A Reduced Error Pruning Technique for Improving Accuracy of Decision Tree Learning

Rinkal Patel, Rajanikanth Aluvalu

Abstract— Decision tree inductions are well thought-out as it is one of the most accepted approaches for representing classifiers. Many researchers from varied disciplines like Statistics, Pattern Reorganization; Machine Learning measured the problem of growing a decision tree from available data. Databases are the rich sources of hidden information that can be used for intelligent decision making. Classification and Prediction techniques of data mining are two form of data analysis that can be used to discovering this type of hidden knowledge. Classification techniques deal with categorical attributes whereas prediction model is the continuous value function. Training data are analyzed by classification algorithm. In decision tree construction attribute selection measure are used to select attributes, that best partition tuples into different classes. The branches of decision tree may reflect noise or outliers in training data. So tree pruning techniques applies to identify and remove those branches which reflect noise with the aim of improving classification accuracy. But still scalability is the issue of decision tree from large database. This paper present implementation of decision tree induction algorithm in java with reduced error pruning(REP) technique for improving accuracy of classifier.

Index Terms— Data Mining, Classification Decision Tree Induction, Information Gain, C4.5, Tree Pruning.

I. INTRODUCTION

The progress of information technology has generated huge amount of databases and huge data in various areas. The research in databases and information technology has specified increase to an approach to store and manipulate this valuable data for further decision making. Data mining [1] is a process of extraction of useful information and patterns from large data. It is also called as knowledge discovery process (KDD) [17, 19], information mining from data, knowledge discovery or data /pattern examination.

A Decision tree [1] is a classifier represented as a recursive separation of the instance space. The decision tree consists of nodes that generate a *rooted tree*, mean that it is a *directed tree* with a node called *root* which has no incoming edges. All the other nodes have exactly one incoming edge. A node with outgoing edges is called an *internal* or *test node* while all other nodes are called *leaves.* Decision tree induction is closely related to rule induction. Each path of a decision tree from root node to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to generate the antecedent part, and taking the leaf's class prediction as the class label value. For example, one of the paths can be transformed into the rule:

Manuscript received on June, 2014.

Rinkal Patel, Computer Engineering, R.K University, Rajkot, India.
Rajanikanth Aluvalu, Computer Engineering, R.K University, Rajkot, India.

"If customer age >=30, and the gender of the customer is "male," then the customer will respond to the mail." The induction of decision trees has been getting a lot of concentration in the field of knowledge discovery in Databases over the past few years. This popularity has been largely due to the efficiency with which decision trees can be induced from large datasets.

Decision tree induction algorithms [6] that automatically construct a decision tree from a given dataset. Classically the objective is to find the optimal decision tree by minimizing the generalization error. However, other target functions can be also defined, for example, minimizing the number of nodes or minimizing the average depth.The database management systems proficiently manage large amount of data and effective and efficient retrieval of meticulous information from a huge collection whenever needed and also contributes to recent massive gathering of all sorts of information. This retrieval of data as and when needed contributes the technology of data mining. Data mining can be viewed as a result of the natural evolution of information technology. This technology provides a wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.

Missing data [16] is a common problem in quantitative social research. In many predictive modeling applications, useful attribute values may be missing or not present. For example, sometimes it is happen that patient data have missing diagnostic tests that may be very helpful for estimating the likelihood of diagnoses or for predicting effectiveness of the treatment; consumer data often do not include values for all attributes useful for predicting buying preferences. The ID3 decision tree algorithm implemented [22] in java but it has limitation of deal with discrete value and can't generate accurate tree. In this paper C4.5 algorithm uses the REP (reduced error pruning) technique for improving accuracy of classifier. The algorithm of these techniques available from [2].

II. DATA MINING

Data mining [12,20] is the process which finds valuable patterns from large amount of data, the process of extracting previously unknown, understandable and precious information from large databases. There are a number of data mining techniques[17] have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns etc., are used for knowledge extraction from databases.

(1) Association

Association is one of the best well-known data mining techniques. In association, a pattern is revealed based on an

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

association of a particular item on other items in the same transaction.

(2) Classification

Classification is a typical data mining technique based on machine learning. Essentially classification is used to classify each item in a set of data into one of predefined set of classes or groups. For Example, Teachers classify students' grades as A, B, C, D, or F, here A,B,C,D and F are class labels [19].

(3) Clustering

Clustering is "the process of arranging objects into groups whose members are similar in some way in term of their characteristics". A *cluster* is therefore a collection of objects that are "similar" to each other and are "dissimilar" to the objects that belonging to other clusters [19].

