Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Performance Evaluation of Decision Tree and Neural Networks for Classification of Hematology Databases

Shrwan Ram
Department of Computer Science and Engineering,
M.B.M Engineering College, J. N.V. University,
Jodhpur, India

Abstract:-Clinical databases are playing the major role for prediction of many types of diseases. Through many types of clinical processes very large volume of pathological datasets are generated for the care of patients. Pathologists analyze these data or test results obtained with the help of many clinical processes and take cares according to the predicted symptoms of disease. These datasets are more helpful for the doctors and health care centers to predict the relevant cause of diseases and to provide better medical treatment. To analyze and classify all these datasets is a very tedious process. Clinical datasets are very complex and require more efficient and accurate algorithms and data analyzing tools. Machine learning algorithms are used in many fields for the classification of large volume of data to generate the rules for building the knowledge base system. These machine learning algorithms are widely used in the medical field to build the disease diagnosis support system. This has become the emerging field of medical research. Many types of machine learning algorithms has been developed and deployed. In this research paper Hematology datasets which are very important for the pathologists to predict symptoms of many diseases. These datasets are classified as normal samples and abnormal samples using decision tree and neural networks. These machine learning algorithms are used for the classification of Hematology datasets. The classification performance of the decision tree and neural networks are evaluated on the basis of classification accuracy and performance matrices.

Index Terms:-Clinical databases, Pathological databases, Hematology datasets, Disease Diagnosis, Classification, Neural Networks, Machine learning

I. INTRODUCTION

Decision tree is a tree-shaped structure that represents sets of decisions. These decisions generate rules for the classification of a dataset. Decision Tree is a popular classifier which is simple and easy to implement [1]. It requires no domain knowledge or parameter setting and can handle high dimensional data. Hence it is more appropriate for exploratory knowledge discovery. It still suffers from repetition and replication. Therefore necessary steps need to be taken to handle repetition and replication. The performance of decision trees can be enhanced with suitable attribute selection. Correct selection of attributes partition the data set into distinct classes. A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes are called leaves (also known as terminal or decision nodes).

Neural networks process information in a similar way the human brain does. The network is composed of a large number of highly interconnected processing elements (neurons) working in parallel to solve a specific problem. Neural networks learn by example. They cannot be programmed to perform a specific task. The examples must be selected carefully otherwise
useful time is wasted or even worse the network might be functioning incorrectly. An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the biological nervous systems, such as the brain, [2].

The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurones) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Artificial Neural Networks (ANN's) have been used widely in many application areas in recent years. Most applications use feed forward ANN's and the backpropagation (BP) training algorithm. There are numerous variants of the classical BP algorithm and other training algorithms. All these training algorithms assume a fixed ANN architecture.

II. DECISION TREE ALGORITHM FOR THE CLASSIFICATION OF HEMATOLOGY DATASETS

Before constructing and using the Decision Tree algorithms to classify the databases, a Relevance Analysis of the features of collected databases is performed. Relevance Analysis aims to improve the classification efficiency by

© 2015, IJARCSSE All Rights Reserved

eliminating the less useful (for the classification) features and reducing the amount of input data to the classification stage. In this paper, the decision tree is a purposed data mining technique used for classification of clinical databases. Attention is focused on the diagnosis of diseases using decision tree as a classification technique. It is one of the widely used data mining technique for the classification of datasets and generating the rules for decision making process.

Data classification using decision tree technique is made up of two steps. The first step is to use the training datasets to train the algorithm. In this step a predetermined training data set is analyzed by a classification algorithm to construct the model. This model is tested with the help of testing datasets. The final step is the classification of the unseen datasets. In this step unseen datasets are used to verify the classification rules or the decision tree. Many algorithms are available for constructing the decision tree , in this thesis work CART (Classification and Regression Tree algorithm) algorithm which is one the best tree generating algorithm is purposed for building decision tree .In a research paper titled as" application of CART algorithm in hepatitis disease diagnosis. " On the basis of the results The CART algorithm is suggested as one of the best algorithm to construct the decision tree [3].

