ISSN:0975-9646

Study and Analysis of Data mining Algorithms for Healthcare Decision Support System

Monali Dey, Siddharth Swarup Rautaray

Computer School of KIIT University, Bhubaneswar ,India

***Abstract*— Data mining technology provides a user oriented approach to novel and hidden information in the data. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. Data mining in healthcare medicine deals with learning models to predict patients' disease. Data mining applications can greatly benefit all parties involved in the healthcare industry. For example, data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. The main aim of this survey is, analysis of the uniqueness of medical data mining, overview of Healthcare Decision Support Systems currently used in medicine, identification and selection of the most common data mining algorithms implemented in the modern HDSS, comparison between different algorithms in Data mining.**

***Index Terms*— Naive Bayes, C4.5, healthcare decision support**,**Neural network**

I. INTRODUCTION

Computer Science is now getting more and more involved in the medicine and health sciences. The branch of computer science which is more actively and efficiently involved in medical sciences is Artificial Intelligence. Various healthcare Decision Support Systems have been constructed by the aid of Artificial intelligence. These systems are now widely used in hospitals and clinics. They are proved to be very useful for patient as well as for medical experts in making the decisions. Different methodologies are used for the development of those systems. The way of gathering the input data and to present output information's is different in different methodologies. Any computer program that helps experts in making healthcare decision comes under the domain of healthcare decision support system. An important characteristic of the Artificial Intelligence is that it can support the creation as well as utilization of the healthcare knowledge. The main objective of this paper is

To present recent trends in healthcare Decision Support Systems.

-To discuss methodologies used in Health Care.

-To use electronic record used in Healthcare .

Data mining is the core step, which results in the discovery of hidden and predictive information from large databases. A formal definition of Knowledge discovery in databases is given as follows: "Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data"[1].

Data mining involves six common classes of tasks:

- Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (Dependency modeling) – 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 technology provides a user-oriented approach to the novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. 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. Following are some of the important areas of interests where data mining techniques can be of tremendous use in health care management [2].

1. Data modeling for health care applications
2. Executive Information System for health care
3. Forecasting treatment costs and demand of resources
4. Anticipating patient's future behaviour given their history
5. Public Health Informatics
6. e-governance structures in health care
7. Health Insurance

II. 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.

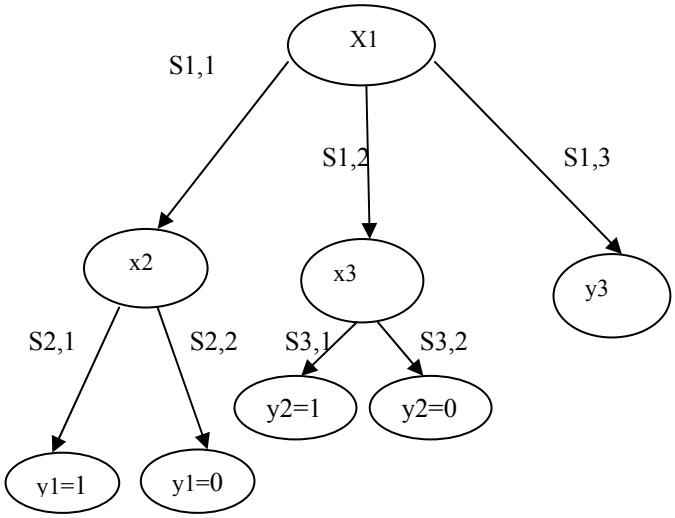

Fig 1 Decision tree applicable in medicine

Naive Bayse- 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:

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

Neural Networks-In medical diagnosis the input to the neural network are the patient's symptoms the set X , 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 have weights which is 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.

