ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

Finding Locally Frequent Diseases Using Modified Apriori Algorithm

Mohammed Abdul Khaleel[1], Sateesh Kumar Pradhan[2], G.N.Dash[3]

Research Scholar, Sambalpur University, India[1]

Post Graduate Department of Computer Science, Utkal University, India[2]

Post Graduate Department of Physics, Sambalpur University, India[3]

Abstract: Data mining is a phenomenon which analyzes large volumes of data and extracts patterns that can be converted to useful knowledge. The data mining techniques can be applied on medical data which has abundant scope to improve Quality of Service in Healthcare industry. Electronic health records and other historical medical data available in textual and graphical formats are a gold mine to researchers in the field. Medical data mining techniques analyze latent medical attributes and the relationships among them to bring about expert decisions in curing diseases. Data mining techniques can also be used toknow the frequently occurring diseases in the local databases. In this paper we present a methodology to discover locally frequent diseases with the help of Apriori data mining technique. We also used visualization techniques to present the trends graphically. We built a prototype application that demonstrates the efficiency of the method. The empirical results revealed that the prototype is useful and can be used in real world Healthcare tools.

Index Terms: Data mining techniques, frequent item sets, medical data mining, locally frequent patterns, Apriori

I. INTRODUCTION

Analyzing medical data has significant utility in modern times. Large volumes of data are being accumulated data by day in healthcare domain. By analyzing this data latten patterns can be discovered. Data mining techniques help to analyze and establish hidden relationships among the attributes of medical data. This sort of knowledge discovery is very useful in many real time applications including disease detection, diagnosis, disease classification, predicting breast cancer survivability, finding whether back surgery fails or succeed, decision support systems in clinical applications. Many algorithms came into existence for general data mining and also for medical data mining. Classification accuracy of diabetic records is explored in [1] using SVM. For examination of laboratory records trajectory mining was explored in [3] using cluster analysis method on hepatitis dataset. In [4], [5], [6], [7], [8], [9], and [10] experiments were made on medical data mining. Their methods have revealed improvement in healthcare maintenance. For high quality information retrieval from medical data sources is made through KDQL (Knowledge Discovery Question Language) [11]. Multi-stage medical diagnosis is explored in [12] using diagnostic taxonomy and rules. A medical diagnostic system was proposed in [15] for effective discovery of knowledge. Heart disease prediction is explored in [17].

The research which is close to our work is in [16] where frequent itemssets were found from medical data sources. The difference between [16] and our approach is that we used pre-processing to associate numerals to all discrete values and we made necessary modifications to the generic Apriori algorithm. Our contributions in this paper are as described below.

- Proposing a modified Apriori algorithm for finding locally frequent diseases. The algorithm also includes pre-processing which associates numeric values with discrete values so as to improve the performance of algorithm.
- The results of our algorithm can help organizations, governments to make well informed decisions with respect to locally frequent diseases. The empirical results revealed that the algorithm is efficient in finding locally frequent diseases. Our visualization graphs help experts in healthcare domain to grasp the facts pertaining to prevailing diseases locally just by a glance.

The remainder of this paper is structured as follows. Section II reviewed literature pertaining to medical data mining, and various prior works that throw light on medical data mining techniques. Section III presents our modified Apriori algorithm which is meant for finding locally frequent diseases. Section IV presents experimental results that reflect locally frequent diseases from various viewpoints while section V concludes the paper.

II. RELATED WORKS

In healthcare domain, many data mining techniques have been used to discover actionable knowledge from medical datasets. They are used to extract the latent relationships which were not exploited earlier for knowledge discovery [2]. Balakrishnan and Narayanaswamy [1] proposed a feature selection approach for classification accuracy which can classify type II diabetes records. The feature selection is associated with SVM ranking for improving classification accuracy.Tsumoto and Hirano [3] studied trajectory mining techniques in order to analyze laboratory examinations. Multiple trajectories are used and compared one with other trajectory to discover data dynamics. To achieve this they proposed a novel cluster analysis method which was applied on hepatitis dataset and discovered interesting results.

