Performance Analysis of Decision Trees

Manpreet Singh
Department of Information Technology, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India

Sonam Sharma
CBS Group of Institutions, New Delhi,India

Avinash Kaur
Department of Computer Science, Lovely Professional University, Phagwara, Punjab, India

ABSTRACT

In data mining, decision trees are considered to be most popular approach for classifying different attributes. The issue of growing of decision tree from available data is considered in various discipline like pattern recognition, machine learning and data mining. This paper presents an updated survey of growing of decision tree in two phases following a top down approach. This paper presents a framework for dealing with the condition of uncertainty during decision tree induction process and concludes with comparison of Averaging based approach and Uncertain Decision Tree approach.

1. INTRODUCTION

With the passage of time, large amount of data is collected in almost every field. The useful data need to be extracted or mined for fulfilling the future needs. Although there are various techniques through which the data can be extracted but the classification technique used for classifying the data for mining is the most useful technique. The classification technique is implemented using different algorithms like decision tree, neural network, genetic algorithm and K-nearest neighbor.

The decision trees are constructed using the data whose values are known and precise. These are implemented using C4.5 algorithm. But during the construction of decision tree major problem of uncertainty arises in the datasets that is responsible for the performance degradation and inaccurate results. The averaging approach and distributed approach are used to handle the problem of uncertainty. These approaches are compared on the same datasets so as to predict which approach reduces the error rate and depict more accurate results.

2 DATA MINING PROCESS

2.1 Data Mining

Data mining is the process which takes data as input and provides knowledge as output. It is non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [1].It is also called as knowledge mining. It is a part of the knowledge discovery process. Advanced Computation methods are applied in data mining for the extraction of information from data [2].

2.2 Data Mining Techniques

Different data mining techniques are used to extract the appropriate data from database. This goal of extraction is divided into two categories: prediction and description. In prediction process, the existing database and the variables in the database are interpreted in order to predict unknown or future values. On the other hand, description focuses on finding the patterns describing the data and subsequent presentation for user interpretation [3]. Every data mining techniqueis based either on the concept of prediction or description. Different Data Mining Techniques are association, classification, regression and clustering [2].

Association rule mining is one of the most and well researched techniques of data mining [4]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [5]. Classification is a process which builds a model based on previous datasets so as to predict future trends with new data sets. The association and classification data mining techniques deals with the categorical attribute. The technique used to predict the numeric or continuous values is known as numeric prediction or regression. It was developed by Sir Frances Galton [2]. It is based on the concept of prediction and is used to predict a model based on the observed data over a period of time [6]. Sometimes, the large amount of data is available but few variables in the datasets do not have class labels. The process of assigning class labels to the variables is known as clustering. In this process, the variables in the data sets are grouped together according to the similarity between each other and dissimilarity from other variables [5].

2.2.1 Classification

Classification is one of the fundamental tasks in data mining and has also been studied extensively in statistics, machine learning, neural networks and expert system over decades [7]. It is a process which builds a model based on previous datasets so as to predict future trends with new data sets. It is a two-step process in which the first step is the learning step which involves building of a classifier or a model based on analyzation of training dataset so as to predict categorical labels [2]. As the class label is provided in the data set, so it is also known as supervised learning [8]. In the second step, the accuracy of the model is predicted using test data set. If the accuracy is acceptable, then the model is used to classify new

data tuples. The different classification techniques are decision trees, neural networks, K-nearest neighbor and genetic algorithms [2]. Among these techniques, the decision trees induction algorithms present several advantages over other learning algorithms, such as robustness to noise, low computational cost for generating the model and ability to deal with redundant attributes. These are fast and accurate among all the classification methods [9].

3 DECISION TREES

Decision tree is a predictive modelling based technique developed by Rose Quinlan .It is a sequential classifier in the form of recursive tree structure [10]. There are three kinds of nodes in the decision tree. The node from which the tree is directed and has no incoming edge is called the root node.A node with outgoing edge is called internal or test node. All the other nodes are called leaves (also known as terminal or decision node) [11]. The data set in decision tree is analyzed by developing a branch like structure with appropriate decision tree algorithm. Each internal node of tree splits into branches based on the splitting criteria. Each test node denotes a class. Each terminal node represents the decision. They can work on both continuous and categorical attributes [2].

