Available online at www.elixirpublishers.com (Elixir International Journal)

Information Technology

Elixir Inform. Tech. 71 (2014) 24666-24669

Elixir ISSN: 2229-712X

Designing Association Models for Disease Prediction using Apriori

N K Kameswara Rao and G P Saradhi Varma
Department of Information Technology, SRKR Engineering College, Bhimavaram, Andhra Pradesh, India.

ARTICLE INFO

Article history:
Received: 7 April 2014;
Received in revised form:
20 May 2014;
Accepted: 3 June 2014;

Keywords
Apriori algorithm,
Association rules,
Data mining,
Item based partitioning,
Multi Dimensional analysis.
Prediction.

ABSTRACT

Data mining is has three major components Clustering or Classification, Association Rules and Sequence Analysis. Association rules used to find interesting relationship between attribute values. Sequence Analysis used to find statistically relevant patterns between data. Data mining techniques have led over various methods to gain knowledge from vast 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 great challenge in candidate generation for large data with low support threshold. Association rules will be effectively worked with the solid data and low support threshold was discussed. By using Apriori algorithm we applied association rules on data set of certain areas to predict the chance of getting the dengue disease, the above data set was collected from some selected areas, so it is the real time data. Three different sets of rules are generated with this dataset and applied the Apriori algorithm to it, find the relation between the parameters in database.

Introduction

Prediction of trend analysis can be done by the process of analyzing data from different perspectives and summarizing it into useful information is called data mining. Unexpected patterns can be discovered by data mining were not under consideration while the mining process was started Prediction, a task of learning a pattern from examples and using the developed model to predict future values of the target variable[1].

Correlation among data items in a transactional database discover by Association rules. These are also involved to discover the rules that are satisfy, defined threshold from tabular database. The occurrence of the rule in the database is known as its frequency which is very important. The process of finding frequent set with minimum support and confidence is known as association rule mining. Finding the frequent generation is the first phase in this process it is also called support counting phase. Effective partitioning method is useful for this.

Creating a border set is very much useful to avoid frequent updating of real time data. The number of frequent sets are very high in real life applications so the number of association rules generated are also very large. The rules which we have interested for dengue disease prediction are only selected in this context. The discovery of frequent itemsets with item constraints is also very much important.

Apriori Algorithm, Partition algorithm, Pincer-Search Algorithm, Dynamic Itemset Counting Algorithm [2], FP-Tree Growth etc. data mining algorithms are used for finding the discovery of frequent sets which are related with association rules. Apriori algorithm is applied to the dataset for finding the frequent sets, with the help of this algorithm predicted the chances of getting dengue disease in the particular areas.

Association Rule

The basic definition of association rule states that Let A={l_1,l_2,l_3,........l_n} 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 l_i, if l_i 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].

Each of the methods assumes 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 are the key feature of association rule algorithm. So the efficient way to make computing the problem in mining association rules must be decomposed into sub problems.

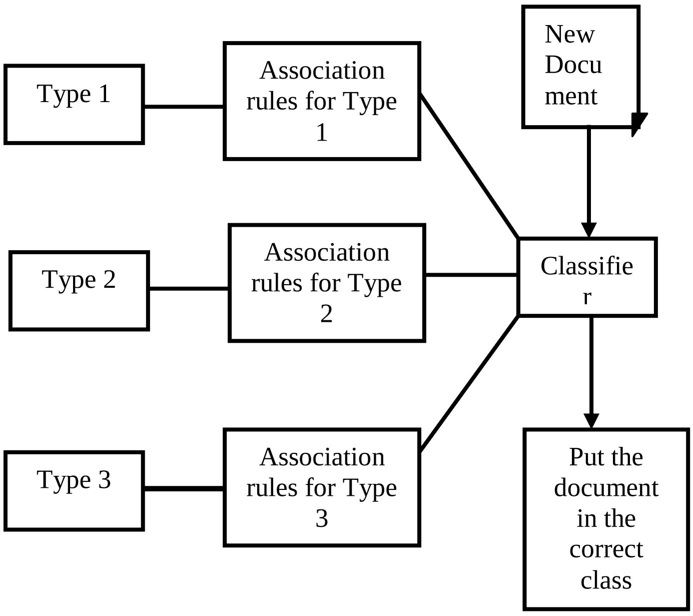
Association rule mining searches for interesting relationships among items in a given set. The interesting rule in Association rule mining is the rule support and confidence which reflect the usefulness and certainty of discovered rules. Finding all the frequent itemsets and generating strong association rules from the frequent itemsets are the two important steps in Association rule mining algorithm [9]. A Boolean association rule is a rule that concerns association between the presence and absence of items. Quantitative association rule is a rule that describes association between quantitative items and attributes. So, the quantitative values for items are partitioned into intervals. The algorithm performance is 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].

