ISSN No. 0976-5697

Volume 5, No. 2, March 2014 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Study on Data Mining Classification Algorithms For Medical Data

P.R.Sudha Rani
Sr.Assistant Professor
Dept. of Computer science and Engineering
Sri Vishnu Engineering College For Women
Bhimavaram, India
sudharani.p@gmail.com

M.R.Narasinga Rao
Professor
Dept. of Computer Science and Engineering
KL University Vijayawada, India.
Ramanarasingarao@kluniversity.in

D.T Vijaya Lakshmi
Dept. of Computer Science and Engineering
Sri Vishnu Engineering College For Women Bhimavaram, India
djoee999@gmail.com

Abstract: This paper briefly describes various classification Algorithms used in medical Data. We introduce four widely used supervised methods. They are Artificial Neural Network, Bayesian classifiers, Decision Trees, Support Vector Machines.

Keywords: Artificial Neural Network(ANN), Bayesian classifiers, Decision Trees(DT), Support Vector Machines(SVM).

I. INTRODUCTION

The process of analyzing data from different perspectives and summarizing it into useful information IS the primary goal of Data mining[1]. Classification is one of the important technique used in data mining.

Classification predicts group membership for data instances.this method used in weather forecasting, health care, medical, financial, homeland security and business intelligence.popular Artificial Neural Networks ,Bayesian classifiers, Decision Trees,Support Vector Machines are the popular Classification Algorithms.

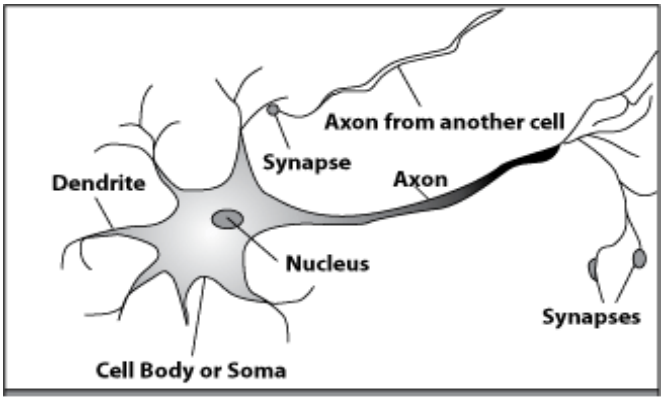
This paper includes problems related to medical domain used by different Classification Algorithms.

II. ALGORITHMS

A. Artificial Neural Network (ANN):

ANN model consist of simple processing units that communicates by sending signals to one another over a large number of weigted connections.They were originally developed from inspiration of human brains on how our own nervous system functions.ANN used in solving problems like facial recognition, which our biological brains can do easily[2].

It works same as neuron cell that processes information in the human brain [3]. The neuron cell body contains the nucleus, the axon and the dendrites. The axon transmits signals to other neurons where as the dendrites receive incoming signals from other neurons. Every neuron is connected and communicates through the pulses as shown in(Figure 3) [3]. The nodes are the artificial neuron and the directed edges represented the connection between output neurons and the input neurons.



Dendrites: Input
Cell body: Processor
Synaptic: Link
Axon: Output as shown in Figure 2. A Sketch of a Neuron in the Human Brain [3].

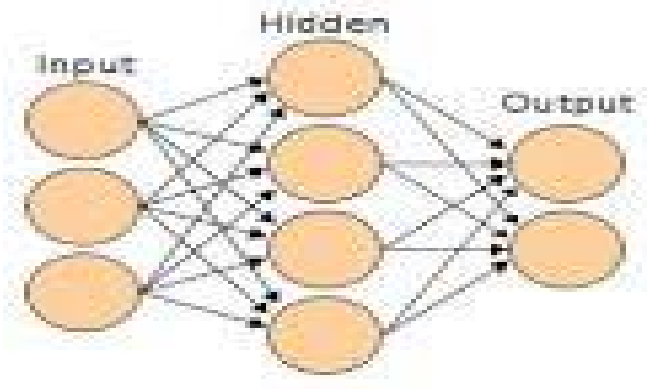


Figure 3. A Neural Network Model. [3].

B. Decision Tree:

The decision tree is one of the popular classification algorithm in the information systems [4]. A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. Decision tree algorithms

CONFERENCE PAPER
Two day National Conference on Advanced Trends and Challenges in Computer Science and Applications
Organized by: Shree Vishnu Engineering College for Women, Bhimavaram A.P.
Schedule: 18-19 March 2014

include Iterative Dichotomiser 3 (ID3), assistant algorithm, C4.5, C5, and CART [4, 5]. Through repeated observations a tree is constructed to predict data. To Identify a variable and threshold the splitting algorithms include ID3,C4.5(InformationGain),CART(gini Index),CHAID(Chi-squared test) and then split two or more sub groups according to the input observation and continues until the tree is built[4] as shown in (figure 4).

