An Algorithm for Better Decision Tree

[1.]T.Jyothirmayi, [2.] Suresh Reddy

[1.] Assistant Professor, Department of Computer Science, GITAM University, Visakhapatnam-530 013, Andhra Pradesh.
[2]MCA Student, Department of Computer Science, GITAM University, Visakhapatnam-530 013, Andhra Pradesh.

Abstract: The present paper aims at constructing the decision tree for a given database which adopts an improved ID3 decision tree algorithm to implement data mining in order to predict the output. The database is generated using the sampling techniques and the classification algorithm is applied on the samples. The obtained results are compared with experimental results in order to verify the validity and accuracy of the developed model.

Keywords- Sampling, Decision Tree, ID3, classification

1. Introduction

Now-a–days the data stored in a database and which is used for applications is huge. This explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge. Hence data mining has become a research area with increasing importance [1]. The problem of classifying objects has been approached in many ways. In [2], Support vector machines (SVM) applied for breast cancer diagnosis. To access the optimal input feature subset for SVM and have a good prediction and less computation time, FScore used to limit the number of input feature. In "An improved decision tree classification algorithm based on ID3 and the application in score analysis" [3] Huang Ming developed an algorithm to implement score analysis. In his paper he made an attempt to make a comparison between ID3 and improved ID3 algorithm in terms of node number, the rule number, the classified precision. Mahesh V**.** Joshi and George Karypis in their paper "ScalParC : A New Scalable and Efficient Parallel Classification Algorithm for Mining Large Datasets"[4] presented ScalParC (Scalable Parallel Classifier), a new parallel formulation of a decision tree based classification process using MPI. Kimiko Matsuoka, Shigeki Yokoyama implemented c4.5 algorithm and statistical techniques in "Mining Rules for Risk Factors on Blood Stream Infection in Hospital Information System"[5]. They used C4.5,chi-square test, odds ratio, logistic regression, and adjusted residual analysis in order to extract certain patterns from hospital clinical microbiology database, whose aim is to analyze the effects of lactobacillus therapy and the background risk factors on blood stream infection in patients.

2. Model Used

As there are many challenging research issues to data mining direct applications of methods and techniques developed in related studies in machine learning, statistics, and database systems, a data mining method techniques are required for efficient and effective data mining.
The predictive models are used to predict or determine the future outcome rather than current behavior.

2.1 Decision Tree:

Decision trees are a popular structure for supervised learning. There are various algorithms for constructing a dec ision tree like c4.5 and cart algorithm.
The basic ID3 algorithm works well for limited number of records in data set and it cannot handle missing values and also when the data set size is increased the tree is not accustomed to the changes. The ID3 algorithm uses the entropy to select a splitting attribute and then construct the tree. There are some other algorithms that consider the gini index and gain ratio to select the splitting criterion and attribute.

Advantages of Decision Trees

- They are easy to use and efficient.
- Rules can easily be generated and are easy to interpret.
- They are suitable for large databases also.
- Each tuple in the data has to be filtered through the tree. This takes the time proportional to height of the tree.

Disadvantages also exist with the Decision Trees. They do not easily handle continuous data. Over fitting may occur.

The issues with Decision Tree algorithms are

- Choosing the splitting attribute.
- Ordering of the splitting attribute
- Splits
- Tree structure
- Stopping criteria
- Training data
- Pruning.

For ID3 decision tree, concept used to quantify information is called entropy. Entropy is used to measure the amount of uncertainty in a set of data. When all data in a set belongs to

a single class the entropy is zero that is there is no uncertainty.

Given the probabilities $p_1,p_2,.....p_n$ where $\sum_{i=1}^{n} p_i=1$,
Entropy I is calculated as

$I(p1,p2,...pn)= \sum_{i=1}^{n} (p_i \log(1/p_i)$ (1)

Gain (D,S)=I(D)- ∑p(Di)I(Di) (2)
Our current algorithm also make use of gain values to construct the decision tree.

2.2 Algorithm:

i) Stratified sampling:

```
Stratified( f, M, N )
  {
     finalEstimate = 0
     for i = 1 .. M
     {
        curEstimate = 0
        for j = 1 .. N
        {
           curEstimate += f(
uniformRandRange( (i-1)/M, i/M ) );
        }
        finalEstimate += curEstimate/N;
     }
     return finalEstimate;
  }
```

ii)Better Decision Tree Algorithm:

```
PROCEDURE BuildTree(Data, ATTRIBUTE)
{
        Build(Data);
        IF (all Risk Class values of sample data in Data are
        the same)
        THEN Return N as a leaf node;
        ELSE
        {
         FOR (each attribute in ATTIBUTE)
        {
        IF (the attribute of the node hasn't been used to be
        a classification attribute before) THEN
                Compute the information gain of the
        attribute of the node;
        }
        IF ( the attribute whose information gain is the
        biggest (>0) is marked as ATT)THEN
        {
         Mark node as the node which needs to be divided
        next step according to ATT ;
        Divide N into Nk, and generate each branch of the
        node N;
        }
        ELSE
         {
        Mark the node as a bad node;
        Return the node as a leaf node;
        }
        FOR (each branch Nk)  BuildTree(N, ATTIBUTE);
        }
}
```

3. Experiment

For the current paper three datasets have been considered. The automobile showroom's customer data is selected as one dataset and second the bank customers data to predict the credit risk and the mpg data of an automobile. Table 1 shows a sample of the automobile customer's data. The automobile showroom needs information to predict whether a customer who approaches them will purchase a car or not or it can be useful to predict what type of car are they willing to purchase. The manager is interested in customers details like name, age, gender , address, phone number whether he owns a car, his marital status, and also his income level. Even though the showroom collects all the information only a part of it is used by the manager. Hence only relevant attributes are selected in this paper to construct a decision tree and predict a new customer's willingness to purchase a car.

