ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 503- 507 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

A New System to Generate Simplified Decision Trees for applicability in the domain of Health Care

Abdul Ahad Md[1], Dr. Suresh Babu Y[2] , Rajesh Chandra G[3]

[1]Asst. Professor, Department of ECM, KL University, Vaddeswaram, INDIA, ahadbabu@gmail.com
[2]Professor, Department of CSE, JKC College, Guntur, INDIA, sureshbabu_y@gmail.com
[3]Asst. Professor, Department of ECM, KL University, Vaddeswaram, INDIA, grajeshchandra@kluniversity.in

***Abstract*— The valuable knowledge can be discovered from application of data mining techniques in the domain of health care. Data mining applications can have more benefit in the healthcare industry. For example, data mining can help to detect fraud and abuse in healthcare insurance and also make customer relationship management decisions, and best practices, and patients receive better and more affordable healthcare services. The large amounts of data maintained by healthcare transactions are too sensible and complex and to be analyzed and processed by traditional methods. Data mining provides the new methodology and technology to transform these complex data into useful information for decision making. The main aim of this paper is, to generate a new system for applicability in the domain of Healthcare Decision Support Systems currently used in medicine.**

Keywords—Hosseinkhah, Fatemeh, et al. "Challenges in Data Mining on Medical Databases."

INTRODUCTION

Now a day, Computer Programming is getting more and more involved in the domain of health and medical sciences. All the healthcare Decision Support Systems have been constructed by the aid of Artificial intelligence. These systems are proved to be very useful for patient as well as for medical experts in making the decisions. Various approaches are used for gathering the input data and to present output data in different methodologies. Any computer program that helps experts in making healthcare decision comes under the domain of healthcare decision support system. The main objective of this paper is

-To present recent methods in HDSS.
-To use E-Records used in Healthcare .
-To discuss methodologies used in Health Care

Data mining results in the discovery of hidden and predictive information from complex and huge databases. A formal definition of Knowledge discovery in databases is the non-trivial extraction of implicit previously unknown and potentially useful information about data.

Data mining involves six common classes of tasks:

Bugs detection– The identification of unusual data records, that might be interesting or data errors that require further investigation.

Association– Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

Clustering– is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Classification – is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as "legitimate" or as "spam".

Regression – Attempts to find a function which models the data with the least error.

Summarization – providing a more compact representation of the data set, including visualization and report generation.

Data mining helps to the novel and hidden patterns in the data. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. The following are some of the important areas where data mining techniques can be in health care management.

1. Public Health Informatics
2. Health Insurance
3. Health Care Information System
4. E-Governance in Health Care
5. Health Care Application for Data Modeling

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 503- 507 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

DATA MINING TECHNIQUES

Data mining technique is most important technique which is used in Knowledge Discovery in Database(KDD).KDD has different types of steps like Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, Knowledge presentation etc. There are different types of techniques used in Data mining project. These include Decision tree, Bayesian networks, Naive bayes, Neural networks etc.

Decision tree-It is the most frequently used techniques of data analysis. It is used to classify records to a proper class and is applicable in both regression and associations tasks. In medical field decision trees specify the sequence of attributes. Such a tree is built of nodes which specify conditional attributes – symptoms X={x1,x2,.....xk},branches which show the values of S ih i.e. the *h*-th
range for *i*-th symptom and leaves which present decisions Y={y1,y2.....yk} and their binary
values Zdk={0,1} . A sample decision tree is presented in the fig1.

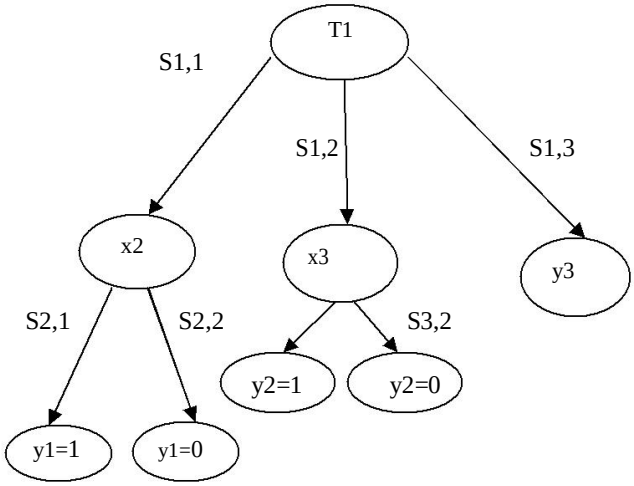

Fig1: Decision tree applicable in medicine

Neural Networks- In medical diagnosis the input to the neural network are the patient's symptoms the set T , and Y is the output of the diagnosis . There are 3 layers in neural networks: input layer, hidden layer, output layer. Hidden layer is the outcomes of the input layer..The condition between neurons has weights which are assigned to them. Their values are calculated with the use of back propagation algorithm. In hidden layers there are some nonlinear features are added to the network..The out layer may have more than one output node which predict the different diseases.