Abe et al. [4] proposed data mining environment for integrated time-series for medical data mining. For extracting useful clinical knowledge from medical databases medical time-series data mining is essential. The environment makes use of inputs from medical experts to obtain quality results. Su et al. [5] studied three data mining techniques for discovering actionable knowledge from medical database. They are Back Propagation Neural Network, C4.5 for decision tree, and Bayesian Network. They used tongue related dataset and cytology dataset for experiments. Out of the three algorithms, the best performance was shown by BPN with 96.0% accuracy. Jiquan et al. [6] studied on heterogeneous medical data and proposed a term-mapping frameworkfor improving medical data mining process. Here the miscall data sources are mapped to various medical terminology for simplifying mining process.Seng et al. [7] studied on medical data mining and sharing of patients' health records securely with respect to telemedicine. They discussed web based data mining techniques for telemedicine. Their methods improved the overall healthcare management.

Abidi and Hoe [8] focused on automatic extraction of symbolic rules from medical data sources. They used a hybrid approach which includes data descretization and data clustering for symbolic rule discovery. To achieve this they proposed a generic workbench for medical data mining. Tsumoto[9] studied on the problems of medical data mining. The problems identified include missing values, data storage issues, coding systems and related problems. Ghannad-Rezaie et al. [10] presented a new medical data mining approach based on advanced swam intelligence. As medical data mining techniques, this approach addresses problems identified in [9] such as missing values, storage and coding problems. They built an application that is meant for surgery candidate selection in temporal lobe epilepsy. Their method has shown highquality in the results.

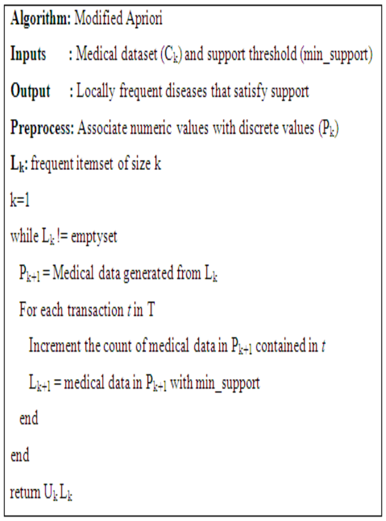
Müller et al. [11]proposed an intelligent data mining technique. They built a query language by name Knowledge Discovery Question Language (KDQL) that is meant for discovery high quality information from medical data sources containing therapeutic and diagnostic measures. Tsumoto [12] presented an approach for multi-stage medical diagnosis through mining of diagnostic rules and diagnostic taxonomy. For grouping medical diseases, they used a measure which has three procedures namely characterization of attributes for decision making, computing similarity between characterization sets, and the concept hierarchy for given classes. Hai Wang and Shouhong Wang [13] studied on medical knowledge acquisition through data mining. They have explored various data mining algorithms for knowledge discovery. They include classification, clustering, association rule mining, regression analysis, sequence mining, and outlier detection.

Khan et al. [14] explored decision tree algorithm for medical data mining. Especially they applied the algorithm for classification of medical images. Expert knowledge discovery is possible with this medical data mining approach which visualized decision tree that can be used in expert decision making systems. Shim and Xu [15] proposed a medical diagnostic system through medical data mining. The system is based on BYY Binary Independent Factor Analysis. They made experiments are diagnosing liver diseases. The results revealed that their system is providing accurate results. Ilayaraja and Meyyappan [16] explored Apriori algorithm for medical data mining. They focused on identifying frequent diseases. The association rule based algorithm could identify frequent diseases. This work is close to our work in this paper. The main difference is that we modified the original Apriori algorithm in order to make it suitable for finding locally frequent diseases. Carlos Ordonex [17] studied the problem of heart disease prediction. This researcher identified constrained association rules for the purpose. The constraints considered include the appearance of attributes only at one side of rule; attribute segregation into uninterested groups, and restriction on number of attributes. With these constraints the number of discovered rules was reducing making the solution more suitable for the prediction of heart diseases.

