Evaluating Performance of Decision Tree Algorithms

[1]Harvinder Chauhan, [2]Anu Chauhan

[1]Assistant Professor, P.G.Dept. of Computer Science, Kamla Nehru College ForWomen, Phagwara (Punjab)
[2]Assistant Professor, P.G.Dept. of Computer Science,

Abstract- Among decision tree classifiers, Bayesian classifiers, k-nearest-neighbor classifiers, case-based reasoning, genetic algorithms, rough sets, and fuzzy logic which are some common classification methods in data mining, decision tree classifier is the most commonly used because it is easier to understand and implement compared to other classification methods. In this study, performance evaluation of two decision tree algorithm named ID3 and C4.5is done in order to discover which one is superior on the basis of accuracy with the help of a software tool using publicly available datasets.

Index Terms- ID3,C4.5,IDE3

I. Introduction

In the past, to extract information by data analysis was a manual and pain staking process because much domain knowledge was required, and understanding of statistical approach is also needed. However such approach will become inappropriate while facing the rapidly growing sizes and dimensions of the data. A community of researchers devoted themselves to the field called "data mining" to solve automating data analysis problem and discover the implicit information within the huge data (Giordana and neri,1995).Data classification is one of data mining techniques used to extract models describing important data classes. Some of the common classification methods used in data mining are: decision tree classifiers, Bayesian classifiers, k-nearest-neighbor classifiers, case-based reasoning, genetic algorithms, rough sets, and fuzzy logic techniques. Among these classification algorithms decision tree algorithms is the most commonly used because of it is easy to understand and cheap to implement.

1.1 Decision tree algorithm ID3-ID3 (Iterative Dichotomiser 3) decision tree algorithm was introduced in 1986 by Quinlan Ross (Quinlan, 1986 and 1987). It is based on Hunt's algorithm and it is serially implemented. Like other decision tree algorithms the tree is constructed in two phases; tree growth and tree pruning. Data is sorted at every node during the tree building phase in-order to select the best splitting single attribute (Shafer et al, 1996). IDE3 uses information gain measure in choosing the splitting attribute.

1.2 Decision tree algorithm C4.5-C4.5 algorithm is an improvement of IDE3 algorithm, developed by Quinlan Ross (1993). It is based on Hunt's algorithm and also like IDE3, it is serially implemented. Pruning takes place in C4.5 by replacing the internal node with a leaf node thereby reducing the error rate (Podgorelec et al, 2002). Unlike IDE3, C4.5 accepts both continuous and categorical attributes in building the decision tree. It has an enhanced method of tree pruning that reduces misclassification errors due noise or too-much details in the training data set.

II. Evaluating Performance

We evaluated performance of classifier ID3 and C4.5 on the basis of accuracy. The parameters are:

a) Number of instances in dataset
b) Number of attributes in dataset

III. Experimental Results

Dataset	Instances	Accuracy rate (%)	
		Id3	C4.5
Iris	150	90	96
Bank_class	300	57	68
Cardata	1728	89	92
mushroom	2074	89	88

Table 3.1 Results for dataset iris, bank, car and mushroom

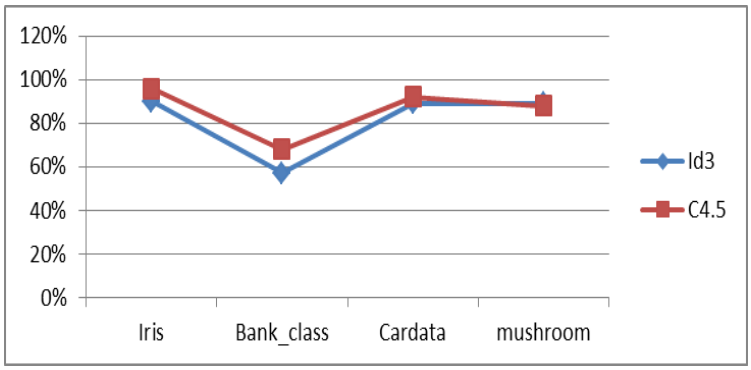

Figure 3.1 Results of classification accuracy in terms of no. of instances. Numeric values for this chart is shown in table 3.1

3.1Performance over the iris dataset- Iris dataset contains 150 instances of type numeric having three types of classes. As shown in table 3.1 the obtained accuracy from both algorithms is good. C4.5 gains a slight classification accuracy advantage over ID3.

3.2 Performance over bank_class- Bank_class dataset consists of 300 instances of numeric as well as categorical type. C4.5 algorithm outperforms decision tree induction on this dataset by 11% classification accuracy. So from this we can say

that C4.5 should be preferred when there are numeric as well as categorical attribute values in a dataset.