In a single neuron there are many input layers and one output layer. The input and output values are issued with the use of combination and activation function.

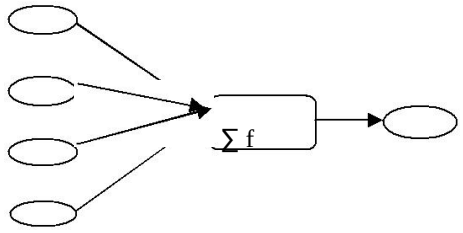

Fig2: Single neuron

A. Advantages of Data mining

Predict future trends, customer purchase habits
Help with decision making
Improve company revenue and lower costs
Market basket analysis
Fraud detection

B. Disadvantages

Great cost at implementation stage
Possible misuse of information
Possible inaccuracy of data

DATA MININ IN HEALTHCARE

Data mining applications are currently being applied to in the following two main branches:

A. Healthcare decision support system

Healthcare decision support system is an interactive decision support system, which is designed to assist physicians and other health professionals with decision making tasks, such as determining diagnosis of patient data. . The main purpose of modern HDSS is to help clinicians at the point of care. It means, a clinician would interact with a HDSS to help determine diagnosis, analysis, etc. of patient data. It is a decision-support system program that offers employees in-depth, objective, personalized, and current information on all healthcare conditions.

An example of how a HDSS might be used by a medicinal comes from the subset of HDSS (Healthcare Decision Support System), DDSS (Diagnosis Decision Support Systems). A DDSS would take the patients data and propose a set of appropriate diagnoses. The doctor then takes the output of the DDSS and point out which are relevant and which are not. Another important classification of a HDSS is based on the timing of its use. Doctors use these systems at point of care to help them as they are dealing with a patient, with the timing of use as either pre-diagnoses, during diagnoses, or post diagnoses. Pre-diagnoses HDSS systems are used to help the physician prepare the diagnoses. HDSS used during diagnoses help review and filter the physician's preliminary diagnostic choices to improve their final results. And post-diagnoses HDSS systems are used to

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 503- 507 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

mine data to derive connections between patients and their past medical history and to predict future events.

B. Characteristics of Healthcare Decision Support Systems

The Healthcare DSS's are the type of computer programs that assist physicians and medical staff in health caredecision making tasks. Most of the healthcare decision support systems (HDSS's) are equipped with diagnostic assistance module, therapy critiquing and planning module, medications prescribing module, information retrieval subsystem (for instance formulating accurate clinical questions) and image recognition and interpretation section (X-rays, CT, MRI scans) Interesting examples of HDSS's are machine learning systems which are capable of creating new healthcare knowledge. By analyzing healthcare cases a Healthcare Decision Support System can produce a detailed description of input features with a unique characteristic of healthcare conditions. It supports may be priceless in looking for changes in patient's health condition. These systems may improve patients' safety by reducing errors in diagnosing. They may also improve medications and test ordering.

IMPORTANCE OF HEALTH CARE

These are the some important features in Healthcare.
Access all the patient records and rapidly detect anomalies, Analyze data using an automated system, which is useful in the case of major and repeated anomalies, Boost productivity and care quality through remote, shorter and more frequent consultations, Interact quickly and easily in a structured way via tools shared between the primary care provider and the nurses responsible for day-to-day patient monitoring, Provide motivational support for patients who desire it, Contribute to biomedical research through the tool's healthcare database.

Hospital Infection Control

No socomial infections affect 2 million patients each year in the United States, and the number of drug-resistant infections has reached unprecedented levels14. Early recognition of outbreaks and emerging resistance requires proactive surveillance. Computer-assisted surveillance research has focused on identifying high-risk patients ,expert systems, and possible cases and detecting deviations in the occurrence of predefined events. The system uses association rules on culture and patient care data obtained from the laboratory information management systems and generates monthly patterns that are reviewed by an expert in infection control .Developers of the system conclude enhancing infection control with the data mining system is more sensitive than traditional infection control surveillance, and significantly more specific.

