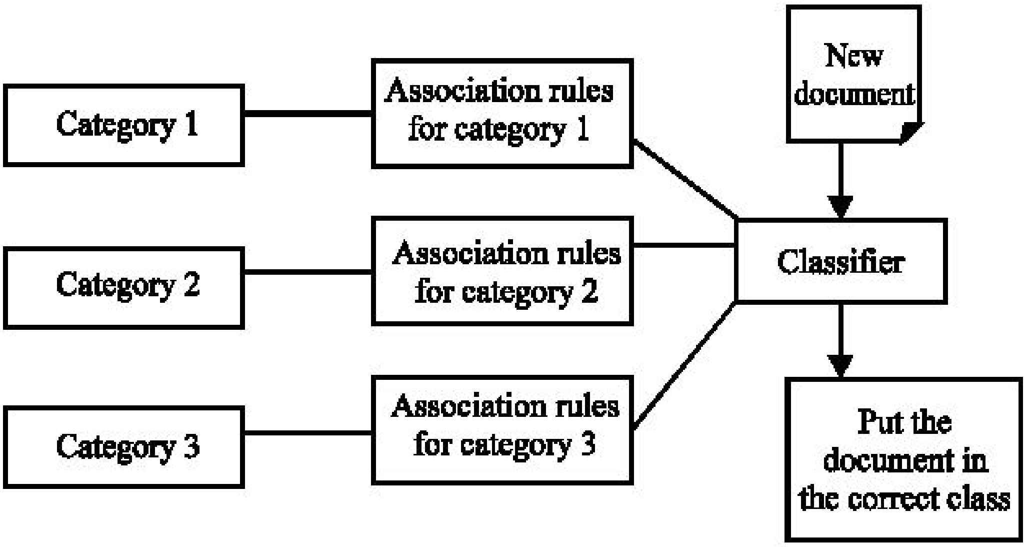
Category 1
Category 2
Category 3
Association rules for category 1
Association rules for category 2
Association rules for category 3
New document
Classifier
Put the document in the correct class

Figure 1. - Architecture of Associative Classifier

1.2 RELATED WORKS IN THIS AREA

Many works related in this area have been going on An article “Item based partitioning approach of soybean data for association Rule mining” ,the authors applied classification technique in data mining in Agriculture land soil. The article on “ A study on effective mining of association rules from huge data base “ by V. Umarani, [20] It aims at finding interesting patterns among the databases. The paper also provides an overview of techniques that are used to improvise the efficiency of Association Rule Mining (ARM) from huge databases. In another article “ K-means v/s K-medoids: A Comparative Study” Shalini S Singh explained that portioned based clustering methods are suitable for spherical shaped clusters in medium sized datasets and also proved that K-means are not sensitive to noisy or outliers.[21]. There are many research works carrying out related with data mining technology in prediction such as financial stock market forecast, rainfall forecasting, application of data mining technique in health care, base oils biodegradability predicting with data mining technique etc,[23].

II. DISCOVERY OF ASSOCIATION RULES

The problems of mining association rules can be decomposed into different sub problems.

Then find the frequent items set by selecting the all the item set whose support is greater than the minimum support specified by the user and then use that frequent item sets to generate the desired rule.

The frequent set can be determined by the following rule.

Let T be the transaction database and σ be the user specified minimum support, then the item set X€A is said to be frequent ser in T with respect to σ if S(X) r>= σ. We cannot establish a definite relationship between the set of maximal frequent sets and the set of border sets. [3]

Consider the example of a dataset which contains information of inhabitants in an area. Some of the attributes includes in the database are age, income, education, family history of any disease, sex, environmental condition, area of the house, hygenity, source of water, income etc. With the help of association rules algorithm, we will be able to discover some of the association and sequential tool to predict the disease hit in that area.

The male person who is the age between 30-60 and living in urban area with poor drinking water facility has a chance to hit typhoid.

Each rule has a left hand side and a right hand side. The left hand side is called the antecedent and the right hand side is called consequent. Both left hand side and right hand side can contain multiple items.

The association rule has two measures called confidence and support. [6]

Let T consists of 1000 data. 250 data contains the value 0 for "disease history" and 750 data contains the value 1 for the same parameter. Similarly suppose 380 data contains the value 0 for hygenity and 620 data contains the value 1 for the same attribute. By applying association rule algorithm, we will be able to predict what type of people is affecting the disease. Ie, how the attributes are co related? Or whether there is a co-relation among the parameters disease history and hygenity in the case of disease prediction.

