Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 13 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Analysis of Data Mining Classification with Decision tree Technique

By Dharm Singh, Naveen Choudhary & Jully Samota
Maharana Pratap University of Agriculture and Technology, India

Abstract- The diversity and applicability of data mining are increasing day to day so need to extract hidden patterns from massive data. The paper states the problem of attribute bias. Decision tree technique based on information of attribute is biased toward multi value attributes which have more but insignificant information content. Attributes that have additional values can be less important for various applications of decision tree. Problem affects the accuracy of ID3 Classifier and generate unclassified region. The performance of ID3 classification and cascaded model of RBF network for ID3 classification is presented here. The performance of hybrid technique ID3 with CRBF for classification is proposed. As shown through the experimental results ID3 classifier with CRBF accuracy is higher than ID3 classifier.

Keywords: data mining, classification, decision tree, ID3, attribute selection.

GJCST-C Classification : H.2.8

Strictly as per the compliance and regulations of:

Analysis of Data Mining Classification with Decision Tree Technique

Dharm Singh [α], Naveen Choudhary [σ] & Jully Samota [ρ]

Abstract- The diversity and applicability of data mining are increasing day to day so need to extract hidden patterns from massive data. The paper states the problem of attribute bias. Decision tree technique based on information of attribute is biased toward multi value attributes which have more but insignificant information content. Attributes that have additional values can be less important for various applications of decision tree. Problem affects the accuracy of ID3 Classifier and generate unclassified region. The performance of ID3 classification and cascaded model of RBF network for ID3 classification is presented here. The performance of hybrid technique ID3 with CRBF for classification is proposed. As shown through the experimental results ID3 classifier with CRBF accuracy is higher than ID3 classifier.

Keywords: data mining, classification, decision tree, ID3, attribute selection.

I. Introduction

With the rapid development of information technology and network technology, different trades produce large amounts of data every year. The data itself cannot bring direct benefits so need to effectively mine hidden information from huge amount of data. Data mining deals with searching for interesting patterns or knowledge from massive data. It turns a large collection of data into knowledge. Data mining is an essential step in the process of knowledge discovery (Lakshmi & Raghunandhan, 2011).The data mining has become a unique tool in analyzing data from different perspective and converting it into useful and meaningful information.Data mining has been widely applied in the areas of Medical diagnosis, Intrusion detection system, Education, Banking, Fraud detection.

Classification is a supervised learning. Prediction and classification in data mining are two forms of data analysis task that is used to extract models describing data classes or to predict future data trends. Classification process has two phases; the first is the learning process where the training data sets are analyzed by classification algorithm. The learned model or classifier is presented in the form of classification rules or patterns. The second phase is the use of model for classification, and test data sets are used to estimate the accuracy of classification rules.

Authors α σ ρ: Department of Computer Science and Engineering, College of Technology and Engineering, Maharana Pratap University of Agriculture and Technology, Udaipur, Rajasthan, India. e-mails: dharm@mpuat.ac.in, naveenc121@yahoo.com, jullysamota304@gmail.com

With the rising of data mining, decision tree plays an important role in the process of data mining and data analysis. Decision tree learning involves in using a set of training data to generate a decision tree that correctly classifies the training data itself. If the learning process works, this decision tree will then correctly classify new input data as well. Decision trees differ along several dimensions such as splitting criterion, stopping rules, branch condition (univariate, multivariate), style of branch operation, type of final tree (Han, Kamber & Pei, 2012).

The best known decision tree induction algorithm is the ID3. ID3 is a simple decision tree learning algorithm developed by Ross Quinlan. Its predecessor is CLS algorithm. ID3 is a greedy approach in which top-down, recursive, divide and conquer approach is followed. Information gain is used as attribute selection measure in ID3. ID3 is famous for the merits of easy construction and strong learning ability. There exists a problem with this method, this means that it is biased to select attributes with more taken values, which are not necessarily the best attributes. This problem affects its practicality. ID3 algorithm does not backtrack in searching. Whenever certain layer of tree chooses a property to test, it will not backtrack to reconsider this choice. Attribute selection greatly affects the accuracy of decision tree (Quinlan, 1986).

In rest of the paper, a brief introduction to the related work in the area of decision tree classification is presented in section 2. A brief introduction to the proposed work is presented in section 3. In section 4 we present the experimental results and comparison. In section 5, we conclude our results.