(4) Prediction

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and also relationship between dependent and independent variables [12].

III. CLASSIFICATION

Classification technique [5] is the grouping of data in given classes. Also it is known as supervised classification, the classification uses given class labeled data in the form of training set to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with previously known class labels. The classification algorithm constructs a model which learns from the training set. The model is used to classify new objects.

In classification [14], we make the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are most likely to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay". And then we can ask our data mining software to classify all the employees into each group.

A. Classification Techniques

(1) Naivy Bayes classification: It calculates explicit probabilities for supposition, among the most practical approaches to definite types of learning problems. Even when Bayesian techniques are computationally difficult, they can provide a standard and optimal decision making against all other methods of classification [15].

(2) Decision Trees induction: A decision tree is flow-chart like tree structure, where internal node represent test on attribute and branch shows result of test. It is very simple and easily understandable classifier.

(3) Neural Networks: prediction accuracy of neural network is generally high, robust, and also works when training examples contain errors. Output may be discrete, real-valued, or a vector of several discrete or real-valued attributes, but it takes long training time and its learning function is difficult to understand.

(4) Case Base reasoning: Instances represented by affluent symbolic descriptions (e.g. function graphs), multiple retrieved cases may be combined using tight coupling between case retrieval, knowledge-based reasoning, and problem solving.

(5) Genetic Algorithm: it is based on a correlation to biological evolution; each rule is represented by a string of bits. Here an initial population is produced consisting of randomly generated rules. For e.g. IF B1 and Not B2 then C2 can be encoded as 100 [15].

Based on the concept of survival of the fittest, a new population is created and consists of only fittest rules and their offspring, the fitness of a rule is represented by its classification accuracy on a set of training examples [8].

IV. DECISION TREE INDUCTION

Decision tree [11] is a predictive model that generates a tree from the given training samples. The tree construction process is heuristically guided by selecting the most significant attribute at each step, whose aim is to minimizing the number of tests that is needed for classification.

Decision tree induction [3] is define as learning a decision tree from class labeled training tupples.it is a flow chart like structure where each internal node of a tree denote a test on attribute, branch represent outcome of the test, and each leaf node contain a class label, the top most node in the tree is called root node. Decision tree learning is a method for resembling discrete-valued functions that is robust to noisy data. Decision tree has many advantages, such as its fast speed, high accuracy as well as the easy mode of production, which attracts many researchers in data mining.

A. Information gain

Information gain [11, 22] is an impurity based criteria that uses the Entropy measure as the impurity measure, Select the attribute with the highest information gain .Assume there are two classes, P1 and N1 [8].

Let us take set of examples S contain p elements of class P1 and n elements of class N1,The amount of information, needed to come to a decision if an random example in S belongs to P1 or N1 is defined as,

$$I(p, n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

If S_i contains p_i examples of *P1* and n_i examples of *N1*, then the entropy, or the expected information needed to classify objects in all sub trees S_i is [3],

$$E(A) = \sum_i^v \frac{pi+ni}{p+n}\, I(pi, ni)$$

A. Algorithm of C4.5

C4.5 algorithm is an enhancement of ID3 algorithm [6, 13], developed by Quinlan Ros. It is based on Hunt's algorithm and also similar to ID3, it is serially implemented. Pruning takes place in C4.5 by replacing the internal node with a leaf node in that way reducing the error rate. Different then ID3, C4.5 [5, 21] accepts both continuous and categorical attributes in generating the decision tree. It has an improved method of tree pruning that reduces misclassification errors due noise or too-much details in the training Data set [1].

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

C4.5 (R1: a set of non-categorical attributes: the categorical attributes: a training set)

returns a decision tree;

begin

If S is empty, return a single node with value Failure;

If S consists of records all with the same value for the categorical attribute,

return a single node with that value;

If R1 is empty, then return a single node with as value

the most frequent of the values of the categorical attribute that are found in records of S; [note that then there will be errors, that is, records that will be improperly classified];

Let D1 be the attribute with largest Gain (D1, S), among attributes in R1;

Let {dj| j=1,2, .., m} be the values of attribute D1;

Let {Sj| j=1,2, .., m} be the subsets of S consisting respectively of records with value dj for attribute D1;

return a tree with root labeled D1 and arcs labeled d1, d2... dm going respectively to the trees

C4.5 (R1-{D1}, C, S1), C4.5 (R1-{D1}, C, S2)... C4.5 (R1-{D1}, C, Sm);

Figure 1. C4.5 Algorithm

V. TREE PRUNING

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of over fitting the data. Such methods typically use statistical measures to remove the least reliable branches there are two common approaches to tree pruning: pre-pruning and post-pruning. Key motivation of pruning [2] is “trading accuracy for simplicity”. There are various techniques for pruning decision trees. Most of them perform top down or bottom up traversal of the nodes. A node is pruned if this operation improves a certain criteria.

