International Journal of Computing Science and Information Technology, **2015, Vol.*, 14-19**
ISSN: 2278-9669, April 2015 (http://ijcsit.org)

A Pattern Categorization Based Association Mining for Medical Databases

Mohit Sachan[1], Priyanka Dhasal[2]
[1]Mtech Scholar
[2] Assistant Professor
PCST Indore
E-mail:- [1]mohitsachan91@gmail.com

Abstract

Data mining is an effective method to find frequent patterns. In this paper we proposed, pattern categorization based Association mining for medical databases. Cancer data symptoms are consider for experimental purpose. We are using association rule for identifying valid and potentially useful patterns of symptoms for the medical database. Applying association rule we can generate individual and combine support of the symptom. Based on the support we find most frequent symptom in any patient after applying the association on 3 -30 days observation. Then we apply categorization so that we find the exact position of the symptom and calculate the diseases support. By considering the disease support we predict the disease. According to our experimental results we can achieve better prediction.

Keywords: Data Mining, Apriori, Association Rule, Medical Diagnosis

1. INTRODUCTION

The last few years Data Mining has become more and more popular. Together with the information age, the digital revolution made it necessary to use some heuristics to be able to analyze the large amount of data that has become available. Data Mining has especially become popular in the fields of forensic science, fraud analysis and healthcare, for it reduces costs in time and money. In this paper we use data mining to emphasize to discover knowledge that is not only accurate, but also comprehensible for the user [1], [2], [3]. According to Yang Jianxiong et al. [4] Comprehensibility is important whenever discovered knowledge will be used for supporting a human decision. After all, if discovered knowledge is not comprehensible for a user, it will not be possible to interpret and validate the knowledge. The Healthcare industry is among
the most information intensive industries. Medical information, knowledge and data keep growing on a dailybasis.

It has been estimated that an acute care hospital may generate five terabytes of data a year [5]. 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” [6].
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 lifes by helping the experts showing the s ecret relationships and correlations in the large databases [7][8].
Data mining has been played an important role in the intelligent medical systems[9][10]. 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. Recent Scenario

In 2011, M. Chaudhary et al. [11] 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 rulesalso 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. [12] 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. [13] analyzes that use association rule learning to Process statistical data ofprivate 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. [14] proposed an extendable and improved itemset 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.

In 2012, AshutoshDubey et al. [16] proposes an efficient method for knowledge discovery which is bas ed on subset and superset approach. In this approach we also use dynamic minimum support so that we reduce the execution time. A frequent superset means it contains more transactions then the minimum support. It utilize the concept that if the item set is not frequent but the superset may be frequent which is. It utilizes the behavior that the less count may be frequent if we attached the less count with the higher order set.

Here we also provide the flexibility to find multiple minimum supports which is useful for comparison with associated items and dynamic support range. Our algorithm provides the flexibility for improved association and dynamic support. Comparative result shows the effectiveness of our algorithm.

In 2011, AshutoshDubey et al. [17] proposed a novel algorithm named Wireless Heterogeneous Data Mining (WHDM).

The entire system architecture consists of three phases:

1) Reading the Database.

2) Stores the value in Tbuf with different patterns.

3) Add the superset in the list and remove the related subset from the list.

Finally they find the frequent pattern patterns or knowledge from huge amount of data. They also analyze the better method or rule of data mining services which is more suitable for mobile devices.

In 2012, DevashriRaich et al. [18] introduce various intelligent computing techniques used for the medical diagnosis of diseases and a brief description about Nephritis and how its diagnosis could be done.

They suggested that computer science is getting more and more involved in medicines and health services. Various AI techniques and soft computing techniques are used for the diagnosis of particular diseases for the betterment of patient health. Various clinical decision support systems are also been devised by the help of AI.

In 2015, MohitSachan et al. [15] proposed a study on data mining. They suggest that 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 and focus on the drawback.

To remove the above drawback, they 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.