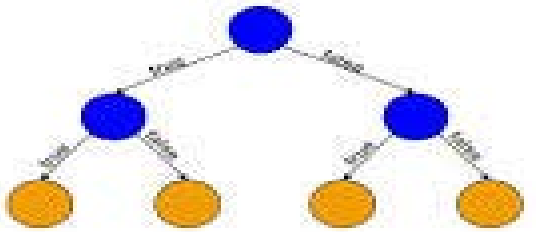


Figure 4. A Structure of a Decision Tree.

C. *Bayesian Classifiers:*

The Bayesian classifiers model consists of the set of conditional probabilities [6]. As shown in (Figure 5) is represented as a directed graph where the nodes represent attributes and arcs represent attribute dependency.

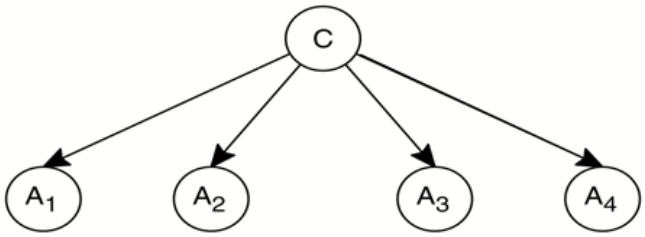


Figure 5. A Representation of a Bayesian Classifier Structure [6].

Classifiers are constructed in classification by using a set of training examples.The classifier of a general Bayesian network includes :

$$c(E) = \arg\max_{c \in C} P(c)P(a_1, a_2, \ldots, a_n|c).$$

The above formula includes the nodes or attributes from a1 to an, and variable C represents the class node and the value of C and c(E), where c(E) is the class of E [7].

All the attributes have to be assumed as independent in Navie bayes. the definition of Naïve Bayes is as shown below:

$$c(E) = \arg\max_{c \in C} P(c) \prod_{i=1}^{n} P(a_i|c).$$

The attributes a are independent attributes [7]. For example, if child X to be determined from class. Then, child X going to be predicted from class H.To calculate P (H|X) using Bayesian networks the formula is as follows:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} .$$

P(H) is the probability of class H, P(X|H) is the posterior probability that X is conditioned on H, and P(X) is the probability of X [7]. The training data obtains from these values.with the assumption that all the attributes are independent of each other for data X is as follows:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k, C_i) .$$

P(X1|Ci), P(X2|Ci),…,P(Xn|Ci) are calculated from the training samples [7].

D. *Support Vector Machine:*

In statistical learning theory the Support Vector Machine (SVM) is a classification algorithm [8]. It analyze data and recognize patterns, used for classification. It can capture nonlinearity in the data by providing similar model.By maximizing the margin and separating both classes and minimizing the classification errors are performed in classification tasks[8]. SVM training set includes optimization of a convex cost function where the learning process is not complicated by local minima [9]. The testing Includes to classify a test dataset and the performance is based on error rate determination [14]. For a training set of l samples, the learning procedure is as follows[7]:

$$\min_{\alpha} : \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{l} \alpha_j .$$

$$\text{s.t.} \quad 0 \le \alpha_i \le C, i = 1, \cdots, l .$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 .$$

For the ith sample xi ,the label is yi[7].The Langrangian multiplier of xi is ai. K(xi,xj) is the kernel and C is the upper bound of ai. If the samples having a > 0 are called support vectors [7]. The decision function is as shown in below, where the number of support vectors(ns) [7] :

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n_s} y_i \alpha_i^* K(x_i, x) + b^*\right) .$$

In this section, we have introduced four famous data mining techniques. In the subsequent section, the importance of data mining in medical systems are briefly discussed.

III. APPLICATION OF DATAMINING IN MEDICAL DATA

Physicians, Medical researchers and health care providers face the problem to use stored data effectively when large database stored in the medical information system such as in patient records, physician diagnosis, and monitoring information [10].

The process of making decisions are done by the help of medical decision support in the field of medical systems such as Clinical Decision Support Systems (CDSS), medical imaging, and Bioinformatics [10]. It helps in reducing medical errors, costs, earlier disease detection, and helps to increase to achieve preventive medicine [10]. One of the advantage of using computerized CDSS is to help in managing overloaded data and turn them into knowledge, by reducing the complexity in automatic complex workflows, and to identify obese children by reducing the errors, time, and variety of practices [11].

