Elegant Decision Tree Algorithm for Classification in Data Mining

B. Chandra, Sati Mazumdar , Vincent Arena and N. Parimi
Indian Institute of Technology, University of Pittsburgh, NSIT,
New Delhi, India Pittsburgh, USA New Delhi.
bchandra104@yahoo.co.in maz1@pitt.edu

Abstract

Decision trees have been found very effective for classification especially in Data Mining. This paper aims at improving the performance of the SLIQ decision tree algorithm (Mehta et. al,1996) for classification in data mining The drawback of this algorithm is that large number of gini indices have to be computed at each node of the decision tree. In order to decide which attribute is to be split at each node, the gini indices have to be computed for all the attributes and for each successive pair of values for all patterns which have not been classified. An improvement over the SLIQ algorithm has been proposed to reduce the computational complexity. In this algorithm, the gini index is computed not for every successive pair of values of an attribute but over different ranges of attribute values. Classification accuracy of this technique was compared with the existing SLIQ and the Neural Network technique on three real life datasets consisting of the effect of different chemicals on water pollution, Wisconsin Breast Cancer Data and Image data It was observed that the decision tree constructed using the proposed decision tree algorithm gave far better classification accuracy than the classification accuracy obtained using the SLIQ algorithm irrespective of the dataset under consideration. The classification accuracy of this algorithm was even better compared to the neural network classification technique. Overall, it was observed that this decision tree algorithm not only reduces the number of computations of gini indices but also leads to better classification accuracy.

1. Introduction

Classification of data is one of the important tasks in data mining. Various methods for classification have been proposed such as decision tree induction and neural networks. CART (Breiman et. al, 1984) is one of the popular methods of building decision trees but it generally does not do the best job of classifying a new set of records because of overfitting. The ID3 decision tree algorithm was proposed by Quinlan in 1981 and there have been several enhancements have suggested to the original algorithm which include C4.5 (Quinlan,1993).The focus of ID3 is on how to select the most appropriate attribute at each node of the decision tree. A measure called *Gain* is defined and each attribute's gain is computed. The attribute with the largest gain is chosen as the splitting attribute.

One of the main drawbacks of ID3 is regarding the attribute selection measure used. The splitting attribute selection measure *Gain* used in ID3 tends to favour attributes with a large number of distinct values. This drawback was overcome to some extent by introducing a new measure called *Gain Ratio*. Gain Ratio takes into account the information about the classification gained by splitting on the basis of a particular attribute. The *Gini* index is an alternative proposed to Gain ratio and is used in in SLIQ algorithm (Mehta et. al.,1996). ID3 was also inadequate in handling missing or noisy data. Several enhancements to ID3 have since been proposed to overcome these drawbacks. These drawbacks are dealt with in SLIQ.

In the case of SLIQ algorithm has to be computed for each attribute at every node. In this paper, an Elegant decision tree algorithm has been proposed which is an improvement over the SLIQ algorithm. It is shown that this algorithm not only reduces the number of computations of gini indices drastically but also increases the classification accuracy over the original SLIQ algorithm and the Neural Network classification technique.

Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops)
0-7695-1754-3/02 $17.00 © 2002 IEEE

2. SLIQ Decision Tree Algorithm

SLIQ (Supervised Learning In Quest) and SPRINT (Shaefer et.al, 1996) was developed by the Quest team at IBM. SPRINT basically aims at parallelizing SLIQ. The tree is generated by splitting one of the attributes at each level. The basic aim is to generate a tree which is the most accurate for prediction. The *gini Index*, the measure of diversity of a tree, was introduced. With the help of the gini index it is decided which attribute is to be split and what is the splitting point of splitting of the attribute. The gini index is minimised at each split, so that the tree becomes less diverse as we progress. The class histograms are built for each successive pairs of values of attributes. At any particular node, after obtaining all the histograms for all attributes, the gini index for each histogram is computed. The histogram that gives the least gini index gives us the splitting point P for the node under consideration. The method to calculate gini index for a sample histogram with 2 classes namely A and B is defined as follows:

