ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Major Disease Diagnosis and Treatment Suggestion System Using Data Maining Techniques

[I]Mariammal.D, [II]Jayanthi.S, [III]Dr. P.S.K.Patra
[I]Dept. of CSE, Agni College of Technology, Thalambur,Chennai,Tamilnadu, India
[II]Asst.Prof, Dept. of CSE, Agni College of Technology, Thalambur, Chennai, Tamilnadu, India
[III]Head of the Dept., Dept. of CSE, Agni College of Technology, Thalambur, Chennai, Tamilnadu, India

Abstract
Disease diagnosis is one of the applications where data mining tools are providing successful results. All the disease(like Heart Disease, Cancer, HIV) are the leading cause of death all over the world. Some of the diagnostic and laboratory procedures are invasive, costly and painful to patients. Single Data Mining Technique in the diagnosis of all disease has been comprehensively investigated showing acceptable levels of accuracy. Still, using data mining techniques to identify a suitable treatment for all disease has received less attention. The gaps in the research on all disease diagnosis and treatment are identified in this paper and proposes a model to systematically close those gaps to discover if applying single and multiple data mining techniques to all disease treatment data can provide as reliable performance as that achieved in diagnosing disease. Using multiple data mining techniques the accuracy also improved.

Keywords
Disease Diagnosis, Data mining Techniques, accuracy, Treatment,performance

I. Introduction

Diversity of information made that useful data processing and acquisition to become very ample processes, this being the main cause for appearing and developing of data mining concept. As we know, data mining represents an analytical process that explore a very large data sets seeking for new patterns and relationships between variables, generalizing this relationships in a new model, formula, or decision tree. It is capable to discover new knowledge without previous hypothesis, the goal being to discover new, unexpected, unintuitive knowledge, analyzing data from different point of views and summarize them in new and useful information. Data mining has become a tool for improving the classical statistical tools used in future tendency's prediction. There have already been some tries to use this tool in Medicine. Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Applications of Data mining in healthcare include analysis of health care centers for better health policy-making and prevention of hospital errors, early detection of disease, prevention and preventable hospital deaths, more value for money and cost savings, and exposure of fraudulent insurance claims. Recent times researchers are using data mining techniques in the diagnosis of several diseases such as diabetes, stroke,cancer, HIV and heart disease.

Heart disease is the leading cause of death all over the world. Using Single Data Mining Technique in the diagnosis of heart disease has been comprehensively investigated showing acceptable levels of accuracy. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Detection of hidden patterns and relationships often goes unexploited. The advanced data mining techniques can help remedy this situation. This research has developed a prototype called Major Disease Prediction System (MDPS) using data mining techniques, namely, Association Rule Decision Trees, Naïve Bayes and Neural Network. The research results shows that each technique has its unique strength in realizing the objectives of the defined mining goals. MDPS can answer complex "what if" queries which traditional decision support systems cannot. Medical profiles such as age, sex, blood pressure and blood sugar are used to predict the likelihood of patients getting heart disease. This approach enables significant knowledge, e.g. patterns, relationships between medical factors related to the heart disease, to be established. MDPS is a Web-based, user-friendly, scalable, reliable and expandable.

Cancer disease is a class of diseases characterized by out-of-control cell growth. There are 100 different types of cancer, and each is categorized by the type of cell that is initially affected. Cancer Disease injuries the body when damaged cells divide uncontrollably to form lumps or masses of tissue called tumors (except in the case of leukemia where cancer prohibits normal blood function by abnormal cell division in the blood stream). Tumors can also grow and interfere with the digestive, nervous, and circulatory systems and they can release hormones that alter body functions. Tumors may stay in one spot and demonstrate limited growth are generally considered to be benign. The causes of cancer are diverse, complex, and only partially understood by all. Many things are harm to increase the risk of cancer, including tobacco use, dietary factors, certain infections, exposure to radiation, lack of physical activity, and obesity. These factors can directly damage genes or combine with existing genetic faults within cells to cause cancerous mutations. Breast cancer represents the second leading cause of cancer deaths in women today and it is the most common type cancer in women. Cancer detection based on the application of data mining techniques to proteomic data has received a lot of attention in recent years.