CART Algorithm: CART [3] is an acronym for Classification and Regression Trees, a decision-tree procedure introduced in 1984 by world-renowned UC Berkeley and Stanford Statisticians, Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone [4]. The CART methodology solves a number of performance, accuracy, and operational problems that still plague many current decision-tree methods.

CART innovations includes:-

(a) Solving the "how big to grow the tree", problem (b) Using strictly two-way (binary) splitting.

(c) Incorporating automatic testing and tree validation (d) Providing a completely new method for handling missing values.

Features of CART Algorithm [4]:

1. The visual display enables users to see the hierarchical interaction of the variables;
2. Further, because simple if then rules can be read right off the tree, models are easy to grasp and easy to apply to new data.
3. CART uses strictly binary, or two-way, splits that divide each parent node into exactly two child nodes by posing questions with yes/no answers at each decision node.
4. CART is unique among decision-tree tools. CART- proven methodology is characterized by:

(a). Reliable pruning strategy - CART developers determined definitively that

(b). no stopping rule could be relied on to discover the optimal tree,

(c). Powerful binary-split search approach – CART binary decision trees are more sparing with data and detect more structure before too little data is left for learning.

(d). Automatic self-validation procedures - in the search for patterns in databases it is essential to avoid the trap of over fitting

III. NEURAL NETWORK FOR THE CLASSIFICATION OF HEMATOLOGY DATASETS

Artificial Neural Network is inspired by the biological neural network system, in the biological neural network the information or signals are received by dendrites [5] .*The* neuron sends out spikes of electrical activity through a long, thin stand known as an *axon*, as shown in figure 1, which splits into thousands of branches. At the end of each branch, a structure called *a* synapse converts the activity from the axon into electrical effects that inhibit or excite activity from the axon into electrical effects that inhibit or excite activity in the connected neurons. When a neuron receives excitatory input that is sufficiently large compared with its inhibitory input, it sends a spike of electrical activity down its axon. Learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another changes.

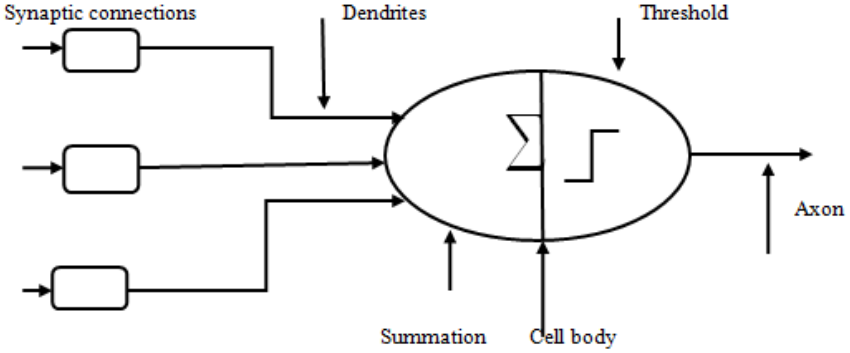


Figure 1 Biological Neuron Model [5]

A. A Simple Neurons Architecture

An artificial neuron is a device with many inputs and one output. The neuron has two modes of operation; the training mode and the using mode. In the training mode, the neuron can be trained to fire (or not), for particular input patterns. In the using mode, when a taught input pattern is detected at the input, its associated output becomes the current output. If the input pattern does not belong in the taught list of input patterns, the firing rule is used to determine whether to fire or not.

