ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

Comparative Study of Decision Tree Classifier with and without GA based feature selection

Mrs. Shanta Rangaswamy[1], Dr. Shobha G[2], Sandeep R V[3], Raj Kiran[4]

Department of Computer Science and Engineering, R.V. College of Engineering, Bangalore, India

Abstract: Machine Learning techniques like Genetic Algorithms and decision trees have been applied to the field of classification for more than a decade. It can learn normal and anomalous patterns from training data and generate classifiers, which can be used to classify samples of unknown class. In general, the input data to classifiers is an extremely large set of features, but not all of features are relevant to the classes to be classified. Hence, the learner must generalize from the given examples in order to produce a useful output in new cases. In this paper, a comparison of decision tree with Genetic Algorithm based feature selection and a decision tree without Genetic Algorithm is carried out on different datasets.

Keywords: Decision Tree, Genetic Algorithm, ID3.

I. INTRODUCTION

1. DECISION TREES

A decision tree is made of decision (internal) nodes and leaf nodes. Each decision/internal node corresponds to a test X over a single attribute of the input data and has a number of branches, each of which handles an outcome of the test X. Each leaf node represents a class that is the result of decision for a case. The process of constructing a decision tree is basically a divide and conquer process. A set T of training data consists of k classes (C_1,C_2,..,C_k). If T only consists of cases of one single class, T will be a leaf. If T contains cases of mixed classes (i.e. more than one class), a test based on some attribute a_i of the training data will be carried and T will be split into n subsets (T_1, T_2;,..,T_n), where n is the number of outcomes of the test over attribute a_i .The same process of constructing decision tree is recursively performed over each T_j , where 1<j<n , until every subset belongs to a single class. The problem here is how to choose the best attribute for each decision node during construction of the decision tree. The criterion that ID3 chooses is Gain Ratio Criterion. The concept in this criterion is, at each splitting step, choose an attribute which provides the maximum information gain while reducing the bias in favor of tests with many outcomes by normalization. Once a decision tree is built, it can be used to classify testing data that has the same features as the training data. Starting from the root node of decision tree, the test is carried out on the same attribute of the testing case as the root node represents. The decision process takes the branch whose condition is satisfied by the value of tested attribute. This branch leads the decision process to a child of the root node. The same process is recursively executed until a leaf node is reached. The leaf node is associated with a class that is assigned to the test case [1][2][3][5].

2. GENETIC ALGORITHM

Genetic Algorithms have been successfully applied to solve search and optimization problems. The fundamental concept of this algorithm is to search a hypothesis space to find the best hypothesis. A pool of initial hypotheses called a population is randomly generated and each hypothesis is evaluated with a fitness function. A GA generally has four components: First, A population of individuals where each individual in the population represents a possible solution. Second, a fitness function which is an evaluation function, decides if an individual is a good solution or not. Third, a selection function, decides how to pick good individuals from the current population for creating the next generation, and fourth, genetic operators such as crossover and mutation which explore new regions of search space while retaining some of the current information at the same time [4][6].

ID3 ALGORITHM

ID3 algorithm is a greedy algorithm that selects the next attributes based on the information gain associated with the attributes. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node[3].

ID3 Algorithm.
FUNCTION ID3(R: a set of non-goal attribute, C: the goal attribute , S: a training set)
RETURN a decision tree,
BEGIN
IF S is empty,
Return a single node with value failure;
IF R is empty,
Return a single node with as value the most frequent of the values of the
goal attribute that are found in records
LET D be the attribute with largest
gain(D,S) aming attribute in R;
LET { dj | j=1,2,.....,m} be the value of
atrribute D;
LET{ sj | j=1,2,.....,m} be the subsets of
S consisting repectively of records

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

with value dj for atrribute D;
RETURN a tree with root labeled D and
arcs labeled d1,d2,....,dm going
respective to the trees ID3 (R- {D} ,C,S1) , ID3 (R-{D}, C,S2),...., ID3 (R-{D},C, Sm;
END ID3

The genetic algorithm and decision tree hybrid learning system was able to outperform the decision tree algorithm which was based on manual feature selection. It is due to the fact that the hybrid approach is able to focus on relevant features and eliminate unnecessary or distracting features. This initial filtering was able to improve the classification abilities of the decision tree [3][5]. The algorithm does take longer to execute than the standard decision tree; however, its non-deterministic process is able to make better decision trees. The training process needs to be done only once.