Human immunodeficiency virus (HIV) is a lent virus that causes acquired immunodeficiency syndrome (AIDS),a condition in humans in which progressive failure of the immune system allows life-threatening opportunistic infections and cancers to thrive. Every 24 hours, an estimated more than 7,000 people are infected with HIV, and more than 1 million contract a STI (sexually transmitted infection). Currently, an estimated 33.3 million people are living with HIV, and 23% of all people living with HIV are under age 24, while 35% of all new infections happen among people between 15 to 24 years of age.

The main objective of this paper is by applying the predictive data mining techniques (both single and multiple) to diagnosing the disease and also for the treatment. This paper proposes a system for diagnosis the disease and providing the suitable treatment

using data mining predictive techniques and also the accuracy is improved than other predictive systems. The rest of the paper is divided as follows: section II provides a related work on using data mining techniques to help health care professionals in the diagnosis of disease like heart disease, cancer, and HIV, section III shows the data mining techniques which are used in this diagnosis system, section IV explains about proposed work and performance improvement and section V concludes this paper.

II. Related Work

Knowledge of the risk factors associated with the major disease helps health care professionals to identify patients at high risk of having the disease. All the health care professionals store significant amounts of patients' data .Analyze these datasets are very important to extract useful knowledge. Data mining approach is an effective tool for analyzing data to extract useful knowledge. Various data mining techniques have been used to help health care professionals in the diagnosis of heart disease. The following works are showed some examples for disease diagnosis using major data mining techniques.

Jyoti Soni et al[1] proposed three different supervised machine learning algorithms. They are Naïve Bayes, k-NN, and Decision List algorithm. These algorithms have been used for analyzing the heart disease dataset. Data mining tool is used for classifying these data. These classified data is evaluated using 10 fold cross validation and the results are compared.

Mohammad Taha Khan et al[2] proposed a prototype model for the breast cancer as well as heart disease prediction using data mining techniques. The data used is the Public- Use Data available on web, consisting of 297 records for heart disease and 297 for breast cancer. There are two decision tree algorithms C4.5 and the C5.0 have been used on these datasets for prediction and performance of both algorithms is compared.

D. Chen, K. Xing et al [3] investigates the statistical analysis of the SEER data and computes survival percentage based on gender, race, geographic area, cancer stage, etc. used SEER data containing records of lung cancer patients diagnosed from 1988 through 1998. They examined the following attributes: AJCC stage, grade, histological type and gender. For each of the first three attributes, they considered four popular values that are generally used in lung cancer studies. The attribute gender had two values: male and female. This gave them 128 (4_4_4_2) possible combinations of values. They applied ensemble clustering on those combinations to get seven clusters and studied survival patterns of those clusters.

SEER data is used by D. Chen, K. Xing et al [4], for patients diagnosed of cancer of lung or bronchus from the year 1988 through 2001. They studied 8 months survivability of lung cancer. They compared penalized logistic regression and SVM for survival prediction of lung cancer, and found that logistic regression resulted in better prediction performance (in terms of <sensitivity, specificity> pair). They also note that SVM modeling is significantly slow, taking hours to train.

Vararuk et al[5]. Have studied the application of data mining techniques on HIV/AIDS data with the purpose of utilizing the data mining results for the management of HIV/AIDS. For the study a total of 2,50,000 records from HIV/AIDS patients in Thailand are used. IBM's Intelligent Miner is used for clustering and association rule discovery. As the researchers indicated, clustering is used in order to identify characteristics of categories of people affected by the disease whereas association rule mining is to identify symptoms that may follow a set of existing ones. The findings of Vararuk et al. Have showed clustering of patients with common characteristics and errors in the data.

A study conducted by Teklu[6] has attempted to investigate the application of data mining techniques on Anti-Retroviral Treatment (ART) service with the purpose of identifying the determinant factors affecting the termination/continuation of the service. This study applied classification and association rules using, J48 and apriori algorithms respectively, on 18740 ART patients' datasets. The methodology employed to perform the research work is CRISP-DM. Finally the investigator proved the applicability of data mining on ART by identifying those factors causing the continuation or termination of the service.

Maria-Luiza Antonie et al[7] analysed the Breast cancer is the second leading cause of cancer deaths in women today and it is the most common type of cancer in women. This paper also presents some experiments for tumour detection in digital mammography. They examine the use of different data mining techniques, neural networks and association rule mining, for anomaly detection and classification.Results are shows that the two approaches are performed well, obtaining a classification accuracy reaching over 70% percent for both techniques. Additionally, the experiments we conducted demonstrate the use and effectiveness of association rule mining in image categorization.

