International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-2 Issue-4 June-2012

An Improved Single and Multiple Association Approach for Mining Medical Databases

Sachin Sohra[1], Narendra Rathod[2]
M.Tech Scholar, Computer Science, Shri Vaishnav Institute of Technogy & Science, Indore[1]
Assistant Professor, Computer Science, Shri Vaishnav Institute of Technogy & Science, Indore[2]

Abstract

The main aim of data mining is to extract useful patterns from huge amount of data. For this purpose some effective techniques like Apriori algorithm is presented. The major drawback of Apriori algorithm is Generate huge candidate sets: 10^4 frequent 1-itemset will generate 10^7 candidate 2-itemsets.To discover a frequent pattern of size 100, e.g., {a1, a2, …, a100}, one needs to generate 2^{100} which is approx. 10^{30} candidates. Candidate Test incurs multiple scans of database: each candidate. To remove the above drawback, we present an improved non candidate single and multiple association approach for mining medical databases. The developed approach generates association rules for determining the relationships among the diseases observed synchronously. The generated association rules are too significant for making early diagnosis for the correlated diseases. Some types of diseases can have triggering effects on different kinds of diseases. The symptoms and diseases which have stronger effect on each other can be determined and interpreted by the constructed system and the large and extended databases can be scanned effectively with the pruning property of the developed system.

Keywords

Data Mining, Apriori, Association Rule, Medical Diagnosis

1. Introduction

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. It has been estimated that an acute care hospital may generate five terabytes of data a year [1]. The ability to use these data to extract useful information for quality healthcare is crucial. Medical informatics plays a very important role in the use of clinical data. In such discoveries pattern recognition is important for the diagnosis of new diseases and the study of different patterns found when classification of data takes place. It is known that "Discovery of HIV infection and Hepatitis type C were inspired by analysis of clinical courses unexpected by experts on immunology and hepatology, respectively" [2].

This lead to the use of data mining in medical informatics, the database that is found in the hospitals, namely, the hospital information systems (HIS) containing massive amounts of information which includes patients information, data from laboratories which keeps on growing year after year. With the help of data mining methods, useful patterns of information can be found within the data, which will be utilized for further research and evaluation of reports. The other question that arises is how to classify or group this massive amount of data. Automatic classification is done based on similarities present in the data. The automatic classification technique is only proven fruitful if the conclusion that is drawn by the automatic classifier is acceptable to the clinician or the end user.

Data mining explores the hidden relationships and secret knowledge that cannot be observed and evaluated by the human beings easily and improves the quality of our life's by helping the experts showing the secret relationships and correlations in the large databases [3, 4].

Data mining has been played an important role in the intelligent medical systems [5, 6]. The relationships of disorders and the real causes of the disorders and the effects of symptoms that are spontaneously seen in patients can be evaluated by the users via the constructed software easily. Large databases can be applied as the input data to the software by using the extendibility of the software. The effects of relationships that have not been evaluated adequately have been explored and the relationships of hidden knowledge laid among the large medical databases have been searched in this study by means of finding frequent items using candidate generation. The sets

of sicknesses simultaneously seen in the medical databases can be reduced by using our non-candidate approach.

The remaining of this paper is organized as follows. We discuss Association Rule Mining in Section 2. In Section 3 we discuss about preparation and stages. In section 4 we discuss about Recent Scenario. In section 5 we discuss about the proposed approach. Conclusions are given in Section 6. Finally references are given.

2. Association Rule Mining

Association Rules
The process of finding association rules is also known as market basket analysis. This is the case since it is widely applied in retail stores; however it might as well have applications in many other cases.
Association Rule Mining finds the set of all subsets of items/attributes that frequently occur in database records /transactions and extracts rules on how a subset of items influences the presence of another subset.

Association Rule can be expressed as:
$A \Rightarrow B$ [S, C] where A and B are sets of items; S is the support of the rules, defined as the rate of the transactions containing all items in A and all items in B i.e. Support $(A \Rightarrow B) = P (A \cup B)$ and C is the confidence of the rule, defined as the ratio of S with the rate of transactions containing A i.e. P (B / A). Support and confidence are measures of the interestingness of the rule. A high level of support indicates that the rule is frequent enough for a business to be interested in it. A high level of confidence shows that the rule is true often to justify a decision based on it. Minimum support / confidence required for a rule to be reported is its threshold value.