II. RESULTS AND INTERPRETATION

Three datasets, namely, groundwater, horse and wine dataset were used to carryout experiments.Each of them included various features that determined the class of the sample. The features used in all the samples had continuous values, equal width interval binding algorithm was used to discretize the samples and the decision construction was based on this discretized values.
The decision tree construction for a data set was done once, and was used for classification of all the samples in the test set.
From the comparative study of Decision Tree with GA based classification and traditional Decision Tree classification, it was concluded that GA based classification shows a higher average increase in accuracy then the traditional one. Ground water, horse and wine datasets samples were given as an input to training set model for experimentation.
The maximum training set accuracy achieved by using a decision tree built with GA based feature was found to be 98.67% for water dataset sample.

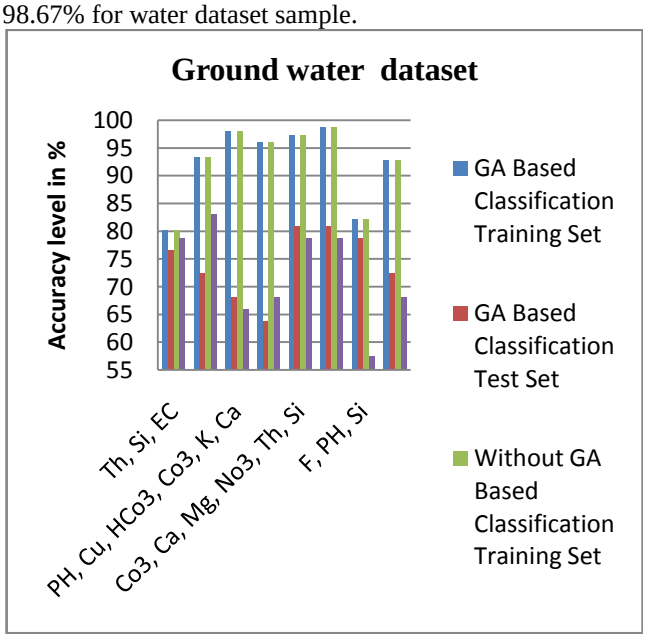

Figure 1: Graphical representation of sample attributes Vs Accuracy level for groundwater dataset sample

Likewise the proportion of K,Na,Cl,HCO_3, endotoxin, aniongap,PLA_2,SDH, GLDH,TPP, breath rate, PCV, pulse rate fibrinogen, dimer fibPerDim found in the blood sample of a horse is required to understand the health of the horse. The class labels for these attributes are either colic or healthy. The maximum training set accuracy achieved by using a decision tree built with GA based feature was 100%.

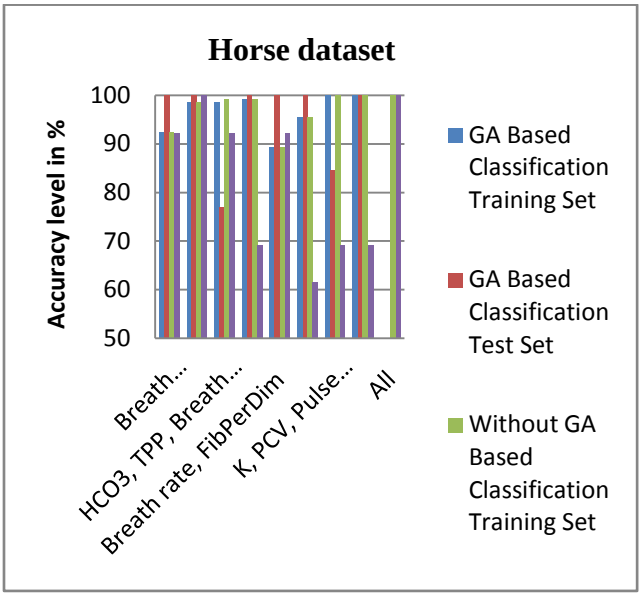

Figure 2: Graphical representation of sample attributes Vs Accuracy level for horse dataset sample

Eleven components in a wine sample that distinguishes the quality of wine as poor or excellent are fixed-acidity, volatile-acidity, citric-acid, residual-sugar, chlorides, free-sulphur-dioxide, sulphur-dioxide, density, pH, sulphates and alcohol. In this case, the accuracy level of 99.33% was found .
It was observed that the accuracy of the training set for the selected features for the classification with and without GA is same, with hardly little variation at times. The accuracy of training dataset without GA being at a lower side than that of training set accuracy with GA. On the other hand the test dataset accuracy level has a larger variation. In most of the cases the test accuracy level for the classification with GA was found to better than that achieved without GA. A genetic algorithm used for optimization of choosing the best features had a major role to play in and increasing the accuracy or efficiency level of the tool. Figure 1, figure 2 and figure 3 show the graphical representation of the comparison of accuracy levels for the ground water, horse and wine datasets.