Table 1 Sample Class Histogram

Attribute Value<P	**A**	B
L	a_1	a_2
R	b_1	b_2

P Splitting value for an attribute
a_1 denotes the number of attributes which are less than P and belong to class A
a_2 denotes the number of attributes which are less than P and belong to class B
b_1 denotes the number of attributes which are greater than P and belong to class A
b_2 denotes the number of attributes which are greater than P and belong to class B

$$\text{Gini Index} = (a_1+a_2)/n * [1-\{a_1/(a_1+a_2)\}^2 - \{a_2/(a_1+a_2)\}^2] + (b_1+b_2)/n * [1-\{b_1/(b_1+b_2)\}^2 - \{b_2/(b_1+b_2)\}^2]$$

where n = total number of records = $a_1 + a_2 + b_1 + b_2$

Using this formula, the gini indices are computed for each histogram for each attribute. Once the gini index for each histogram is known, the split attribute is chosen to be the one whose class histogram gives the least gini index, and the split value equals the splitting point P for that histogram. The stopping criterion is when the gini index at a node becomes zero, as this implies that all data records at that node have been classified completely.

3. Elegant Decision Tree Algorithm

SLIQ proposes to scan each attribute, and each value corresponding to an attribute in order to ascertain the best split. However, in the case of large databases, with numerous data records, this process of scanning each attribute and each corresponding pair of successive values leads to a huge number of computations. In the modified SLIQ algorithm, it is proposed that instead of performing calculations for the gini index at each value in an attribute, the calculation be performed at a certain fixed interval within the range of values of an attribute. For instance, the computation may be performed at intervals of 10% of the number of records at each node. This would result in the total number of computations being significantly lesser than that in SLIQ. In SLIQ, the total number of computations per node is the product of the number of attributes (n) and the number of records at that node (m). In case of the modified algorithm, the total computations at a node is the product of the number of attributes (n) and the number of groups formed (g=m/k where k is the interval/group size). It is to be noted that since the number of unclassified patterns varies at each progressive node of the tree, the computation to divide these unclassified patterns into various groups has to be performed separately at each node of the tree. An optimum value of k needs to be chosen to generate the best results. Thus, depending on the value of k, a significant reduction is observed in the modified algorithm.

Algorithm

Input : Training set, samples consisting of attributes $A_1 \ldots A_n$ and records $R_1 \ldots R_m$. num is the no. of groups into which each attribute has to be divided into for computation of gini index.

```
attr stores the splitting attribute
ginival stores the final gini index
splitval stores the final splitting value of attribute attr
Assumed to be c no. of classes
//initializations
ginival = infinity // high value
splitval = 0
attr = 0
Copy records R1…Rm from samples to data.
Maketree(data)
{
    1.  construct attribute lists alist1…alistn such that alisti contains all records sorted on Ai.
    2.  //make histograms for each alisti
    3.  for i = 1 to n
    4.      val = (alist[i][m] – alist[i][0])/num
    5.      count=0
    6.      while count<num
    7.          split = alist[i][0] + count*val
    8.          let l[c] contain all records that have value in Ai<split. Order these
            records into l[0],l[1]…l[c] according to the class to which they belong
    9.      Repeat step 8 for records that contain value in Ai>=split to obtain r[c].
    10.         //calculate gini index
    11.         gini = calc(l[c],r[c])
    12.         if ginival>gini
    13.             ginival = gini
    14.             splitval = split
    15.             attr = I
    16.         count++
    17. Construct a new node with (attr,splitval,ginival)
    18. if gini = 0
    19.     return //classification complete
    20. Put all records with value in Aattr<splitval into data1
    21. Put all records with value in Aattr>=splitval into data2
    22. maketree(data1)
    23. maketree(data2)
}
//function to compute gini index
calc(l[c],r[c])
{
   let k be total no. of records in l[c] and r[c].
   //let ans be the return value
    1.  ans = 0, lval = 1, rval = 1.
    2.  lsum = sum of no. of records in l[c].
    3.  rsum = sum of no. of records in r[c].
    4.  for i = 1 to c
    5.      lval -= (l[i]/lsum)^2
    6.      rval -= (r[i]/rsum)^2
    7.  lval *= lsum/k.
    8.  rval *= rsum/k.
    9.  ans = lval + rval.
    10. return(ans).
}
```

4. Results

Classification accuracy using original SLIQ algorithm, the Elegant decision tree algorithm and the Neural Network technique was compared using three datasets viz. the pollution dataset, Wisconsin Breast Cancer Dataset and the Image Dataset.