Literature review:

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

Tele:
E-mail addresses: nkkamesh@gmail.com

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].


Type 1
Association rules for Type 1
New Docu ment
Type 2
Association rules for Type 2
Classifie r
Type 3
Association rules for Type 3
Put the document in the correct class

Discovery of Association Rules:

In association rules, the mining problem decomposed into different sub problems. Then the frequent items can be find by selecting the all the itemset whose support is greater than the minimum support specified by the user, then using that frequent itemsets the desired rule can be generated. 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 itemset X€A is said to be frequent set 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].

Age, Income, Education, Hereditary (family history of disease), Gender, Environmental Condition, Area of the house, Hygienist, Source of water etc. information collected from the people in a selected areas, consider these attributes from the database for disease prediction. We can discover some of the association and sequential tools to predict the change of getting disease in those areas with the help of association rule mining algorithm.

The people having age between 0-20 and above 60 years living in tribal and hill areas with poor water disposal system have a change to getting dengue disease. Every rule has two sides' left hand side and right hand side. Both sides can contain multiple items. Confidence and support are two measures in the association rule [6].

Let T consists of 2424 records. 1191 records contain the value no for "Hereditary" and 1233 records contain the value yes for the same parameter. Similarly suppose 843 records contain the value yes for "hygienist" and 1581 records contain the value no for the same attribute. We can predict which people are affected by dengue disease by applying association rule algorithm. We can understand how the attributes are correlated and where there is a correlation between the parameters Hereditary and hygienist in dengue prediction.

We measured the confidence and support from the above dataset. The pruning step eliminates the itemset which are not found in frequent from being considered for counting support [13].

Apriori Algorithm for Candidate Generation And Pruning:

The Apriori frequent set discovery itemset 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].

The Apriori frequent set discovery itemset 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]. The candidate generation method algorithm is as follows

Gen-itemsets with the given L_{k-1} as follows:

```
Ck=Ǿ
For all itemset $l_1$€$L_{k-1}$ do
  For all itemset $l_2$€$L_{k-1}$ do
        If $l_1[1]=l_2[1]$^$l_1[2]=l_2[2]$^…..^$l_1[k-1] <l_2[k-1]$ then
        C=$l_1[1],l_1[2]……l_1[k-1],l_2[k-1]$
        $C_k$= $C_k$ U{C}………………….Eq(1)
```