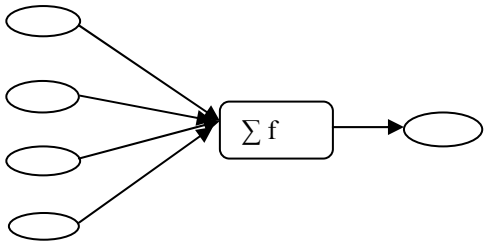

Fig 2 Single neuron

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

III. DATA MININ IN HEALTHCARE

Data mining applications are currently being applied to two main branches in health care and medicine: Healthcare decision support system, and policy planning/decision making. [5][6]

A. Healthcare decision support system

HDSS is an interactive Decision support system(DSS) Computer Software, 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. Employees receive the information, tools, and support they need from integrated web, phone, and print-based materials. This helps employees make more informed healthcare decisions while working with their own physician.[7] There are two main types of HDSS.

- Knowledge-Based
- Non Knowledge-Based

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

Features of a Knowledge-Based HDSS

Most HDSS consist of three parts, the knowledge base, inference engine and mechanism to communicate. The knowledge base contains the IF-THEN rules. The inference engine combines the rules from the knowledge base with the patient's data . The communication mechanism will allow the system to show the results to the user as well as have input into the system.

Features of a non-Knowledge-Based HDSS

Two types of non-knowledge-based systems are neural networks and genetic algorithm. Neural networks use nodes and weighted connections between them to analyze the patterns found in the patient data to derive the associations between the symptoms and a diagnosis. Genetic Algorithms are based on simplified evolutionary processes using directed selection to achieve optimal HDSS results. The HDSS features associated with success include the following:

- it is integrated into the health care workflow rather than as a separate log-in or screen.
- it is electronic rather than paper-based templates.
- it provides decision support at the time and location of care rather than prior to or after the patient encounter.
- it provides (active voice) recommendations for care, not just assessments.

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. [8]

- 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.
- Furthermore, the quality of care gets better due to the lengthening of the time clinicians spend with a patient. It may be an effect of application of proper guidelines, up-to date healthcare evidence and improved documentation. Moreover, the efficiency of the health care delivery is improved by reducing costs through faster order processing or eliminated duplication of tests.

C. Examples of Healthcare Decision Support Systems

These are the examples of HDSS

- CADUCEUS
- Diagnosis Pro
- DX mate
- DX plain
- ESAGIL
- MYCIN
- RODIA
- HELP
- ERA

There exist several Healthcare Decision Support Systems (HDSS's). They help in early detection of diseases. In this survey a few of the most important systems are presented. They are utilized in hospitals. To present the idea of Healthcare Decision Support Systems three sample ones are described: HELP, DX plain and ERA. [9]

HELP

One of the most popular and advanced Healthcare Decision Support System is called HELP. It helps the clinicians in interpreting healthcare information, diagnosing the disease of patients, maintaining healthcare protocols and other tasks .In 2003 a new version was released, called HELP II. It is equipped with a knowledge database which stores about 32000 emergency cases and a health care decision support engine. This system contains two assistants called antibiotic assistant and pneumonia diagnostic assistant. The purpose of the former is to find the pathogens causing the infection and to suggest the cheapest therapy for patients with e.g. allergies or renal functions.

DX plain

It is a Healthcare Decision Support System (HDSS) available through the World Wide Web that assists clinicians by generating stratified diagnoses based on user input of patient signs and symptoms, laboratory results, and other healthcare findings Each healthcare finding entered into DX plain is assessed by determining the importance of the finding and how strongly the finding supports a given diagnosis for each disease in the knowledge base. Using this criterion, DX plain generates ranked differential diagnoses with the most likely diseases yielding the lowest rank.

ERA (Early Referrals Application)

The Early Referrals Application (ERA) is one of the newest and most promising Healthcare Decision Support Systems. This solution is dedicated to detection of different types of cancers in their early stage. The application has been developed in Great Britain by GP's associated with the university hospitals of Leicester NHS Trust since 2001.

IV. 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.

A. Application of Data Mining in Healthcare

Business and marketing organizations may be ahead of healthcare in applying data mining to derive knowledge from data. This is quickly changing. Successful mining applications have been implemented in the healthcare arena, three of which are described below.[11]

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 would also be important for meaningful comparisons across hospitals.