3.1 Decision Tree Induction

Decision Tree induction is an important step of segmentation methodology. It acts as a tool for analyzing the large datasets. The response of analyzation is predicted in the form of tree structure [12]. The Decision tree are classified using the two phase's tree building phase and tree pruning phase [2]

a) Tree Building Phase: In this phase the training data is repeatedly partitioned until all the data in each partition belong to one class or the partition is sufficiently small. The form of the split depends on the type of attribute. The splitting for numeric attributes are of the form $A \leq s$, where s is a real number and the split for categorical attributes are of the form $A \in c'$, where c' is a subset of all possible values of A [13]. The alternative splits for attribute are compared using split index. The different tests for choosing the best split are Gini Index (Population Diversity), Entropy (Information Gain), Information Gain Ratio [11].

i) The Gini Index measures the impurity of a data partition or a set of training tuples as

$$Gini(t) = 1 - \sum_j [p(j \mid t)]^2$$

wherep (j | t) is the relative frequency of class j at node t.

It is used in CART, SLIQ and SPRINT decision tree algorithms. The node with the least computed Gini index is chosen [14].

ii) Entropy is an information-theoretic measure of the 'uncertainty' contained in a training set, due to the presence of more than one possible classification. If there are K classes, we can denote the proportion of instances with classification i by pi for i = 1 to K. The value of pi is the number of occurrences of class i divided by the total number of instances, which is a number between 0 and 1 inclusive The Entropy is measured in bits. It is given by

$$Entropy(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

Where p (j | t) is the relative frequency of class j at node t[15]

iii) Information Gain is the change in information entropy from prior state to a state that takes some information [19]. It measures reduction in entropy achieved because of split. It is given by

$$Gain_{spit} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

where Parent Node, p is split into k partitions and n_i is number of records in partition i[15].

The node with the highest information gain is chosen as best split. It is used in ID3 and C4.5 algorithms .The disadvantage of information gain is that it is biased towards the splits that result in large number of partitions, each being small but pure [16].

iv) Gain Ratio is the variant introduced by the Australian academician Ross Quinlan in his influential system C4.5 in order to reduce the effect of the bias resulting from the use of information gain. Gain Ratio adjusts the information gain for each attribute to allow for the breadth and uniformity of the attribute values [17]. This split method doesn't select a test with an equal or below average information gain and a low split info, because a low split info doesn't provide any intuitive insight into class prediction [18].The gain ratio is given by

$$GainRATIO_{split} = \frac{GAIN_{split}}{SplitINFO}$$

$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

where Parent Node, p is split into k partitions n_i is the number of records in partition i.[15]

This technique is used in C4.5 algorithm [19].

b) Tree Pruning Phase: This phase avoids model overfitting [9]. The issue of overfitting arises due to random errors, noise in data or the coincidental patterns, which can lead to strong performance degradation. There are two approaches that fulfill the task of tree pruning. In the Pre-Pruning Approach (Forward Pruning) top-down construction of decision tree is stopped at a point when the sufficient data is no longer available to make the accurate decisions. In the Post-Pruning Approach (Backward Pruning) the tree is first fully grown and then the sub trees are removed which lead to unreliable results. Further these approaches are classified into Reduced Error Pruning and Rule Post Pruning [7].

i) Reduced error pruning: In this process all the internal nodes in the tree are considered. For each node in the tree it is checked that if removing it along the subtree below it does not affect the accuracy of the validation set [11]. This process is repeated recursively for a node until no more improvements on a node are possible. The reduced error pruning is used by ID3 decision tree algorithm. The biggest disadvantage of

reduced error pruning is that it cannot be used on the small training data set [14].

ii) Rule Post Pruning: In this kind of pruning the tree structure is converted into rule based system and the redundant conditions are removed by pruning each rule .At the end the rules are sorted by accuracy. The Rule post pruning is used by C4.5 decision tree algorithm [15]. The advantage of Rule post pruning is that pruning becomes flexible and improves interpretability. This yields higher accuracy [14].

3.2 Decision Tree Algorithms

The different decision tree algorithms are ID3, C4.5 and CART.

a) ID3 (Iterative Dichotomiser)

ID3 is a greedy learning decision tree algorithm introduced in 1986 by Quinlan Ross [17]. It is based on Hunts algorithm [20] .This algorithm recursively selects the best attribute as the current node using top down induction. Then the child nodes are generated for the selected attribute. It uses an information gain as entropy based measure to select the best splitting attribute and the attribute with the highest information gain is selected as best splitting attribute. The accuracy level is not maintained by this algorithm when there is too much noise in the training data sets. The main disadvantage of this algorithm is that it accepts only categorical attributes and only one attribute is tested at a time for making decision. The concept of pruning is not present in ID3 algorithm. [21]

b) C4.5 algorithm

C4.5 is an extension of ID3 algorithm developed by Quinlan Ross [20].This algorithm overcomes the disadvantage of overfitting of ID3 algorithm by process of pruning. It handles both the categorical and the continuous valued attributes. Entropy or information gain is used to evaluate the best split. The attribute with the lowest entropy and highest information gain is chosen as the best split attribute. The pessimistic pruning is used in it to remove unnecessary branches indecision tree [7].

d) C5

C5 algorithm is the successor of C4.5 proposed by Quinlan [22].It uses the concept of maximum gain to find the best attribute. It can produce classifiers which can be represented in the form of rule sets or decision trees.

