An Empirical Comparison Of Inductive Learning Algorithms On Large Dataset

S. Sivakumari[1], R. Praveena Priyadarsini[2], S. Chandralekha[3]

Abstract

Data mining is the extraction of hidden predictive information from large database. Classification is a data mining task that takes a large collection of examples from multiple groups as inputs and identifies the characteristics patterns or property for each group. One common approach to classification is to use decision tree. Decision tree classification method has emerged as the essential knowledge acquisition procedure which follows the machine learning strategy, 'learning from examples'. In this paper we perform a comparative study of the performance of the decision tree classification algorithms C4.5 and C5. C4.5 algorithm constructs the decision tree with a 'divide and conquer' strategy. C5 algorithm uses the concept of gain to produce a classifier in the form of decision tree according to the previously chosen classification. These two algorithms are applied to the large data set 'adult', obtained from the UCI Machine Learning Repository, which is used to predict the individual's income.

The result indicates that C5 algorithm has a higher performance rate when compared with C4.5 algorithm.

Keywords : Data mining, Classification, Large dataset, Decision tree, C4.5, C5.

1. Introduction

Data mining automates the detection of relevant patterns in a database. It uses well established statistical and machine learning techniques to build models that predict some behavior of the data. It is the extraction of hidden predictive information from large databases [3]. Some of the functionalities of data mining are characterization, discrimination, classification, prediction, mining frequent patterns, association and correlations. Classification is a process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown[2].

The main objective of this paper is to compare C4.5 and C5 algorithms on the large data set 'adult' obtained from the UCI Machine Learning Repository, which is used to predict the individual's income. The classification model begins with the categorization of the classes of data objects. It is in the form of classification rules, decision trees or formulae. The model used in this paper predicts the percentage of U.S citizen who are above 30 years, with master or professional school or doctorate degree and work in the private or self-employed have income more than 50K or less.

In section 2 we provide an introduction to decision tree. Description of C4.5 and C5 algorithms are given in section 3 and 4 respectively, description of the dataset is given in section 5, evaluation performance of the classifier is provided in section 6, result of the paper is elaborated

[12] Professor & Head, Senior Lecturer, PG Scholar, Department of Computer Science and Engineering, Faculty of Engineering, Avinashilingam University for Women, Coimbatore. email : hod cse au@yahoo.co.in

in section 7, and finally the conclusion of the paper is provided in section 8.

2. Decision Tree

Decision tree is a classification scheme which generates a tree and a set of rules representing the model of different classes, from a given dataset. It is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The construction of the decision tree does not require any domain knowledge [2]. It can handle high dimensional data, and provide good accuracy.

One of the decision tree construction algorithms is ID3 (Iterative Dichotomizer 3). It is conceptually simple but powerful classification algorithm [7]. In ID3 each node corresponds to a splitting attribute and each arc is a possible value of that attribute. At each node the splitting attribute is selected to be the most informative among the attributes not yet considered in the path from the root. This algorithm uses the criterion of information gain to determine the goodness of a split. The attributes with greatest information gain is taken as the splitting attributes and the dataset is split for all distinct values of the attributes [1]. The main component of the ID3 algorithm is selecting which feature to test at each node in the tree. The resulting tree is used to classify future samples [3]. The leaf nodes of the tree contain the class name and the non-leaf node is the decision node, which is an attribute test with each branch (to another decision tree) being a possible value of the attribute. ID3 uses a statistical property, called *gain*, which helps to select the attribute as a decision node.

Information gain is defined in terms of entropy which is given by the following equation

$$\text{Entropy or Info } (S) = -p_1\log_2 p_1 - p_2\log_2 p_2 \qquad (1)$$

Where p_1 denotes the proportion of the positive examples and p_2 denotes the proportion of the negative examples in S. We can find the expected information as weighted sum over the subsets, as

$$\text{Info } (S, A) = |S_i| / |S| \text{x Info } (S_i) \text{ where } i = 1..\, n \qquad (2)$$

The term gain is given by

$$\text{Gain } (S, A) = \text{Info}(S) - \text{Info}(S, A) \qquad (3)$$

ID3 algorithm forms the basis for both C4.5 and C5 algorithms.

3. C4.5 Algorithm

C4.5 algorithm constructs the decision tree with a 'divide and conquer' strategy. It is an extension of ID3. It eliminates the problem of unavailable values, continuous attributes value ranges, pruning of decision trees and rule derivation. In C4.5, each node in a tree is associated with a set of cases. Also, these cases are assigned weights to take in to account unknown attribute values. At the beginning, only the root is present and it is associated with the whole training set, and all the weights are equal to one. At each node the divide and conquer algorithm is executed, trying to exploit the locally best choice with no backtracking allowed. In building a decision tree, we deal with training set that have records with unknown attributes by considering only those records where those attribute values are available. We can classify records that have unknown attribute values by estimating the probability of the various possible results. C4.5 produces tree with variable branches per node. When a discrete variable is chosen as the splitting attribute in C4.5 there will be one branch for each value of attributes [4,6].