One of the main attributes needed in an Association Rule-mining algorithm is Scalability, the ability to handle massive data stores. As a result fast and efficient Association rules are required to handle increasing number of transactions in real world databases. These rules are able to discover related items occurring together in the same transaction. Since these transaction databases contain extremely large amount of data, current Association Rule discovery techniques try to prune the search space according to the support for the items under consideration [7] and [8].

In the application domain, items correspond to web resources, while transactions correspond to user sessions. Thereby, a rule such as res1 ⇒ res2; mean that if res1 appears in a user session, res2 is expected to appear in the same session, though possibly in reverse order and not consecutively. The basic algorithm called Apriori Algorithm for finding the association rules was proposed and later modified uses Breadth First Search, Bottom Up Approach and performs well when the Frequent Items are short and thus it is easy to implement when the support required is high as it leads to a smaller number of frequent items. But for larger number of frequent items it generates huge set of candidate items leading to high memory requirement and more searching time. Our work therefore focuses on developing an efficient algorithm of mining frequent item sets for association rules with Procreation count .We modify Apriori by using Procreation Count of Frequent Item sets at a level, which is related to Support Count of Candidate Item sets at a next level. Our modification leads to reduction of total number of Candidate Item sets by reducing number of rows in a transaction database that reduces the Cardinality of candidate item sets to improve efficiency of finding frequent item sets. As an example of an association rule, we can think of the case of a super market. An association rule might, then, be in the following form: 'If a customer buys pork steaks, he buys at least two bottles of Coke as well'. In other words, it is in the form X => Y, where X and Y are items or set of items from the super market's database. In the case of City University, an example of an association rule might be that if a student wishes to do business studies, then with a probability of 90% chooses City.

The problem of identifying association rules was first introduced in (Agrawal, 1993). In (Hipp and Guntzer and Nakhaeizadeh, 2000) the formal description of the problem is given as follows: "Let ;={x1,...,xn} be a set of distinct literals, called items. A set X⊆X; with k=|X| is called a k-item set or simply an item set. Let a database D be a multi-set of subsets. Each T∈ D is called a transaction. We say that a transaction T∈ D supports an item set X⊆X ; if X ⊆T holds. An association rule is an expression X =>Y, where X,Y are item sets and X ∩Y = Ø holds. The fraction of transactions T supporting an item set X with respect to database ' is called the support of X, supp(X)= |{T∈D | X ⊆T}| / |D|. The support of a rule X =>Y is defined as supp(X =>Y) = supp(X UY). The confidence of this rule is defined as conf(X =>Y) = supp(X U Y)/ supp(X)". The latter implies that we

are looking at the fraction of transactions that contain the X item set to see how many contain the Y item set as well. That is, while the support of the rule examines the fraction of item sets that contain both X and Y within the database, the confidence of the rule deals with the proportion of item sets that contain Y with respect to those that already contain X [9].

Association Rules and Frequent Itemsets
The market-basket problem assumes we have some large number of items, e.g., \bread," \milk." Customers will their market baskets with some subset of the items, and we get to know what items people buy together, even if we don't know who they are. Marketers use this information to position items, and control the way a typical customer traverses the store.

In addition to the marketing application, the same sort of question has the following uses:
1. Baskets = documents; items = words. Words appearing frequently together in documents may represent phrases or linked concepts. It can be used for intelligence gathering.
2. Baskets = sentences, items = documents. Two documents with many of the same sentences could represent plagiarism or mirror sites on the Web.

Goals for Market-Basket Mining
1. Association rules are statements of the form (X1; X2; …….;Xn) =>Y , meaning that if we find all of X1; X2;……..; Xn in the market basket, then we have a good chance of finding Y . The probability of finding Y for us to accept this rule is called the confidence of the rule. We normally would search only for rules that had confidence above a certain threshold. We may also ask that the confidence be significantly higher than it would be if items were placed at random into baskets.

2. Causality. Ideally, we would like to know that in an association rule the presence of X1;……..;Xn actually \causes" Y to be bought. However, \causality" is an elusive concept. For market-basket data, the following test suggests what causality means. If we lower the price of diapers and raise the price of beer, we can lure diaper buyers, who are more likely to pick up beer while in the store, thus covering our losses on the diapers. That strategy works because \diapers causes beer." However, working it the other way round, running a sale on beer and raising the price of diapers, will not result in beer buyers buying diapers in any great numbers, and lose money.