© 2015, IJARCSSE All Rights Reserved

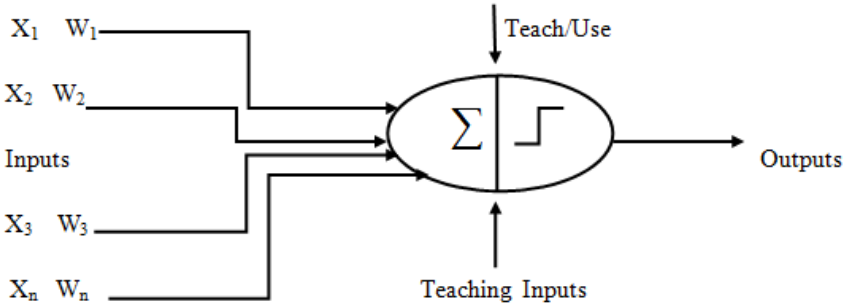


Figure 2. Simple Neuron [6]

IV. LITERATURE REVIEW

Different data classification techniques are used for classification of large volume of data for decision making purpose. During the literature survey it is found that there are many types of data classification techniques are used to build the knowledge base system. Many software tools are available for research and commercial purpose. Many enterprises are using these classification techniques to provide the quality of services to their customers because the time and quality of service is more important. The literature reviews are followings:

1. Brijesh Verma, "A Neural Learning Algorithm for the Diagnosis of Breast Cancer", 2006 International Joint Conference on Neural Networks, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21, 2006.
2. Philip de Chazal and Branko G. Celler, **"**Selecting a Neural Network Structure for ECG Diagnosis", Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 20, No 3,1998.
3. Ms. Ishtake S.H and Prof. Sanap S.A, "Intelligent Heart Disease Prediction System Using Data Mining Techniques**",** International J. of Healthcare & Biomedical Research, Volume: 1, Issue: 3, April 2013, Pages 94-101.
4. Serhat Özekes and A.Yilmaz Çamurcu, "Classification and Prediction in a Data Mining Application", Journal of Marmara for Pure and Applied Sciences, 18 (2002) 159-174.
5. Markos G. Tsipouras, Themis P. Exarchos, Dimitrios I. Fotiadis, Anna P. Kotsia, Konstantinos V. Vakalis, Katerina K. Naka, and Lampros K. Michalis, "Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling", IEEE Transactions on Information Technology in Biomedicine, Vol. 12, No. 4, July 2008.
6. Filippo Amato, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara, Aleš Hampl and Josef Havel, "Artificial neural networks in medical diagnosis", J Appl Biomed. **11**: 47–58, 2013.
7. Minas A. Karaolis, Joseph A. Moutiris, Demetra Hadjipanayi and Constantinos S. Pattichis, "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees", IEEE Transactions on Information Technology in Biomedicine, Vol. 14, No. 3, MAY 2010.
8. Anchana Khemphila and Veera Boonjing, "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients", International Conference on Computer Information Systems and Industrial Management Applications (CISIM), 2010, pp 193-198.

V. RESEARCH METHODOLOGIES

A. Proposed Classification Model

Different data classification techniques are used for the classification of large volume of medical datasets and other databases. In this research paper Decision tree and Neural networks are used for the classification of Hematology datasets which are collected from the health care centers in Jodhpur, Rajasthan (India). Classification Models are developed and deployed in matlab. In this paper the proposed classification model takes inputs as Hematology datasets and on the basis of components of blood samples it classifies the datasets in two categories. Normal datasets are classified and category 1 and abnormal datasets are classified as category 0. Decision rules are generated on the basis of classification accuracy