Table 2 shows the bank customers data used to predict customer's credit risk. The bank needs different types of information in order to manage risk through capital allocation for Value at Risk coverage. The bank is concerned about the customer's details and his credit risk. While constructing the data warehouse bank may collect various information about the customer, but only a part of it is used to predict the risk.

Table 3 shows the automobile MPG data set. Based on the attributes a car model can be predicted. The third data set consists of different attributes of a car and type of car. So based on the properties of a car like mpg, number of cylinders, displacement, horsepower and other we can predict the class to which a car belongs to. In all the three cases the data obtained is very large and entire data is not needed for validation hence some sampling techniques were used to generate the samples and the algorithm is applied on the samples.

In our proposed algorithm some sampling techniques were applied on datasets and samples were generated then algorithm proceeds by calculating gain and gini values for the attributes in samples. Based on the gini values obtained corresponding attribute is selected as best split attribute and decision tree is constructed.

For each data set two samples were generated using two sampling techniques random sampling and stratified sampling. 20 records from those samples are shown in this paper.

sno	age	edu	sal	noc	occ	std	ME	C
1	24	2	3	3	2	3	0	1
2	45	1	3	10	3	4	0	1
3	43	2	3	7	3	4	0	1
..								

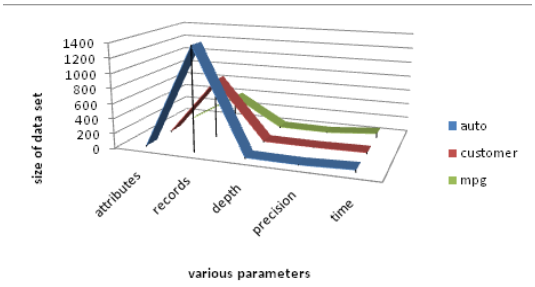
Table1. Automobile Customer Data

sno	married	owns car	age	income level	gender	credit risk
1	no	Yes	30	30000	male	c2
2	no	No	35	20000	female	c1
3	no	Yes	28	18000	female	c2
4	yes	Yes	44	19000	male	c1
....	...	...	..	..	..	..

Table 2. Customer credit risk

Data set	*No of Attr*	*Sample 1*			
		Number of records	Max Depth of tree	Precision	No of nodes
Automobile customer	8	1400	10	0.98	69
Customer credit risk	6	800	7	0.74	17
Car MPG	8	400	10	0.14	64

Table 4. showing the the sample details like number of records, attributes and number of nodes generated in tree



Data set Sample1 using random sampling

	Error rate	*Importance of variable with 100%*	*Value of std error*
Automobile customer	0.8206	education	0.0288
Customer credit risk	0.7260	education	0.0
Car MPG	0.1406	displacement	0.0003

Table 5. showing the error rate and standard error and 100% importance variable

sno	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Car name
1	18	8	307	130	3504	12	chevrolet chevelle Malibu
2	15	8	350	165	3693	11.5	buick skylark 320
3	18	8	318	150	3436	11	Plymouth satellite
4	16	8	304	150	3433	12	amc rebe sst
....	..	..	..				

Table 3. MPG dataset

Data set Sample2 using stratified sampling

	Error rate	*Importance of variable with 100%*	*Value of std error*
Automobile customer	0.79	education	0.025
Customer credit risk	0.719	education	0.0
Car MPG	0.1407	displacement	0.0029

Table 6. showing the error rate and standard error and 100% importance variable

4. Conclusion

This paper uses one of the sampling techniques to generate the samples from the original datasets and then apply the improved decision tree algorithm which overcomes the limitations of the ID3 algorithm. The results have shown that the samples generated are identical and the values are very accurate when compared to two samples. Hence the proposed algorithm can be implemented for various datasets where size of data is large.

References

[1] UM Fayyad, G Piatetshy-Shapiro, P Smyth, and R Uthurusamy,Advances in Knowledge Discovevy and Data Mining AAAI/MIT Press, 1996

[2] Akay.M. "Support vector machines combined with feature selection for breast cancer diagnosis". Expert Systemswith Applications, 2009, 36, pp.3240–3247.

[3] An improved decision tree classification algorithm based on ID3 and the application in score analysis by Huang Ming, NiuWenying, Liang Xu.

[4] ScalParC : A New Scalable and Efficient Parallel Classification Algorithm for Mining Large Datasets by Mahesh V. Joshi, George Karypis, Vipin Kumar Department of Computer Science University of Minnesota, Minneapolis, MN.

[5] Mining Rules for Risk Factors on Blood Stream Infection in Hospital Information System Kimiko Matsuoka *Osaka Prefectural General Medical Center, Osaka 558-8558, Japan,*

[6] A Data Mining Approach to Credit Risk Evaluation and Behaviour, Sara C. Madeira , Arlindo L. Oliveira , Catarina S. Conceição

[7] "Lan Huang, Chun-guang Zhou, Yu-qin Zhou, Zhe Wang, "Research on Data Mining Algorithms for Automotive Customers' Behavior Prediction Problem," icmla, pp.677-681, 2008 Seventh International Conference on Machine Learning and Applications, 2008