3. Frequent itemsets. In many (but not all) situations, we only care about association rules or causalities involving sets of items that appear frequently in baskets. For example, we cannot run a good marketing strategy involving items that no one buys anyway. Thus, much data mining starts with the assumption that we only care about sets of items with high support; i.e., they appear together in many baskets. We then find association rules or causalities only involving a high-support set of items (i.e., {X1;X2….. ;Xn; Y }) must appear in at least a certain percent of the baskets, called the support threshold [10] and [11].

Framework for Frequent Itemset Mining
Use the term frequent item set for a set S that appears in at least fraction s of the baskets," where is some chosen constant, typically 0.01 or 1%. We assume data is too large to fit in main memory. Either it is stored in a RDB, say as a relation Baskets (BID; item) or as a at file of records of the form (BID; item1; item2… item n). When evaluating the running time of algorithms:

→ Count the number of passes through the data. Since the principal cost is often the time it takes to read data from disk, the number of times we need to read each datum is often the best measure of running time of the algorithm. There is a key principle, called monotonicity or the a-priori trick that helps us find frequent itemsets.
→ If a set of items S is frequent (i.e., appears in at least fraction s of the baskets), then every subset of S is also frequent.

3. Preparation and Stages

A general view of the stages involved in DM is shown in Figure 1. The process starts with a clear definition of the problem - stage 1, followed by stage 2, which is the selection process aimed at identifying all the internal and external sources of information and selecting the sub -group of data necessary for the application of Data Mining, to deal with the problem. Stage 3 consists of preparing the data, which includes pre -processing, the activity that involves the most effort.

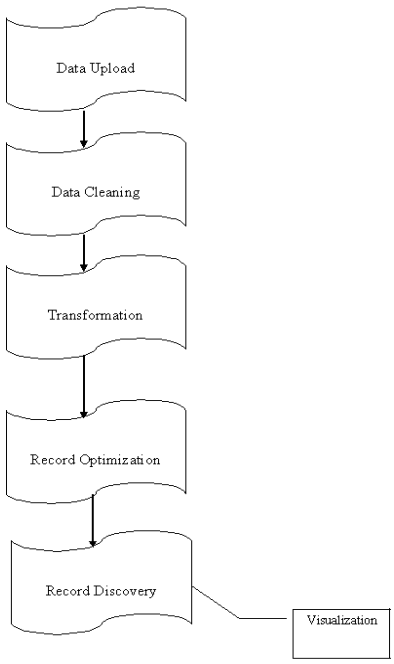

Figure 1: Stages of Data Mining

4. Recent Scenario

In 2011, jinwei Wang et al. [7] proposed to conquer the shortcomings and deficiencies of the existing interpolation technique of missing data, an interpolation technique for missing context data based on Time-Space Relationship and Association Rule Mining (TSRARM) is proposed to perform spatiality and time series analysis on sensor data, which generates strong association rules to interpolate missing data. Finally, the simulation experiment verifies the rationality and efficiency of TSRARM through the acquisition of temperature sensor data.

In 2011, M. Chaudhary et al. [8] proposed new and more optimized algorithm for online rule generation. The advantage of this algorithm is that the graph generated in our algorithm has less edge as compared to the lattice used in the existing algorithm. The Proposed algorithm generates all the essential rules also and no rule is missing. The use of non-redundant association rules help significantly in the reduction of irrelevant noise in the data mining process. This graph theoretic approach, called adjacency lattice is crucial for online mining of data. The adjacency lattice could be stored either in main memory or secondary memory. The idea of adjacency lattice is to pre store a number of large item sets in special format which reduces disc I/O required in performing the query.

In 2011,Fu et al. [9] analyzes Real-time monitoring data mining has been a necessary means of improving operational efficiency, economic safety and fault detection of power plant. Based on the data mining arithmetic of interactive association rules and taken full advantage of the association characteristics of real-time test-spot data during the power steam turbine run, the principle of mining quantificational association rule in parameters is put forward among the real-time monitor data of steam turbine. Through analyzing the practical run results of a certain steam turbine with the data mining method based on the interactive rule, it shows that it can supervise stream turbine run and condition monitoring, and afford model reference and decision-making supporting for the fault diagnose and condition-based maintenance.

In 2011, Xin et al. [10] analyzes that use association rule learning to process statistical data of private economy and analyze the results to improve the quality of statistical data of private economy. Finally the article provides some exploratory comments and suggestions about the application of association rule mining in private economy statistics.