4. C5 Algorithm

The basic idea of C5 is same as ID3 and uses the concept of gain to produce a classifier in the form of decision tree according to the previously chosen classification. Instead of using gain C5 uses the gain ratio. Gain ratio is the ratio between the gain and the splitting information of the training set. The splitting information is the information due to split of training set on the basis of the attributes. The attribute with highest information gain is chosen as the splitting attribute. It helps us to choose the attributes at different level, deals with the unavailable values, continuous attribute value ranges, pruning of decision trees, rule derivation etc. C5 generates set of production rules from a decision tree. These rules better express the classification model than trees [5].

5. Dataset Description

In this paper for the comparison of C4.5 and C5 algorithms, we have used one of the large dataset 'adult' which is obtained from the UCI machine-learning library [8]. This dataset include ?the census data that are used to predict whether an individual's income is greater than $50k. For preprocessing the adult data set, we have eliminated all the tuples that have the missing values. This reduces the size of the dataset from 48842 to 32,560 tuples. The dataset has 15 attributes. They are age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country and salary. In this paper we use the attributes, age, education, work class and salary for the classification of the dataset.

6. Performan⌐e Evaluation

Classifier performance depends greatly on the characteristics of the data to be classified. Various empirical tests can be performed to compare the classifier performance like holdout, random sub-sampling, k-fold cross validation and bootstrap method. In this study have selected k-fold cross validation for evaluating the classifiers.

6.1 Cross Validation

In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subset or 'folds' d_1, d_2, …, d_k, each of approximately equal size. Training and testing is performed k times. In iteration I, partition d_i is reserved as the test set, and the remaining partitions are collectively used to train the model. In the first iteration, subsets d_2, …, d_k collectively serve as the training set inorder to obtain a first model, which is tested on d_1; the second iteration is trained in subsets d_1,d_3,…d_k and tested on d_2; and so on [2]. The accuracy estimate is the overall number of correct classifications from the k iterations, divided by the total number of tuples in the initial data.

7. Experimental Results and Discussions

In this paper 3-fold cross validation is applied for evaluating the performance of the classifiers. The rule used for building the classification tree is that if *age* is greater than or equal to 30 and *education* is greater than or equal to masters and if *work_class* is equal to private or self employed, then their income is greater than 50k. This rule is given in Table 1.

Table 1 : Rule Set for the Classification of the Adult Data Set

Attribute	Age	Education	Work class	Salary
Value	>=30	>=Masters	Private or self employed	>50k

The classification tree for the rule is given in Figure 1. From the total 32,560 records, 22,850 people were of age greater than or equal to 30. Out of that, 2482 people are with master or professional schooling or doctorate degree. For this set of record we have applied 3 fold cross validation and predicted that 67% of the self employed or private employee earn more than 50k.

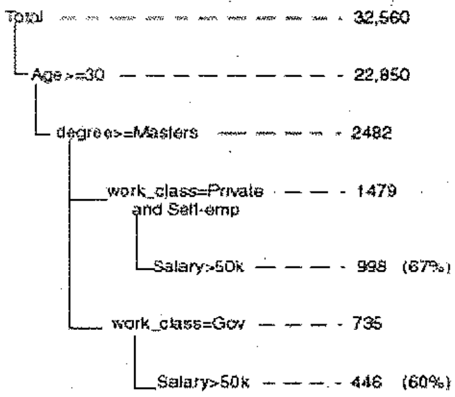

Figure 1 : Classification Tree

The comparison chart of C4.5 and C5 algorithms based on accuracy, error rate and time is given in Figures 2-4 respectively.