II. Related Work

The structure of decision tree classification is easy to understand so they are especially used when we need to understand the structure of trained knowledge models. If irrelevant attribute selection then all results suffer. Selection space of data is very small if we increase space, selection procedure suffers so problem of attribute selection in classification. There have been a lot of efforts to achieve better classification with respect to accuracy.

Weighted and simplified entropy into decision tree classification is proposed for the problem of multiple-value property selection, selection criteria and property value vacancy. The method promotes the

efficiency and precision (Li & Zhang, 2010).A comparison of attribute selection technique with rank of attributes is presented. If irrelevant, redundant and noisy attributes are added in model construction, predictive performance is affected so need to choose useful attributes along with background knowledge (Hall & Holmes, 2003). To improve the accuracy rate of classification and depth of tree, adaptive step forward/decision tree (ASF/DT) is proposed. The method considers not only one attribute but two that can find bigger information gain ratio (Tan & Liang, 2012). A new heuristic technique for attribute selection criteria is introduced. The best attribute, which have least heuristic functional value are taken. The method can be extended to larger databases with best splitting criteria for attribute selection (Raghu, Venkata Raju & Raja Jacob, 2012).Interval based algorithm is proposed. Algorithm has two phases for selection of attribute. First phase provides rank to attributes. Second phase selects the subset of attributes with highest accuracy. Proposed method is applied on real life data set (Salama, M.A., El Bendary, N., Hassanien, Revett & Fahmy, 2011).

Large training data sets have millions of tuples. Decision tree techniques have restriction that the tuples should reside in memory. Construction process becomes inefficient due to swapping of tuples in and out of memory. More scalable approaches are required to handle data (Changala, R., Gummadi, A., Yedukondalu & Raju, 2012).An improved learning algorithm based on the uncertainty deviation is developed. Rationality of attribute selection test is improved. An improved method shows better performance and stability (Sun & Hu, 2012).

Equivalence between multiple layer neural networks and decision trees is presented. Mapping advantage is to provide a self configuration capability to design process. It is possible to restructure as a multilayered network on given decision tree (Sethi, 1990).A comparison of different types of neural network techniques for classification is presented. Evaluation and comparison is done with three benchmark data set on the basis of accuracy (Jeatrakul & Wong, 2009).

The computation may be too heavy if no preprocessing in input phase. Some attributes are not relevant .To rank the importance of attributes, a novel separability correlation measure (SCM) is proposed. In input phase different subsets are used. Irrelevant attributes are those which increase validation error (Fu & Wang, 2003).

III. Proposed Method

The input processing of training phase is data sampling technique for classifier. Single layer RBF networks can learn virtually any input output relationship (Kubat, 1998). The cascade-layer network has connections from the input to all cascaded layers. The additional connections can improve the speed. Artificial neural networks (ANNs) can find internal representations of dependencies within data that is not given. Short response time and simplicity to build the ANNs encouraged their application to the task of attribute selection.

IV. Process Method

1. Sampling of data from 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.
8. Decode the feature variable in result class