3. Proposed Approach

In this study an improved association approach is used with single and multiple associations on symptoms observed on cancer diseases. The developed algorithm shows the relationships of the symptoms observed together by generating the item sets and constructing association rules using the frequent generation approach. The algorithm for the proposed concept in finding significant patterns for cancer prediction is presented in this section. In this framework administrator add the database, admin first add the disease and then the symptoms of the disease. According to the added value the parameter for that particular cancer is decided. This aim of our research is to apply the Association Rule Mining algorithm on patient symptoms for an efficient detection of Medical databases . In this we took cancer symptoms as medical database.

Figure 1: Flowchart of Association with Pattern Categorization

The flowchart of figure 1 shows the actual phenomena of our work. In this paper we took cancer symptoms as the cancer database. User selects the symptoms as he/she observed as shown in Figure 2. User enters the observation as per the days. of observation selected which is shown in figure 3. User enters 1 if he/she observe the symptom otherwise 0 as shown in figure 4. This can be converted in the tabular form as shown in figure 5.

Figure 2: Symptoms

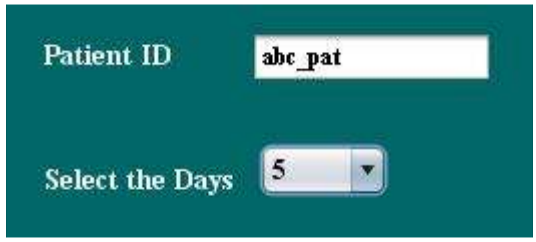

Figure 3: Observation Period

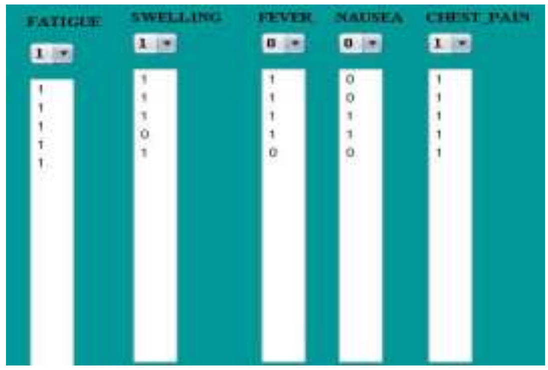

Figure 4: Observed Symptoms by the Patient

TNo	FATIGUE	SWELLING	FEVER	NAUSEA	CHEST_...
D1	1	1	1	0	1
D2	1	1	1	0	1
D3	1	1	1	1	1
D4	1	0	1	1	1
D5	1	1	0	0	1

Figure 5: Observed Symptoms

Initialize: K: = 1, C1 = all the 1- item sets;

Step 1:C1 to determine L1.

Step 2:L1:= {frequent 1- item sets}; k:=2; //k represents the pass number// Step 3: while (Lk-1 ≠ Æ) do

begin

Step 4: Ck: = gen_candidate_itemsets with the given Lk-1 Step 5:Prune (Ck)

Step 6:for all candidates in Ck do

count the number of transactions by using intersect method that are common in each item Î Ck

Step 7: Lk := All candidates in Ck with minimum support ; k := k + 1;

end

Step 8: Frequent pattern = ÈkLk ;

Step 9: General strategy: for each set find rule set that covers all instances in it (excluding instances not in the class). This approach is called a categorization approach because at each stage a rule is identified that covers some of the instances.

Step 10: General to specific rule induction : For each class C

Initialize E to the instance set

While E contains instances in class C

Create a rule R with an empty left-hand side that predicts class C

While R covers instances from classes other than C do:

For each attribute A not mentioned in R, and each value v,

Select A and v to maximize the accuracy Add (A = v) to R

Remove the instances covered by R from E Step 11: Specific to general rule induction:

Pick up an instance and generalize it by repeatedly dropping conditions. Stop when all further generalizations lead to covering instances from other classes. Save the generalized instance as a rule.

Remove all instances covered by R and continue until all instances are covered.