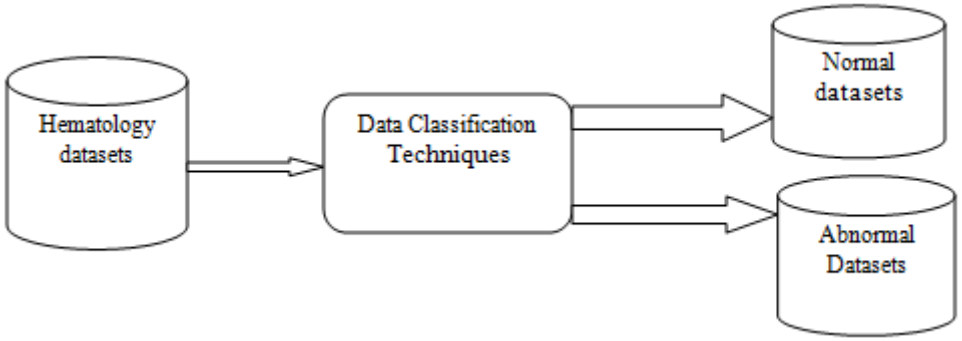


Figure 3. Classification Model

© 2015, IJARCSSE All Rights Reserved

B. Empirical Data Collection

Hematology datasets are collected from the Different Healthcare Centers in Jodhpur which is one of the city in Rajasthan (India). I have personally collected the datasets and build an .xls database. Datasets consists of 16 attributes and the normal values of the different attributes are provided with datasets.

Table 1. Hematology Database Attributes

S.NO	ATTRIBUTE NAME	DESCRIPTION
1	Age	age in years
2	Sex	1 = male ; 0 = female
3	White Blood Cells (WBC)	4.0 to 10.0 m/mm^3
4	Lymphocytes	20% to 40%
5	Mix	1.0% to 10%
6	Neutrophils	30% to 70%
7	RBC	Men: 4.7 to 6.1 million mcL, Women: 4.2 to 5.4 million mcL, Children: 4.0 to 5.5 million mcL, Newborn: 4.8 to 7.1 million mcL (m/mm^3)
8	Mean Corpuscular Volume (MCV)	80 to 95 fL To calculate the MCV, expressed in femtoliters (fL, or 10^{-15}L)
9	Hematocrit (HCT, packed cell volume)	33.0 to 54.0 %
10	Mean Corpuscular Hemoglobin (MCH)	Normal range: 26 to 34 pg(picograms/cell)
11	Mean corpuscular hemoglobin concentration (MCHC)	32 to 36 g/dL (gram per deciliter)
12	Red Blood Cell Distribution width (RDW)	8.0 to 12.0% RDW = (Standard deviation of MCV ÷ mean MCV) × 100
13	Hemoglobin (Hb)	Women: 12.1 to 15.1 gm/dl Men: 13.8 to 17.2 gm/dl Children: 11 to 16 g/dl Pregnant women: 11 to 12 g/dl
14	THR	100 to 450 m/mm^3
15	MPV	6.0 to 13.0
16	Platelet Distribution Width (PDW)	6.0 to 10.0

VI. DECISION TREE CLASSIFICATION ALGORITHM

Decision tree is a tree-shaped structure that represents sets of decisions. These decisions generate rules for the classification of a dataset. Decision Tree is a popular classifier which is simple and easy to implement [7]. It requires no domain knowledge or parameter setting and can handle high dimensional data. Hence it is more appropriate for exploratory knowledge discovery. It still suffers from repetition and replication. Therefore necessary steps need to be taken to handle repetition and replication. The performance of decision trees can be enhanced with suitable attribute selection. Correct selection of attributes partition the data set into distinct classes.

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes are called leaves (also known as terminal or decision nodes).

A. Basic Decision Tree Building Algorithm.

Algorithm Learn Decision Tree (examples, attributes, default) returns a decision tree
Inputs: examples, a set of examples
Attributes, a set of attributes
Default, default value for goal attributes
if examples is empty then return default
else if all examples have same value for goal attribute then return value
else
Best = Choose Attribute (attributes, examples)
Tree = a new decision tree with root test best
for each value vi of best do
Examples = {elements of examples with best = vi}
Subtree = Learn Decision Tree(examples, attributes – best,
Majority Value (examples))
add a branch to tree with label vi and subtree , return tree.

B. Results Generated Using Decision Tree Classification Aalgorithm.

MATLAB Version: 8.0.0.783 (R2012b) supports the entire data analysis and classification process, with acquiring data from external data sources, databases through pre-processing, normalizing data if we set the data in a particular range e.g. Values between 0 and 1. With visualization and numerical data analysis capability of MATLAB, it

© 2015, IJARCSSE All Rights Reserved

produces best presentation and quality outputs. All the data classification are carried out with the help of latest MATLAB Version: 8.0.0.783 (R2012b), in this version many types of data classification and analyzing algorithms are implemented. Through the MATLAB programming proposed data mining techniques are implemented.

