DECISION TREE BASED PATTERN ANALYSIS WITH CANCER DATASET

Shiv Shakti Shrivastava

(Research Scholar, Computer Science & Engg., Mewar University, Chittorgarh. (Raj.)

Dr. V.K.Choubey, Dr. Anjali Sant

1. INTRODUCTION

In this paper discuss Cascaded Model of RBF Network for ID3 classification. The great advantage of RBF network is single layer processing unit and target output independent with input data. In the process of cascading input feature passes through margin of classifier, margin classifier function separates data into layers such as positive and negative in data space domain. The part of positive and negative used as input in cascaded model. The neural networks trained on this set should make more favorable decision to the minority class with the minimization of misclassification of the majority class and have increased generalization capability. Let us assume the input vector is M dimensional. Here various methods discussed like support ID3, feature correlation and ID3 and finally discuss a proposed hybrid technique for ID-CRBF for classification.

2. CLASSIFIER (ID3)

ID3 is a simple decision tree learning greedy algorithm which generates decision trees from a fixed set of data in a top-down, recursive, divide-and-conquer manner. The major disadvantages of this algorithm are that the attributes must be categorical and continuous-valued attributes must be decentralized in advance.

3. CONSTRUCTION OF THE TREE

For the construction of the decision tree, the classical tree and some modifications of it are used. The classical ID3 construct a collection of trees. For the construction of each tree, a bootstrap sample of the dataset is selected. The tree is built to the maximum size without pruning. The tree is just grown until each terminal node contains only members of a single class. The Gini index is used to determine the best split of each node. Only a subset m of the total set of M features is employed as the candidate splitters of the node of the tree. The number of the selected features (m) remains constant throughout the construction of the tree. In ID3 with ReliefF, the Gini index is replaced by ReliefF. ReliefF evaluates the partitioning capability of attributes according to how well their values distinguish between similar instances.

4. CASCADED RADIAL BASIS FUNCTION NETWORK (CRBF)

A CRBF Network which is linear in the parameters provided all the RBF centers are prefixed. Given fixed centers i.e. no adjustable parameters the first layer or the hidden layer peID3orms a fixed nonlinear transformation, which maps the input space onto a new space [20]. With n inputs and m hidden neurons is shown in Figure 6.1.

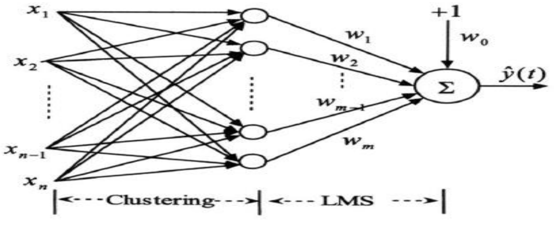

Figure 6.1 Neuron structure of CRBF

Such as network can be representd as

$$\hat{y}(t) = w_0 + \sum_{j=1}^{m} w_j \exp\left(-\frac{\left\|x - x_{cj}\right\|^2}{\sigma_j^2}\right)$$ ……………………6.2.1

in which x is the input vector, xc; is the center and a2 is the width of the j th Gaussian hidden unit. Parameters wj are weights of the connections that feed the output unit. The output layer then implements a linear combiner on this new space and the only adjustable parameters are the weights of this linear combiner. These parameters can therefore be determined using the linear least mean square algorithm (LMS), which is an important advantage of this approach. This is basically how an RBF network works. Our CCRBF networks are as follows. The training begins with minimal structure (no hidden units), and then more connections, neurons are added to the network according to some predefined rule. Adding the hidden units one by one is started by first creating a set of candidate hidden units, which can be done in many existent ways.

5. PROPOSED METHOD FOR CRBF-ID3

CRBF models are creating for data training for minority and majority class data sample for processing of tree classification. The input processing of training phase is data sampling technique for classifier. While single-layer RBF networks can potentially learn virtually any input output relationship, RBF networks with single layers might learn complex relationships more quickly. The function neID3 creates cascade-forward networks. For example, a cascaded layer network has connections from layer 1 to layer 2, layer 2 to layer 3, and layer 1 to layer 3. The cascade-layer network also has connections from the input to all cascaded layers. The additional connections might improve the speed at which the network learns the desired relationship. CRBF artificial intelligence model is similar to feed-forward back-propagation neural network in using the back-propagation algorithm for weights updating, but the main symptom of this network is that each layer of neurons related to all previous layer of neurons. Tan-sigmoid transfer function, log - sigmoid transfer function and pure linear threshold functions were used to reach the optimized status.