Some important features of C5 algorithm are:

- Accuracy: C5 rule sets are small in size thus they are not highly prone to error
- Speed : C5 has a great speed efficiency as compared to C4.5
- Memory: Memory utilization of C4.5 was very high as compared to C5 thus C4.5 was called memory hungry algorithm while it is not true in case of C5.
- Decision Trees: C5 produces simple and small decision trees
- Boosting: C5 adopts a boosting technique to calculate accuracy of data. It is a technique for creating and combining multiple classifiers to generate improved accuracy.
- Misclassification costs: C5 calculates the error costs separately for individual classes then it constructs classifiers to minimize misclassification costs.
- It supports sampling and cross validation [23].

3.3 Uncertainty in Decision Tree

Although the different algorithms have been devised for formulation of accurate results but a major problem of uncertainty arises in data sets in decision tree algorithms due to various factors like the missing tuples value, noise. There are two approaches to handle uncertain data.

First approach is the averaging approach (AVG) which is based on greedy algorithm and induces top down tree. In this expected value replaces each probability density function (probability that a random variable x falls in a particular range) in uncertain information and thus converts the data tuples into point valued tuples. This function is also denoted by pdf. Thus the decision tree algorithms ID3 and C4.5 can be reused.

Another approach is the distributed approach which considers the complete information carried by probability distribution function to build a decision tree. The challenge in this approach is that a training tuples can now ‘pass’ a test at tree node probabilistically when its pdf properly contains split point of test. This induces a decision tree which is known as Uncertain Decision Tree (UDT) [14].

To test the accuracy of both the approaches and decide which approach is better than the other the experiments are conducted on the 5 real data sets as shown in Table 1

Table 1 Selected Data sets from UCI Machine Learning Repository [14]

Data Set	Training Tuples	No.of Attributes	No. of classes	Test Tuples
Satellite	4435	36	6	2000
Vehicle	846	18	4	10-fold
Ionosphere	351	32	2	10-fold
Glass	214	9	6	10-fold
Iris	150	4	3	10-fold

For the different datasets Gaussian distribution and the Uniform Distribution is considered as an error model to generate the probability density function (pdf). Gaussian distribution is chosen for the measure that involves noise and uniform distribution is chosen when the digitization of measured values introduces noise. The results of applying AVG and UDT to different datasets are shown in table 2.

Table 2 Accuracy Improvement by Considering Distribution

Data Set	AVG	UDT		
		Best Case	Gaussian Distribution	Uniform Distribution
Satellite	84.48	87.73	87.73	87.2
Vehicle	71.03	75.09	75.09	71.62
Ionosphere	88.69	91.69	91.69	N/A
Glass	66.49	72.75	72.75	N/A
Iris	94.73	96.13	96.13	N/A

In the experiments each PDF is represented by 100 sample points (i.e. s=100).

For most of the data sets Gaussian distribution is assumed as an error model but for the datasets "Satellite" and "Vehicle" uniform distribution error model has been used. The second and third columns in the table depict the percentage accuracy achieved on a particular data set by the AVG and UDT approach consequently. Comparing the second and third column of table 2, it is seen that UDT approach builds a more accurate decision tree then the AVG approach. For instance for the data set "Satellite" accuracy is improved from 84.48 to 87.73 that is reducing the error rate from 15.52% to 12.27%. For the dataset "Ionosphere" accuracy is improved from 71.03 to 75.09 thus reducing the error rate from 28.97% down to 24.91%.Similarly in other datasets there is reduction in error rate.

The Figure 1 depicts that UDT is more accurate than AVG for same data sets.

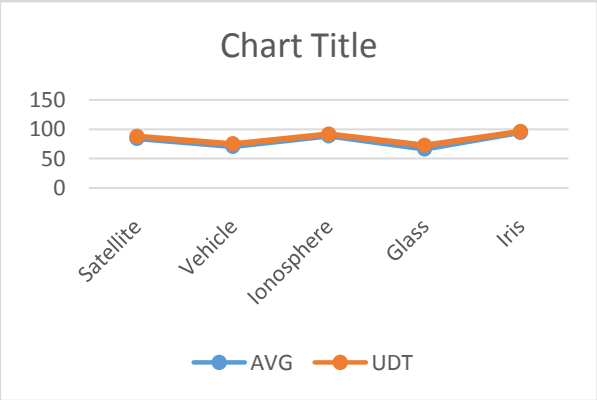

Figure 1: Comparison of UDT and AVG

4 CONCLUSION

The concept of data mining and different techniques used for mining of the data are discussed. The data mining technique classification solves a purpose of classifying the data sets and reaching a decision with different algorithms of decision tree. The performance is the issue in decision trees which arises due to data uncertainty. The uncertainty arises due to missing data tuples and noise in data. To handle the problem of uncertainty, two kind of approaches are followed: averaging based approach and distributed approach. The decision trees are induced based on both the approaches. According to the experiments conducted on different data sets, the UDT builds a more accurate decision tree then averaging approach. The experimental results are depicted graphically.