Thus we are measuring the confidence and support from the dataset. The pruning step eliminates the item set which are not found in frequent from being considered for counting support.[13]

The A Priori frequent set discovery item set uses the functions candidate generation and pruning at every iteration. It moves upward in the lattice starting from level 1 till level k, where no candidate set remain after pruning.[8]

2.1 APRIORI ALGORITHM FOR CANDIDATE GENERATION AND PRUNING

The APriori frequent set discovery item set uses the functions candidate generation and pruning at every iteration. It is also known as the level wise algorithm which is used to find all the frequent sets. It uses a bottom up approach and moving upward level wise in the lattice. In each level the data sets has to be pruned to take the frequent sets.

The candidate generation method algorithm is as follows

```
    Gen-itemsets with the given $L_{k-1}$ as follows:
            $C_k$=Ǿ
            For all itemset $l_1$€$L_{k-1}$ do
            For all itemset $l_2$€$_{Lk-1}$ do
            If $l_1[1]=l_2[1]$^$l_1[2]=l_2[2]$^…..^$l_1[k-1]<l_2[k-1]$ then
            C=l1[1],l1[2]……l1[k-1],l2[k-1]
            $C_k$= $C_k$ U{C}…………………equ(2)
```

The pruning set eliminates the extension of (k-1) item sets which are infrequent from the counting support.[10]

The pruning algorithm is as follows:

```
          Prune(Ck)
                For all c€ Ck
                 For all(k-1) subsets d of c do
            If d€Lk-1
                Then Ck=Ck\{c}…………Equ(3)
```

It is known as the level wise algorithm which is used to find all the frequent sets. It uses a bottom up approach and moving upward level wise in the lattice. In each level the data sets has to be pruned to take the frequent sets. [25]

III. Models used in Predictive Association Rule Mining

Association rules allows the analysts to identify the behavior pattern with respect to a particular event where as frequent items are used to find how a group are segmented for a specific set. Clustering is used to find the similarity between entities having multiple attributes and grouping similar entities and classification rules are used to categorize data using multiple attributes.[13]

3.1 APRIORI ALGORITHM BY EXAMPLE

We have applied out data set to work with the Apriori algorithm to check its reliability.
Initially k:=1 . Read the database to count the support of 1-itemsets and found the frequent item set and their support. Find L1 with k=1 then change k=2 and find the candidate generation step and find the value of C2.Check the pruning step and check whether there is any change in C2.Read the data base to count the support of elements in C2. Then Assign k to 3 and find c3. Read the database to count the support of itemsets in C3 to get L3.Find the set of frequent sets along with their respective support values and qpply it to the association rules.[22]

Table 1: To read the database to count the support of L –item sets

{1}	2
{2}	6
{3}	6
{4}	4
{5}	8
{6}	5
{7}	7
{8}	4
{9}	2

K:=1
L1:= ({2}->6,{3}->6, {4}->4, {5}->8, {6}->5, {7}->7, {8}->4, {9}->2}.
L1 contains 8 elements.
K:=2, calculate L2 and C2.
L2:= {{2,3}->3,{2,4}->3,(3,5)->3, (3,7)->3, {5,6)->3, (5,7)->5, (6,7)->3}
K:=3 calculate L3 and C3.
C3={3,5,7},{5,6,7}}and . L3:={(3,5,7}->3K:=4
As L3 contains only one element candidate C4 is empty. So algorithm can stop
L:= L1UL2UL3…………………..Equ(4)

3.2 Generating 1-itemset Frequent Pattern

If the database consists of 900 patterns, Calculate minimum support count
Minimum support count =200/900= 2%.
Let the minimum confident required is 70%. So we have to find the frequent item set using apriori algorithm and generate the association rule with minimum support and maximum confidence.
So scan the data set and count each candidate. Then compare candidate support count with minimum support count.

Table 2: Generating 1-itemset Frequent Pattern

itemset	Support count
{ 11}	6
{12}	7
{13}	6
{14}	2
{15}	2

itemset	Support count
{ 11}	6
{12}	7
{13}	6
{14}	2
{15}	2

In the first iteration of the algorithm, each item is a member of the set of candidate then generate 2-item set frequent pattern.

Step 2: Generating 2-itemset Frequent Pattern

To discover the set of frequent 2-itemsets, L2, the algorithm uses L1 *Join* L1to generate a candidate set of 2-itemsets, C2.,•Next, the transactions in D are scanned and the support count for each candidate itemset in C2is accumulated .•The set of frequent 2-itemsets, L2, is then determined, consisting of those candidate 2-itemsets in C2having minimum support.

Table 3: Generating 2-itemset Frequent Pattern

C2

Item set
{11,12}
{11,13}
{11,14}
{11,15}
{12,13}
{12,14}
{12,15}
{13,14}
{13,15}
{14,15}

C2 L2

itemset	Support count	**itemset**	**Support count**
{11,12}	4	**{11,12}**	**4**
{11,13}	4	**{11,13}**	**4**
{11,14}	1	**{11,14}**	**1**
{11,15}	2	**{11,15}**	**2**
{12,13}	4	**{12,13}**	**4**
{12,14}	2	**{12,14}**	**2**
{12,15}	2	**{12,15}**	**2**
{13,14}			
{13,15}			
{14,15}			

STEP 3: Generating 3 itemset Frequent Pattern

This step involves the use of Apriori algorithm.