Ranking Hospitals

Organizations rank hospitals and healthcare plans based on information reported by healthcare providers. There is an assumption of uniform reporting, but research shows room for improvement in uniformity. Data mining techniques have been implemented to examine reporting practices. With the use of International Classification of Diseases, 9th revision, codes (risk factors) and by reconstructing patient profiles, cluster and association analyses can show how risk factors are reported.16 Standardized reporting is important because hospitals that underreport risk factors will have lower predications for patient mortality. Even if their success rates are equal to those of other hospitals, their ranking will be lower because they reported a greater difference between predicted and actual mortality.16 Standardized reporting

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 503- 507 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

Identifying High-Risk Patients
American Health ways provides diabetes disease management services to hospitals and health plans designed to enhance the quality and lower the cost of treatment of individuals with diabetes. To augment the company's ability to prospectively identify high-risk patients, American Health ways uses predictive modeling technology. Extensive patient information is combined and explored to predict the likelihood of short-term health problems and intervene proactively for better short-term and long-term results. A robust data mining and model-building solution identifies patients who are trending toward a high-risk condition .This information gives nurse care coordinators a head start in identifying high-risk patients so that steps can be taken to improve the patients' quality of healthcare and to prevent health problems in the future.

Treatment effectiveness
Data mining applications can be developed to evaluate the effectiveness of medical treatments. By comparing and contrasting causes, symptoms, and courses of treatments, data mining can deliver an analysis of which courses of action prove effective. For example, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective.

Healthcare management
To aid healthcare management, data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims.

Customer relationship management
While customer relationship management is a core approach in managing interactions between commercial organizations—typically banks and retailers—and their customers, it is no less important in a healthcare context. Customer interactions may occur through call centers, physicians' offices, billing departments, inpatient settings, and ambulatory care settings.

Fraud and abuse
Data mining applications that attempt to detect fraud and abuse often establish norms and then identify unusual or abnormal patterns of claims by physicians, laboratories, clinics, or others. Among other things, these applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims.

505

DATA SETS

There are different types of data sets are used like patients record datasets, disease data sets and anthropometry datasets. Four UCI medical datasets: hepatitis, heart disease, dermatology disease, diabetes, lung cancer.

Heart disease database-According to statistics heart disease is a leading reason of death in 2007 . The most common heart diseases are coronary heart disease, ischaemic heart disease, cardiovascular disease, cor pulmonale, hereditary heart disease, hypertensive heart disease and valvular heart disease. There may be a number of symptoms of the disease. Finding patterns in heart disease data may help diagnose future cases of this illness. The heart disease database was collected by the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation in 1988.

Hepatitis database-The Hepatitis database comes from Jozef Stefan Institute in Yugoslavia. The data was gathered in 1988. The hepatitis is induced by a dangerous virus called hepatitis B virus (HBV). If the disease is not eliminated in its initial infection it in 15% cases it cause chronic hepatitis.

Diabetes database-The diabetes disease also can have a large number of symptoms. While diagnosing a plasma glucose level is measured. Such examination shows whether patient is in risk of diabetes or not. It is extremely important to diagnose diabetics as quickly as possible. Unrecognized disease may lead to hypertension, shock, amputation or even death. The Pima Indians Diabetes Database was created in National Institute of Diabetes and gestive and Kidney Diseases and shared in 1990 in . e database contains information about diabetes among adult women (the youngest one is 21 years old, the oldest one 81 years old). The data was gathered with the use of unique algorithm called ADAP.

Dermatology database-The database was created while diagnosing six dermatologic diseases: soriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. The most part of the features concerns the biopsy examinations. At the beginning only twelve clinically symptoms were specified. On the basis of various analyses of skin samples another twenty two observations are added. Furthermore, there are features called scaling and erythema whose values do not differ fundamentally, however these symptoms are important while diagnosing some diseases.

Lung cancer Database-The IARC (International Agency for Research on Cancer) air pollution to Group 1 carcinogenic — the same category under which tobacco, UV radiation and plutonium come. Air pollution was known be among the causes for heart and lung diseases. There is sufficient evidence that exposure to outdoor air pollution causes lung cancer with a positive association with an increased risk of bladder cancer.

506

DISCUSSION

In this survey we are discussing that, now a days the

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 503- 507 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

doctors are unable to detect the disease like cancer, tumor etc, so death ratio is increasing day to day. Basically the heart disease is the common disease among the patients, it

is very much dangerous, so we are using some modern technologies like Data mining, Data warehouse etc. By using this technique we can easily find out the hidden information from the disease. We are using different Data mining techniques such as classification, naive bayes, bayesian technique, neural network ,multilayer perceptron etc.

