# A Survey on Various Classification Techniques for Clinical Decision Support System

Chaitali Vaghela
Student, M. Tech
Computer Engineering
CSPIT, Changa

Nikita Bhatt
Assistant Professor
U and P.U. Patel Dept. of Computer Engineering
CSPIT, Changa

Darshana Mistry
Technical Associate
Software Training
eiTRA, Ahmedabad

## ABSTRACT

Clinical Decision Support systems link health observations with health knowledge to determine health option by clinicians for improved health care. The main idea of Clinical Decision Support System is a set of rules derived from medical professionals applied on a dynamic knowledge. Data mining is well suited to give decision support for the healthcare. There are several classification techniques available that can be used for clinical decision support system. Different techniques are used for different diagnosis. In this paper, various classification techniques for clinical decision support system are discussed with example**.** .

## Keywords

Clinical Decision Support System, Data Mining, Classification.

## 1. INTRODUCTION

Clinical Decision Support System is a healthcare system, which is intended to assist physicians and other health professionals on decision making tasks. It can be also defined as “A computer system that uses two or more patient data to generate case specific or encounter specific advice.” [1]

Most CDSS comprises of three parts, the knowledge base, inference engine, and mechanism to communicate. [2] The knowledge base comprises of compiled information that is often, but not always, in the form of if–then rules. The inference engine comprises the expressions for combining the rules or associations in the knowledge base with actual patient data. Then a communication mechanism is used for bringing the patient data into the system and supplying the output of the system to the user who will make the actual decision

## 2. DATA MINING FOR CDSS

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not reinforced to determine hidden information for effective decision making. It is very difficult to discover hidden patterns and relationships. Data mining techniques can help in this situation.

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs.[3] The data generated in healthcare organization are too complex and copious. So it is very difficult to process and analyze those data by traditional methods. Data mining offers the methodology and technology to transform these piles of data into useful information for decision making. Healthcare providers can utilize data mining to uncover previously unknown patterns from vast data stores and then apply this information to build predictive models.

## 3. DATA MINING TECHNIQUES

Data mining techniques have been widely used in clinical decision support systems for prediction and diagnosis of various diseases with better accuracy. These techniques have been very effective in developing clinical support systems because they are able to detect hidden patterns and relationships in medical data.

Classification is a data mining techniques that assigns items in a collection to target categories or classes. The aim of classification is to predict the target class for each case in the data accurately. Classification is important when a data repository contains samples that can be used as the basis for future decision making. There are several classification techniques which can be used for clinical decision support system. In this section various classification techniques are discussed.

### 3.1 Bayesian Belief Network

The Bayesian network is a knowledge-based graphical representation that shows a set of variables and their probabilistic relationships between diseases and symptoms. Bayesian network is utilized to find the probability of the presence of possible diseases given their symptoms. Its advantage is that it requires the knowledge and conclusions of experts in the form of probabilities. But it is not practical for large complex systems given multiple symptoms.

Iliad system developed by University of Utah School of Medicine's Dept. of Medical Informatics is a CDSS based on Bayesian network which applies Bayesian reasoning to calculate posterior probabilities of possible diagnoses depending on the symptoms provided. Iliad was developed primarily for diagnosis in Internal Medicine now covers about 1500 diagnoses in this domain, based on several thousand findings. [4]

DXplain is also a CDSS that uses a modified form of the Bayesian logic. It produces a list of ranked diagnoses associated with the symptoms. It is very useful for the physician who has no computer expertise. It also serves as a clinician reference with a searchable database of diseases and clinical manifestations. [5]

Another example is SimulConsult, which uses Bayesian systems to input data in a scalable way and determine probabilities, accomplishing it by focusing specialty by specialty. It applies a statistical pattern-matching approach that considers the age of onset and offset of the findings in each disease. [6]

### 3.2 Neural Network

Neural Networks is a non-knowledge-based CDSS that is adaptive. It allows the systems to learn from existing knowledge and experiences. Neural Network has three main layers: Input, Output and Hidden layer. Neural Network is made of nodes called neurons. And there is weighted connection between nodes of different layers, which is used to transfer signals between the nodes. Neural Network is able to continue with incomplete data that gives educated guesses about missing data and get improved with every use due to its adaptive system learning.