Antonie et al. [18] studied data mining techniques for tumor detection. The experiments are related to digital mammography. The data mining techniques explored to achieve this are association rule mining and neural networks. With the two algorithms they could achieve classification accuracy up to 70%. Barati et al. [19] applied data mining techniques on skin diseases. They also explored other diseases. For instance association rule mining was found suitable for cancer diagnosis. They have focused on the skin disease classification with respect to dermatology. They used techniques like Genetic algorithms, fuzzyclassification, andneural networks. They found that clustering medical images has very important utility in medical data mining. Bellaachia et al. [20]studied the problem of prediction of survivability of Brest cancer patients. They used three different data mining techniques to discover the survivability probability of breast cancer patients. They include decision tree algorithm such as C4.5, Naïve bayes and back propagation neural network. Out of them they found C4.5 provides better performance. Heart disease prediction system and heart attack prediction system were presented by Subbalakshmi et al. [21] and Deepika et al. [22] respectively. Both the systems work on similar datasets for detectingheart diseases.

III. MODIFIED APROIRI ALGORITHM

The generic Apriori algorithm has been altered in order to find locally frequent diseases through medical data mining. This algorithm is used to find frequent itemsets from which association rules can be generated. The aim of this paper is to find locally frequent diseases from the medical dataset. Frequent itemsets are the set of items that have minimum support in the given dataset. As per Apriori a subject of a frequent itemset should also be a frequent itemset. For instance consider the itemset {A, B} in which both {A} and {B} must be frequent itemsets. The algorithm generates frequent itemsets iteratively for 1 to k. The pseudocode of the algorithm used in this paper is presented in listing 1.



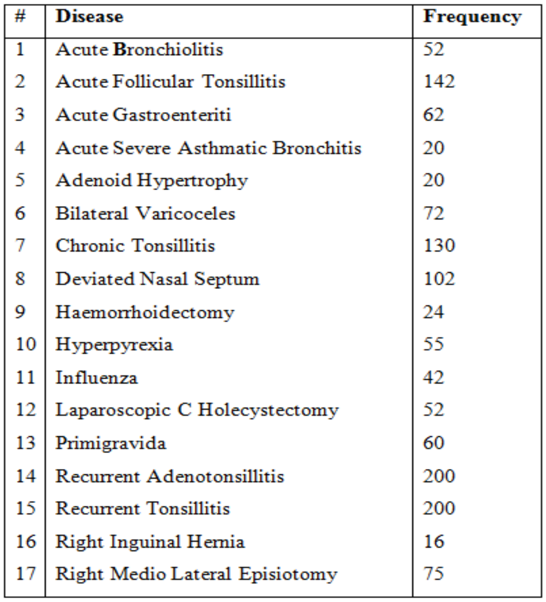
```
**Algorithm:** Modified Apriori
**Inputs**     : Medical dataset ($C_k$) and support threshold (min_support)
**Output**     : Locally frequent diseases that satisfy support
**Preprocess:** Associate numeric values with discrete values ($P_k$)
$L_k$: frequent itemset of size k
k=1
while $L_k$ != emptyset
  $P_{k+1}$ = Medical data generated from $L_k$
  For each transaction *t* in T
    Increment the count of medical data in $P_{k+1}$ contained in *t*
    $L_{k+1}$ = medical data in $P_{k+1}$ with min_support
  end
end
return $U_k L_k$
```

Listing 1 – Modified Apriori algorithm for finding locally frequent diseases

The modified apriori algoritmtakes medical dataaset and minimum support as inputs and generated locallyfrequent diseases from given medical data. Prior to applying the algorithm, the data set is pre-processed to associate numerals to discrete values for easy processing. The algorithm generates statistics pertaining to frequnent itemsets that satisfy minimum support provided.

IV. EXPERIMENTAL RESULTS

We built a prototype application to evaluate the proposed algorithm for finding locally frequent diseases. The application is built in Java platform using Net Beans IDE. The environment used to build and test application includes a PC with 4 GB RAM, core 2 dual processor running Windows 7 operating system. The dataset is collected from Abha Private Hospital in Saudi Arabia. The training dataset contains electronic records of around 2000 patients. The dataset contains attributes like name, address, age, sex, disease, date, symptoms and so on for the year 2012-13. Number of diseases in the dataset are 17. The modified aproiri algorithm is employed to discover locally frequent diseases. The results are presented in table 1. These results reflect all diseases without minimum support value from domain expert.