C4.5 is a well-known classification technique in decision tree induction which has been used by Abdelghani Bellaachia and Erhan Gauven[8] along with two other techniques i.e. Naïve Bayes and Neural Network. They conduct the investigation of the prediction of survivability rate of breast cancer patients using above data mining techniques and it is used in the new version of the SEER Breast Cancer Data. However, the author found the model generated by C4.5 algorithm for the given data has a much better performance than the other two techniques.

III. Data Mining Techniques Used in Diagnosis System

Classification technique is the most frequently used data mining task with a majority of the implementation of Bayesian classifiers, neural networks, and Association Rule. Myriad of quantitative performance measures were proposed with a high proportion of accuracy, sensitivity, specificity, and ROC curves. The latter are usually associated with qualitative evaluation.

Classification methods maps the data in to predefined targets. This approach is a supervised learning as targets are predefined. The goal of the classification method is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, these classifiers are used to predict the group attributes of new cases from the domain based on the values of other attributes. The following techniques are used in the disease diagnosis.

A. Decision Trees (DT's)

Decision tree is a tree where each non-terminal node represents a test or decision on the considered data item. Selection of a certain branch depends upon the outcome of the test. To classify a particular test data item, we start at the root node and follow the assertions down until we reach a terminal node or leaf node. Decision is made when a terminal node is approached. Decision trees that use recursive data partitioning can also be interpreted as a special form of a rule set [9]. The Decision Tree algorithm, is based on conditional probabilities and unlike naïve Bayes ,decision trees generate rules. Rules are the conditional statement that can

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

easily be understood by humans and easily used within a database to identify a set of records.

In some other applications of data mining, the accuracy of a prediction is the only thing that really matters. It may not be very important to know how the model works. In others, the ability to explain the reason for a decision can be crucial. For eg, a Marketing professional would need complete descriptions of customer segments in order to launch a successful marketing campaign. The Decision Tree algorithms are ideal for this type of application.

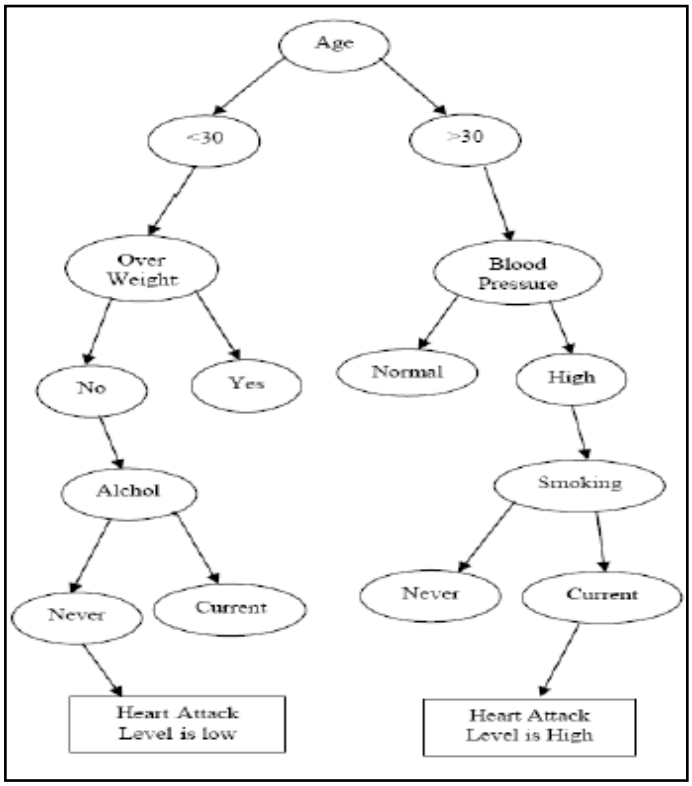

Fig. 3.1: Decision Tree

B. Neural Networks

Neural networks were recently the most popular artificial intelligence-based data modeling algorithm used in clinical medicine. Neural networks (NN) are those systems modeled based on the working of human brain. The power and speed of modern digital computers is truly astounding. Human cannot ever hope to compute a million operations a second. Still, there are some tasks for which even the most powerful computers cannot compete with the human brain, perhaps not even with the intelligence of an earthworm. This is a simple model and consists of a single `trainable' neuron. The term 'Trainable' means that its threshold and input weights are modifiable. Inputs are presented to the neuron and each input has a desired output (determined by us). If the neuron doesn't give the desired output, then it has made a mistake. To rectify this, its threshold and/or input weights must be changed. How this change is to be calculated is determined by the learning algorithm .Artificial Neural networks may be able to model complex non-linear relationships, comprising an advantage over simpler modeling methods like the Naïve Bayesian classifier or logistic regression.