Mr. P. A. Kharat et al 2011 proposed clinical decision support system based on Jordan/Elman neural network for the diagnosis of epilepsy and they obtained relatively high overall accuracy for training data 99.83% and for cross-validation data and testing data 99.92% [7]

Mrudula Gudadhe and Kapil Wankhade et al 2010 designed a decision support system based on neural network for heart diseases classification and they classified the data into 5 categories of heart disease with 97.5% accuracy by using multilayer perceptron with back propagation training algorithm. [8]

R.R.Janghel et al 2009 developed a CDSS using artificial neural network to predict the fetal delivery to be done normal or by surgical procedure. In that system, they used three different training algorithms to train the neural network, which are Back Propagation algorithm, Radial Basis function and Learning vector quantization Network and they were able to gain accuracy of 93.75%, 99% and 87.5% respectively. [9]

## 3.3 Decision Tree

Decision tree is the most often used techniques of data analysis. It is applied to classify records to a proper class. In medical field decision trees determine the sequence of attributes. First it makes a set of solved cases. Then the whole set is divided into training set and testing set. A training set is used for the induction of a decision tree. A testing set is used to find the accuracy of an obtained solution.

AY AI-Hyari et al 2013 developed a CDSS for diagnosing patients with Chronic Renal Failure using various classification methods like neural network, naïve bays and decision tree. They proved that Decision tree algorithm is the most accurate CRF classifier (92.2%) when compared to all other algorithms/implementations involved in their study. [10]

M. Maity et al 2012 developed a system to assist pathologist by giving support of automated anemia diagnosis and computerized report generation. They applied advanced image processing algorithm and data mining approach to analyze patient medical information. They applied supervised decision tree classifier C4.5 to classify image samples with sensitivity of 98.1% and specificity of 99.6%. [11].

## 3.4 Naïve Bays

Naïve Bays allows selecting the kernel estimator for numeric attributes rather than a normal distribution and utilized Supervised Discretization while converting numeric attributes to normal ones. Naïve Bayes classifier gives output in text form. Advantages of Naïve bays are that it is simple and efficient and it gives better classification performance.

Abeer Y. Al-Hyari et al 2013 designed a CDSS for prediction and diagnosis of Chronic Renal Failure (CRF) using naïve bays. The implemented CDSS can be used to observe the progression stage of the disease in patient. They were able to achieve accuracy of 88.2% using naïve bays algorithm. [12]

Mrs.G.Subbalakshmi et al 2011 developed a CDSS for heart disease prediction. Their system extracts hidden knowledge from a historical heart disease database. They claimed that it is the most effective model to predict patients with heart disease. [13]

## 3.5 Support Vector Machine

Support vector machine (SVM) has become more and more popular tool for machine learning tasks involving classification, regression etc. SVM is supervised learning model that is applied for classification. SVM serves as the linear separator between two data points to identify two different classes in the multidimensional environment. SVM separate the data into two categories of performing classification and constructing an N-dimensional hyper plane. SVM algorithms are binary; therefore in the case of multi-class problem one must reduce the problem to a set of multiple binary classification problems.

Hui-Ling Chen et al 2011 proposed a rough set based support vector machine classifier to diagnosis breast cancer. They applied rough set for feature selection and SVM for classification. They attained very high classification accuracy of 99.41% for 50–50% of training-test partition, 100% for 70–30% of training-test partition, and 100% for 80–20% of training-test partition. And they were also able to discover a combination of five informative features, which can be important to the physicians for breast diagnosis. [14]

Mrudula Gudadhe et al 2010 presented a DSS for heart disease classification based on SVM. They classified the heart disease data into two classes that indicates presence of heart disease or absence of heart disease with 80.41% accuracy. [8]

## 3.6 Fuzzy Set Approach

Fuzzy set theory is useful for data mining systems performing rule-based classification. It gives operations for combining fuzzy measurements. The Fuzzy Logic Rule based classifier is very effective in high degree of positive predictive value and diagnostic accuracy. Fuzzy Logic is a type of multi-valued logic derived from fuzzy set theory to deal with approximate reasoning. Aniele C. Ribeiro et al 2014 proposed fuzzy breast cancer system to map two controlled and two non-controlled input variable into the risk of breast cancer occurrence. It can provide health support to predict measurement of developing breast cancer to the female population and the health authorities, to reduce both the outcomes and mortality rate. [15]