From the comparative study of Decision Tree with GA Based Classification and traditional Decision Tree classification, it can be concluded that GA Based Classification shows a higher average increase in accuracy then the traditional one.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

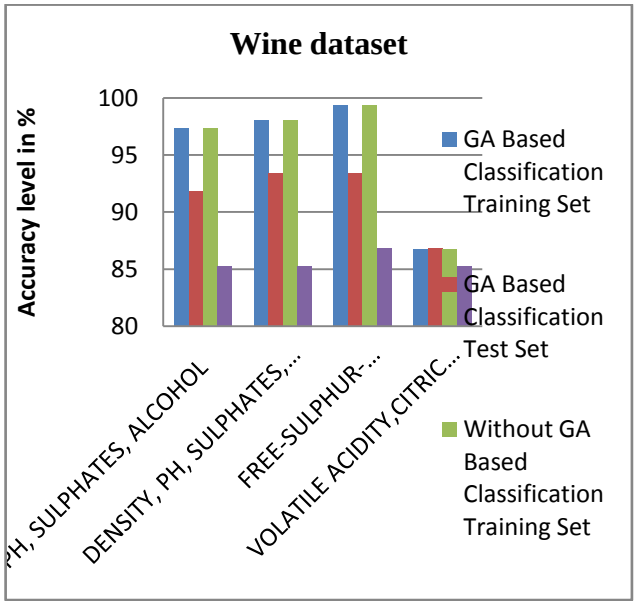

Figure 3: Graphical representation of sample attributes Vs Accuracy level for wine dataset sample

III. FUTURE ENHANCEMENTS

At present decision tree construction with GA Based classification takes longer time then the traditional one although it eliminates many of the unnecessary features; this is one area where more research can be done, to improve the response time.

Currently the GA based feature selection is considering the features entirely by the information through GA operations, this can be changed by allowing the user to specify certain features which might be a dominating factor for the sample classification from user perspective, and using the GA based selection for the remaining subset of features, comparing this with entirely selecting features based on GA is one more aspect to look into.

REFERENCES

[1] Gary Stein, Bing Chen, Annie S. Wu, Kien A. Hua "Decision tree classifier for network intrusion detection with GA based feature selection" ACM-SE 43: Proceedings of the 43rd annual Southeast regional conference- Vol. 2, 136-141, March [2005].

[2] J. Bala, J. Huang, H. Vafaie, K. DeJong and H. Wechsler "Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification". IJCAI conference, Montreal, August 19-25, 1995

[3] J. R. Quinlan, "Induction of decision tree," Machine Learning, Vol. 1, No. 1, pp. 81-106,1986.

[4] Zhiwei Fu, Bruce L. Golden, ShreevardhanLele, S. Raghavan , Edward A. Wasil "A Genetic Algorithm-Based Approach for Building Accurate Decision Trees". Vol. 15, No. 1, 1526-5528 electronic ISSN INFORMS Journal on Computing © 2003 INFORMS , Winter 2003.

[5] Hua; Kien A., Wu; Annie S.; Chen; Bing, Stein; Gary Decision Tree Classifier For Network Intrusion Detection With GA-based Feature Selection, in 43rd annual southeast regional conference, Vol 2, pp 136-141, 2005.

[6] Sung-Hyuk Cha Charles Tappert, "A Genetic Algorithm for Constructing Compact Binary Decision Trees", Journal of Pattern Recognition Research Vol 1, pp 1-13, 2009.

BIOGRAPHIES

Ms.Shanta Rangaswamy, Assistant Professor, Department of CSE, R.V. College of Engineering, Bangalore, is pursuing her PhD from Kuvempu University. Her research areas of interest are Autonomic computing, Data mining, Machine learning techniques, Performance Evaluation of systems,Cryptography and Steganography, and System Modeling and Simulation.

Dr. Shobha G., Professor and Head, Department of CSE, R V College of Engineering is associated with the college, since 1995.She has received her Masters degree from BITS, Pilani and Ph.D (CSE) from Mangalore University. Her research areas of interest are Database Management Systems, Data mining, Data warehousing, Business analytics, Image Processing and Information and Network Security.

Sandeep R V is currently pursuing B.E in Computer Science from R V College of Engineering, Bangalore, Karnataka, India. He is a programming enthusiast and constantly involves himself in various programming competitions. His areas of interest includes Data Mining, Big Data and Mobile Application Development. He is currently working in PayPal on Data analytics

Raj Kiran is currently pursuing B.E in Computer Science from R V College of Engineering, Bangalore, Karnataka, India. He has good management abilities and good with data analytics. His areas of interest includes Database Management and Data Mining. He is currently working on Canopy Clustering using MapReduce.