S.J Gnanasoundhari *et al*, International Journal of Computer Science and Mobile Applications,
Vol.2 Issue. 2, February- 2014, pg. 72-77 **ISSN: 2321-8363**

# A Survey on Heart Disease Prediction System Using Data Mining Techniques

**S.J Gnanasoundhari[1], G.Visalatchi[2], Dr.M.Balamurugan[3]**

**[1]School of Computer Science and Engineering, Bharathidasan University**
**Tiruchirappalli-620023, India**
**anubaburajan@gmail.com**

**[2]School of Computer Science and Engineering, Bharathidasan University**
**Tiruchirappalli-620023, India**
**Visamphil88@gmail.com**

**[3] School of Computer Science and Engineering, Bharathidasan University**
**Tiruchirappalli-620023, India**
**mmbalamurugan@gmail.com**

## Abstract
This paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques which will be useful for medical practitioners to take effective decision. The objective of this research work is to predict more accurately the presence of heart disease with reduced number of attributes. Four classifiers like Naive Bayes, Neural network, WAC algorithms are used to predict the diagnosis of patients, whereas WAC provides the accurate result when compared to other algorithms.

***Keywords:*** Naïve bayes, WAC, Neural Network, Support Vector Machine

## I. INTRODUCTION

Data mining used to gather, store, analyze and integrate biological information which can then be used for discovery and development. Medical Data mining in healthcare is regarded as an important yet complicated task that needs to be executed accurately and efficiently. Healthcare data mining attempts to solve real world health problems in diagnosis and treatment of diseases. This survey paper aims to analyze the several data mining techniques proposed in recent years for the diagnosis of heart disease. Many researchers used data mining techniques in the diagnosis of diseases such as tuberculosis, diabetes, cancer and heart disease, in which several data mining techniques are used in the diagnosis of heart disease such as KNN, Neural Networks, and Bayesian classification. Classification based on clustering, Decision Tree, Genetic Algorithm, Naive Bayes, Decision tree, WAC which are showing accuracy at different levels. Using medical profile such as age, sex, blood pressure and blood sugar we can easily predict the likelihood of patients getting heart disease. In this paper we have evaluated the performance of new classification approach that uses the experienced Doctor's knowledge to assign the weight to each attribute. More weight is assigned to the attribute having high impact on disease prediction.

## II. DATA MINING TECHNIQUES

### 1. NAÏVE BAYES

Naïve Bayes Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. A naive Bayes classifier is a term dealing with a simple probabilistic classification based on

 

applying Bayes' theorem. In simple terms, a naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting [1]. Naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Here independent variables are considered for the purpose of prediction or occurrence of the event. The algorithm is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. For example, a patient may be observed to have certain symptoms. Based on the observation, Bayes' theorem can be used to compute the probability that a proposed diagnosis is correct.

Baye's Theorem finds the probability of an eventoccurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes′ theorem can be st ated as follows.

$$`P (B given A) = Prob(A and B) / Prob (A)$$

To calculate the probability of B given A,
the algorithm counts the number of cases where A and B occur together and divides it by the number of cases where A occurs alone.

The algorithm works as follows:

- Each data sample is represented by an n dimensional feature vector, X = (x1, x2….. xn), depicting n measurements made on the sample from n attributes, respectively A1, A2, An.

- Suppose that there are m classes, C1, C2……Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive probability assigns an unknown sample X to the class Ci if and only if:

$$P (Ci/X)>P (Cj/X) \text{ for all } 1<=j<=m \text{ and } j!=i$$

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem,

$$P (Ci/X) = (P(X/Ci)P(Ci))/P(X)$$

- As P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1) = P(C2) = …..= P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci) = si/s , where Si is the number of training samples of class Ci, and s is the total number of training samples.

## 2. WEIGHTED ASSOCIATIVE CLASSIFIER(WAC)

Weighted Associative Classifier is a new concept that uses Weighted Association Rule for classification. Weighted ARM uses Weighted Support and Confidence Framework to extract Association rule from data repository. The WAC has been proposed as a new Technique to get the significant rule instead of flooded with insignificant relation. The steps are as follows:

- Initially, the heart disease data warehouse is pre -processed in order to make it suitable for the mining process.
- Each attribute is assigned a **weight** ranging from o to 1 to reflect their importance in prediction model .Attributes that have more impact will be assigned a high weight(nearly 0.9)and attributes having less impact are assigned low weight(nearly 0.1) .
- Once the pre-processing gets over, Weighted Association Rule Mining (WARM) algorithm is applied to generate interesting pattern. This algorithm uses the concept of Weighted Support and Confidence framework instead of tradition support and confidence. Rules generated in this step are known as CAR (Classification Association Rule) and is represented as X →Class label where X is set of symptoms for the disease.

 

- These rules will be stored in **Rule Base.**
- Whenever a new patient's record is provide, the CAR rule from the rule base is used to predict the class label.