Chang-Shing Lee et al 2011 presented a novel five-layer fuzzy ontology to model the domain knowledge with uncertainty and extend the fuzzy ontology to the diabetes domain. They proved that the proposed method works more effectively for diabetes application than previously developed ones. [16]

Markos G. Tsipouras et al 2007 presented a methodology for the automated development of fuzzy expert systems (FES). They proposed methodology is tested by applying it to problems related to cardiovascular diseases. The FES indicates significant improvement of the initial classification system, which is based on expert’s knowledge and has the form of a set of rules. The obtained results are also fully interpretable, which is a major advantage compared to other approaches. [17]

## 3.7 Genetic Algorithm

Genetic Algorithm is proposed in the 1940s at the Massachusetts Institute of Technology based on Darwin’s evolutionary theories that dealt with the survival of the fittest.

In this algorithm, the system goes through an iterative process to get an optimal solution. Similar to neural networks, the genetic algorithms derive their information from patient data. Because of its lacks of transparency in the reasoning involved for the decision support systems, it is unwanted by physicians. But still it is being used in clinical decision support with other algorithm for improved result.

S.U. Amin et al 2013 presented a hybrid system of genetic algorithm and neural network for prediction of heart disease using major risk factors. They used the global optimization benefit of genetic algorithm to initialize the weight of neural network. [17]

AZ Shabgahi et al 2011 proposed cancer detection on Global Cancer Map dataset by creating fuzzy rule with genetic algorithm. [18]

## 3.8 Rough Set Approach

A Rough Set is determined by a lower and upper bound of a set. Rough set theory provides mathematical tools to determine hidden patterns in data that can be used in data mining. The lower and upper bound is chose based on selection of attributes. Therefore it may not be applicable for some application. It does not need any preliminary or extra information concerning data.

## 3.9 K-Nearest Neighbor

K-Nearest neighbor classifies item based on nearest training data in the feature space. It is a type of instance base learning or lazy learning. It is very simple but its accuracy can be affected by noisy or irrelevant features.

# 4. RESULTS

Different Classification technique for different dataset is applied for better comparison. The WEKA tool is used for that.

## 4.1 Breast Cancer Data

Instances: 286

Attributes: 10

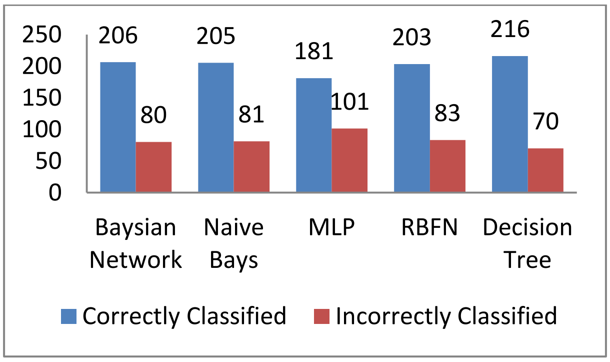

**Figure 1: Classification of Breast Cancer Data**

For Breast cancer data, Decision Tree gives better result, while neural network gives poor result. And all other classifier gives average result. We can say that for smaller dataset decision tree algorithm is better.

## 4.2 Heart Disease Data

Instances: 303

Attributes: 14

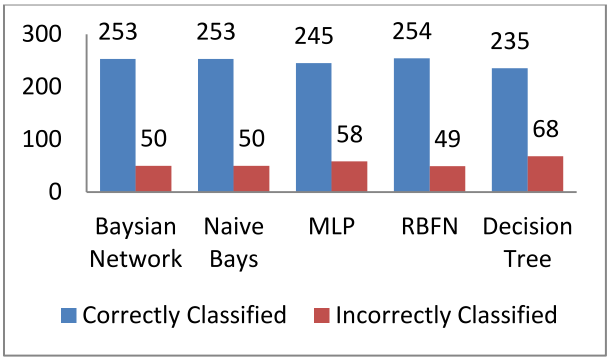

**Figure 2: Classification of Heart Disease Data**

For Heart disease data, Neural Network (Radial Base Function Network) gives good result, Decision tree gives poor result and all other classifier gives average result. We can say that Neural Network gives better result with more number of attributes.