When dropping conditions choose the ones that maximize rule coverage.

Problems: rule overlapping, rule subsumption.

Step 12: Then we find support based on the category , that is sup(x) of an itemset x is defined as the proportion of transactions in the data set which contain the item set.

Step 13: Discovery the percentage in the medical data set Step 14: End

Figure 6: Algorithm for Association with Pattern Categorization

After the observed symptoms as obtained in figure 5. We apply our algorithm that is association with pattern categorization[Figure 6]. First we find the category of the individual as well as in the combination as shown in Figure 7, Figure 8, Figure 9 ,Figure 10 and figure 11. So that we achive the frequent pattern list according to the support as shown in Figure 12.

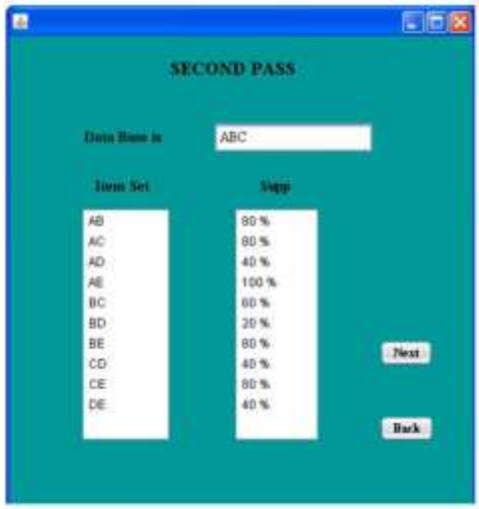

Figure 8: Cancer Detection Pass2

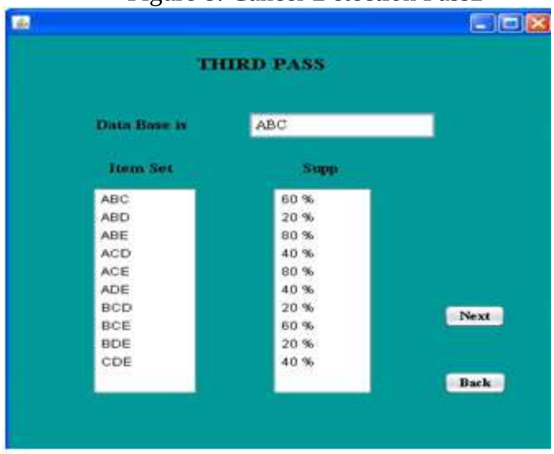

Figure 9: Cancer Detection Pass3

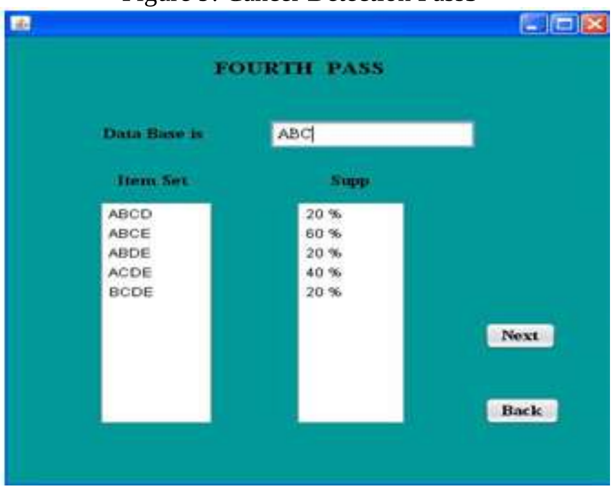

Figure 10: Cancer Detection Pass4

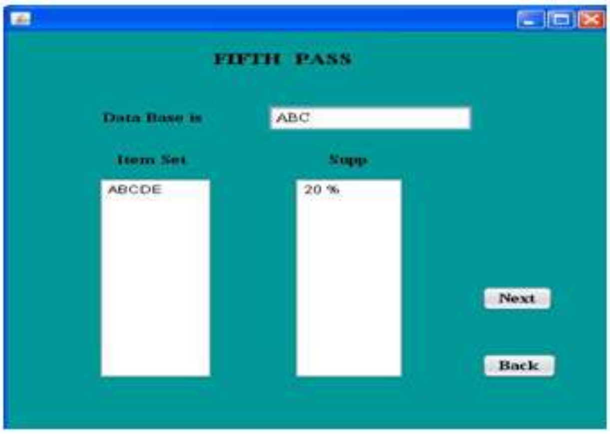

Figure 11: Cancer Detection Pass5

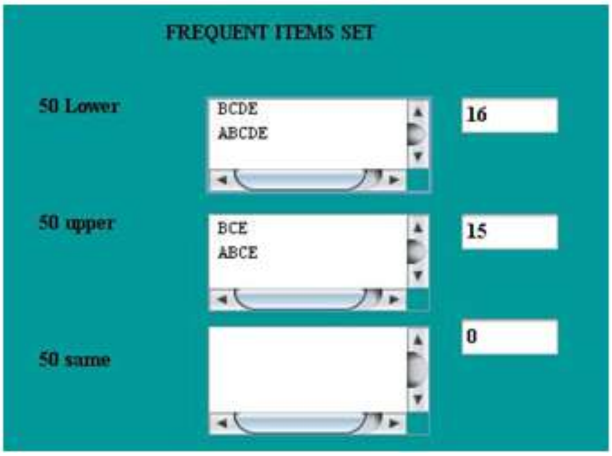

Figure 12: Frequent Item Set

4. Result analysis

Then we apply categorization accoring to three diseases having highest support. In this we catgorize by creating these disease as a seprateclass.Then we generatel to specific rule induction For each class C. We Initialize E to the instance set While E contains instances in class C. Then we create a rule R with an empty left-hand side that predicts class C.While R covers instances from classes other than C then iterate it.For each attribute A not mentioned in R, and each value v. Select A and v to maximize the accuracy Add (A = v) to R

Remove the instances covered by R from E. Then we Pick up an instance and generalize it by repeatedly dropping conditions. Stop when all further generalizations lead to covering instances from other classes. Save the generalized instance as a rule. We also remove all instances covered by R and continue until all instances are covered. When dropping conditions choose the ones that maximize rule coverage.

Then we find support based on the category , that is sup(x) of an itemset x is defined as the proportion of transactions in the data set which contain the item set.