C. Decision Tree for the Classification of Hematology Database.

Now the Hematology datasets are classified Using Matlab and the Results are shown below. Figure 4 shows the Decision Tree generated for the Hematology database. The terminal nodes show the classification results as 1 and 2.

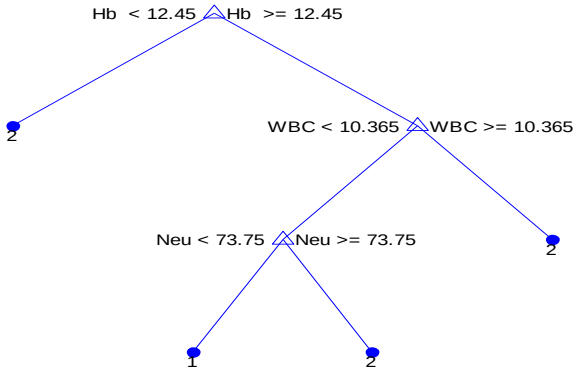


Figure 4. Decision Tree for Hematology Database

D. Decision Tree Before Pruning:

Decision tree classification rules:
1. if Hb <12.45 then node 2 elseif Hb >=12.45 then node 3 else 2
2. class = 2
3. if WBC<10.365 then node 4 elseif WBC>=10.365 then node 5 else 1
4. if Neu<73.75 then node 6 elseif Neu>=73.75 then node 7 else 1
5. class = 2, 6 class = 1,7 class = 2

E. Decision Tree After Pruning:

Decision tree for classification
1. if Hb <12.45 then node 2 elseif Hb >=12.45 then node 3 else 2
2. class = 2
3. class = 1

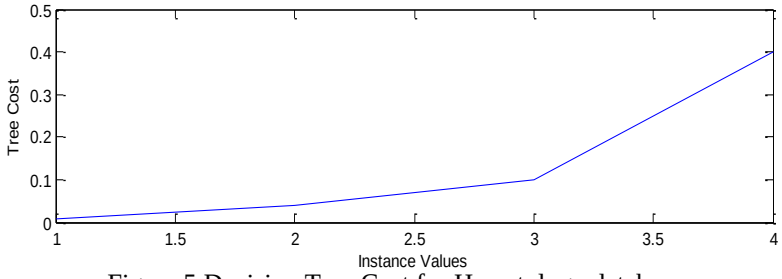


Figure 5 Decision Tree Cost for Hematology database

F. Confusion Matrix Generated by Decision Tree for the Classification.

Table 2 Confusion Matrix Generated for Hematology Database

		1 (Predicted)	2 (Predicted)
1 (Actual)	100	99 True Positive(TP) 99%	1 False Negative(FN) 1%
2 (Actual)	150	0 False Positive(FP) 0%	150 True Negative(TN) 100

According to the values predicted in the confusion matrix it shows that the Classification Accuracy of the Decision tree model is 99.5% and the misclassification is equal to 0.5.

VII. NEURAL NETWORK FOR THE CLASSIFICATION OF HEMATOLOGY DATABASE

The Neural Network model as shown in figure 5 is used for the classification of Hematology database. The 26 hidden layer Neurons with one hidden layer is used for the data classification and the target values are normalised as 0 and 1 in place of 1 and 2 respectively.