#	Disease	Frequency
1	Acute Bronchiolitis	52
2	Acute Follicular Tonsillitis	142
3	Acute Gastroenteriti	62
4	Acute Severe Asthmatic Bronchitis	20
5	Adenoid Hypertrophy	20
6	Bilateral Varicoceles	72
7	Chronic Tonsillitis	130
8	Deviated Nasal Septum	102
9	Haemorrhoidectomy	24
10	Hyperpyrexia	55
11	Influenza	42
12	Laparoscopic C Holecystectomy	52
13	Primigravida	60
14	Recurrent Adenotonsillitis	200
15	Recurrent Tonsillitis	200
16	Right Inguinal Hernia	16
17	Right Medio Lateral Episiotomy	75

Table 1 – Experimental results

As seen in table 1, the frequency of occurrence of various diseases over a period of 12 months is presented. The least locally frequent disease is Right Lnguinal Hernia while the most frequent diseases are Recurrent Adenotonsillitis and Recurrent Tonsillitis. Acute Follicular Tonsillitis and Chronic Tonsillitis have 2nd and 3rd most locally frequent diseases. Figure 1 shows the frequencies of all diseases graphically.

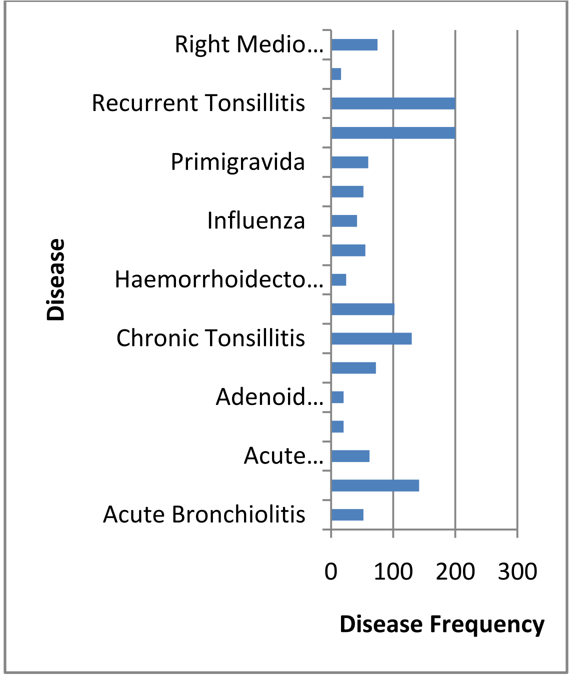


Fig. 1 – Illustrates the trend of occurrence of locally frequent diseases

Figure 1 visualized the frequency of various diseases. The diseases that have highest occurrence include Recurrent Adenotonsillitis and Recurrent Tonsillitis while the least occurred disease is Right Inguinal Hernia. Table2 shows the occurrence of diseases over a period of time based on the support value mentioned by domain expert.

DISEASE	SUPPORT		
	0.2	0.3	0.5
Acute Bronchiolitis	X		
Acute Follicular Tonsillitis	X	X	X
Acute Gastroenteriti	X		
Acute Severe Asthmatic Bronchitis			
Adenoid Hypertrophy			
Bilateral Varicoceles	X	X	
Chronic Tonsillitis	X	X	X
Deviated Nasal Septum	X	X	
Haemorrhoidectomy			
Hyperpyrexia	X		
Influenza	X		
Laparoscopic C Holecystectomy	X		
Primigravida	X		
Recurrent Adenotonsillitis	X	X	X
Recurrent Tonsillitis	X	X	X
Right Inguinal Hernia			
Right Medio Lateral Episiotomy	X	X	

Table 2 – Illustrates locally frequent diseases based on support

Table 2 shows all diseases considered in dataset and occurrence with respect to given support value. As the support value increases, many diseases that do not satisfy the support values are omitted from the results. For this reason, when compared with figure 1, the figure 2 has shown less number of diseases as some of the diseases do not come under the given support. For instance Acute Severe Asthmatic Bronchitis, Adenoid Hypertrophy, Haemorrhoidectomy, and Right Inguinal Hernia do not satisfy support 0.2. In the same fashion, some diseases do not satisfy support value 0.3 and 0.5.