A. Cost-Complexity Pruning

Cost complexity pruning (also famous as weakest link pruning or error complexity pruning) takings in two stages. In the first stage, sequences of trees are built on the training datasets, where the original tree before pruning is the root tree. In the second stage, one of these trees is chosen as the pruned tree, based on its generality of error estimation.

B. Pessimistic Pruning

Pessimistic pruning avoids the need of pruning set or cross validation and it uses the pessimistic statistical association test in its place. The basic idea is that the error ratio estimated using the training set is not consistent sufficiently. Instead a more practical measure known as “continuity correction” for binomial allocation should be used.

C. Reduced-Error Pruning

As traversing over the internal nodes from the bottom to the top of a tree, the REP [5] procedure Checks for each internal node, whether replacing it with the most repeated class that does not reduce the accuracy of trees. In this case, the node is pruned. The procedure continues until any further pruning would decrease the accuracy. In order to estimate the accuracy Quinlan provides to use a pruning set. It can be shown that this procedure ends with the smallest accurate sub- tree with respect to a given pruning set [2, 3].

VI. EXPERIMENTAL RESULT

This experiment performs on implementation of C4.5 algorithm in java with Net-beans. It takes inputs from the Datasets given in the example of decision tree in chapter-6 [9].Here we have 4 attributes names as outlook, Humidity, Temperature, and Wind. The attribute Outlook has highest information gain, so it is selected as root node of tree. Then next humidity attribute is selected for partitioning root node "overcast”. If outlook is "sunny" then we check possibility of Humidity. Else if outlook is "rainy" then check for attributes "wind”. Then accordingly tree is generated.

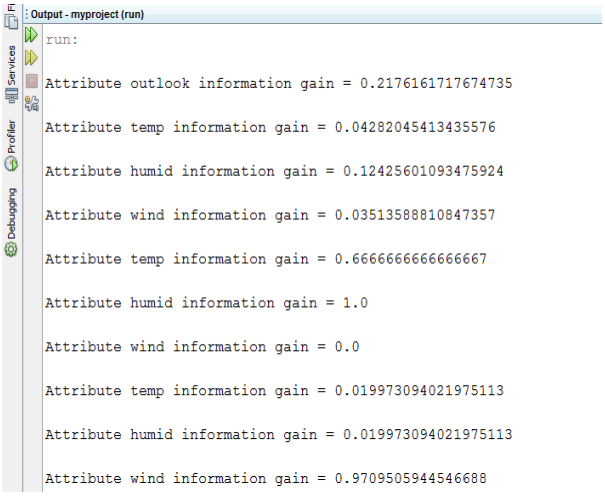

Figure 2. Information gain of Attributes

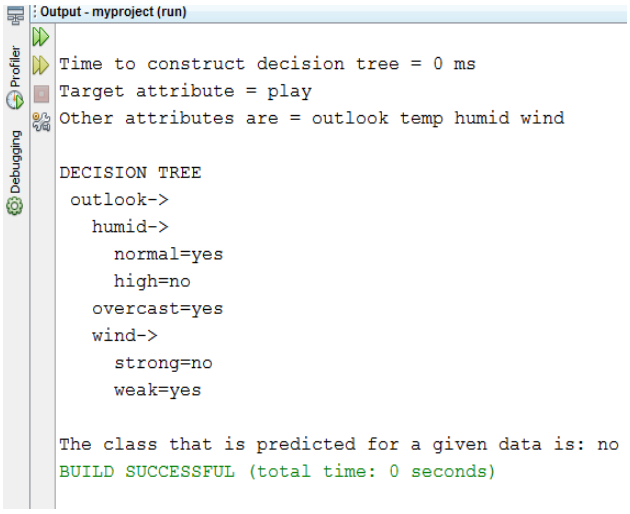

Figure 3. Decision Tree

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

VII. CONCLUSION

Decision tree induction is the learning of decision trees from class-labeled training Tuples.. The individual tuples making up the training set are referred to as training tuples and they are selected from the database under analysis. The main objective of research is associated to improve accuracy and generate small decision tree. To achieve this proposed system is developed on the bases of C4.5 algorithm. It uses reduced error pruning technique for pruning tree so its complexity is reduced and optimal decision tree is generated. In this situation it is possible to prove that reduced error pruning fulfills its proposed task and produces an optimal pruning of the given tree. The REP algorithm evaluates the cost at each decision tree node to determine whether to convert the node into a leaf, prune the left or the right child, or leave the node intact The algorithm proceeds to prune the nodes of a branch as long as both sub-trees of an internal node are pruned and stops immediately if even one sub-tree is kept.