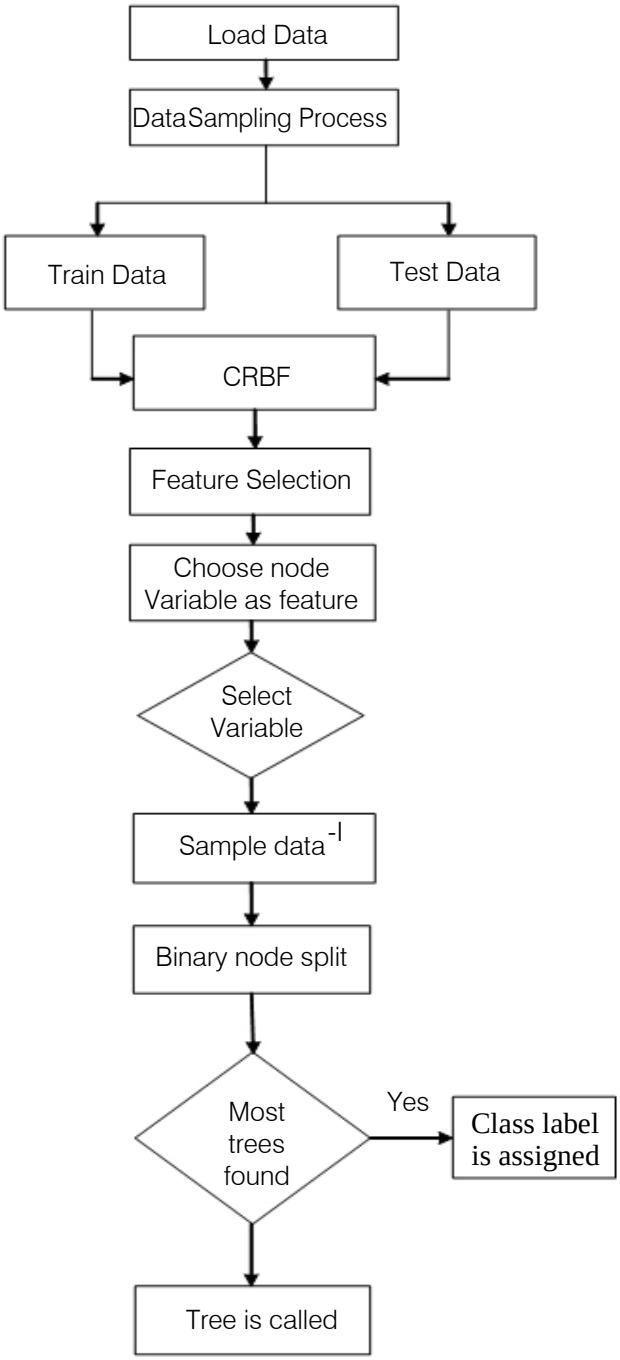

Figure 1 : Process block diagram of modified ID3-CRBF

V. Experimental Results

For the performance evaluation cancer dataset from UCI machine learning repository is used.

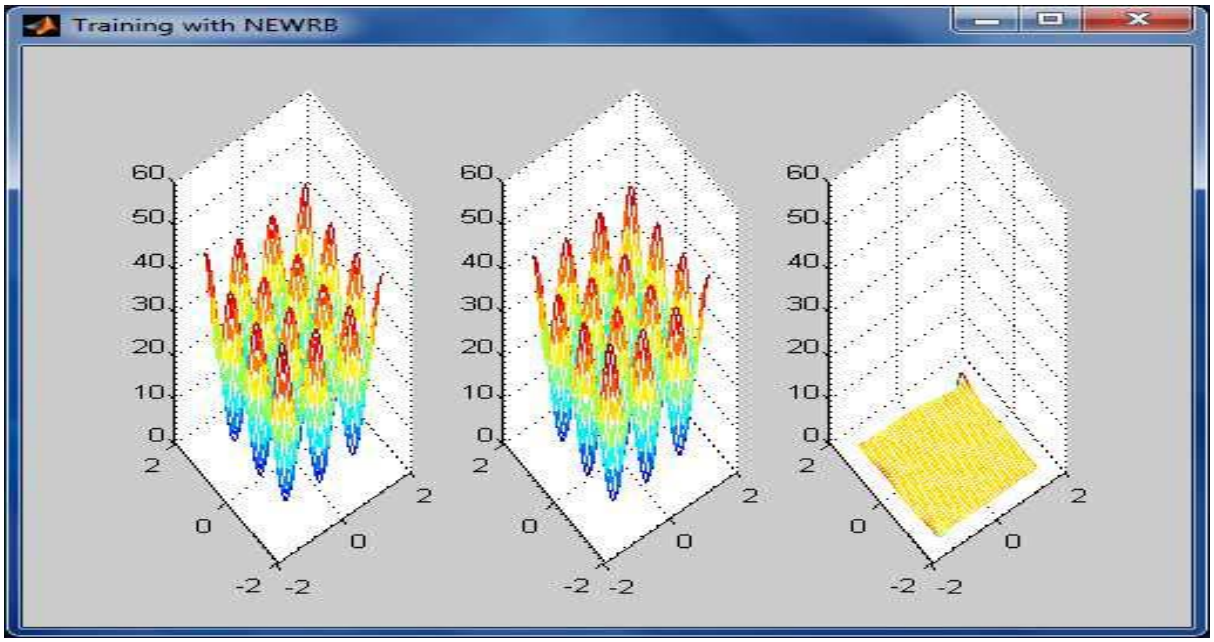

Figure 2 : Trained data, test data and unclassified region classification

Accuracy based on cross fold ratio of classifier.