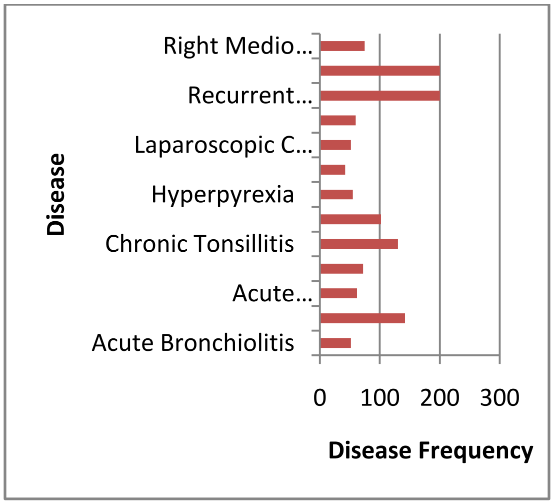


Fig. 2 – Illustrates the trend of occurrence of locally frequent diseases (0.2 support)

Figure 2 visualizes the frequency of various diseases that satisfies the support value 0.2. The diseases that have highest occurrence include Recurrent Adenotonsillitis and Recurrent Tonsillitis while the least occurred disease is Influenza. The diseases that do not have 0.2 support are Acute Severe Asthmatic Bronchitis, Haemorrhoidectomy, Adenoid Hypertrophy and Right Inguinal Hernia.

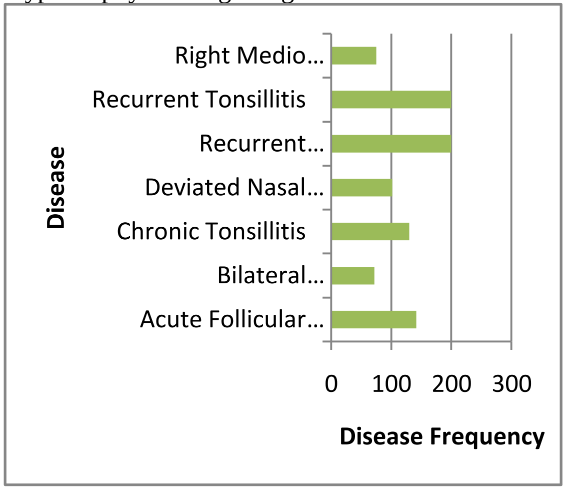


Fig. 3 – Illustrates the trend of occurrence of locally frequent diseases (0.3 support)

Figure 3 visualizes the frequency of various diseases that satisfies the support value 0.3. The diseases that have highest occurrence include Recurrent Adenotonsillitis and Recurrent Tonsillitis while the least occurred disease is Bilateral Varicoceles. The diseases that do not have the given support include Acute Bronchiolitis, Acute Gastroenteriti, Acute Severe Asthmatic Bronchitis, Adenoid Hypertrophy, Haemorrhoidectomy, Hyperpyrexia, Influenza, Laparoscopic C Holecystectomy, Primigravida, and Right Inguinal Hernia.

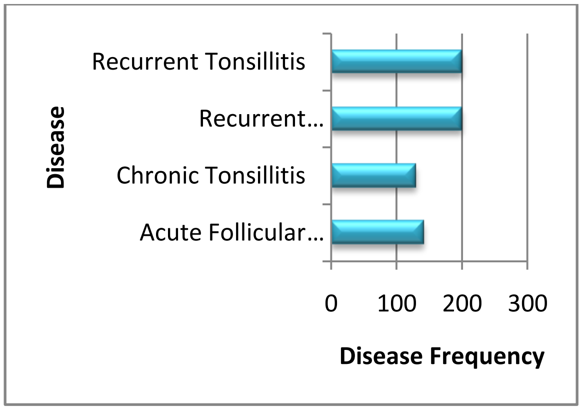


Fig. 4 – Illustrates the trend of occurrence of locally frequent diseases (0.5 support)

Figure 4 visualizes the frequency of various diseases that satisfy the support value 0.5. The diseases that have highest occurrence include Recurrent Adenotonsillitis and Recurrent Tonsillitis while the least occurred disease is Chronic Tonsillitis. The diseases that have the given support include Acute Follicular Tonsillitis, Chronic Tonsillitis, Recurrent Adenotonsillitis, and Recurrent Tonsillitis.