Figure 13: Disease Wise List

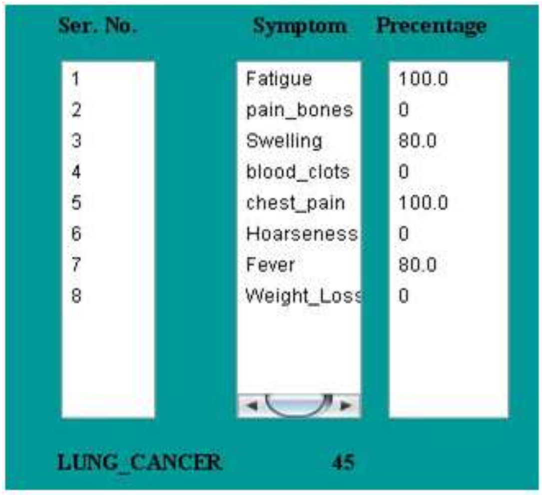

Figure 14: Chances of Cancer detection

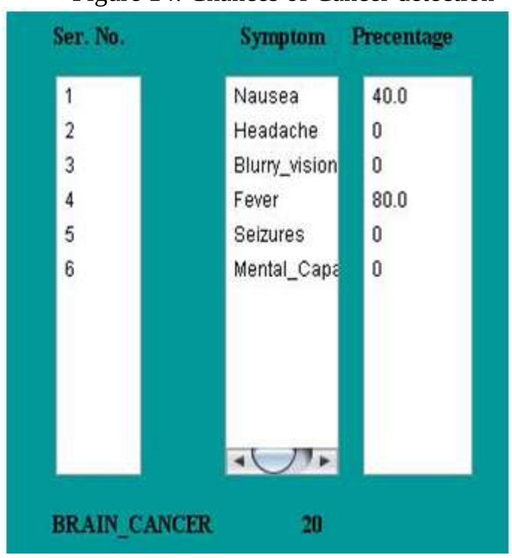

Figure 15: Chances of Cancer detection

As shown in figure 14 and figure 15 we find the support weight as generated from the association. And according to the weight we then calculate support based on the symptom because it is dynamic in naturre. And we achive the cahnce of the symptom as shown in the figure 14 and figure 15. According to the final detection we show the result as shown in figure 16. As per our observation our detection is better , because we comprise all individual items.

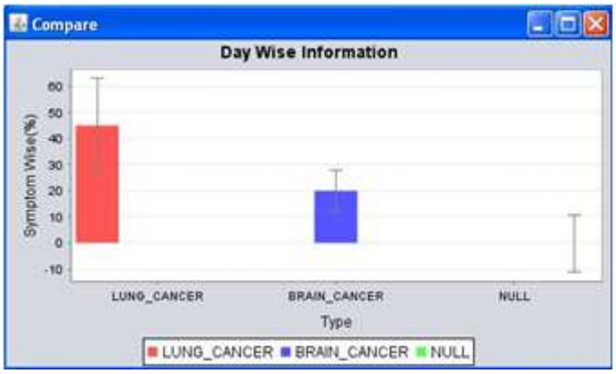

Figure 16: Cancer Chart

5. 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. In this paper we find the better association on medical databases. For experimental analysis we taken the cases of cancer symptoms. Based on our observations and result shown in the result analysis we provide better prediction of medical databases.

References

[1] M. Dorigo, G. Di Caro, and L. M. Gambardella, "Ant algorithms for discrete optimization," Artif. Life, vol. 5, no. 2, pp. 137–172,1999.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," in Advances in Know Discovery & Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth,and R. Uthurusamy, Eds. Cambridge, MA: MIT Press, 1996, pp. 1–34.

[3] JunzoWatada, Keisuke Aoki, Masahiro Kawano, Muhammad SuzuriHitam, Dual Scaling Approach to Data M Journal of Advanced Computational Intelligence Intelligent Informatics (JACIII), Vol. 10, No. 4, pp. 441-447, 2006.12.

[4] Yang Jianxiong and JunzoWatada, "Wise Mining Method through Ant Colony Optimization", Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA - October 2009.

[5] 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.

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

[7] 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.

[8] E. Lazcorreta, F. Botella, A. Fernández-Caballero, "Towards personalized recommendation by two -step modified Apriori data mining algorithm" Expert Systems with Applications, Volume 35, Issue 3, October 2008, pp. 1422-1429

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

[10] 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.

[11] 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.

[12] Fu Jun ,Yuan Wen-hua, Tang Wei-xin ,PengYu,"study on Monitoring Data Mining of Steam Turbine Based on Interactive Association Rules ",IEEE 2011, Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM). 13] 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.

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

[15] Ashutosh Kumar Dubey, Animesh Kumar Dubey, Vipul Agarwal, YogeshverKhandagre, "Knowledge Discovery with a Subset-Superset Approach for Mining Heterogeneous Data with Dynamic Support",Conseg-2012.

[16] Ashutosh Kumar Dubey, SmritiPandey, Nitesh Gupta," A Novel Wireless Heterogeneous Data Mining (WHDM) Environment Based on Mobile Computing Environments",International Conference on Communication Systems and Network Technologies, CSNT 2011.

[17] DevashriRaich,P.S.Kulkarni, "Application of Intelligent Computing Techniques for the Interpretation and Analysis of Biological and Medical Data for Various Disease diagnosis: Review", International Journal of Advanced Computer Research (IJACR) ,Volume-2 Number-4 Issue-6 December-2012.