Weighted Associative Classifiers is another concept that assigns different weights to different features and can get more accuracy in predictive modelling system like medical field etc. In any prediction model all attributes do not have same importance in predicting the class label. So different weights can be assigned to different attributes according to their predicting capability. . A weighted associative classifiers consists of training dataset T={r1, r2, r3…. ri…} with set of weight associated with each {attribute, attribute value} pair. Each ith record ri is a set of attribute value and a weight wi attached to each attribute of ri tuple / record. In a weighted framework each record is set of triple {ai, vi, wi} where attribute ai is having value vi and weight wi, 0<wj<=1. Weight is used to show the importance of the item.

**Attribute set weight:** Attribute weight is assigned depending upon the domain. Weight of attribute set X is denoted by W(X) and is calculated as the average of weights of enclosing attribute. And is given by

$$W(X) = \frac{\sum_{i=1}^{|X|} \text{weight}(a_i)}{\text{Number of attribute in X}}$$

**Record weight/Tuple Weight:** The tuple weight or record weight can be defined as type of attribute weight. It is average weight of attributes in the tuple. If the relational table is having n number of attribute then Record weight is denoted by $W(r_k)$ and given by

$$W(r_k) = \frac{\sum_{i=1}^{|rk|} \text{weight}(ai)}{\text{Number of attributes in a record}}$$

**Weighted Confidence:** Weighted Confidence of a rule X→Y where Y represents the Class label can be defined as the ratio of Weighted Support of (X∪Y) and the Weighted Support of (X).

$$\text{Weighted confidence} = \frac{\text{Weighted support}(X \cup Y)}{\text{Weighted support}(X)}$$

## 3. NEURAL NETWORK (NN)

A neural network (NN) is a parallel, distributed information processing structure consisting of multiple numbers of processing elements called node, they are interconnected via unidirectional signal channels called connections. Each processing element has a single output connection that branches into many connections; each carries the same signal i.e. the processing element output signal. The NN can be classified in two main groups: Supervised learning and unsupervised learning. In medical field, decision making is done by neural network because they provide more accurate results. Decision support system is developed for predicting heart disease of a patient. The prediction is done based on historical heart disease database. The technique used to develop system is Multilayer Perceptron Neural Network (MLPNN) with Back propagation algorithm (BP).

The working of multilayer perceptron neural network is summarized in steps as mentioned below:

- Input data is provided to input layer for processing, which produces a predicted output.
- The predicted output is subtracted from actual output and error value is calculated.
- The network then uses a Back propagation algorithm which adjusts the weights.
- For weights adjusting it starts from weights between output layer nodes and last hidden layer nodes and works backwards through network.
- When back propagation is finished, the forwarding process starts again.
- The process is repeated until the error between predicted and actual output is minimized.

 

The most widely used training algorithm for multilayer and feed forward network isBackpropagation. The name given is back propagation because, it calculates the difference between actual and predicated values is propagated from output nodes backwards to nodes in previous layer. This is done to improve weights during processing.

The working of Back propagation algorithm is summarized in steps as follows:

- Provide training data to network.
- Compare the actual and desired output.
- Calculate the error in each neuron.
- Calculate what output should be for each neuron and how much lower or higher output must be adjusted for desired output.
- Then adjust the weights.

MLPNN model proves the better results and helps the domain experts and even person related with the field to plan for a better diagnose and provide the patient with early diagnosis results as it performs realistically well even without retraining. Neural networks are known to produce highly accurate results in practical applications. Neural networks have been successfully applied to a variety of real world classification tasks in industry, business and science [2]. Also they have been applied to various areas of medicine, such as diagnostic aides, medicine, biochemical analysis, image analysis, and drug development**.** They are used in the analysis of medical images from a variety of imaging modalities. Artificial neural networks provide a powerful tool to help doctors analyze, model, and make sense of complex clinical data across a broad range of medical applications [3-8].As the volume of stored data increases, data mining techniques assume an important role in finding patterns and extracting knowledge to provide better patient care and effective diagnostic capabilities. Neural networks can be used to extract rules from a disease classification. From the rules system so discovered, we can predict if someone will have a particular stage of a particular disease.

## 4. SUPPORT VECTOR MACHINE(SVM)

**S**upport Vector Machine (SVM) is a set of related supervised learning method used in medical diagnosis for classification and regression. SVM can be used for pattern classification and nonlinear regression. Support Vector Machines (SVM's) are a relatively new learning method used for binary classification. SVM simultaneously minimize the empirical classification error and maximize the geometric margin [9]. SVM is called Maximum Margin Classifiers and it can be efficiently perform non-linear classification using kernel trick. The RBF (Radial Basis Function) kernel of SVM is used as the Classifier, as RBF kernel function can analyze higher-dimensional data. RBF kernel function is used because [10] RBF kernel nonlinearly maps samples into a higher dimensional space and also it has less numerical difficulties. The values fed to the SVM classifier are normalized initially to improve the accuracy. Test data sets were used to assess the performance of the SVM model. Validation using the test data sets avoid potential bias of the performance, estimate due to over-fitting of the model to training data sets. The SVM classifier with RBF kernel is used for classification. An automated classifier for the discrimination between the person with heart disease and without heart disease has been developed using supervised learning algorithm named SVM. The support vector machine can provide good generalization performance on pattern classification problem[11]. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data.