It is also possible to see graphical (more intuitive) form of the tree, here the Builds models which can be easily interpreted, it is easy to implement, it can use both categorical and continuous values, it does not work very well on a small training set. In Naive bayes algorithm, It is a simple probabilistic classifier, which is based on an assumption about mutual independency of attributes. The probabilities which is applied in the Naïve Bayes algorithm are calculated according to the Bayes Rule, the probability of hypothesis *H* can be calculated on the basis of the hypothesis *H* and evidence about the hypothesis *E* according to the following formula:

$$P(H|E)=\frac{P(E|H)*P(H)}{P(E)}$$

The naive Bayes classifier's beauty is in its simplicity, computational efficiency, and good classification performance. Three important issues are, First the naive Bayes classifier requires a very large number of records to obtain good results. Second, where a predictor category is not present in the training data, naive Bayes assumes that a new record with that category of the predictor has zero probability. When it classifies, performance does not show significant improvement. In Neural network, there are 3 layers in neural networks: input layer, hidden layer, output layer.

This method is difficulty in understanding the predictions. In this survey we have used a most modern technique which is designed to assist physicians and other health professionals with decision making tasks, such as determining diagnosis of patient data .It is called Healthcare decision support system(HDSS). The main purpose of modern HDSS is to help clinicians at the point of care. It means, a clinician would interact with a HDSS to help determine diagnosis, analysis, etc. of patient data. It offers opportunities to reduce medical errors as well as to improve patient safety. One of the most important applications of such systems is in diagnosis and treatment of heart diseases (HD) because statistics have shown that heart disease is one of the leading causes of deaths all over the world.ERA, HELP, DX plain are the different examples of HDSS. ERA is one of the newest and most promising Healthcare Decision Support Systems, which is dedicated to detection of different types of cancers in their early stage.

Here we are using different diseases databases: Heart disease database, Hepatitis databases, Diabetes database, Dermatology database

Summary of Data mining Techniques

TECHNIQUES	UTILITY	DISEASE
Decision Tree Algorithms such as ID3, C4.5, C5, and CART.	Decision support	Heart Disease
Neural Networks	Extracting patterns, detecting trends	Heart Disease
Naive Bayesian	Improving classification accuracy.	Coronary Heart Disease

Fig 3: Data mining techniques

Here c4.5 is better than the naive bayes technique, there is a detailed description of the data and the required pre-processing activities.c4.5 yields highly accurate results within few folds of cross validation considering the attribute with high performance gain for classification while the Naive bayes classifies performance does not show much significant improvement.

CONCLUSION

The main goal of this survey was to identify the most common data mining algorithms, implemented in modern Healthcare Decision Support Systems, and evaluate their performance on several medical datasets. Three algorithms were chosen: C4.5, Multilayer Perceptron and Naïve Bayes,and different disease database are taken.There are several Healthcare Decision Support Systems utilized in medical centers all over the world.

FUTURE WORK

The plans of future work include the evaluation of chosen algorithms on the basis of other medical datasets. The experiments would be conducted for the wider range of medical records what make the evaluation even more precise. The good idea is taking also other algorithms to the experiments and compares their performance in medical field. This would develop a new ranking and help in designing Medical Decision Support Systems by the choice of the most suitable algorithms. We can also take other techniques which are not included in this survey for comparison purpose and can find the best one by evaluating the advantages and limitations of the existing one.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 503- 507 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

REFERENCES

1. Ozer, Patrick. "Data Mining Algorithms for Classification." (2008).
2. Hosseinkhah, Fatemeh, et al. "Challenges in Data Mining on Medical Databases." (2009): 1393-1404.
3. Miller, Randolph A. "Medical Diagnostic Decision Support Systems—Past, Present, And Future A Threaded Bibliography and Brief Commentary." *Journal of the American Medical Informatics Association* 1.1 (1994): 8-27.
4. Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare."*Journal of Healthcare Information Management— Vol* 19.2 (2011): 65.
5. Walus, Y. E., H. W. Ittmann, and L. Hanmer. "Decision support systems in health care." *Methods of information in medicine* 36.2 (1997): 82.
6. Mangiameli, Paul, David West, and Rohit Rampal. "Model selection for medical diagnosis decision support systems." *Decision Support Systems* 36.3 (2004): 247-259.
7. Lemke, Frank, and Johann-Adolf Mueller. "Medical data analysis using self-organizing data mining technologies." *Systems Analysis Modelling Simulation*43.10 (2003): 1399-1408.
8. Baylis, Philip. "Better health care with data mining." *SPSS White Paper, UK* (1999).
9. Jenn-Lung Su, Guo-Zhen Wu, I-Pin Chao (2001). The Approach Of Data Mining Methods For Medical Database. *IEEE*. p1-3.
10. Abbasi, M. M., and S. Kashiyarndi. "Clinical Decision Support Systems: A discussion on different methodologies used in Health Care." (2006).