4.1 Classification using SLIQ and Elegant decision tree algorithm

Pollution Dataset

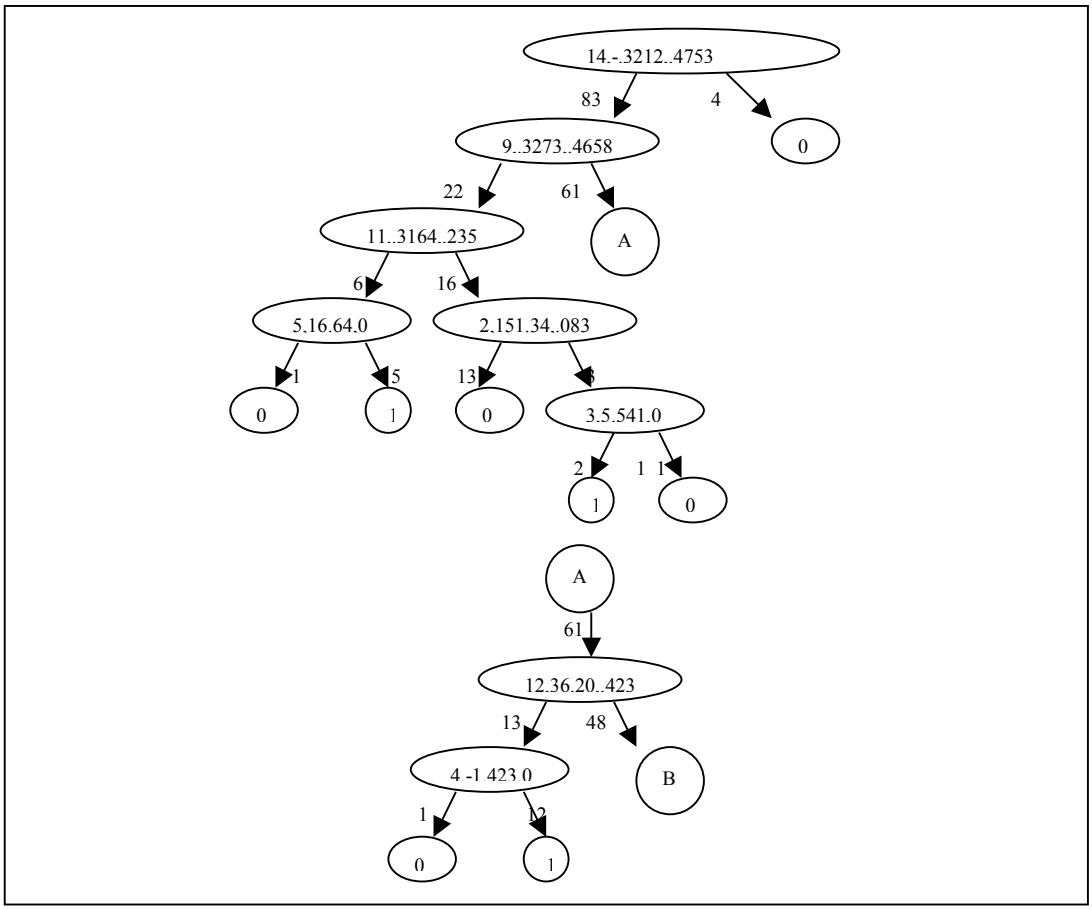
The pollution dataset comprises of data samples (107 records or patterns) for testing the chemical pollution of water based on a set of 15 attributes which signify the effect of 15 different types of chemicals. The output reveals whether or not the water sample is polluted, denoted by 1 or 0. Decision trees were built both by SLIQ and the modified SLIQ algorithm using 87 records for training. Remaining 20 records were used for testing the classification accuracy. Classification on test data using SLIQ algorithm is given below. Coding of the algorithm was done in C++.

Target classes: 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1

Output classes: 0 0 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 1 1

Classification Accuracy = 12/20 = 60%.

The decision tree generated using the modified form of SLIQ algorithm is shown in Fig.1. Each attribute was divided into k groups. In this case k was equal to 9. The first value in the elliptical boxes denotes the attribute number and the second value denotes the splitting value of the attribute. At the first step, four records have been classified as class 1 perfectly and the remaining 83 are yet to be classified.