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.

V. EXPERIMENT AND ANALYSIS OF ALGORITHMS IN WEKA

Data mining algorithms which are indentified to be very common in MDSS are implemented in WEKA environment. The aim of this chapter is to familiarize one with the WEKA algorithms' implementation details, describe important parameters and show the ways of the result presentation.

A. C4.5 algorithm

In WEKA environment the algorithm C4.5 is called J48 and it is the newest version of this algorithm's implementation. The parameters of C4.5 algorithm allows changing confidence threshold responsible for tree pruning, minimum number of instances which are permitted at a leaf. It is also possible to set the size of pruning set which is the number of data parts from which the last is used for tree pruning. Furthermore, WEKA's C4.5 decision tree may be pruned with the reduced error pruning. To achieve this it is essential to turn on reduced Error Pruning (set True instead default False). The generated decision tree may be presented in the text form. It is also possible to see graphical (more intuitive) form of the tree. The decision tree leafs have values in brackets like for instance (15.0/1.0) what means that 15 instances followed this formula correctly and 1 was misclassified.

Advantages & disadvantages:

The advantages of the C4.5 are:

- Builds models that can be easily interpreted
- Easy to implement
- Can use both categorical and continuous values
- Deals with noise

The disadvantages are:

- Small variation in data can lead to different decision trees (especially when the variables are close to each other in value)
- Does not work very well on a small training set

C4.5 is used in classification problems and it is the most used algorithm for building DT.

It is suitable for real world problems as it deals with numeric attributes and missing values. The algorithm can be used for building smaller or larger, more accurate decision trees and the algorithm is quite time efficient.

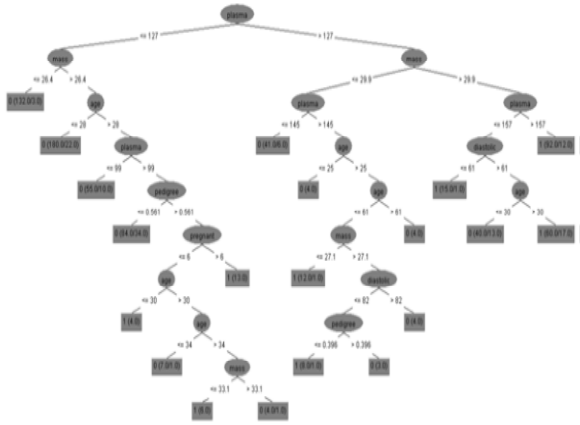

Figure 3 Graphical form of the C4.5 tree in WEKA

B. Naïve Bayes

Naïve Bayes classifier has quite simple interface in WEKA environment . It allows one to select the kernel estimator for numeric attributes rather than a normal distribution and used Supervised Discretization while converting numeric attributes to normal ones. The output of Naïve Bayes classifier has text form. Following the algorithm the dependences between each conditional attribute and decision attribute are verified and full statistics are presented.

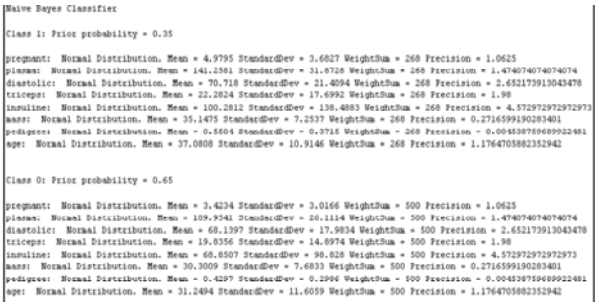

Figure 4 Dependences between conditional attributes and decisional attribute generated by WEKA's Naïve Bayes algorithm

Advantages & disadvantages:

