The Application of Apriori Algorithm in Analysis Usage of Medical Diseases

Pulicherla Murendra[1] Dr. K Venkataramana[2]
[1]Student [2]Professor
[1,2]Department of Computer Applications
[1,2]KMM Institute of PG Studies, Tirupati, India

Abstract— The data mining is a process of analyzing a huge data from different perspectives and summarizing it into useful information. The information can be converted into knowledge about historical patterns and future trends. Data mining plays a significant role in the field of information technology. Health care industry today generates large amounts of complex data about patients, hospitals resources, diseases, diagnosis methods, electronic patients records, etc. The data mining techniques are very useful to make medicinal decisions in curing diseases. The healthcare industry collects huge amount of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. In this paper, authors developed a method to identify frequency of diseases in particular geographical area at given time period with the aid of association rule based Apriori data mining technique.

Key words: Frequent Diseases; Data Mining; Medical Data; Association Rule; Apriori Algorithm

I. INTRODUCTION

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The difference between data analysis and data mining is that data analysis is to summarize the history such as analyzing the effectiveness of a marketing campaign, in contrast, data mining focuses on using specific machine learning and statistical models to predict the future and discover the patterns among data With the progress of the technology of information and the need for extracting useful information of business people from dataset, data mining and its techniques is appeared to achieve the above goal. Data mining is the essential process of discovering hidden and interesting patterns from massive amount of data where data is stored in data warehouse, OLAP (on line analytical process), databases and other repositories of information. This data may reach to more than terabytes. Data mining is also called (KDD) knowledge discovery in databases, and it includes an integration of techniques from many disciplines such as statistics, neural networks, database technology, machine learning and information retrieval, etc.

Interesting patterns are extracted at reasonable time by KDD's technique. KDD process has several steps, which are performed to extract patterns to user, such as data cleaning, data selection, data transformation, data preprocessing, data mining and pattern evaluation. The architecture of data mining system has the following main components: data warehouse, database or other repositories of information, a server that fetches the relevant data from repositories based on the user's request, knowledge base is used as guide of search according to defined constraint, data mining engine include set of essential modules, such as characterization, classification, clustering, association, regression and analysis of evolution. Pattern evaluation module that interacts with the modules of data mining to strive towards interested patterns. Finally, graphical user interfaces from through it the user can communicate with the data mining system and allow the user to interact. The problem of identifying constrained association rules for heart disease prediction was studied by Carlos Ordonez. The assessed data set encompassed medical records of people having heart disease with attributes for risk factors, heart perfusion measurements and artery narrowing. Three constraints were introduced to decrease the number of patterns. First one necessitates the attributes to appear on only one side of the rule. The second one segregates attributes into uninteresting groups. The ultimate constraint restricts the number of attributes in a rule. Experiments illustrated that the constraints reduced the number of discovered rules remarkably besides decreasing the running time. Two groups of rules envisaged the presence or absence of heart disease in four specific heart arteries.

II. LITERATURE SURVEY

A. Applications of Data Mining Techniques in Pharmaceutical Industry

[1]Almost two decades ago, the information flow in the pharmaceutical industry was relatively simple and the application of technology was limited. However, as we progress into a more integrated world where technology has become an integral part of the business processes, the process of transfer of information has become more complicated. Today increasingly technology is being used to help the pharmaceutical firms manage their inventories and to develop new product and services. The implications are such that by a simple process of merging the drug usage and cost of medicines (after completing the legal requirements) with the patient care records of doctors and hospitals helping firms to conduct nationwide trials for its new drugs. Other possible uses of information technology in the field of pharmaceuticals include pricing (two-tier pricing strategy)

and exchange of information between vertically integrated drug companies for mutual benefit. Nevertheless, the challenge remains though data collection methods have improved data manipulation techniques are yet to keep pace with them.[2] Data mining fondly called patterns analysis on large sets of data uses tools like association, clustering, segmentation and classification for helping better manipulation of the data help the pharma firms compete on lower costs while improving the quality of drug discovery and delivery methods. A deep understanding of the knowledge hidden in the Pharma data is vital to a firm's competitive position and organizational decision-making. The paper explains the role of data mining in pharmaceutical industry.

B. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction

[3] The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still „information rich" but „knowledge poor". There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. This research paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction. [4] Number of experiment has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and some time Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

C. Improving Heart Disease Prediction Using Constrained Association Rules

The Healthcare industry is generally "information rich", which is not feasible to handle manually. These large amounts of data are very important in the field of Data Mining to extract useful information and generate relationships amongst the attributes. The doctors and experts available are not in proportion with the population. Also, symptoms often be neglected. Heart disease diagnosis is a complex task which requires much experience and knowledge. Heart disease is a single largest cause of death in developed countries and one of the main contributors to disease burden in developing countries. In the health care industry the data mining is mainly used for predicting the diseases from the datasets. The Data Mining techniques, namely Decision Trees, Naive Bayes, Neural Networks, Associative classification, Genetic Algorithm are analyzed on Heart disease database.