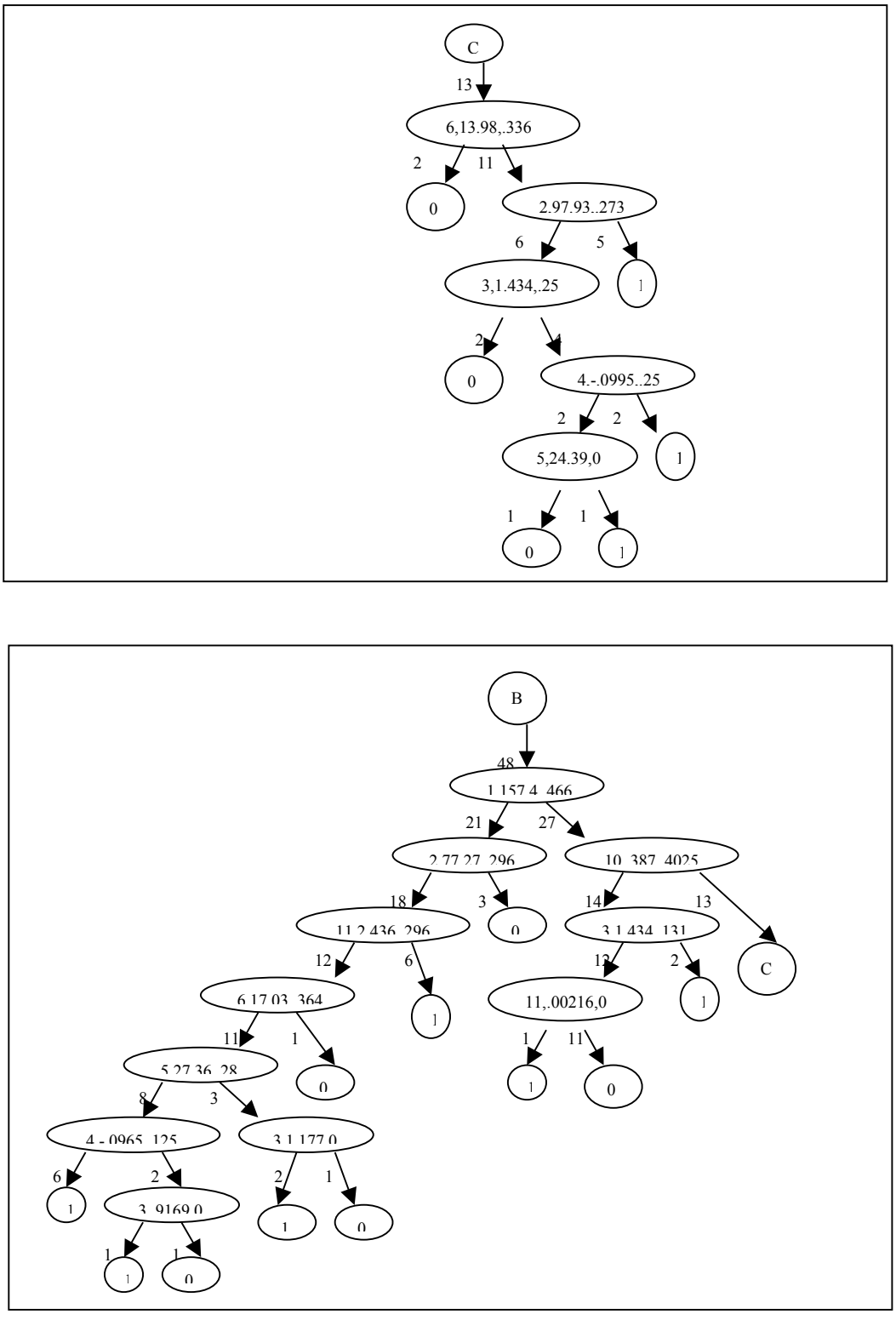


Figure 1 . Decision Tree using Elegant Decision Tree Algorithm (Pollution Data)

Classification accuracy for the pollution dataset using the Elegant decision tree algorithm is given below:

Target classes: 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
Output classes:1 0 0 1 0 0 1 1 0 0 1 1 1 1 1 1 1 1 1 1

Classification Accuracy = 80%.

Wisconsin Breast Cancer Data

250 patterns with 9 attributes.were considered for training from Wisconsin breast cancer data. The attributes are as follows:

#	Attribute	Domain
1.	Clump Thickness	1-10
2.	Uniformity of Cell Size	1-10
3 .	Uniformity of Cell Shape	1-10
4.	Marginal Adhesion	1-10
5.	Single Epithelial Cell Size	1-10
6.	Bare Nuclei	1-10
7.	Bland Chromatin	1-10
8.	Normal Nucleoli	1-10
9.	Mitoses	1-10

There were two classes namely Benign and Malignant denoted by 0 and 1. The decision tree is shown in figure 2. 20 patterns were used for testing. Classification accuracy using original SLIQ algorithm is:

Target classes: 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
Output classes: 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 1 1 0 0

Classification Accuracy = 75%.