6. PROCESS OF METHOD

1. Sampling of data of sampling technique
2. Split data into two parts training and testing part
3. Apply CRBF function for training a sample value
4. Using 2/3 of the sample, fit a tree the split at each node
   For each tree. .
   - Predict classification of the available 1/3 using the tree, and calculate the misclassification rate = out of CRBF.
5. For each variable in the tree
6. Compute Error Rate: Calculate the overall percentage of misclassification
   - Variable selection: Average increase in CRBF error over all trees and assuming a normal division of the increase among the trees, decide an associated value of feature.
7. Resulting classifier set is classified
   - Finally to estimate the entire model, misclassification

Decode the feature variable in result class

7. EXPERIMETNAL RESULT PROCESS

Method Name	Elapsed Time	Confusion Error	Error	Accuracy
DT	35.82	40.00	24.15	87.08

ID3	36.17	41.00	24.84	86.18
ID3 CRBF	70.62	**39.50**	28.80	**97.18**

Table 1. Comparative performance evaluation for classification of dataset using DT, ID3 and ID3-CRBF method.

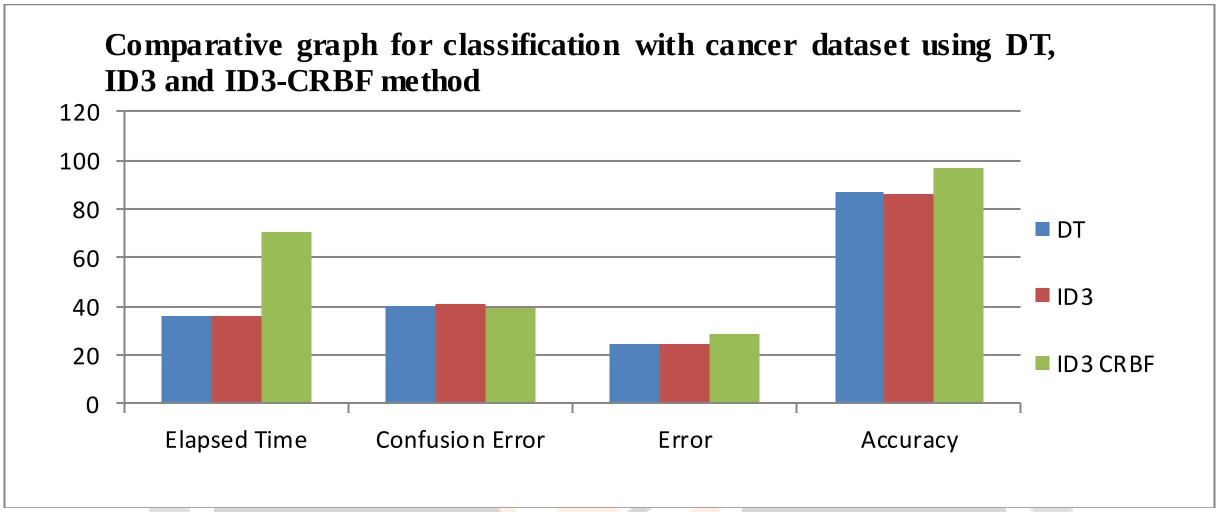

Figure 1. Comparative performance evaluations for classification of dataset using DT, ID3 and ID3-CRBF method, here find the value of Elapsed time, Confusion error, error and Accuracy.