III. PROPOSED ALGORITHM

[4] Data mining tools have been developed for effective analysis of medical information to help the clinician in making better diagnosis. In this research work, the researcher can collect data from Hospital Information System (HIS) which has the sufficient details of patient including patient's name, age, disease, location, district, date from laboratories which keeps on growing year after year. Having collected the data from hospital information system, this research can find the frequent disease with the help of association techniques. This research work helps to mine the data about the frequent diseases with the help of weka tool applied over training data set. In the medical field, the hospital information system is used to receive different kinds of medical records of diseases and its patients who hail from different areas. It is a herculean task to identify the frequent diseases and its causes from the large data set. Patients from different locations approach different hospitals. They do not converge in a same place. Their records are maintained by the hospitals where they get treated. Collecting information about the frequently occurring diseases is not an easy job. The data collection regarding these sorts of diseases can be done through association rule. Apriori of the Association rule is adopted for the mining of data.

A. Apriori Algorithm:

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as analysis. The Apriori algorithm was proposed by Agrawal and Srikant in 1994. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation or IP addresses[2]). Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing). Each transaction is seen as aset of items (an itemset). Given a threshold , the Apriori algorithm identifies the item sets which are subsets of at least transactions in the database.[5] Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length from item sets of length . Then it prunes the candidates which have an infrequent sub pattern. According tothe downward closure lemma, the candidate set contains all frequent -length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates. The pseudo code for the algorithm is given below for a transaction data base , and a support threshold of . Usual set theoretic notation is employed; though note that is a multi set. Is the candidate set for level. At each step, the algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma. Accesses a field of the data structure that represents candidate set , which is

initially assumed to be zero. Many details are omitted below, usually the most important part of the implementation is the data structure used for storing the candidate sets, and counting their frequencies. Apriori algorithm, a classic algorithm, is useful in mining frequent item sets and relevant association rules. Usually, you operate this algorithm on a database containing a large number of transactions. One such example is the items customers buy at a super market. It helps the customers buy their items with ease, and enhances the sales performance of the departmental store. This algorithm has utility in the field of healthcare as it can help in detecting adverse drug reactions (ADR) by producing association rules to indicate the combination of medications and patient characteristics that could lead to ADRs. Apriori figuring (Agrawal et al. 1993) is the maximum everyday and essential calculation for mining standard itemsets. Apriori is utilized to find all dynamic item sets in a given database DB. In context of the Apriori rule any subset of a dynamic item set must in like way be go to. Perspective: if XY is a dynamic item set, each A and B ought to be visit item sets.[6] The key thought of Apriori tally is to make exceptional overlooks the database. It utilizes an iterative method known as an expansiveness first intrigue (level-sensible pursue) through the demand space, in which okay-itemsets are utilized to explore (k+1)- item sets. Toward the begin, the approach of consistent 1-itemsets is found. The approach of that carries a particular something, which fulfill the assist side, is inferred by L1. In each ensuing bypass, we begin with a seed set of item sets noticed to be liberal inside the past pass. This seed set is utilized for making new possibly expansive item sets, called sure item sets, and take a look at the veritable assist for these contender item sets amidst the disregard the facts. Around the entire of the skip, we recognize which of the contender item sets are while doubtful enormous (perpetual), and they trade into the seed for the going with skip. Thusly, L1 is utilized to discover L2, the strategy of unending 2-itemsets, which is utilized to find out L3, etc, until no constantly regular okay-item sets.

Ck: Medical data item set of size k.
Lk: frequency item set of size k.
L1={frequent items};
For(k=1;Lk!= Φ ;k++) do begin
Ck+1=Medical data generated from Lk;
Each transaction t in database do
Increment the count of all medical data in Ck+1 that are c.
Obtained in t
Lk+1=medical data in Ck+1 with min_support.
End
Return$\cup_k L_k$.

IV. RESULT &ANALYSIS

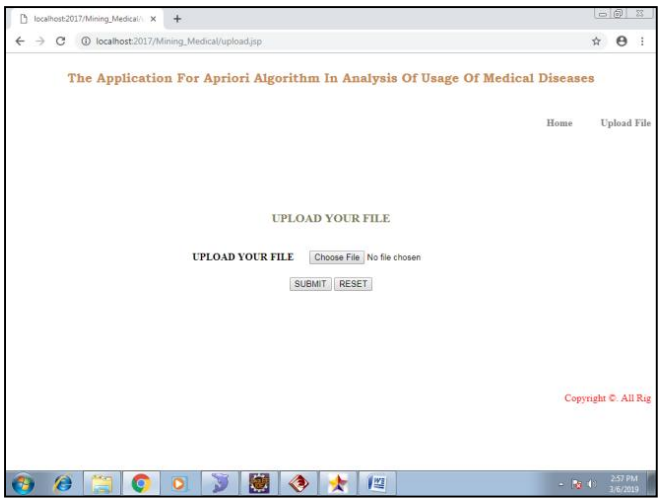

Fig. 1: Upload File

Here to upload the patients description and treatment files. If you see the all uploaded files and click on top right Upload button.