Data mining tools with advanced algorithms are popular for the advantage of pattern discovery in biological data [12]. The medical data includes biological problems like protein interactions, sequence and gene expression data analysis, drug discovery, discovering homologous sequences or structure, construction of phylogenetic trees, gene finding, gene mapping, and sequence alignment [12].

CONFERENCE PAPER
Two day National Conference on Advanced Trends and Challenges in Computer Science and Applications
Organized by: Shree Vishnu Engineering College for Women, Bhimavaram A.P.
Schedule: 18-19 March 2014

Patient record data includes three groups they are structured data, semi-structured data, and unstructured data (Figure 6) [13]. As shown in below figure.

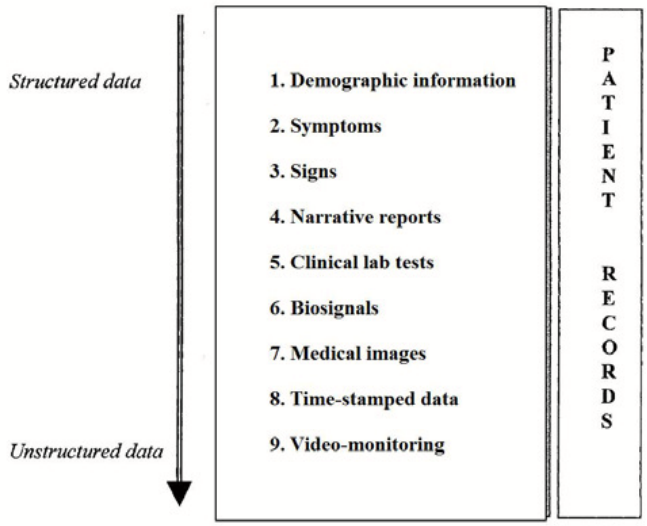


Figure 6. Data That Can be Captured From a Patient Record [6].

IV. UTILIZATION OF DATAMINING FOR PREDICTIONS IN MEDICAL DOMAIN

This section includes different classifications and predictions that was applied to medical related problems. they are discussed as follows:

A. Coronary Heart Diseases:

Coronary heart disease (CHD) is a serious disease that caused to increase the death rate especially found in china [8].In This study according to survey on 5 clinical centres collected 1069 CHD located in two provinces. 80 symptoms that are related closely to CHD and always appear in the literatures of CHD were selected.

Classification algorithms used are: Bayesian network (BN), decision tree (C4.5), neural network mode (MLP), and SVM. The performance is determined by considering their sensitivity, specificity, and accuracy. They are as follows

sensitivity = TP / (TP + FN);

specificity = TN / (TN + FP);

and the accuracy = (TP + TN) /(TP + FP + TN + FN) [12, 13, 18, 24].

The number of samples classified as true while they were actually true(TP);

The number of samples classified as false while they were actually false(FN);

The number of samples classified as false while they were actually false(TN); and

The number of samples classified as true while they were actually false(FP).

The results are as follows:

The highest accuracy is SVM (82.5%)

and the second highest is Bayesian model (82.0%).

Decision tree (C4.5) has the lowest accuracy (80.4%) when compared. From these results, SVM have showned good accuracy on predicting the coronary heart diseases.

B. Breast Cancer:

The study on breast cancer survey involve 1035 breast cancer patient [15]. At the time of surgery 22 medical patient features were recorded and 10 more features were recorded through follow-up more than 10 years.

Classification algorithms used are:

Naïve Bayes (NB), decision tree, SVM, multilayer ANN.

The results are :

NB have shown the highest accuracy (0.7).

The decision tree showed slightly lower accuracy (0.67 and 0.68 respectively). The models were also tested on binary datasets.NB have the highest mean accuracy (0.68 and 0.678 respectively) ,decision tree (0.674) and ANN (0.608).

When compare from the above results The Naïve Bayes classification has shown good and consistent accuracy in this study.

C. Diabetes:

When a person has the medical condition called *diabetes*, the body can't produce enough insulin to process the glucose in the blood and the body cannot make proper use of carbohydrate which greatly affects the patient lifestyle [16, 17]. The study on diabetes prediction invoved clinical information containing 2017 diabetic patients information[16].

The Classification algorithms used are:

Decision tree C4.5, and Naïve Bayes. The performance was done based on the specificity and sensitivity.

The results are:

Discretized C4.5 was the best in classifying bad blood glucose control patients (sensitivity).and

Naïve Bayes was the best in classifying good blood glucose control patients (specificity).

In comparing the differences in both sensitivity and specificity, Naïve Bayes has the least differences.