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($C_k$)
   For all c€ $C_k$
     For all(k-1) subsets d of c do
             If d€$L_{k-1}$ Then
             $C_k$=$C_k$\{c}…………Eq(2)
```

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].

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 is 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]

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 is segmented for a specific set. Finding the similarities between entities having multiple attributes and grouping similar entities using by clustering and categorizing data using multiple attributes are using by classification rules [13].

Apriori Algorithm by Example

We have applied out data set to work with the Apriori algorithm to check its reliability. To check the reliability of the Apriori algorithm we applied a dataset to it. Initialize k value with 1. Retrieved the database to count the support of 1-itemsets and find the frequent itemsets and their support. Find L_1 with k=1 than increase the k value to 2, find the candidate generation step and find the value of C_2. Checked the pruning step whether is there any change in C_2. Retrieved the data base to count the support of elements in C_2 and then increased the k value to 3 and find C_3. Retrieved the data base to count the support of itemsets in C_2 to get L_3, Find the set of frequent sets along with their respective support values and applied it to the association rules [22].

K:=1

L_1:= ({2}->6,{3}->6, {4}->4, {5}->8, {6}->5, {7}->7, {8}->4, {9}->2}.

L1 contains 8 elements.

K:=2, calculate L_2 and C_2.

L_2:= {{2,3}->3,{2,4}->3,(3,5)->3, (3,7)->3, {5,6)->3, (5,7)->5, (6,7)->3}
K:=3 calculate L_3 and C_3.
C_3={3,5,7},{5,6,7}}and . L_3:={(3,5,7}->3
K:=4
As L_3 contains only one element candidate C_4 is empty. So algorithm can stop
L:= $L_1UL_2UL_3$…………………..**Eq(3)**

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

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

Step 1: Generating 1-itemset Frequent Pattern

If the database consists of 3300 patterns, Calculate minimum support count Minimum support count =667/3300= 2%.

Let the minimum confident required is 70%. So we have to find the frequent itemset 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

Each item is a member of the set of candidate then generate 2-itemset frequent pattern in the first iteration of the algorithm.

Step 2: Generating 2-itemset Frequent Pattern

To discover the set of frequent 2-itemsets, L_2, the algorithm uses L_1 Join $L_1$to generate a candidate set of 2-itemsets, C_2., Next, the transactions in D are scanned and the support count for each candidate itemset in $C_2$is accumulated. The set of frequent 2-itemsets, L_2, is then determined, consisting of those candidate 2-itemsets in C_2having minimum support.

Table 3: Generating 2-itemset Frequent Pattern

C_2

Itemset
{11,12}
{11,13}
{11,14}
{11,15}
{12,13}
{12,14}
{12,15}
{13,14}
{13,15}
{14,15}

C_2 L_2

Step 3: Generating 3 itemset Frequent Pattern
This step involves the use of Apriori algorithm.
Find C_3 by computing L_2 join L_2.

C_3= L_2 JoinL_2 = {{I_1, I_2, I_3}, {I_1, I_2, I_5}, {I_1, I_3, I_5}, {I_2, I_3, I_4}, {I_2, I_3, I_5}, {I_2, I_4,I_5}}…….**Eq(4)**

Now, Join step is complete and Prune step will be used to reduce the size of C_3.

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 {I_1, I_2, I_3}.

The 2-item subsets of it are {I_1, I_2}, {I_1, I_3} & {I_2, I_3}.

Since all 2-item subsets of {I_1, I_2, I_3} are members of L_2, We will keep {I_1, I_2, I_3} in C_3. {I_3, I_5} is not a member of $L_2$and hence it is not frequent violating Apriori Property.

Thus We will have to remove {I_2, I_3, I_5} from C_3. Therefore, C_3= {{I_1, I_2, I_3}, {I_1, I_2, I_5}} after checking for all members of result of Join operation for Pruning. Now, the transactions in D are scanned in order to determine L_3, consisting of those candidates 3-itemsets in C_3having minimum support.[24]

Step 4: Generating 4-itemset Frequent Pattern

The algorithm uses L_3 Join$L_3$to generate a candidate set of 4-itemsets, C_4. Although the join results in {{I_1, I_2, I_3, I_5}}, this itemset is pruned since its subset {{I_2, I_3, I_5}} 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 itemsets 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

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}			

Support-count(l)/support-count(s)>=min-conf where min-conf is minimum confidence threshold.
Let minimum confidence threshold is, say 70%.
The resulting association rules are shown below, each listed with its confidence.

–R_1: I_1 ^ I_2 ℵI_5
Confidence=sc{I_1,I_2,I_5}/sc{I_1,I_2}=2/4= 50%... ....**Eq(5)**
R_1 is rejected.
R_2: I_1 ^ I_5 ℵI_2
Confidence=sc{I_1,I_2,I_5}/sc{I_1,I_5}=2/2= 100%.................**Eq(6)**
R_2 is selected.
–R_3: I_2 ^ I_5 ℵI_1
Confidence=sc{I_1,I_2,I_5}/sc{I_2,I_5}=2/2= 100%...............**Eq(7)**
R_3 is selected.

Step 5: Generating Association Rules from Frequent Itemsets

R_4: I_1 ℵI_2 ^ I_5
Confidence=sc{I_1,I_2,I_5}/sc{I_1}=2/6=33%........................**Eq(8)**
R_4 is rejected.
–R_5: I_2 ℵI_1 ^ I_5
Confidence=sc{I_1,I_2,I_5}/{I_2}=2/7= 29%.........................**Eq(9)**

R_5 is rejected.

$-R_6$: I_5 ℵ$I_1 \wedge I_2$

Confidence=sc$\{I_1,I_2,I_5\}/\{I_5\}$=2/2= 100%........................**Eq(10)**

R6 is selected.

Three strong association rules were founded by the above way.

Conclusion:

Different three strong association rules are generated with the data set by applying Apriori algorithm. From the study it reveals 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.

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.

Acknowledgements:

Sincere thanks to management, Principal and faculty members of our institute (Sagi Ramakrishnam Raju Engineering College)

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]. N K Kameswara Rao, Dr. G P S Varma, "Classification Rules Using Decision Tree for Dengue Disease", , International Journal of Research in Computer and Communication Technology vol-3, Issue-3, March 2014, PP-340-343.

[8]. N K Kameswara Rao, Dr. G P S Varma, "Knowledge Discovery from Realtime Data using Data Mining", IJRSAT April 2014.

[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.

Authors:

N K Kameswara Rao did M Tech in Software Engineering from JNT University-Hyderabad. Presently he is working as Associate Professor in Information Technology Department, SRKREC – Bhimavaram.

Dr. G P. Saradhi Varma did Ph. D in Computer Science & Systems Engineering from Andhra University-Visakhapatnam. Presently he is working as Director PG Courses, Professor and HOD in Information Technology Department, SRKREC – Bhimavaram. He Guided 15 Ph D. scholars. He published 26 National Journals, 37 International Journals and 6 books.