## 4.3 Diabetes Data

Instances: 768

Attributes: 9

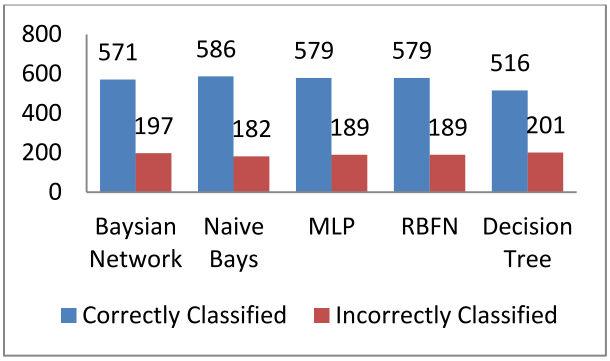

**Figure 3: Classification of Diabetes Data**

For diabetes data, Naïve Bays gives better result, Decision tree gives poor result and all other classifier gives an average result. Naïve bays give better result for large number of instances.

# 5. CONCLUSION

The main goal of this survey was to describe the most common classification algorithms of data mining, utilized in Clinical Decision Support Systems, and assess their performance on several medical datasets. Here we observed that result of different classifiers is depend on various factor like dataset, number of attributes in dataset, number of class variable, number of missing values in dataset. The selection of classifier for clinical decision support is application dependent.

# 6. REFERENCES

[1] Wyatt JC, Liu JL. J Epidemiol Community Health 2002; 56(11): 808-812, "Basic concepts in medical informatics"

[2] Characteristics of Clinical Decision Support System, en.wikipedia.org/wiki/Clinical_decision_support_system

[3] David Crockett, Ryan Johnson, and Brian Eliason, Analytics in health care, "What is Data mining in HealthCare?"

[4] Decision Support System, Iliad, www.openclinical.org/aisp_iliad.html

[5] Barnett GO, Cimino JJ, Hupp JA, Hoffer EP, JAMA. July 3, 1987. “DXplain. An evolving diagnostic decision-support system”.

[6] A simultaneous consult on your patient’s diagnosis, Simulconsult, www.simulconsult.com/

[7] Mr. P. A. Kharat, Dr. S. V. Dudul, IEEE 2011, Clinical Decision Support based on Jordan/Elman Network”

[8] Mrudula Gudadhe, Kapil Wankhade and Snehlata Dongre, IEEE 2010, Decision Support System for heart disease using support vector machine and artificial neural network”

[9] R. R. Janghel, Anupam Shukla and Ritu Tiwari, IEEE 2009, “Clinical Decision Support System for Fetal Delivery using artificial neural network”

[10] AY AI-Hyari, A. M. Al-Taee and M. A. Al-Taee, IEEE 2013, “Clinical Decision Support for diagnosis and management of Chronic Renal Failure”

[11] M. Maity, P. Sarkar and C. Chakraborty, IEEE 2012, “Computer-assisted approach to anemic erythrocyte classification using blood pathological information”

[12] Abeer Y. Al-Hyari, Ahmad M. Al-Taee and Majid A. Al-Taee, IEEE 2013, “Clinical Decision Support System for Diagnosis and Management of Chronic Renal Failure”

[13] G.Subbalakshmi, K. Ramesh and M. Chinna Rao, IJCSE 2011, “Decision Support in Heart Disease Prediction System using Naive Bayes”

[14] Hui-Ling Chen, Bo Yang, Jie Liu and Da-You Liu, Springer 2011, “A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis”

[15] Aniele C. Ribeiro, Deborha P. Silva and Ernesto Araujo, IEEE 2014, “Fuzzy Breast Cancer Risk Assessment”

[16] Chang-Shing Lee and Mei-Hui Wang, IEEE 2011, “A Fuzzy Expert System for Diabetes Decision Support Application”

[17] S. U. Amin, K. Agarwal and R. Beg, IEEE 2013, “Genetic neural network based data mining in prediction of heart disease using risk factors”

[18] AZ Shabgahi and MS Abadeh, IEEE 2011, “Cancer Tumor Detection by Gene Expression Data Exploration using genetic Fuzzy System”