3.3 Performance over cardata dataset- Cardata dataset contains 1728 categorial instances. C4.5 proved to generalize to the car dataset with greater accuracy than ID3. As cardata contains four classes that's why ID3 provides just 3% less accuracy as compared to C4.5 algorithm.

3.4 Performance over mushroom dataset- Both algorithm performed very well on mushroom dataset which consists of 2074 instances of categorical type. Here ID3 provides greater accuracy as compared to C4.5due to its categorical nature and because number of classes in this dataset is two.

Dataset	Attributes	Accuracy rate (%)	
		Id3	C4.5
Segment	20	89	93
Cmc	10	86	93
Soybean	36	90	92
Bank_data	12	50	62

Table 3.2 Results for dataset segment, cmc, soybean, bank_data

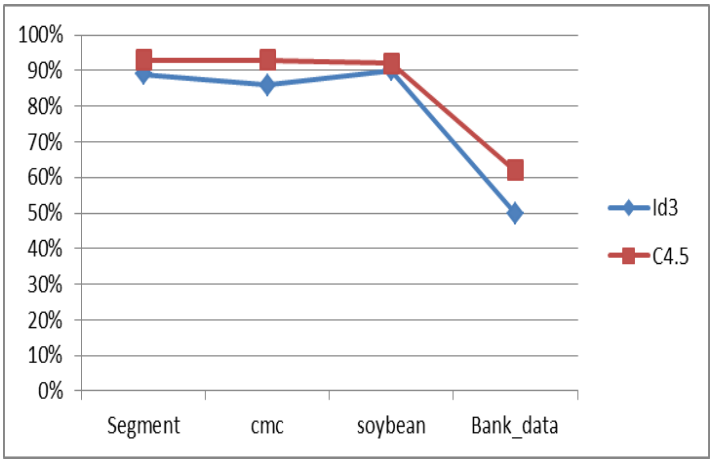

Figure 3.2 Results in terms of no. of attributes. Numeric values shown in table 3.2

3.5 Performance over segment dataset- Segment dataset consists of twenty attributes of type numeric having seven classes. C4.5 provides greater accuracy as compared to ID3 with slight difference of 4%.

3.6 Performance over cmc dataset- Contraceptic method used (cmc) dataset consists of ten attributes of type numeric, binary and categorical as well having three classes. For this dataset also C4.5 outperform ID3 by 7%.

3.7 Performance over soybean dataset- Soybean dataset consists of thirty six attributes of type nominal or categorical having nineteen classes. For this dataset C4.5 provides higher accuracy as compared to ID3.

3.8 Performance over bank_data- Bank data consists of twelve attributes of type categorical and numeric and have two classes. Algorithm C4.5 outperform ID3 by 12% accuracy.

IV. Conclusion

Our experimental analysis of performance evaluation of the commonly used decision tree algorithms ID3 and C4.5 using publicly available datasets shows, for datasets with many instances and for datasets with many attributes C4.5 outperformed ID3. This suggests that C4.5 algorithm outperformed ID3 in terms of classification accuracy.

References

[1] Anyanwu, M., and Shiva, S. (2009). Application of Enhanced Decision Tree Algorithm to Churn Analysis. 2009 International Conference on Artificial Intelligence and Pattern Recognition (AIPR-09), Orlando Florida

[2] Jiaweihan and michelinekamber. Data mining concepts and techniques, second edition,285-291

[3] Matthew N.anyanwu, Sajjan g. shiva. Comparative analysis of serial decision tree classification algorithms

[4] Mehdi piroozma, Youpingdeng, jack y yang and maryqu yang. A comparative study of different machine learning methods on microarray gene expression data, BMC genomics

[5] Tzung-I tang,GangZheng, Yalouhuang,GuangfuShu,Pengtaowang. A comparative study of medical data classification methods based on decision tree and system reconstruction analysis. IEMS vol.4,no.1,pp-102-108,june 2005

[6] Xu,M, wang, J. Chen, T. (2006). Improved decision tree algorithm: ID3+, intelligent computing in signal Processing and pattern recognition, Vol. 345, PP.141-149

Authors

First Author – Harvinder Chauhan, Assistant Professor, P.G.Dept. of Computer Science, Kamla Nehru College ForWomen, Phagwara (Punjab), harrymit21@yahoo.com
Second Author – Anu Chauhan, Assistant Professor, P.G.Dept. of Computer Science, Anu.chauhan711@yahoo.com