© 2015, IJARCSSE All Rights Reserved

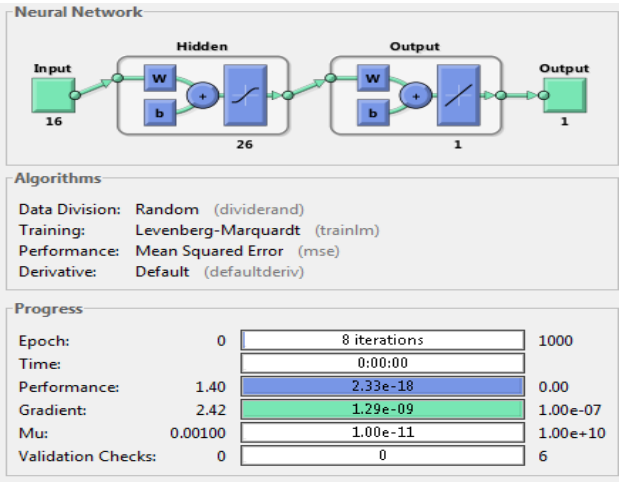


Figure 5 Neural Network for the classification of Hematology Database

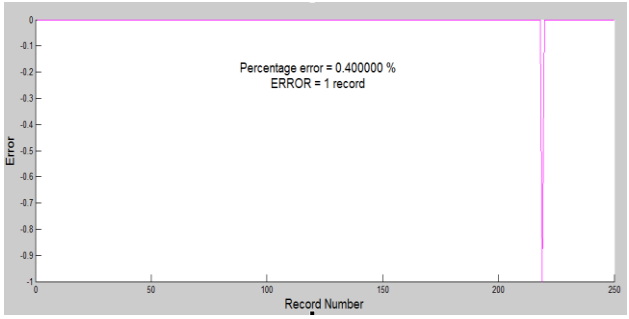


Figure 6 Classifications of the Data Records

The classification of the data shows that only one record is found as the unclassified record and the percentage of the error is 0.40000% which is the quite good performance as comparative to the previously implemented models found through the literature survey.

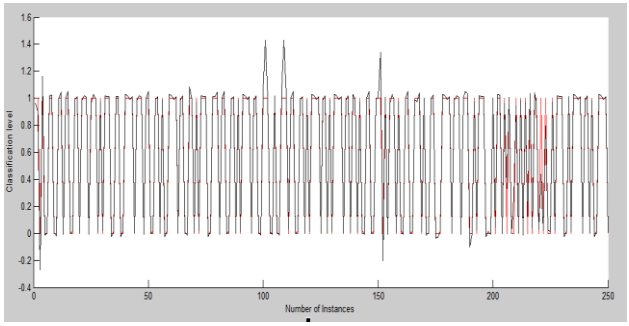
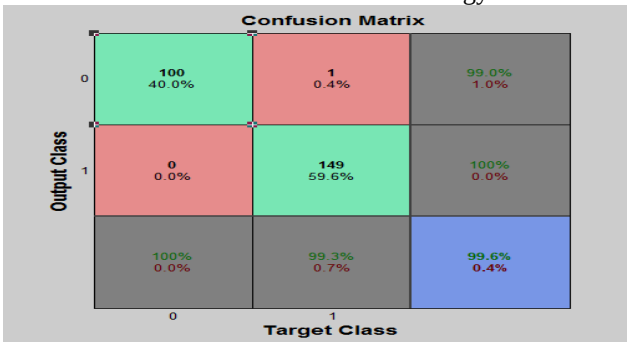


Figure 7 Classification Level with 250 Instances

Figure 7 shows that the data are classified as 0 and 1 and the classification values are lies in between 0 and 1 but some values are the outside of the range 0 and 1,these values show the Misclassification of the datasets.

A. Confusion Matrix Generated for the Classification

Table 2 Confusion Matrix for Hematology Database



According to the values predicted in the confusion matrix shows that the Classification Accuracy of the Neural Network model is 99.6% and the misclassification is rate is 0.4%.

© 2015, IJARCSSE All Rights Reserved

VII. COMPARISION OF THE PERFORMENCE

Comparative table shown below shows that the classification performance of the Decision tree and Neural Networks are all most similar but Neural Networks are better then the Decision tree classifier. The classification performance also depend on the purity of the datasets collected, if we have the noisy datasets then the performance of the machine learning algorithms are not good therefore it is required that the datasets collected from the different sources should be accurate.