The new methodologies can be introduced to improve the accuracy of decision trees which may lead to highly accurate decisions.

.

5. REFERENCES

[1] F[1] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (2006, March). "From Data Mining to Knowledge Discovery in Databases". The Knowledge Engineering Review,Vol 21, No.1, pp. 1-24.

[2] Han, J., & Kamber, M. (2006). "Data Mining: Concepts and Techniques" (2nd ed.). Morgan Kaufmann Publishers.

[3] Pujari, A. K. (2001). "Data Mining Techniques". Universites Press India Private Limited.

[4] Agarwal, R., Imieliński, T., & Swami, A. (1993, June). "Mining association rules between sets of items in large databases". ACM SIGMOD Record, Vol 22,No.2, pp. 207-216.

[5] Zhao, Q., & Bhowmick, S. S. (2003). "Association Rule Mining: A Survey". Retrieved fromhttp://sci2s.ugr.es/keel/pdf/specific/report/zhao03ars.pdf

[6] Oracle® Data Mining Concepts 11g. (2008, May). Retrieved from oracle.com: http://docs.oracle.com/cd/B28359_01/datamine.111/b28129.pdf

[7] Lavanya, D., & Rani, U. k. (2011, July). "Performance Evaluation of Decision Tree Classifiers on Medical Datasets". International Journal of Computer Applications, Vol 26,No. 4,pp. 1-4.

[8] Kotsiantis, S. B. (2007). "Supervised Machine Learning : A review of classification techniques",Vol 160,No. 3,Frontiers in Artificial Intelligence and Applications

[9] Barros, R. C., Basgalupp, M. P., Carvalho, A. C., &Freitas, A. A. (2010, Jan). "A Survey of Evolutionary Algorithms for DecisionTree Induction". IEEE Transactions on Systems,Mans and Cybernetics, Vol. 10,No. 10,pp. 1-22.

[10] Quinlan, J. R. (1987). "Generating production rules from decision trees". Proceedings of the 10th international joint conference on Artificial intelligence , pp. 304-307.

[11] Maimon, O., & Rokach, L. (2010). "Data Mining and Knowledge Discovery Handbook". (2nd, Ed.) Springer.

[12] De ville, B. (2006). "Decision trees for Business Intelligent and Data Mining using SAS Enterprise Miner". NC, USA: SAS Institute Inc.

[13] Apers, P., Bouzeghoub, M., & Gardarin, G. (1996). Advances in Database Technology. 5th International Conference on Extending Database Technology Avignon,Vol. 1057.

[14] Tsang, S., Kao, B., Yip, K. Y., Ho, W.-S., & Lee, S. D. (2011, Jan). "Decision Trees for Uncertain Data". IEEE Transactions on Knowledge and Data Engineering, Vol 23,No.1.

[15] Bramer, M. (2007). *Principles of data Mining.* London: Springer.

[16] Quinlan, J. R. (1993). "C4.5 programs for machine learning". *Morgan, .*

[17] Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning, Vol 1,No.*1,pp. 81-106

[18] Harris, E. (2002). "Information Gain Versus Gain Ratio: A Study of Split Method Biases". *AMAI.*

[19] Das, S., & Saha, B. (2009). "Data Quality Mining using Genetic Algorithm". *International Journal of Computer Science and Security, Vol 3,No.*2,pp. 105-112.

[20] Anyanwu , M. N., & Shiva, S. G. (2009). "Comparative Analysis of Serial Decision Tree Classification". *International Journal of Computer Science and Security, Vol 3*,No. 3,pp. 230-239.

[21] Venkatadri, & Lokanatha. (2011). "A Comparative Study of Decision Tree Classification Algorithms in Data Mining.". *International Journal of Computer Applications in Engineering, Technology and Sciences, Vol 3,No.*3.,pp. 230-240

[22] Ruggieri, S. (2002, April). "Efficient C4.5 [classification algorithm]". *IEEE Tansactions on Knowledge and Data Engineering*,pp. 438-444.

[23] *C5 Algorithm*. (2012). Retrieved fromhttp://rulequest.com/see5-comparison.html