Table 1 : Comparison of Accuracy

Cross Fold Ratio	Accuracy	
	ID3	ID3_CRBF
5	86.18	97.18
6	81.95	92.95
7	81.95	92.95

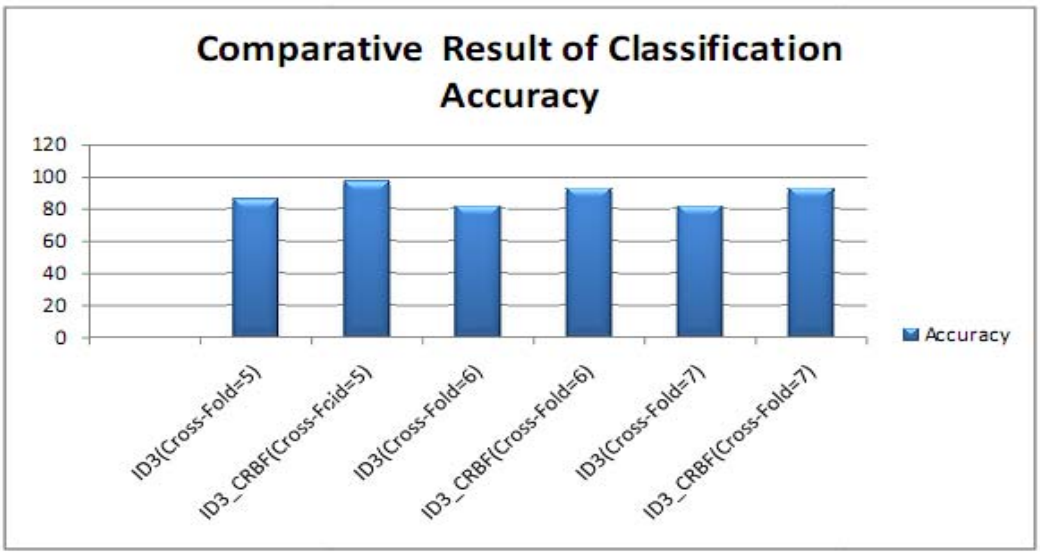

Figure 3 : Graphical representation of Table 1

VI. Conclusion

In this paper, we have experimented cascaded model of RBF with ID3 classification. The standard presentation of each attribute on selected ID3 is calculated and the Classify the given data. We can say from the experiments that the cascaded model of RBF with ID3 approach provides better accuracy and reduces the unclassified region. Increased classification region improves the performance of classifier.

References Références Referencias

1. Lakshmi, B.N., Raghunandhan, G.H. (2011). A conceptual overview of data mining. Proceedings of the National Conference on Innovations in Emerging Technology, 27-32.
2. Han, J., Kamber, M., Pei, J. (2012) DATA MINING: Concepts and Techniques. Elsevier. pp. 327-346.
3. Li, L., Zhang, X. (2010). Study of data mining algorithm based on decision tree. International Conference On Computer Design And Applications Vol. 1.
4. Quinlan, J.R. (1986). Induction of Decision Trees. Machine Learning, 1 (1): 81-106.
5. Hall, M.A., Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. IEEE transactions on knowledge and data engineering, Vol. 15, No.6.
6. Tan, T.Z., Liang, Y.Y. (2012). ASF/DT, Adaptive step forward decision tree construction. Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition.
7. Raghu, D., Venkata Raju, K., Raja Jacob, Ch. (2012). Decision tree induction-A heuristic problem reduction approach. International Conference on Computer Science and Engineering.
8. Salama, M.A., El-Bendary, N., Hassanien, A.E., Revett, K., Fahmy, A.A. (2011). Interval-based attribute evaluation algorithm. Proceedings of the Federated Conference on Computer Science and Information Systems, 153-156.
9. Changala, R., Gummadi, A., Yedukondalu, G., Raju, U. (2012). Classification by decision tree induction algorithm to learn decision trees from the class-labeled training tuples. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2. Issue 4.
10. Sun, H., Hu, X. (2012). An improved learning algorithm of decision tree based on entropy uncertainty deviation.
11. Sethi, I.K. (1990). Layered neural net design through decision trees.
12. Jeatrakul, P., Wong, K.W. (2009). Comparing the performance of different neural networks for binary classification problems. Eighth International Symposium on Natural Language Processing.
13. Kubat, M. (1998). Decision trees can initialize radial-basis function networks. IEEE transactions on neural networks, vol. 9, no. 5.
14. Fu, X., Wang, L. (2003). Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance. IEEE transactions on systems, man, and cybernetics vol. 33, no. 3.

This page is intentionally left blank