C. Naive Bayes

The Microsoft Naive Bayes algorithm is a classification algorithm provided by Microsoft SQL Server Analysis Services for use in predictive modeling. The term(name) Naive Bayes derives from the fact that the algorithm uses Bayes theorem but does not take into account dependencies that may exist, and therefore its assumptions are said to be naive.

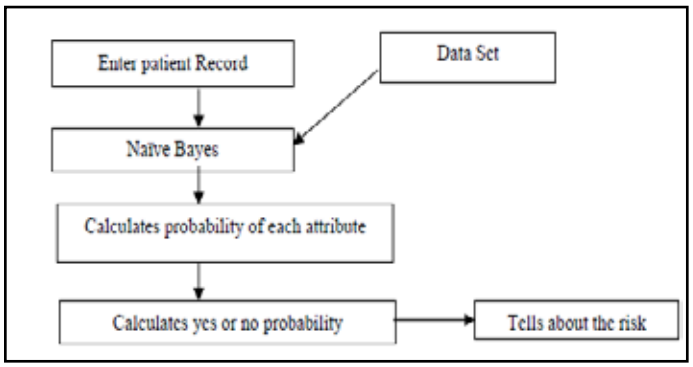

Fig. 3.2: Implementation of Naïve Bayes on the patient data

This algorithm is less computationally intense than other Microsoft algorithms, and therefore is useful for quickly generating mining models to discover relationships between input columns and predictable columns. You can use this algorithm to do initial explorations of data, and then later you can apply the results to create additional mining models with other algorithms that are more computationally intense and more accurate. The Microsoft Naive Bayes algorithm calculates the probability of every state of each input, given each possible state of the predictable column. You can use the Microsoft Naive Bayes Viewer in Business Intelligence Development Studio to see a visual representation of how the algorithm distributes states.

D. Association Rule

In the process of generating the class association rules, instead of considering all the attributes, apply PCA and rank all attributes. The attribute with highest ranking is used to generate the class association rules. This approach generates n*m rules for a single test instance with n non class attributes and m classes in the entire data set. If t Test cases are to be predicted the no. Of rules generated will be t*n*m. After identifying the principle 43 component attribute, the subsets are generated. For each generated subset, probability values are calculated. The decision of which class will be assigned to test instance X is based on the analysis of the subsets of attributes values with the highest posterior probabilities. Find the accuracy of the data set.

$$\text{Accuracy} = \frac{\text{no. of correctly predicted test data}}{\text{Total no. of test data}}$$

IV Proposed Work

The following steps are involved in the heart disease diagnosis and treatment suggestion system The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease is the major cause of casualties in the world. Record set with medical attributes was obtained. With the help of the dataset, the patterns significant to the heart attack prediction are extracted.

The attribute "Diagnosis" is identified as the predictable attribute with value "1" for patients with heart disease and value "0" for patients with no heart disease. In this module we analyze the records which are stored in the database. Originally 13 attributes

were involved in predicting the heart disease like age ,sex ,cp, chol, Trestbps, Fbs ,Restecg and so on. The data mining algorithms are applied to the training dataset. Naïve Bayes, neural networks, decision tree, regression, association rule are the data mining algorithms are used in this paper. Each and every algorithm is applied to the dataset to produce the patterns for the disease analysis.The architecture diagram for the proposed work is given below.