Classification accuracy using Elegant decision tree algorithm is as follows:

Target classes: 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
Output classes: 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1

Classification Accuracy is 100%.

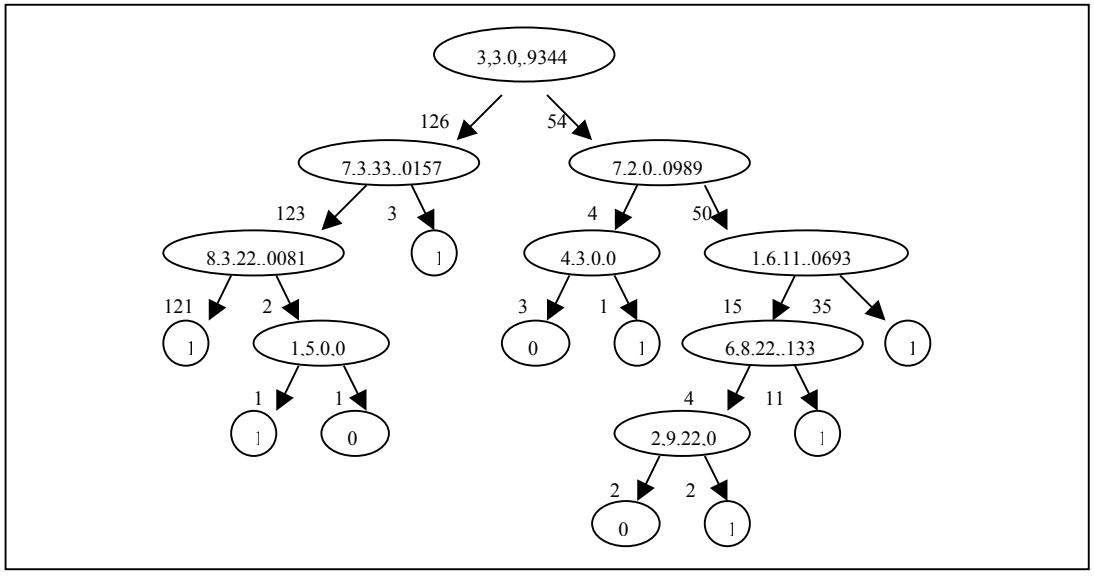


Figure 2. Decision Tree using Elegant Decision Tree Algorithm (Breast cancer data)

Image dataset

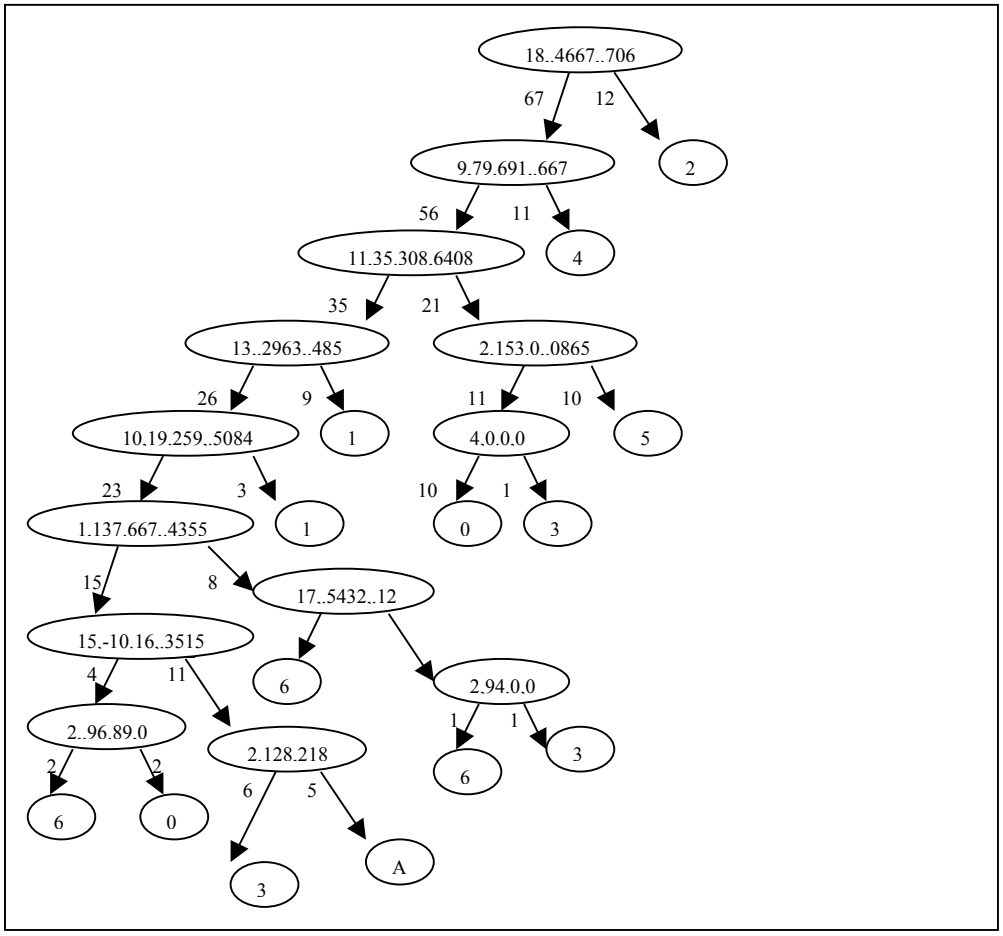
The dataset classifies a given image into one of seven class labels:
Cement, Brickface, Grass, Foliage, Sky, Path and Window. Classification is based on 18 features of each image-