Find C3 by computing L2 join L2.

C3= L2 *Join*L2 = {{I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5}}…….equ(4)

Now, Join step is complete and Prune step will be used to reduce the size of C3 .

Based on the Apriori property that all subsets of a frequent itemset must also be frequent, we can determine that four latter candidates cannot possibly be frequent.

Consider the data {I1, I2, I3}.

The 2-item subsets of it are {I1, I2}, {I1, I3} & {I2, I3}.

Since all 2-item subsets of {I1, I2, I3} are members of L2, We will keep {I1, I2, I3} in C3. {I3, I5} is not a member of L2and hence it is not frequent violating Apriori Property.

Thus We will have to remove {I2, I3, I5} from C3.•Therefore, C3= {{I1, I2, I3}, {I1, I2, I5}} after checking for all members of result of Join operation for Pruning. Now, the transactions in D are scanned in order to determine L3, consisting of those candidates 3-itemsets in C3having minimum support.[24]

Step 4. Generating 4-itemset Frequent Pattern

The algorithm uses L3 *Join*L3to generate a candidate set of 4-itemsets, C4. Although the join results in {{I1, I2, I3, I5}}, this itemset is pruned since its subset {{I2, I3, I5}}is not frequent. Thus, C4= φ, and algorithm terminates, having found all of the frequent items. This completes our Apriori Algorithm. These frequent itemsets will be used to generate strong association rules which satisfy both minimum support & minimum confidence.[22]. Generate association rules from frequent item sets for 4 items as follows.

For each frequent itemset ***"l",*** generate all nonempty subsets of ***l.***

For every nonempty subset ***s*** of ***l***, output the rule **"s->l-s"** if **support_count(l)/support_count(s)>=min_conf** where min_confis minimum confidence threshold.

Let minimum confidence threshold is , say 70%.

The resulting association rules are shown below, each listed with its confidence.

–R1: I1 ^ I2 □I5

Confidence = sc{I1,I2,I5}/sc{I1,I2} = 2/4 = 50%......................Equ(5)

R1 is Rejected.

R2: I1 ^ I5 □I2

Confidence = sc{I1,I2,I5}/sc{I1,I5} = 2/2 = 100%........................Equ(6)

R2 is Selected.

–R3: I2 ^ I5 □I1

Confidence = sc{I1,I2,I5}/sc{I2,I5} = 2/2 = 100%.......................Equ(7)

R3 is Selected.

Step 5:Generating Association Rules from Frequent Itemsets

R4: I1 □I2 ^ I5

Confidence = sc{I1,I2,I5}/sc{I1} = 2/6 = 33%................................Equ(8)

R4 is Rejected.

–R5: I2 □I1 ^ I5

Confidence = sc{I1,I2,I5}/{I2} = 2/7 = 29%...............................Equ(9)

R5 is Rejected.

–R6: I5 □I1 ^ I2

Confidence = sc{I1,I2,I5}/ {I5} = 2/2 = 100%..............................equ(10)

R6 is Selected.

In this way, We have found three strong association rules.

IV. RESULT AND DISCUSSION

Different three strong association rules are generated with the data set by applying Apriori algorithm. From the study it revels that there are certain associations between different parameters in the database such as age, sex, environmental conditions and humidity, for the prediction of disease of an area. The study reveals the prediction that male person at the age between 30-60 having poor environmental condition have a tendency to hit the contagious disease. Study also reveals that family history of the disease is not an important factor for hitting contagious disease.

V. FUTURE ENHANCEMENT

Without the candidate generation process also we can apply the same mining technique to the data set. In this candidate generation process we should have to apply the database scan. So to avoid costly database scan, we can do frequent pattern tree structure. The same algorithm can also be applied with different datasets.

REFERENCES

[1]. Arijay Chaudhry and DrP.S.Deshpande. Multidimensional Data Analysis and data mining,Black Book

[2]. Oulbourene G, Coenen F and Leng P, "Algorithms for Computing Association Rules using a Partial support Tree" Knowledge Based System 13(2000)pp-141-149.

[3]. R.Agarwal, T.Imielinski and A.Swamy "Mining association Rules between Set of Items in Large Database".In ACM SIGMO international conference on Management of Data .

[4]. en.wikipedia.org/wiki/Data_mining

[5]. David Hand,Heikki Mannila, Padhraic Smyth,"principles ofData Mining".

[6]. Smitha.T ,Dr.V.Sundaram"Case study on High Dimensional Data Analysis using Decision Tree model", , International journal of computer science issues Vol9,Issue 3, May 2012.