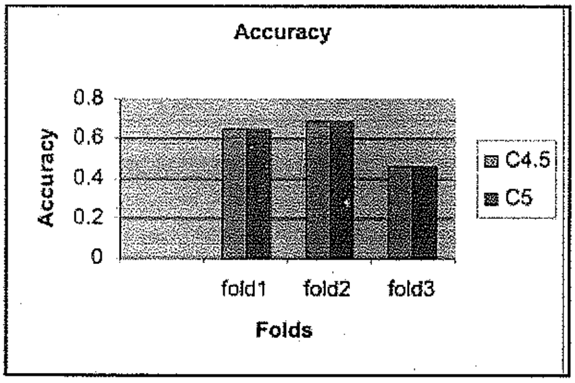

Figure 2 : Comparison Chart of C4.5 and C5 Based on Accuracy

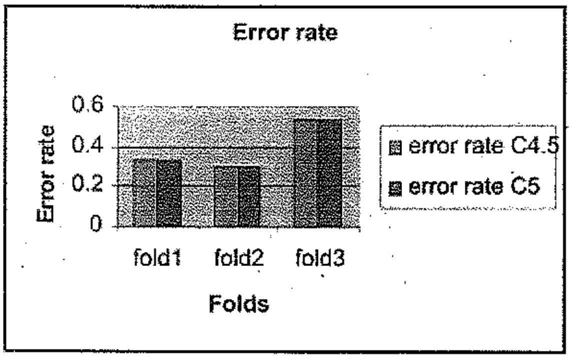

Figure 3 : Comparison Chart of C4.5 and C5 Based On Error Rate

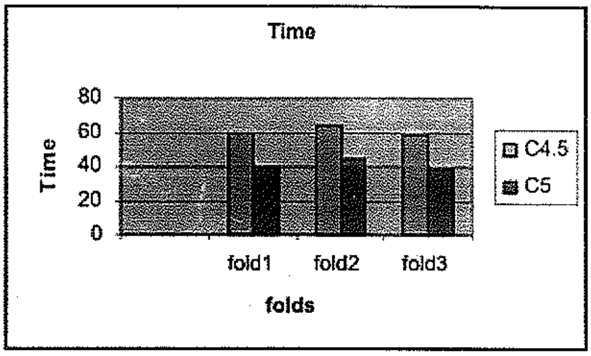

Figure 4 : Comparison Chart of C4.5 and C5 Based On Time

The comparison results of C4.5 and C5 algorithms are tabulated in Table 2.

Table 2 : Comparison Result of C4.5 and C5 Based on Adult Dataset

	Fold1		Fold2		Fold3	
Algorithm	C4.5	C5	C4.5	C5	C4.5	C5
Accuracy	0.65	.65	0.69	.69	0.46	0.46
Error Rate	0.35	0.35	0.31	0.31	0.54	0.54
Time	60 sec	40 sec	64 sec	45 sec	58 sec	38 sec

In this comparison, we can observe that accuracy rate of both C4.5 and C5 algorithms are equal in every fold of cross validation. In the second fold of cross validation, classification accuracy of both algorithms are maximum. When time factor is considered, C5 out performs C4.5 in all the three folds and provides minimum execution time in fold 3.

8. Conclusion and Future Development

In this paper, the performance of two decision tree classification algorithms C4.5 and C5 are compared. The experiments were conducted on the large benchmark dataset 'adult'. Classification accuracy is validated by 3-fold cross validation method. Our study reveals that C5 outperforms C4.5 when time factor is considered. Maximum accuracy of 69% is obtained in fold 2 of cross validation for both the classifiers. Possible extension of this work will be on applying bagging / boosting techniques on the decision tree classifiers to improve their accuracy and also the dataset can be cross validated to more number of folds.

Acknowledgement

The authors would like to thank Dr. S.C. Sharma, Principal, R. V. College of Engineering, Bangalore, for his valuable guidance to carry out this work.

References

[1] A.K. Pujari, *"Data mining Techniques"*, University Press, India 2001.

[2] J. Han and M. Kamber, *"Data Mining Concepts and Techniques"*, Morgan Kauffman Publishers, USA, 2006.

[3] Kietikul Jearanaitanakij, *"Classifying Continuous Dataset by ID3 Algorithm"*, Proceedings of fifth International Conference on Information Communication and Signal processing, PP. 1048-1051, 2005.

[4] J. R. Quinlan, *"C4.5: Programs for Machine Learning"*, Morgan Kaufman Publishers, San Mateo, Calif, 1993.

[5] J. R. Quinlan, *"Data Mining Tools See5 and C5.0"*, 2000. [http://www.rulequest.com/see5-info.html]

[6] Salvatore Ruggieri, *"Efficient C4.5 Proceedings of IEEE transactions on knowledge and data Engineering"*, Vol. 14, 2, No.2, PP. 438-444, 2002.

[7] S.N. Sivanandam and S. Sumathi, *"Data Mining Concepts, Tasks and Techniques"*, Thomson, Business Information India Pvt. Ltd, India, 2006.

[8] UCI Machine learning repository [http://archive.ics.uci.edu/beta/datasets/Adult]

Author's Biography

S. Sivakumari, graduated from Madurai Kamaraj University, in Electronics and Communication Engineering during the year 1988. She obtained her Master degree in Applied Electronics from PSG College of Technology, Bharathiyar University, Coimbatore in the year 1995 and currently pursuing PhD degree in the area of data mining. At present she is a professor and Head of the Department of Computer Science and Engineering, Faculty of Engineering, Avinashilingam University for Women, Coimbatore, India. She has published many research papers in the International/National conferences. Her research areas include Data mining and Softcomputing. She has more than 19 years of experience in teaching and research. She is a member of ISTE and Computer Society of India.

R. Praveena Priyadarshini has graduated her B.E in Computer Science and Engineering in the year 1994 from Madras University, and received M.E degree in the year 2007 from Vinayaka University. She also belongs to the same department. She has seven years of teaching experience in the field of Computer science. Her research interests include Data mining and Software Engineering. She has published several papers in the National/International conferences. She is a member of ISTE and Computer Society of India.

Ms. S. Chandralekha has completed her MCA degree in Madras University in the year 2004 and presently doing her M.E degree with the specialization of Computer Science and Engineering in Avinashilingam University for Women, Coimbatore