centre pixel column of the region
centre pixel row of the region
low contrast line count
high contrast line count
mean horizontal contrast
standard deviation of horizontal contrast
mean vertical contrast
standard deviation of vertical contrast
average color intensity
average red intensity
average blue intensity
average green intensity
excess red intensity
excess blue intensity
excess green intensity
3-d nonlinear transformation of RGB
mean saturation
mean hue

The decision tree is shown in figure 3.
Classification accuracy using original SLIQ algorithm is as follws:

Target classes: 3 4 5 6 0 1 2 3 4 5 5 6 0 1 2 3 4 5 6 0
Output classes: 3 4 5 3 0 1 2 3 4 5 5 3 3 1 2 0 4 5 3 0

Classification Accuracy = 75%.



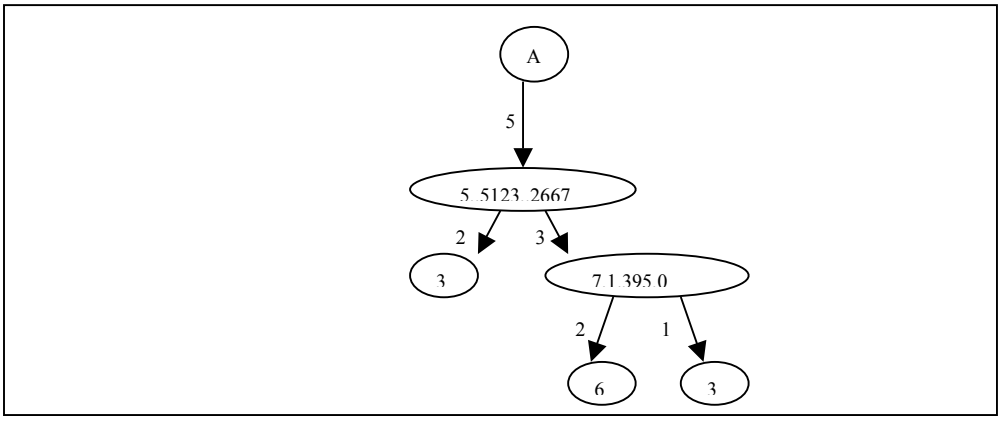


Figure 3. Decision Tree using Elegant Decision Tree Algorithm (Image data set)

Target classes: 3 4 5 6 0 1 2 3 4 5 5 6 0 1 2 3 4 5 6 0
Output classes: 3 4 1 6 0 1 2 3 4 5 5 6 0 1 2 3 4 5 6 0

Classification accuracy = 95%.

4.2 Classification using Neural Networks

All the data sets were also for trained using Neural Network technique. Back propagation algorithm (Rumelhart, 1986) was used for classification. . Classification results for the pollution test data using Neural Network are as follows. The model comprised of 15 input neurons and 2 output neurons.