S.No	Name	Age	Disease	Months	Location	District	Date
1	Voddula Sanjay Kumar	78	HIV	April	Jagitial	Jayashankar Bhupalpally	14/4/2017
2	Penubarti Vishnu Sravya	90	Lenukemia	December	Kothagudem	Jangaon	22/10/2016
3	Pannem Gautami	94	Allergies	May	Charminar	Bhadradri Kothagudem	26/12/2017
4	Shaik Abdul Adil	48	Allergies	September	Jayashankar	Adilabad	11/2/2016
5	Mamidi Sumikendar	59	smoking	April	Jagitial	Bhadradri Kothagudem	26/12/2017
6	Nagoor Rakesh	64	human papilloma virus	September	Jagitial	Jangaon	4/3/2016
7	Sai Sneha Mulpuru	48	Insomnia	May	Kothagudem	Bhadradri Kothagudem	14/4/2017
8	Voddula Sanjay	64	Pregnancy	November	Kothagudem	Jagitial	22/10/2016

Fig. 2: View File

After click on the upload file Button and it shows all Uploaded Patients Details Like Name , Age , Disease , Location etc,.

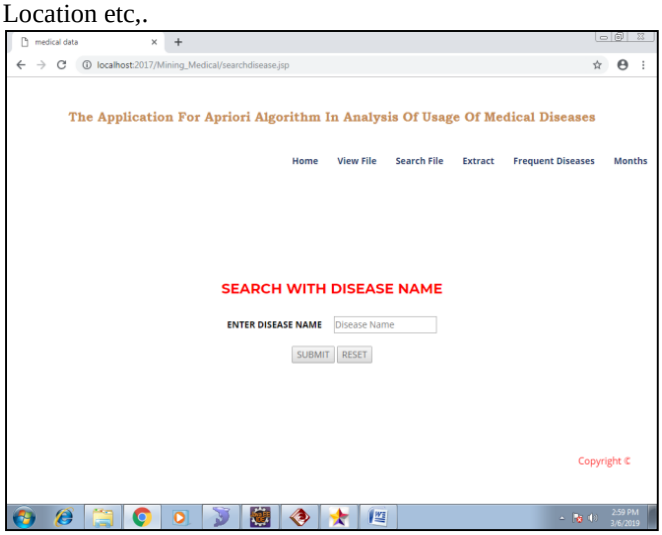

Fig. 3: Search File

If you Choose any one Disease of the patients details and click on Search File to enter the Disease name.

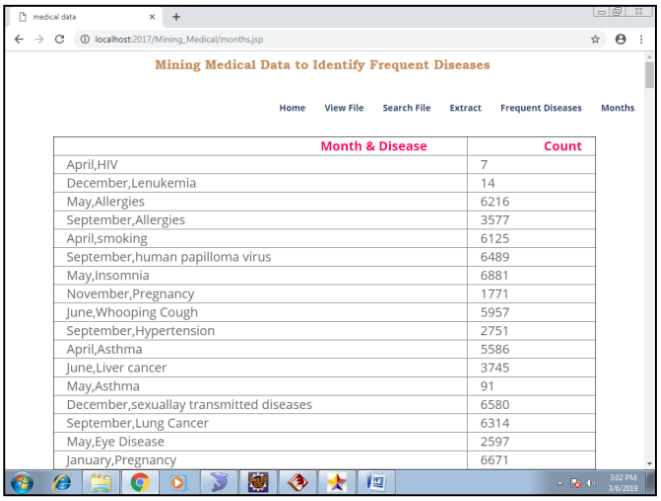

Fig. 4: Months

If you want to see which Diseases will comes in which month and click on the Month Button and it shows Monthly Disease And count of particular month.

V. CONCLUSION

This research work proposes an association rule based apriori data mining technique that finds the frequency of diseases affecting patients. The study is made on patients from various geographical locations and at various time periods. Existing electronic medical records obtained from hospitals are used as training data set for the study. Totally 1216 patient records affected by 29 different diseases during the year 2012 are analyzed. WEKA data mining tool is employed to identify the frequency of the diseases that are recurring in people living in various geographical locations during different time periods. The analysis revealed the fact that four different diseases affected the patients frequently at various geographical locations during the year 2012.

REFERENCES

[1] Arun K Pujari "Data Mining Techniques", Edition 2001.
[2] Jayanthi Ranjan, "Applications of Data Mining Techniques in Pharmaceutical Industry", Journal of Theoretical and Applied Information Technology.
[3] Jyoti Soni, et al., "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887), Volume 17– No.8, March 2011.
[4] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules", Seminar Presentation at University of Tokyo, 2004.
[5] 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.
[6] 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.