In 2011, K. Zuhtuogullari1 et al. [11] proposed an extendable and improved item set generation approach which has been constructed and developed for mining the relationships of the symptoms and disorders in the medical databases. The algorithm of the developed software finds the frequent illnesses and generates association rules using Apriori algorithm. The developed software can be usable for large medical and health databases for constructing association rules for disorders frequently seen in the patient and determining the correlation of the health disorders and symptoms observed simultaneously.

5. Proposed Approach

In this study an improved Non Candidate approach is used with single and multiple associations on symptoms observed on several diseases. The developed algorithm shows the relationships of the symptoms observed together by generating the itemsets and constructing association rules using the non-candidate generation approach. In the developed approach, the algorithm can be stopped by the user

according to the itemset number (generation) determined by the user in addition to the classical type Apriori itemset generation approaches. The developed approach can finalize itemset generation process before reaching the last itemset, this approach gives the opportunity to construct different association rules according to the user defined itemset number parameter. For this we can consider any of the diseases from Figure 2.

Diseases / Ailments	
AIDS	Gingivitis
Acne	Gum Disease
Allergies	Headaches
Altitude Sickness	Hepatitis
Alzheimer's Disease	Herpes
Anemia	Herpes Simplex
Angina	Herpes Zoster
Arrhythmia	HIV Infection
Arteriosclerosis	Influenza
Arthritis	Insect bites
Asthma	Leg ulcers
Bacterial Infections	Leukemia
Bronchitis	Lupus Erythematosis
Burns	Lymphoma
Cancer	Metastatic Carcinoma
Candidiasis	Migraine headaches
Cardiovascular Disease	Mononucleosis
Cerebral Vascular Disease	Multiple Sclerosis
Cholesterol (High)	Open sores and wounds
Chronic Pain	Parasitic infections
Cirrhosis of the liver	Parkinson's Disease
Cluster headaches	Periodontal Disease
Colitis	Proctitis
COPD	Prostatitis
Cystitis	Rheumatoid Arthritis
Diabetes Type II	Shingles
Diabetic Gangrene	Sinusitis
Diabetic Retinopathy	Sore Throat
Digestion Problems	Temporal Arteritis
Eczema	Trichomoniasis
Emphysema	Ulcers
Epstein-Barr infection	Vascular Diseases
Food allergies	Vascular headaches
Fungal infections	Viral infections
Fungus	Warts
Gangrene	Yeast infection

Figure 2: Diseases List

Algorithm:

Output : Frquent Itemset in the database

1. Scan data and and find support for each item.
2. Discard infrequent items.
3. Sort frequent items in increasing\decreasing order based on their support.
4. For any itemset X in R, if X's number of occurrences in
R is not equal to K * (K-l) / 2, then X will not frequent. R
represent k-itemset which is generated by joining any two
itemsets in Lk-1 .
5. Procedure: apriori_gen change(Lk_1)
6.Input: Lk-1
7.for each itemset 11 is in Lk-1 {
8.for each itemset 12 {
9.C=ll join 12 when the Item Set is Frequent otherwise skip
10.if (length (c>k) then
11.continue;
12.Else {
13.if (exists (RESULT,c)) then
14.count + +;
15.Else {
16.add c to RESULT;
17.c.count = 1
18.F or each itemset which is greater than min-support
20.add to the list;
21.return the list;

For Better understanding of the algorithm we consider mouth cancer disease. We taken five symptoms of mouth cancer leukoplakia,erythroplakia,lump on the lip,Speech problems and Weight loss. We symbolize the symptoms in A1,A2,A3,A4 and A5. Then we taken a sample database of Table 1, where T1,T2….T12 are the person who suffers from the disease. If the symptom value is 1 means the symptom is present in the patient. The symptoms and the transaction are shown in table 1. The transactions are belongs to patient.

Table 1: Symptoms and patient Transcations

T	A1	A2	A3	A4	A5
T1	1	1	0	0	1
T2	0	1	0	1	0
T3	0	1	1	0	0
T4	1	1	0	1	0
T5	1	0	1	0	0
T6	0	1	1	0	0

T7	1	0	1	0	0
T8	1	1	1	0	1
T9	1	1	1	0	0
T10	0	0	0	1	1
T11	1	1	1	1	0
T12	1	0	1	0	0

If the Minimum Support Count is 3 then we only get those symptoms which are found in at least three patients. [Table 2]

Table 2: Support Count : 3

A1 : 8
A2 : 8
A3 :8
A4 : 4
A5 : 3

We can calculate the confidence by the below formula: Confidence (A→B) = P (B/A) = P (AUB) / P (A)
Association:
T1: A1, A2, A5
Then we find the association, based on the above rule which is shown in Table 3.