[7]. Smitha.T,Dr.V.Sundaram"Classification Rules By Decision Tree for disease Prediction", ,International journal of Computer Applications vol-43, No-8, April 2012.

[8]. “Smitha.T, Dr.V.Sundaram” Knowledge Discovery from Real TimeDatabase using Data Mining Technique, IJSRP vol 2, issue 4, April 2012.
[9]. Hyndman R and Koehler A”Another Look at Measures of Forecast Accuracy” (2005).
[10]. S. Weng I C Zhang I Z. Lin I X. Zhang 2 **“**Mining the structural knowledge of high-dimensional medical data using Isomap”
[11]. Bhattachariee.A 'Classification of human lung carcinomasby mRNA expression profiling reveals distinct adenocarcinomasubclasses', Proc. Nat. Acad. Sci. USA, 98, pp. 13790 13795 BLAKE, C. L. and Merz, C. J. 2001
[12]. Borg.T and Groenen.P.): 'Modern multidimensional scaling: theory and application' (Springer-Verlag, New York,Berlin, Heidelberg, 1997).
[13]. Adomavicius G,TuzhilinA2001 “ Expert-driven validation of rule-based user models in personalization
[14]. Applications”. Data Mining Knowledge Discovery 5(1/2): 33–58.
[15]. Shekar B, Natarajan R “ A transaction-based neighbourhood-driven approach to quantifying interestingness of association rules.”Proc. Fourth IEEE Int. Conf. on Data Mining (ICDM 2004)(Washington, DC: IEEE Comput. Soc. Press) pp 194–201
[16]. Mohammed J. Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and WeiLi. Parallel algorithms for discovery of association rules. Data Mining and Knowledge Discovery: An International Journal, special issue on Scalable High-Performance Computing for KDD, 1(4):343–373, December 2001.
[17]. Refaat, M. “Data Preparation for Data Mining Using SAS,Elsevier”, 2007.
[18]. El-taher, M.” Evaluation of Data Mining Techniques”, M.Sc thesis (partial-fulfillment), University of Khartoum, Sudan,2009.
[19]. Lee, S and Siau, K. A review of data mining techniques, Journal of Industrial Management & Data Systems, vol 101,no 1, 2001, pp.41-46.
[20]. Moawia Elfaki Yahia1, Murtada El-mukashfi El-taher2 “A New Approach for Evaluation of Data Mining Techniques”, ,IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.
[21]. V.Umarani “A study on effective mining of association rules from huge database” al. / IJCSR International Journal of Computer Science and Research, Vol. 1 Issue 1, 2010.
[22]. Shalini S Singh “ K-means v/s K-medoids: A Comparative Study”, National Conference on Recent Trends in Engineering & Technology, May 2011.
[23]. C. MÁRQUEZ-VERA” Predicting School Failure Using Data Mining” IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010
[24]. K.Srinivas et al. “ Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks” International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255.
[25]. Smitha.T, Dr.V.Sundaram “Comparative study of data mining algorithm for high dimensional data analysis” International journal of advances in Engineering & Technology, Vol 4, issue 2, ISSN. 2231-1963, Sept-12, pp. 173-178.
[26]. Arun K Pujari “Data mining Techniques” Arun K Pujari.
[27]. Jie Tang,Hang Li,Yunbo Cao and Zhaohui Tang,2005.Email datacleaning.KDD’05,Chigago,USA.
[28]. G.SenthilKumar “online message categorization using Apriori algorithm” International Journal of Computer Trends and Technology- May to June Issue 2011.
[29]. Han, J.and M.Kamber,2001.Data Mining:Concepts and Techniques,Morgan Kanfmann publishers

AUTHOR’S BIOGRAPHY

Smitha.T.: She has acquired her Post Graduate Degree in Computer Application and M.Phil in Computer science from M. K. University. Now doing PhD in Computer Science at Karpagam University under Dr. V. Sundaram. .She has 10 years of teaching experience and 4 years of industrial and research experience. She has attended many national and international conferences and workshops and presented many papers, regarding data mining,. She has also published many articles regarding data mining techniques in international journals with high impact factor. Now working as an Asst. Professor–MCA department of Sree Narayana Guru Institute of Science and Technology, N. Paravoor, Kerala. Her area of interest is Data mining and Data Warehousing.

V. Sundaram: He is a postgraduate in Mathematics with PhD in applied mathematics. He has 45 years of teaching experience in India and abroad and guiding more than 15 scholars in PhD and M. Phil at Karpagam and Anna University. He has organized and presented more than 40 papers in national as well as international conferences and have many publications in international and national journals. . He is a life member in many associations. His area of specialization includes fluid Mechanics, Applied Mathematics, Theoretical Computer Science, Data mining, and Networking etc.