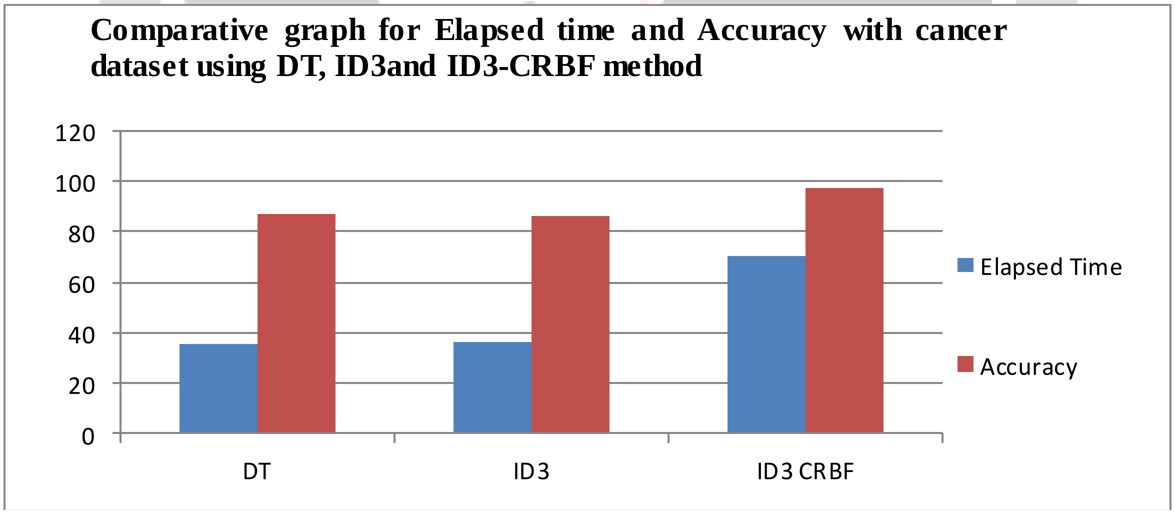

Figure 2. Comparative performance evaluation for classification of dataset using DT, ID3 and ID3-CRBF method, here find the value of Elapsed time and Accuracy

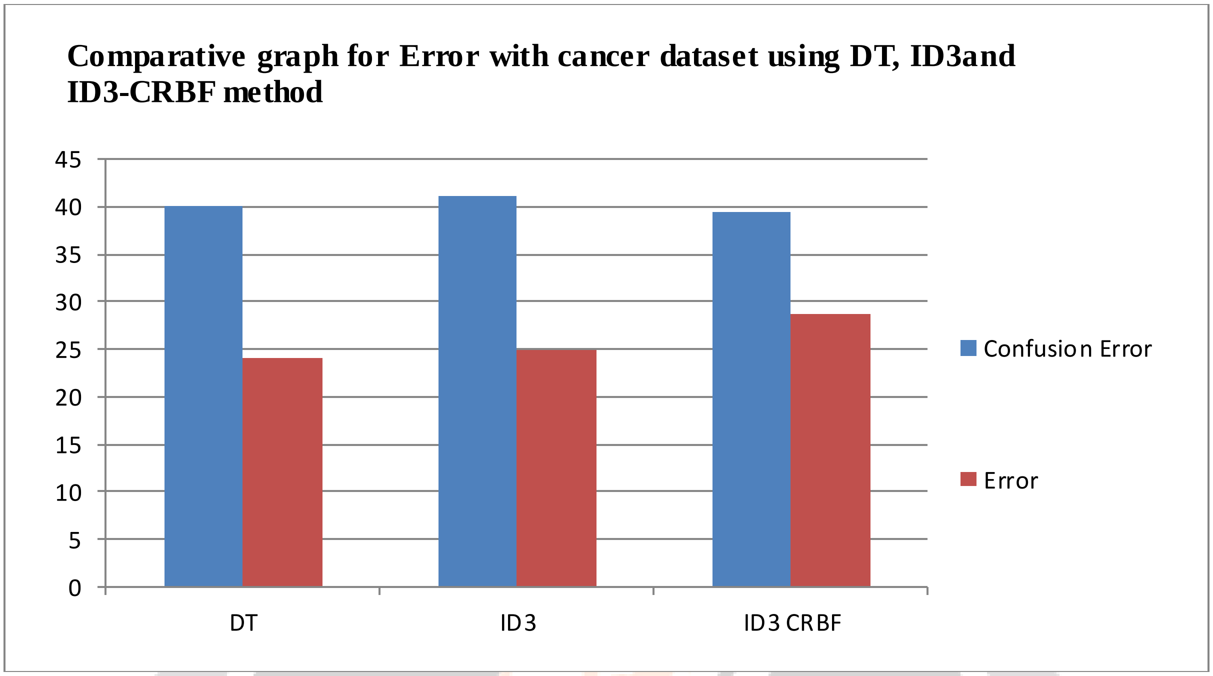

Figure 3. Comparative performance evaluation for classification of dataset using DT, ID3 and ID3-CRBF method, here find the value of Confusion error and error.

