ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Analysis of Clinical Databases Using Data Mining Techniques

L. Jagjeevan Rao*
School of Computing
KL University Green fields, Vaddeswaram
Andhra Pradesh India
jeevan@kluniversity.in

N. V. S. Pavan Kumar
School of Computing
KL University Green fields, Vaddeswaram
Andhra Pradesh India
nvspavankumar@kluniversity.in

M.Srinivas
School of Computing
KL University Green fields, Vaddeswaram
Andhra Pradesh India
srinu_cse@kluniversity.in

Abstract: Recent advances in high throughput data acquisition, digital storage, and communications technologies have made it possible to gather very large amounts of data in many scientific and commercial domains. Much of this data resides in relational databases. Over the last decade, we have seen the emergence of Data mining techniques that cater to the analysis of these databases. These techniques are typically upgraded from well-known and accepted.

Clinical databases have accumulated large quantities of information about patients and their medical diagnosis reports which describe their condition. Relationships and patterns within this data could provide new medical knowledge. Many methodologies have been developed and applied to discover this hidden knowledge. In this study, the techniques of data mining were used to search for relationships and multi dimensions in a large medical database

Key words – Classification, Association Rules, Clustering, K-means, Hyperilimedia

I. INTRODUCTION

Many authors did tremendous work in data mining and published their research which has enormous applications and use in many domains. In this paper we analyzed our study on few of the above data mining techniques on the medical diseases and related complications data. This survey depends on applying different mining techniques on medical reports of different patents with different kinds of medical complications. Researchers from different areas are continuously trying to apply these data mining techniques to medical data. The ways of solving their problems in an easy way is the main intension of this paper.

II. WHAT IS DATA MINING?

According to Arun K Pujari [1], Data mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Great scholars have developed so many techniques, as part of data mining. Here we introduce our study on them.

III. ASSOCIATION RULE MINING

Association rule mining is one of the best known fundamental data mining technique. In this, a pattern is discovered based on a relationship of a particular item with other items in the same transaction. Sengul Dogan [2] et.al proposed projection from biochemistry blood parameters in diagnosis of Hyperlipidemia. The basic parameters like LDL, Triglyceride, Cholesterol, HDL and VLDL etc were used in the process to identify Hyperlipidemia (T) and Hyperlipidemia (F).

IV. CASE DESCRIPTON

The term Hypermedia defines the Lipids in the bloodstreams. Lipids mean fats consisting of Vitamins and Minerals helpful to produce energy for human body. The abnormal level of these Lipids causes hypermedia. Extra producing of these lipids effects changing of serum levels, this is the base for Hypermedia diagnosis. Cholesterol, LDL, Triglyceride, HDL and VLDL are basic parameters to conform the hypermedia, these parameters produces no of related data stacks. Discovering knowledge from these data stacks is challenge for the medical science. Technically Lipid stands for fats[2]. Certain fats are useful to the body. The measurement of plasma lipid levels helps in finding hyperlipidemia patients. In real-life cases different types of data which is related to various researches is available [3]. Some of them are DNA [4], Graphs [5], measurements [6] etc. It is difficult for medical science people to work with this enormous data and discover the knowledge.

A. Classification:

Classification is one good technique of Data mining, could be used in order to get some information. This technique is based on machine learning. Basically classification is used to classify each item in a set of data [6] into one of Predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we make the software that can learn how to classify the data items into groups. Few medical cases decide the patients diseases based on the lab reports. Doctors have pre defined measurements to decide it. Example Hypertension has been

classified into three categories based on the severity of the disease into Normal, Mild, Moderate and Sever.

Normal	-	120/80
Mild	-	120/90
Moderate	-	140/100
Severe	-	160/110

B. Association Rules:

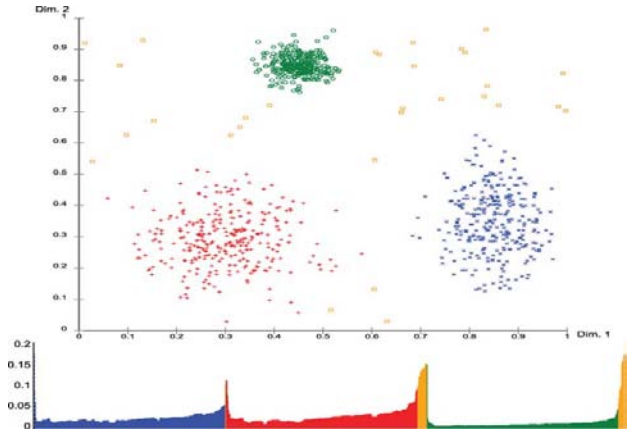
There are several Association rule algorithms which are mainly useful to summarize and identify the patterns. They also use correlation along with support and confidence in order to find the right patterns. The general form of an association rule is of the form X→ Y meaning that the transactions containing X, tend to contain Y. One simple example for this is the customers who purchase a mobile phone tend to purchase accessories for mobile like case etc.

Using this they have successfully developed a decision support system for Hyperlipidemia diagnosis.

C. Clustering:

Clustering is another data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes.

A process of grouping a set of physical or abstract object into a class of similar objects is called clustering. Clustering Technique is unsupervised learning. Clustering is of two types hierarchical clustering and partition clustering.



Generally a patient suffering with fever is treated initially by giving medication against general signs and symptoms of fever. After verifying the blood reports, doctor continues initial treatment along with specific treatment for the type of fever. Here different fevers (viral, malaria, typhoid…. etc..) are clustered with common features.

Some diagnostic and laboratory procedures are invasive, costly and painful to patients. An example of this is conducting a biopsy in women to detect cervical cancer. Thangavel et al [7] used the K-means clustering algorithm to analyze cervical cancer patients and found that clustering gives better predictive results than existing medical opinion. They found a set of interesting attributes that could be used by doctors as additional support on whether or not to recommend a biopsy for a patient suspected of having the cervical cancer.