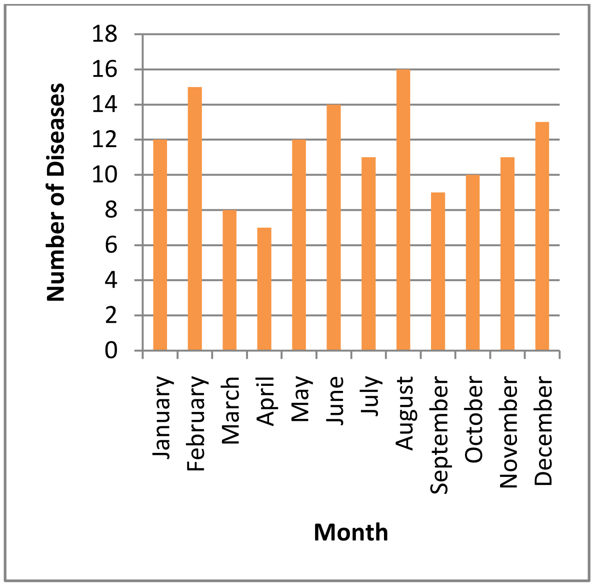


Fig. 6 – Month wise count of occurrence of diseases

Having understood the frequency of diseases, it is useful to know the number of diseases that occur each month. Figure 6 shows the details of the same. As per the results in April there is less number of diseases occurred. In August highest number of diseases occurred. The total number of diseases in the dataset is 17. Their occurrence is spread across the months. On average the percentage of diseases occurred is 11.5.

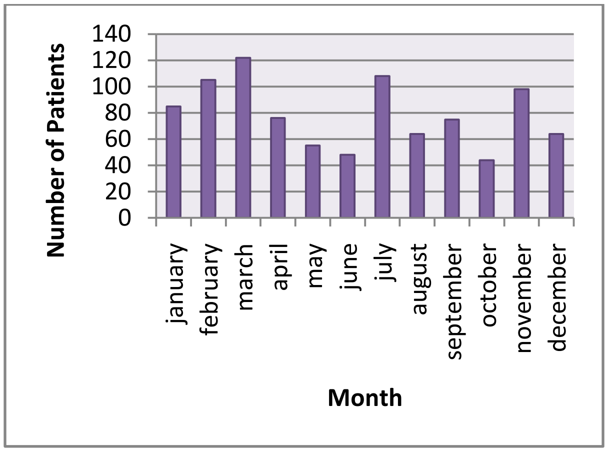


Fig. 7 – Number of patients victimized by the diseases

Having understood the number of disease spread across the months, it is useful to know the number of patients victimized by the diseases in each month. Figure 6 shows the details of the same. As per the results in October there is least number of patients are victimized. In March highest number of patients is victimized. Average number of patients victimizes is 78.6667.

V. CONCLUSIONS AND FUTURE WORK

In this paper we proposed and implemented a modified Apriori algorithm that is meant for discovering locally frequent patterns from medical data sources. The dataset is collected from Abha Private Hospital in Saudi Arabia. Various data mining techniques were used earlier for medical data mining. However, for finding locally frequent disease we thought of adapting Aproiri as it is suitable for discovering frequent patterns. We modified the Apriori algorithm with pre-processing step that makes the algorithm work efficiently. Our algorithm can generate locally frequent diseases and visualize the experimental results in various view points. We built a prototype application to demonstrate the proof of concept. The empirical results reveal that our algorithm has plenty of scope to improve the Quality of Service in healthcare industry. Our work is significant in the context where electronic health records and other historical medical data available in textual and graphical formats are a gold mine to researchers in the field. Our prototype is useful and can be incorporated into real world Healthcare tools.