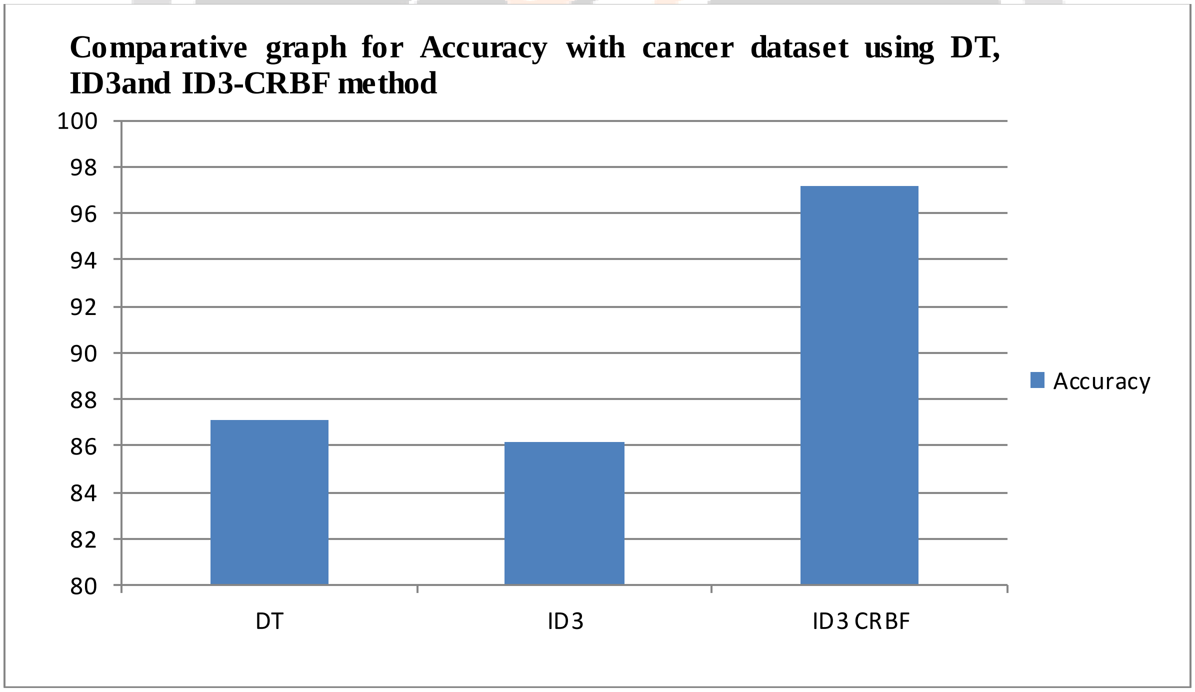

Figure 4. Comparative performance evaluations for classification of dataset using DT, ID3 and ID3-CRBF method, here find the value Accuracy.

REFERENCES

[1] D. P. Shukla, Shamsher Bahadur Patel, Ashish Kumar Sen "A Literature Review in Health Informatics Using Data Mining Techniques", international journal of software & hardware research in engineering 2014 PP 123-129.

[2] Mihaela Gheorghe, Ruxandra Petre "Integrating Data Mining Techniques into Telemedicine Systems" Informatica Economică 2014 PP 120- 130.

[3] Li Jiang, Stefan M Edwards, Bo Thomsen, Christopher T Workman, Bernt Guldbrandtsen , Peter Sørensen "A random set scoring model for prioritization of disease candidate genes using protein complexes and data-mining of GeneRIF, OMIM and PubMed records" BMC Bioinformatics 2014 PP 1-13.

[4] Saurabh Pal, Vikas Chaurasia "Data Mining Approach to Detect Heart Diseases" IJACSIT 2013 PP 56-66

[5] Salim Diwani ,Suzan Mishol , Daniel S.Kayange ,Dina Machuve ,Anael Sam "Overview Applications of Data Mining In Health Care: The Case Study of Arusha Region" International Journal of Computational Engineering Research 2013 PP 73-77.

[6] Divya Tomar , Sonali Agarwal "A survey on Data Mining approaches for Healthcare" International Journal of Bio-Science and Bio-Technology 2013 PP 241-266.

[7] V.Krishnaiah , Dr.G.Narsimha, Dr.N.Subhash Chandra "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" (IJCSIT) 2013, PP 39 – 45.

[8] Ashish Kumar Sen, Shamsher Bahadur Patel, Dr. D. P. Shukla "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level" International Journal Of Engineering And Computer Science 2013 PP. 2663-2671.

[9] Shweta Kharya "Using Data Mining Techniques For Diagnosis And Prognosis Of Cancer Disease" International Journal of Computer Science, Engineering and Information Technology , 2012 PP 55-66

[10] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg "Data Mining in Clinical Decision Support Systems for Diagnosis, Prediction and Treatment of Heart Disease" International Journal of Advanced Research in Computer Engineering & Technology 2013 PP 218-223.

[11] Duen-Yian Yeh , Ching-Hsue Cheng, Yen-Wen Chen "A predictive model for cerebrovascular disease using data mining" Elsevier2011 PP 8970–8977

[12] M. Durairaj, V. Ranjani "Data Mining Applications In Healthcare Sector: A Study" International Journal Of Scientific & Technology Research 2013 PP 29-35.

[13] Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte "A Data Mining Approach For Prediction Of Heart Disease Using Neural Networks" IAEME 2012, PP. 30-40