Gorunescu et al [2] described how computer-aided diagnosis (CAD) and endoscopic ultrasonographic elastography (EUSE) were enhanced by data mining to create new noninvasive cancer detection.

In the traditional approach, doctors look at the ultrasound movie and decide on whether a patient is to be subjected to a biopsy. The physician's judgment is primarily subjective, depending mostly on his/her interpretation of the ultrasound video. Gorunescu approach to this problem is a different way, using data mining. He did not study patient demographics. Instead his team focused on the ultrasound movies. They first trained a classification algorithm using a multi-layer perception (MLP) on known cases of malignant and benign tumors.

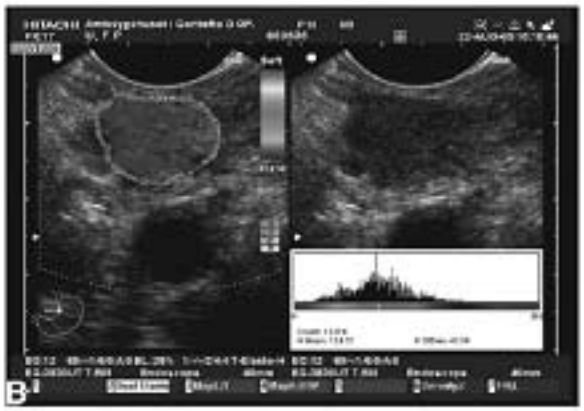


Fig. 1 EUSE sample movie frame with corresponding histogram

V. ADVANTAGES

Cheng et al[2] cited the use of classification algorithms for early detection and prevention of heart disease which is a common health concern round the globe. Pandemic diseases could be successfully managed as proposed by Kellogg et al [2]. They proposed techniques to combine spatial modeling, simulation and spatial data mining regarding out-break of disease. Also Wilson et al [2] worked a lot on drug side effects which are used for a long time. They proposed an algorithm called MGPS Multi-item Gamma Poisson Shirker for this purpose and were able to find around seventy percent success.

VI. DISADVANTAGES

In medical field there is a possibility of having extreme conditions or cases where a minute variation in symptoms like anthrax to flu is difficult to find using regular mining techniques. We need to have more advanced developments in standard mining mechanism to implement this technology in practical scenarios. There are always out liars in medical data which are to be dealt with more attention.

VII. CONCLUSION

Here our intention is to study various data mining techniques with respect to medical data. We have referred few important research articles of implementation of various techniques, and the results with examples. Still there are lot more developments not mentioned here. There is more scope and need for further usage of remaining techniques like prediction and sequential patterns etc.

The prediction as it name implied, discovers relationship between independent variables and relationship between dependent and independent variables

For example generally fever starts with body pains followed by chills, rigors, cough, cold and raise in body temperature, all these symptoms are possible in sequential pattern.

Sequential pattern seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data

Doctor's decision with regard to medical data analysis and their conclusions about dieses is strongly influence the patient's mental and physical environment. Using the appropriate techniques of data mining, improves the decision making and medication process in the medical field. Data mining techniques gives qualitative and quantitative treatments for the people.

VIII. REFERENCES

[1]. Data mining techniques, Arun K pujari, Universities Press, Pages 288, First edition,2001.

[2]. Ruben D. Canlas Jr., MSIT, MBA, Data Mining in the Health Care Current Applications and issues, August 2009.

[3]. J. Barrera, R.M. Cesar-Jr, an Environment Knowledge Discovery in Biology, 34, 427–447, 2003.

[4]. M. M. Yin, J. T. L. Wang, Gene Scout: a Data Mining System or Predicting Vertebrate Genes in Genomic DNA sequence, Information Sciences 163, 201–218,2003

[5]. I.Turkoglu, A. Arslan, An Intelligent System for Diagnosis of Heart Valve Disease with Wavelet Packet Neural Networks, Computer in Biology and Medicine 33(4),319-331,2003

[6]. R. J. Shebuski, Utility of Point-of- care Diagnostic testing in Patients with Chest pain and Suspected acute Myocardial Infarction, Current Opinion in Pharmacology 2, 160-164,2002.

[7]. J. M. Ayub, C. R. Smulski, Protein–Protein Interaction Map of the Trypanosoma Cruzi Ribosomal P Protein Complex, Gene 357, 129 – 136, 2005.

[8]. I. Turkoglu, A. Arslan, An Intelligent System for Diagnosis of Heart Valve Diseases with Wavelet Packet Neural Networks, Computer in Biology and Medicine 33(4), 319-331, 2003.

[9]. R. J. Shebuski, Utility of Point-of-Care Diagnostic Testing in Patients with Chest Pain and Suspected Acute Myocardial Infarction, Current Opinion in Pharmacology 2, 160–164, 2002.12

[10]. L. D'Avolio, G. Rees, Improving the Secondary Utilization of Clinical Data by Incorporating Context, Los Angeles CA, 2000.

[11]. Quinlan, J.R. 1992, C4.5: Program for machine learning, Morgan Kaufmanns

[12]. C. C. Bojarczuk, H. S. Lopes, A Constrained-Syntax Genetic Programming System for Discovering Classification Rules: Application to Medical Data Sets, Artificial Intelligence in Medicine 30, 27–48, 2004.

[13]. G. Lupattelli, S. Marchesi, Mechanisms of High-Density Lipoprotein Cholesterol Effects on the Endothelial Function in Hyperglycemia, Metabolism 52(9), 1191–1195, 2003.

[14]. R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, Very Large Databases 153(8), 487–499, 1994