Output classes:

Columns 1 through 7

0.9997	0.0643	0.0000	0.0000	0.0000	0.0133	0.0000
0.0003	0.9387	1.0000	1.0000	1.0000	0.9878	1.0000

Columns 8 through 14

0.9418	0.0003	0.9769	0.9997	1.0000	0.0000	0.0001
0.0595	0.9997	0.0211	0.0003	0.0000	1.0000	0.9999

Columns 15 through 20

0.9646	0.0000	0.0000	0.0000	0.0000	1.0000
0.0327	1.0000	1.0000	1.0000	1.0000	0.0000

Classification accuracy=14/20=70%.

Classification results for the Breast cancer test data using Neural Network are as follows. The model comprised of 9 input neurons and 2 output neurons.

Output classes :

Columns 1 through 7

0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.0000	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Columns 8 through 14

0.0000	0.0000	0.0002	1.0000	0.0030	1.0000	1.0000
1.0000	1.0000	0.9998	0.0000	0.9969	0.0000	0.0000

Columns 15 through 20

1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Classification accuracy=18/20=90%.

Classification results using Neural Network for the Image test data are as follows. !8 input neurons and 7 output neurons were taken in the model.

Output classes:

Columns 1 through 7

1.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
0.0000	0.9998	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.9999	0.0000	0.0000	0.0001	0.0000
0.0000	0.0000	0.0000	0.9926	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	1.0000	0.0007	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
0.0000	0.0010	0.0001	0.9979	0.0000	0.0000	0.9999

Columns 8 through 14

1.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
0.0055	1.0000	0.0000	1.0000	0.0000	0.0000	0.0000
0.0000	0.0001	0.0000	0.0003	0.9996	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0012	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9989
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0013
0.0001	0.0001	0.0000	0.0000	0.0000	1.0000	0.0000

Columns 15 through 20

0.0004	0.0000	1.0000	0.0000	0.0000	0.0033
0.0000	0.0266	0.0000	1.0000	0.0000	0.0001
0.0000	0.0000	0.0000	0.0003	0.9998	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.9994	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	1.0000	0.0000	0.0000	0.0000	0.0000

Classification Accuracy=90%.

4.3 Comparison of classification accuracy on test data using various classification techniques

Table 2. Classification Accuracy of Datasets using Various Algorithms

Classification Techniques	SLIQ	Elegant Decision Tree	Neural Networks
Pollution data	60 %	80 %	70 %
Breast cancer data	75 %	100 %	90%
Image data	75 %	95 %	90 %

It is observed that for all the data sets the elegant decision tree algorithm gives better classification accuracy compared to the original SLIQ decision tree algorithm. The new algorithm gives better classification accuracy even compare to neural networks.

5. Conclusions

The paper presents an elegant decision tree algorithm which is a modification of SLIQ algorithm. The main advantage of this algorithm is that it helps in drastic reduction of the number of computation of gini indices at each node of the tree and at same time helps in increasing the classification accuracy. The classification results on three popular data sets show that this new algorithm gives far more superior classification accuracy compared to the SLIQ algorithm and the neural networks technique.

6. References

[1] Alis Hadi. Identifying multiple outliers in multivariate data. J.R. Statist. Soc.,1992 B 54, No. 3,pp. 761-771.

[2] Breiman L, Freidman J. , Olshen R. and Stone C.: Classification and Decision Trees , Wadsworth,1984.

[3] Mehta M., Agarwal R. and Rissanen J. SLIQ: A fast scalable classifier for data mining, Proceedings of the International conference on Extending Database Technology,France, March 1996

[4] Quinlan J.R. Introduction of decision tree. Machine Learning,1:81-106,1986

[5] Quinlan J.R C4.5 : Programs for Machine Learning:Morgan Kauffman, 1993

[6] Rocke, M. and David L. Woodruff. Identification of outliers in multivariate data. Journal of the American Statistical Association. September 1996, vol. 91, No. 435, Theory and Methods.

[7] Rumelhart, D.E., Hinton G.E. and R.J. Williams, " Learning internal representation of error propagation",in D.E. Rumelhartand J. McCelland editors, Parallel Distributed processing, Cambridge, MIT press,1986.

[8] Shafer J.C., Agarwal R. and Mehta M.. SPRINT: A scalable parallel classifier for data mining, Proceedings of the 24th International Conference on Very lrge Databases, Mumbai, India September 1996.