The advantages of naive bayes are: The naive Bayes classifier's beauty is in its simplicity, computational efficiency, and good classification performance. Three issues should be kept in mind, however. 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. This can be a problem if this rare predictor value is important. For example, consider the target variable bought high-value life insurance and the predictor category own yacht. If the training data have no records with owns yacht = 1, for any new records where owns yacht = 1, naive Bayes will assign a probability of 0 to the target variable bought high-value life insurance. With no training records with owns yacht = 1, of course, no data mining technique will be able to incorporate this potentially important variable into the classification model—it will be ignored. With naive Bayes, however, the absence of this predictor actively "out votes" any other information in the record to assign a 0 to the target value (when, in this case, it has a relatively good chance of being a 1). The presence of a large training set (and judicious binning of continuous variables, if required) helps mitigate this effect. The disadvantages are: A subtle issue ("disadvantage" if you like) with Naive-Bayes is that if you have no occurrences of a class label and a certain attribute value together (e.g. class="nice", shape="sphere") then the frequency-based probability estimate will be zero. Given Naive-Bayes' conditional independence assumption, when all the probabilities are multiplied you will get zero and this will affect the posterior probability estimate. This problem happens when we are drawing samples from a population and the drawn vectors are not fully representative of the population. Lagrange correction and other schemes have been proposed to avoid this undesirable situation.

C. Neural Network

NN is a non knowledge-based adaptive HDSS that uses a form of artificial intelligence, also known as machine

learning, that allows the systems to learn from past experiences / examples and recognizes patterns in healthcare information. It consists of nodes called neuron and weighted connections that transmit signals between the neurons in a forward or looped fashion. It consists of 3 main layers: Input which is data receiver, Output which communicates results or possible diseases and Hidden which processes data. The system becomes more efficient with known results for large amounts of data.

Advantages & disadvantages:

The advantages of NN include the elimination of needing to program the systems and providing input from experts. The NN HDSS can process incomplete data by making educated guesses about missing data and improves with every use due to its adaptive system learning. Additionally, NN systems do not require large databases to store outcome data with its associated probabilities. A neural network can perform tasks that a linear program can not. When an element of the neural network fails, it can continue without any problem by their parallel nature

Some of the disadvantages are that the training process may be time consuming leading users to not make use of the systems effectively. The NN systems derive their own formulas for weighting and combining data based on the statistical recognition patterns over time which may be difficult to interpret and doubt the system's reliability. The neural network needs training to operate. The architecture of a neural network is different from the architecture of microprocessors therefore needs to be emulated. Examples include the diagnosis of appendicitis, back pain, myocardial infarction, psychiatric emergencies and skin disorders. The NN's diagnostic predictions of pulmonary embolisms were in some cases even better than physician's predictions.

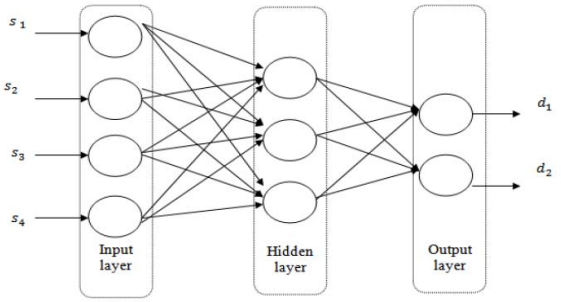

Figure 5 Neural Network for medical diagnosis case

VI. 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.[21]

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.[22]

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 Digestive and Kidney Diseases and shared in 1990 in . The 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 .[24]

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.

VII. DISCUSSION

In this survey we are discussing that, now a days the 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. We discussed 3 techniques in this survey, i-c4.5 algorithm, ii-Naive bayes algorithm, iii-neural network. In c4.5 algorithm, it allows changing confidence threshold

responsible for tree pruning, minimum number of instances which are permitted at a leaf. 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. Hidden layer is the outcomes of the input layer..The condition between neurons have weights which is 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.

When an element of the neural network fails, it can continue without any problem by their parallel nature. 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-6 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.

VIII. 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.

IX.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.