Table 3: Confidence

A1	⟶ A2	5/8	62.5%
A1	⟶ A5	2/8	25%
A2	⟶ A5	2/8	25%
A2	⟶ A1	5/8	62.5%
A5	⟶ A1	2/3	66.6%
A5	⟶ A2	2/3	66.6%

Then we check the accepted and not accepted combination which is shown in table 4.

Table 4: Accepted and Not Accepted Symptoms

A1,A2 : 5	
A1,A5 : 2	Not Accepted
A2,A5 : 2	Not Accepted

Table 5: Accepted Symptoms

A1,A2	⟶ A5	2/5	40 %

The above symptom is found in the patient according to the minimum support.
Medium Stage: Disease in the stage of 60%.
More Spreading Symptom is A1 and A2.

Table 6: According to Support Count: 4

A1 : 8
A2 : 8
A3 :8
A4 : 4
A5 : 3 Not Accepted

T1: A1, A2

Table 7: Confidence for Support Count: 4

A1	⟶ A2	5/8	62.5%
A2	⟶ A1	5/8	62.5%

Two Symptoms are found. By the above algorithm we successfully find the nature of the symptoms.

6. Conclusion

In addition to the classical approaches, the constructed approach can calculate the association rules from the desired item set number and this specification gives the system the opportunity to generate different association rules. This approach is useful for finding the symptoms for health disorders and also the percentage.

References

[1] Huang, H. et al. "Business rule extraction from legacy code", Proceedings of 20th International Conference on Computer Software and Applications, IEEE COMPSAC'96, 1996, pp.162-167.

[2] Anthony S. Fauci, et al 1997. "Harrison's Principles of Internal Medicine ed. New York": McGraw-Hill.

[3] A. Sadanandam, M. L. Varney, R. K. Singh, "Identification of Semaphorin A Interacting Protein by Applying Apriori Knowledge and Peptide Complementarity Related to Protein Evolution and Structure Genomics", Proteomics & Bioinformatics, Volume 6, Issues 3-4, 2008,pp. 163-174.

[4] Lazcorreta, Enrique, Federico Botella, and Antonio Fernández-Caballero. "Towards personalized recommendation by two-step modified Apriori data mining algorithm." Expert Systems with Applications 35, no. 3 (2008): 1422-1429.

[5] C. Aflori, M. Craus, "Grid implementation of the Apriori algorithm Advances in Engineering Software, Volume 38, Issue 5, May 2007, pp. 295-300.

[6] A. J.T. Lee, Y.H. Liu, H.Mu Tsai, H.-Hui Lin, H-W. Wu, "Mining frequent patterns in image databases with 9D-SPA representation",Journal of Systems and Software, Volume 82, Issue 4, April 2009, pp.603-618.

[7] Jinwei Wang and Haitao Li ," An Interpolation Approach for Missing Context Data Based on the Time-Space Relationship and Association Rule Mining " ,Multimedia Information Networking and Security (MINES), 2011,IEEE.

[8] Chaudhary, M. ,Rana, A. , Dubey, G," Online Mining of data to generate association rule mining in large databases ", Recent Trends in Information Systems (ReTIS), 2011 International Conference on Dec. 2011,IEEE.

[9] Fu Jun ,Yuan Wen-hua, Tang Wei-xin ,Peng Yu,"study on Monitoring Data Mining of Steam Turbine Based on Interactive Association Rules ",IEEE 2011, Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM).

[10] Jinguo, Xin; Tingting, Wei, "The application of association rules mining in data processing of private economy statistics", E-Business and E-Government (ICEE), 2011 IEEE.

[11] K. Zuhtuogullari, N. Allahverdi, "An Improved Itemset Generation Approach for Mining Medical Databases", IEEE 2011.

Bachelor of Engineering from Institue of Engineering & Technology ,DAVV in Information Technology with first division Pursuing ME (Computer Science) from Shri Vaishnav institute of Technology and science Indore.