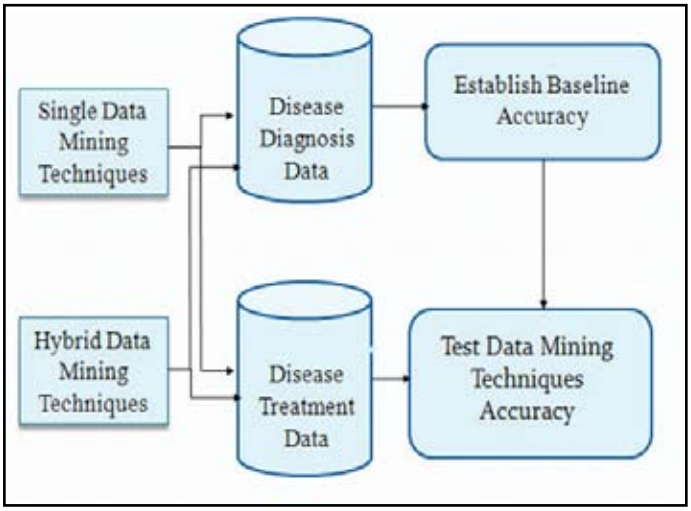

Fig. 4.1: System Architecture Design

After the implementation of algorithms into the dataset it will create the pattern for each disease dataset. Every algorithm will create pattern for all the disease in the database. Pattern formation is the process of creating the certain binary values according to the algorithm which is implemented. The primary goal of pattern recognition is supervised or unsupervised classification. Among the numerous frameworks in which pattern recognition has been traditionally formulated.

Input attributes for disease are get through the Questionnaires. Questionnaires have advantages over some other types of medical symptoms that do not require as much effort from the questioner as verbal or telephone analyses, and often have standardized answers that make it simple to compile data. Though, such standardized answers may frustrate handlers. Questionnaires are also sharply limited by the fact that respondents must be able to read the questions and respond to them.

Then, the input attributes are converted into some pattern by using data mining analysis algorithms. That patterns are matched with the patterns which are generated by the algorithm implementation in dataset by using mining processes. After mining process it will produce some binary values for both dataset and input attributes. The binary values are generated for each algorithm implementation. If the output is maximum number of 1 means that the person having the diseases. If the output is maximum number of 0 means that the person is not having the diseases. By using this algorithm implementation we can find accurate result of diagnosing the disease.

Then the probability of the disease is calculated by using some formula. By using that probability we are giving treatment suggestion for the patients. The probability may be varying for each patient according to that we can suggest the medicine for treatment of the particular disease.

V. Conclusion

Disease prediction is a major challenge in the health care industry. Instead of going for a number of tests, predicting the major disease with less number of attributes is a challenging task in Data Mining. Decision Support in Disease Prediction System is developed using all the five data mining techniques. The Disease diagnosis system extracts hidden knowledge from a historical disease database. This is the most effective model to predict patients with disease. This system could answer complex queries, each with its own strength with respect to ease of model interpretation, access to information and accuracy. Disease Prediction System is expanded for other diseases HIV, Lung cancer, Breast cancer and Stomach cancer also.It can be further enhanced and expanded. For e.g, it can incorporate other medical attributes besides the 15 listed. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules(AR). Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another one major challenge would be to integrate data mining and text mining. In future the disease prediction system may use the EEG signals for predicting the disease like brain disease.

References

[1] *Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer science and Engineering, Vol. 3, No. 6, June 2011.*

[2] *Mohammad Taha Khan, Dr. Shamimul Qamar and Laurent, F. Massin, "A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining", International Journal ofApplied Engineering Research, Vol. 7 No. 11 (2012), pp. 1-6.*

[3] *D.Chen, K. Xing, D. Henson, L. Sheng, A. Schwartz, and X.Cheng. "Developing prognostic systems of cancer patients by ensemble clustering". Journal of Biomedicine and Biotechnology, 2009*

[4] *F. D. "Machine learning methods in the analysis of lung cancer survival data". DIMACS Technical Report 2005-35 February 2006.*

[5] *Ruben, D.C.J.,"Data Mining in Healthcare: Current Applications and Issues". 2009.*

[6] *Maria-LuizaAntonie ,Osmar R. Za¨iane,AlexandruComan," Application of Data Mining Techniques for Medical Image Classification",Auguest26, 2001.*

[7] *Eldon Y. Li," Artificial neural networks and their business applications" Information & Management 27 (1994) 303-313*

[8] *Bellaachia Abdelghani and Erhan Guven,"Predicting Breast Cancer Survivability using Data Mining Techniques", Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining,06*

[9] *A.Shameem Fathima,D.Manimegalai and Nisar Hundewale," A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011.*