REFERENCES

1. Mariscal, Gonzalo, Óscar Marbán, and Covadonga Fernández. "A survey of data mining and knowledge discovery process models and methodologies."*Knowledge Engineering Review* 25.2 (2010): 137.
2. Dr. Lokanatha C. Reddy, A Review on Data mining from the Past to the Future, *International Journal of Computer Applications (0975 – 8887) Volume 15– No.7, February 2011*
3. Ozer, Patrick. "Data Mining Algorithms for Classification." (2008).
4. Ozer, Patrick. "Data Mining Algorithms for Classification." (2008).
5. Hosseinkhah, Fatemeh, et al. "Challenges in Data Mining on Medical Databases." (2009): 1393-1404.
6. Baylis, Philip. "Better health care with data mining." *SPSS White Paper, UK* (1999).

7. Jenn-Lung Su, Guo-Zhen Wu, I-Pin Chao (2001). The Approach Of Data Mining Methods For Medical Database. *IEEE*. p1-3.
8. Abbasi, M. M., and S. Kashiyarndi. "Clinical Decision Support Systems: A discussion on different methodologies used in Health Care." (2006).
9. Walus, Y. E., H. W. Ittmann, and L. Hanmer. "Decision support systems in health care." *Methods of information in medicine* 36.2 (1997): 82.
10. Mangiameli, Paul, David West, and Rohit Rampal. "Model selection for medical diagnosis decision support systems." *Decision Support Systems* 36.3 (2004): 247-259.
11. 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.
12. Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare."*Journal of Healthcare Information Management—Vol* 19.2 (2011): 65.
13. Lemke, Frank, and Johann-Adolf Mueller. "Medical data analysis using self-organizing data mining technologies." *Systems Analysis Modelling Simulation*43.10 (2003): 1399-1408.
14. Bach, Mirjana Pejić, and Dijana Ćosić. "Data mining usage in health care management: literature survey and decision tree application." *Medicinski Glasnik* 5.1 (2008): 57-64.
15. Lu, Zhengwu, and Jing Su. "Clinical data management: Current status, challenges, and future directions from industry perspectives." *Open Access J Clin Trials* 2 (2010): 93-105.
16. Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." *International Journal of Computer Applications* 17.8 (2011): 43-48.
17. Kolçe, Elma, and Neki Frasheri. "A Literature Review of Data Mining Techniques Used in Healthcare Databases." (2012).
18. Rafe, Vahid, and Roghayeh Hashemi Farhoud. "A Survey on Data Mining Approaches in Medicine." (2013).
19. Markov, Zdravko, and Ingrid Russell. "An introduction to the WEKA data mining system." *ACM SIGCSE Bulletin*. Vol. 38. No. 3. ACM, 2006.
20. Hall, Mark, et al. "The WEKA data mining software: an update." *ACM SIGKDD Explorations Newsletter* 11.1 (2009): 10-18.
21. Bouckaert, Remco R., et al. "WEKA Manual for Version 3-7-8." (2013).
22. Bhatla, Nidhi, and Kiran Jyoti. "An Analysis of Heart Disease Prediction using Different Data Mining Techniques." *International Journal of Engineering* 1.8 (2012).
23. Sathyadevi, G. "Application of CART algorithm in hepatitis disease diagnosis."*Recent*
24. *Trends in Information Technology (ICRTIT), 2011 International Conference on*. IEEE, 2011.
25. Sa-ngasoongsong, Akkarapol, and Jongsawas Chongwatpol. "An Analysis of Diabetes Risk Factors Using Data Mining Approach."
26. Breault, Joseph L., Colin R. Goodall, and Peter J. Fos. "Data mining a diabetic data warehouse." *Artificial Intelligence in Medicine* 26.1 (2002): 37-54.
27. Phillips-Wren, Gloria, Phoebe Sharkey, and Sydney Morss Dy. "Mining lung cancer patient data to assess health care resource utilization." *Expert Systems with Applications* 35.4 (2008): 1611-1619.