 

## III. PERFORMANCE OF CLASSIFIERS

Table 1. Performance of Naïve Bayes, WAC, NN, SVM

| DM Techniques | Accuracy |
|---|---|
| Naïve Bayes | 52.33 |
| WAC | 81.51 |
| Neural Network (NN) | 78.43 |
| Support Vector Machine (SVM) | 60.78 |

From Table I, it is been proved that WAC[14]provides an accurate result when compared to other classification techniques Naïve Bayes[12], Neural Network[13] and Support Vector Machine[13].

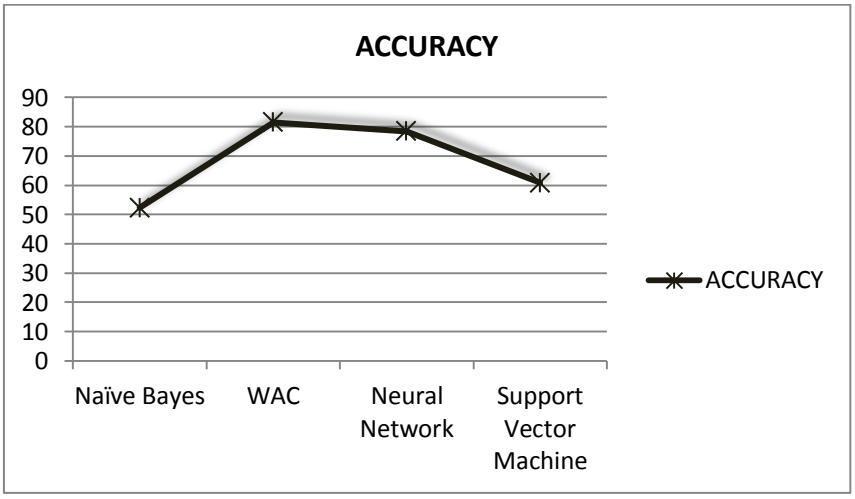


Figure 1: Graph shows accuracy for data mining techniques

## IV. CONCLUSION

This survey paper is developed using four data mining classification modelling techniques. This system extracts hidden knowledge from a historical heart disease database. Here four data mining classification techniques were applied namely Naive Bayes, Neural Networks, Support Vector Machine and Weighted associative Classifier. The models are trained and validated against a test dataset. Classification methods are used to evaluate the effectiveness of the models. . From the result it is been proved that Weighted Associative Classifier provides the accurate result compared to other techniques. The four models are able to extract patterns in response to the predictable state. This system can be further enhanced and expanded. It can also incorporate other data mining techniques like Time Series, Clustering and Association Rules.

## REFERENCES

[1] Ho, T. J.: "Data Mining and Data Warehousing", Prentice Hall, 2005.
[2] B.Widrow, D. E. Rumelhard, and M. A. Lehr, (1994) "Neural networks: Applications in industry, business and science," Commun. *ACM*, vol. 37, pp.93–105.
[3] W. G. Baxt, (1990) "Use of an artificial neural network for data analysis in clinical decision making:The diagnosis of acute coronary occlusion," Neural Comput., vol. 2, pp. 480–489.

 

[4] Dr. A. Kandaswamy, (1997) "Applications of Artificial Neural Networks in Bio Medical Engineering", The Institute of Electronics and Telecommunicatio Engineers, Proceedings of the Zonal Seminar on Neural Networks, Nov 20-21.
[5] A. Kusiak, K.H. Kernstine, J.A. Kern, K A. McLaughlin and T.L. Tseng, (2000) "Data mining: Medical and Engineering Case Studies", Proceedings of the Industrial Engineering Research Conference, Cleveland, Ohio, May21-23,pp.1-7.
[6] H. B. Burke, (1994) "Artificial neural networks for cancer research: Outcome prediction," Sem.Surg. Oncol., vol. 10, pp. 73–79.
[7] H. B. Burke, P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, F. E. Harrell, J. R.Marks, D. P. Winchester, and D. G. Bostwick, (1997) "Artificial neural networks improve the accuracy of cancer survival prediction," Cancer, vol. 79, pp. 857–8621997.
[8] Siri Krishan Wasan1, Vasudha Bhatnagar2 and Harleen Kaur, (2006)" The impact of Data Mining Techniques on Medical Diagnostics", Data Science Journal, Volume 5, 119-126.
[9] Boser, B. E., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin Classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pp.144 -152. ACM Press. 1992.
[10] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer Verlag. 1995.
[11] Christopher J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, Springer, 2(2), pp.121-167, 1998.
[12] A. Rajkumar and G. S. Reena, "Diagnosis of Heart Disease Using Datamining Algorithm", Global Journal of Computer Science and Technology, vol. 10, no. 10, **(2010)**.\
[13] Y. Kangwanariyakul, C. Nantasenamat, T. Tantimongcolwat and T. Naenna, "Data Mining of Magneto cardiograms For Prediction of Ischemic Heart Disease", EXCLI Journal, **(2010)**.
[14] Jyoti Soni et al. "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers" International Journal on Computer Science and Engineering (IJCSE), Vol.3, No.6, June (2011).

 