This study has shown that decision tree may produce high sensitivity. But according to fair distribution between sensitivity and specificity, the Naïve Bayes is better

D. Childhood Obesity:

The classification algorithms used are:

Decision tree (C4.5),Neural Network, Naïve Bayes, SVM. The main aim is to identify obese and overweight children at 3 years old by using the recorded data at birth, 6 weeks, 8 months, and 2 years. The study includes 16653 instances, where only 20% of the samples are obese. By measuring accuracy using the sensitivity and specificity.

The results are:

For overweight predictions at 3 years old, SVM have the highest sensitivity (60% respectively). However, the specificity of SVMs is the lowest compared to other.

The specificity of Bayesian classifiers is very high (93.1%) hence when they have the highest overall accuracy (91.9%), Neural Network is 24.6%;. In this study, the Bayesian classifiers and the SVMs have shown good accuracy for overweight prediction.

V. CONCUSION

This paper presents the study of data mining importance in medical data. The popular data mining Classification Algorithms include the Artificial Neural Network (ANN), decision tree, Bayesian classifiers, Support Vector Machine (SVM). From the studieson above diseases it is important to know that the Data mining utilization is increasing in medical informatics and for improving the decision making such as diagnostic and prognostic problems in oncology,

CONFERENCE PAPER
Two day National Conference on Advanced Trends and Challenges in Computer Science and Applications
Organized by: Shree Vishnu Engineering College for Women, Bhimavaram A.P.
Schedule: 18-19 March 2014

liver pathology, Neuropsychology, and Gynaecology, and also has shown the importance of predicting the best algorithm in each case study on medical related problems.

VI. REFERENCES

[1]. Q. Luo, "Advancing knowledge discovery and data mining," in WKDD '08 Proceedings of the First International Workshop on Knowledge Discovery and Data Mining, Washington, DC, USA, 2008.

[2]. B. Novak and M. Bigec, "Application of artificial neural networks for childhood obesity predicition," in ANNES '95 Proceedings of the 2nd New Zealand Two-Stream International Conference on Artificial Neural Networks and Expert Systems 1995.

[3]. A. K. Jain, et al. Artificial neural network : a tutorial [Online].

[4]. I. H. Witten, et al., Data mining: practical machine learning tools and techniques, 3rd ed.: Morgan Kaufmann, 2011.

[5]. Y. Fu. Data mining.potentials [Online].

[6]. L. Jiang, et al., "A novel bayes model: hidden naive bayes," IEEE Trans. on Knowl. and Data Eng., vol. 21, pp. 1361-1371, 2009.

[7]. S. Zhang, et al., "Comparing data mining methods with logistic regression in childhood obesity prediction," Information Systems Frontiers, vol. 11, p. 51, 2009.

[8]. J. Chen, et al. (2007). A comparison of four data mining models:bayes, neural network, SVM and decision trees in identifying syndromes in coronary heart disease. 4491/2007.

[9]. I. Maglogiannis, et al., "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers,"Applied intelligence, vol. 30, 2007.

[10]. I. Maglogiannis, et al., "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers," Applied intelligence, vol. 30, 2007.

[11]. W. Lord and D. Wiggins, "Medical Decision Support Systems Advances in Health care Technology Care Shaping the Future of Medical." vol. 6, G. Spekowius and T. Wendler, Eds., ed: Springer Netherlands, 2006, pp. 403-419.

[12]. H. Cheng, et al. (2010). Data mining for protein secondary structure prediction. 134.

[13]. M. Kantardzic, Data mining: concepts, models, methods, and algorithms: Wiley-IEEE Press, 2003.

[14]. J. Chen, et al. (2007). A comparison of four data mining models:bayes, neural network, SVM and decision trees in identifying syndromes in coronary heart disease. 4491/2007.

[15]. I. Maglogiannis, et al., "An intelligent system for automated breastcancer diagnosis and prognosis using SVM based classifiers,"Applied intelligence, vol. 30, 2007.

[16]. E. Strumbelj, et al., "Explanation and reliability of prediction models: the case of breast cancer recurrence," Knowl. Inf. Syst., vol. 24, pp. 305-324, 2010.

[17]. Yue Huang, et al., "Evaluation of outcome prediction for a clinical diabetes database ", ed, 2004.

[18]. Y. Huang, et al., "Evaluation of Outcome Prediction for a Clinical Diabetes Database, Knowledge Exploration in Life Science Informatics." vol. 3303, J. López, et al., Eds.,ed: Springer Berlin / Heidelberg, 2004, pp. 181-190 .

CONFERENCE PAPER
Two day National Conference on Advanced Trends and Challenges in Computer Science and Applications
Organized by: Shree Vishnu Engineering College for Women, Bhimavaram A.P.
Schedule: 18-19 March 2014