Table 4.18 Classification of Hematology Dataset Using Matlab

S.No	**Data Mining Technique**	**Data Samples**	**Accuracy (%)**
1.	Decision Tree	250	99.5%
2.	Neural Network	250	99.6%

VIII. CONCLUSION

The Decision tree and Neural network are widely used in many real life applications to find the patterns and building the knowledgebase system. On the basis of classification accuracy it is found that the **Neural Network** and **Decision Tree** data mining techniques are more efficient and useful for developing the knowledge base system for the prediction of hidden and unseen patterns from the large volume of database.

REFERENCES

[1] T. Smitha, and V. Sundaram, "Comparative Study of Data Mining Algorithms for High Dimensional Data Analysis", International Journal of Advances in Engineering & Technology, Vol.4, Issue 2, pp. 173-178, 2012.

[2] Yavuz Del_Can, Lale Özyilmaz and Tülay Yildirim, "Evolutionary Algorithms Based RBF Neural Networks For Parkinson's disease Diagnosis", 7th International Conference on Electrical and Electronics Engineering (ELECO), pp. 311-315, 2011.

[3] G. Sathyadevi, "Application of CART Algorithm in Hepatitis Disease Diagnosis" International Conference on Recent Trends in Information Technology (ICRTIT), pp.1283 – 1287, 2011

[4] Antonia Vlahou, , John O. Schorge, Betsy W. Gregory, and Robert L. Coleman, "Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data", Journal of Biomedicine and Biotechnology, Vol. 5, Issue 5, pp. 308-314, 2003.

[5] Yavuz Del_Can, Lale Özyilmaz and Tülay Yildirim, "Evolutionary Algorithms Based RBF Neural Networks For Parkinson's disease Diagnosis", 7th International Conference on Electrical and Electronics Engineering (ELECO), pp. 311-315, 2011.

[6] Brijesh Verma, "A Neural Learning Algorithm for the Diagnosis of Breast Cancer", 2006 International Joint Conference on Neural Networks, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21, 2006.

[7] T. Smitha, and V. Sundaram, "Comparative Study of Data Mining Algorithms for High Dimensional Data Analysis", International Journal of Advances in Engineering & Technology, Vol.4, Issue 2, pp. 173-178, 012.

[8] M. T. Hagan, H. B. Demuth, and M. H. Beale, Neural network design. PWS Pub, 1996.

[9] Philip de Chazal and Branko G. Celler, **"**Selecting a Neural Network Structure for ECG Diagnosis", Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 20, No 3,1998.

[10] Ms. Ishtake S.H and Prof. Sanap S.A, "Intelligent Heart Disease Prediction System Using Data Mining Techniques**",** International J. of Healthcare & Biomedical Research, Volume: 1, Issue: 3, April 2013, Pages 94-101.

[11] Serhat Özekes and A.Yilmaz Çamurcu, "Classification and Prediction in a Data Mining Application", Journal of Marmara for Pure and Applied Sciences, 18 (2002) 159-174.

[12] Markos G. Tsipouras, Themis P. Exarchos, Dimitrios I. Fotiadis, Anna P. Kotsia, Konstantinos V. Vakalis, Katerina K. Naka, and Lampros K. Michalis, "Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling", IEEE Transactions on Information Technology in Biomedicine, Vol. 12, No. 4, July 2008.

[13] Filippo Amato, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara, Aleš Hampl and Josef Havel, "Artificial neural networks in medical diagnosis", J Appl Biomed. **11**: 47–58, 2013.

[14] Minas A. Karaolis, Joseph A. Moutiris, Demetra Hadjipanayi and Constantinos S. Pattichis, "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees", IEEE Transactions on Information Technology in Biomedicine, Vol. 14, No. 3, MAY 2010.

[15] Anchana Khemphila and Veera Boonjing, "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients", International Conference on Computer Information Systems and Industrial Management Applications (CISIM), 2010, pp 193-198.

© 2015, IJARCSSE All Rights Reserved