REFERENCES

[1] Matthew N. Anyanwu & Sajjan G. Shiva," Comparative Analysis of Serial Decision Tree Classification Algorithms", Department of Computer Science The University of Memphis,Memphis, TN 38152, U.S.A.

[2] Tapio Elomaa ,Matti Kääriäinen," An Analysis of Redued Error Pruning ", Department Journal of Artificial Intelligence Research 15 (2001) 163-187.

[3] Lior Rokach and Oded Maimon, "Top-Down Induction of Decision Trees Classifiers—A Survey", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 35, NO. 4, NOVEMBER 2005.

[4] Arno J. Knobbe, Arno Siebes, Daniël van derWallen," Multi-Relational Decision Tree Induction", 3821 AE Amersfoort the Netherlands.

[5] A. S. Galathiya, A. P. Ganatra and C. K. Bhensdadia," : Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning", A. S. Galathiya et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012,3427-3431.

[6] Chen Jin, Luo De-lin, Mu Fen-xiang," An Improved ID3 Decision Tree Algorithm", Proceedings of 2009 4th International Conference on Computer Science & Education.

[7] Vaibhav Tripathy," A Comparative Study of Multi-Relational Decision Tree Learning Algorithm", International Journal of Scientific and Technology Research Volume 2, Issue 8, August 2013.

[8] Ravindra Changala,Annapurna Gummadi, G Yedukondalu,UNPG Raju, " Classification by Decision Tree Induction Algorithm to Learn Decision Trees from the class-Labeled Training Tuples" , International Journal of Advanced Research in Computer Science and Software Engineering April 2012.

[9] Jaiwei Han, Micheline kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006, pp 360-361.

[10] Maytal Saar-Tsechansky, Foster Provost," Handling Missing Values when Applying Classification Models", Journal of Machine Learning Research 8 (2007) 1217-1250.

[11] Lior Rokach and Oded Maimon,"Decision Tree", Department of Industrial Engineering Tel-Aviv University.

[12] Patel Nimisha R., Sheetal Mehta," A Survey on Mining Algorithms", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[13] Pea-Lei Tu Jen- Yao Chung," A New Decision-Tree Classification Algorithm for Machine Learning", Proc. of the 1992 IEEE Int. Conf. on Tools with AI Arlington, VA, Nov. 1992.

[14] Rodrigo Coelho Barros, M´arcio Porto Basgalupp, Andr´e C. P. L. F. de Carvalho, and Alex A. Freitas," A Survey of Evolutionary Algorithms for Decision-Tree Induction", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS.

[15] Raj Kumar,Dr. Rajesh verma," Classification Algorithms for Data Mining: A Survey", International Journal of Innovations in Engineering and Technology (IJIET).

[16] Liu Peng, Lei Lei," A Review of Missing Data Treatment Methods", Department of Information Systems, Shanghai University of Finance and Economics, Shanghai, 200433, P.R. China.

[17] B.N. Lakshmi,G.H. ,G.H. Raghunandhan," A Conceptual Overview of Data Mining", Proceedings of the National Conference on Innovations in Emerging Technology-2011 Kongu Engineering College, Perundurai, Erode, Tamilnadu, India.17 & 18 February, 2011.pp.27-32.

[18] Chowdhury Farhan Ahmed ,Syed Khairuzzaman Tanbeer ,Byeong-Soo Jeong andYoung-Koo Lee," HUC-Prune: an efficient candidate pruning technique to mine high utility patterns", Springer Science+Business Media, LLC 2009.

[19] Yongjian Fu,"Data Mining: Tasks, Techniques and Applications", Depertment of Computer Science University of Missouri –Rolla.

[20] Kalyani M Raval," Data Mining Techniques" , International Journal of Advanced Research in Computer Science and Software Engineering, October 2012.

[21] Salvatore Ruggieri,"Efficient C4.5", Dipartimento di Informatica, University a di Pisa Corso Italia 40, 56125 Pisa Italy.

[22] Wei Peng, Juhua Chen and Haiping Zhou," An Implementation of ID3 - Decision Tree Learning Algorithm", University of New South Wales, School of Computer Science & Engineering,Sydney, NSW 2032, Australia.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.