One future direction we have in mind is working on discovering temporally frequent diseases in near future.

REFERENCES

[1] SarojiniBalakrishnan, RamarajNarayanaswamy, Nickolas Savarimuthu and Rita Samikannu.(2008). SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases. *IEEE*.0 (0), p2628-2633.

[2]Dilly,Ruth.DataMining.2002http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html.

[3] ShusakuTsumoto and Shoji Hirano.(2008). Mining Trajectories of Laboratory Data using Multiscale Matching and Clustering. *IEEE*.0 (0), p626-631.

[4] Hidenao Abe, Hideto Yokoi, Miho Ohsaki and Takahira Yamaguchi. (2008). Developing an Integrated Time-Series Data Mining Environment for Medical Data Mining. *IEEE*. 0 (0), 127-131.

[5] Jenn-Lung Su, Guo-Zhen Wu, I-Pin Chao. (2001). THE APPROACH OF DATA MINING METHODS FOR MEDICAL DATABASE. *IEEE*.0 (0), p3824-3826.

[6] Liu Jiquan Deng WenliangXudong Lu HuilongDuan.(2008). A Term-mapping Framework for Data Mining in Heterogeneous Medical Data Sources. *IEEE*.0 (0), p767-770.

[7] Wong KokSeng. (2006). Collaborative Support for Medical Data Mining in Telemedicine. *IEEE*.0 (0), p1894-1899.

[8] Syed SibteRazaAbidiKokMeng Hoe. (2002). Symbolic Exposition of Medical Data-Sets: A Data Mining Workbench to Inductively Derive Data-Defining Symbolic Rules. *IEEE*.0 (0), p1-6.

[9] ShusakuTsumoto. (2000). Problems with Mining Medical Data. *IEEE*.0 (0), p467-468.

[10] M. Ghannad-Rezaie, H. Soltanain-Zadeh, M.-R.Siadat, K.V. Elisevich. (2006). Medical Data Mining using Particle Swarm Optimization for Temporal Lobe Epilepsy. *IEEE*.16 (21), p761-768.

[11] Oliver Hogl, Michael Müller, Herbert Stoyan, Wolf Stühlinger, Am Weichselgarten and Anichstrasse.(2001). On Supporting Medical Quality with Intelligent Data Mining. *IEEE*.0 (0), p1-10.

[12] ShusakuTsumoto. (2007). Mining Diagnostic Taxonomy and Diagnostic Rules for Multi-Stage Medical Diagnosis from Hospital Clinical Data. *IEEE*.0 (0), p611-616.

[13] Hai Wang and Shouhong Wang.(2007). Medical Knowledge Acquisition through Data Mining. *IEEE*.0 (0), p777-780.

[14] Safwan Mahmud Khan, Md. Rafiqul Islam and Morshed U. Chowdhury. (n.d). Medical Image Classification Using an Efficient Data Mining Technique. *IEEE*.0 (0), p1-6.

[15] Jeong-Yon Shim and Lei Xu. (2003). MEDICAL DATA MINING MODEL FOR ORIENTAL MEDICINE VIA BYY BINARY INDEPENDENT FACTOR ANALYSIS.*IEEE*. 0 (0), p717-720.

[17] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules", Seminar Presentation at University of Tokyo, 2004

[18] Maria-Luiza Antonie et al., “Application of Data Mining Techniques for Medical Image Classification”, Proceedings of the second international workshop on multimedia Data Mining (MDM/KDD’2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001.

[19] E. Barati et al., “A Survey on Utilization of Data Mining Approaches for Dermatological (Skin) Diseases Prediction”, Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Health Informatics (JSHI): March Edition, 2011.

[20] AbdelghaniBellaachia and Erhan Guven, “Predicting Breast Cancer Survivability Using Data Mining Techniques**”**

[21] G.Subbalakshmi et al., “Decision Support in Heart Disease Prediction System using Naive Bayes”, Indian Journal of Computer Science and Engineering (IJCSE)

[22] N.DEEPIKA et al., “Association rule for classification of Heart-attack patients”, International Journal of Advanced Engineering Sciences and Technologies, Vol No. 11, Issue No. 2, 253 – 257.