International Journal of Information Technology and Knowledge Management
July-December 2010, Volume 2, No. 2, pp. 455-459

IMPLEMENTATION OF MULTIVARIATE DATA SET BY CART ALGORITHM

Sneha Soni

Data mining deals with various applications such as the discovery of hidden knowledge, unexpected patterns and new rules from large Databases that guide to make decisions about enterprise to products and services competitive. Basically, data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. Data Mining, which is known as knowledge discovery in databases has been defined as the nontrivial extraction of implicit, previous unknown and potentially useful information from data. In this paper CART Algorithm is presented which is well known for classification task of the datamining.CART is one of the best known methods for machine learning and computer statistical representation. In CART Result is represented in the form of Decision tree diagram or by flow chart. This paper shows results of multivariate dataset Classification by CART Algorithm. Multivariate dataset Encompassing the Simultaneous observation and analysis of more than one statistical variable.

Keywords: Data Mining, Decision Tree, Multivariate Dataset, CART Algorithm

1. Introduction

In data mining and machine learning different classifier are used for classifying different dataset for finding optimal classification. In this paper Classification and Regression Tree or CART Algorithm is implanted on multivariate dataset. In Data mining Classification is a major task for in classification systematic distribution of objects is done according to their attribute like, shape, species play, or by any other. In a classification problem, we have a number of cases (examples) and wish to predict which of several classes each case belongs to. Each case consists of multiple attributes, each of which takes on one of several possible values. The attributes consist of multiple predictor attributes (independent variables) and one target attributes (dependent variable). Each of the target attribute's possible values is a class to be predicted on the basis of that case's predictor attribute values.[19]. While in the Regression we defines dependencies between input and output parameters that belong to object set. Regression is the oldest and most well-known statistical technique that the data mining community utilizes. Basically, regression takes a numerical dataset and develops a mathematical formula that fits the data. The major limitation of this technique is that it only works well with continuous quantitative data (like length, weight, speed or age). If you're working with categorical data where order is not significant (like color, name or gender) you're better off choosing another technique [19].

Samrat Ashok Technological Institute, Vidisha, India
Email: Soni.snehaa@gmail.com

2. CART Algorithm

CART stands for Classification and Regression Trees a classical statistical and machine learning method introduced by Leo Breiman, Jerome Friedman, Richard Olsen and Charles Stone in 1984.it is a data mining procedure to present the results of a complex data set in the form of decision tree, diagram or flow chart. In classification tree and categorical outcomes is obtained while in the regression tree continuous outcome is obtained. In CART Algorithm following concept is simply used for making a decision tree [3]

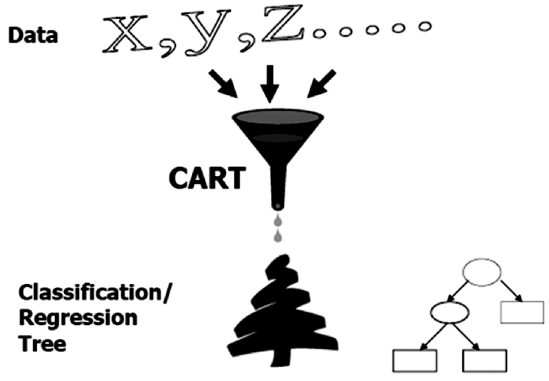


Fig 1: CART Concept

Decision tree structure in CART is as follows [4] as in the fig Top most node is called Root node and each root node is get subdivided into sub nodes called child node, again each child is treated as parent node as splitting criteria are performed until some stopping criteria is not reached, then terminal node is declared and each nodes are assigned to some classes as in this paper terminal node is species of the iris flower dataset.

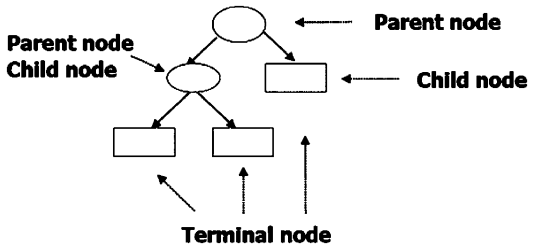


Fig. 2: Concept of Root and Child Node

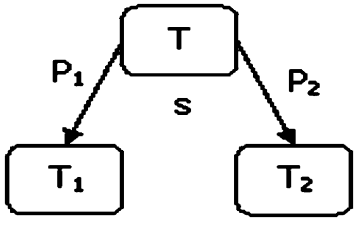
Main Steps for making a decision tree using CART Algorithm:

1. The first is how the splitting attribute is selected.
2. The second is deciding upon what stopping rules need to be in place.
3. The last is how nodes are assigned to classes.

Step1: Splitting a Node [4] [5]: The goal of splitting up a sample is to get sub-samples that are more pure than the original sample. If there are M attributes, there will be a total of M splits to consider. For numerical attributes the splits are binary in nature and the test is of the form {is X_m <= c?}. Commonly used technique is to choose a split that will create the largest and purest child nodes by only looking at the instances in that node. This technique is referred to the 'greedy' or the 'local optimization' approach.

In Greedy Approach following steps are used (i) Search each attribute to find the best split for it (ii) Each of these splits is a candidate split, and there will be M candidate splits for the M attributes being considered (iii) Compare the M splits and pick the best split (iv) Some algorithms keep the second and third best splits as surrogate splits in reserve. These splits are ranked in order in how close they resemble the behavior of the primary splitting rule. These surrogate splits are used when predicting new instances that have missing values for the splitting attribute. Splitting attributes are chosen based on a goodness of split. If we define an impurity function I(t) where t is any given node, then the goodness of split is defined to be the decrease in impurity resulting from the split.

The next picture shows a candidate split s that will create child node T_1 and T_2 from T. The goodness of split will be the difference between the impurity of node T and the sum of the impurities for the child nodes of T (in this case T_1 and T_2). The goal is to find the split with the greatest reduction in impurity.



In the split shown above, the goodness of split for split s defined as:

$$\Delta I(S, T) = I(t) - P_1 I(t_1) - P_2 I(t_2) \quad (1)$$

P_1 and P_2 = Proportions of the instances of t that go into t_1 and t_2 respecpectively

I(t) = Impurity Function.

Impurity function can be defined as by using the concept of conditional probability p(j|t). If there are j classes in all, the conditional probability p(j|t) is the probability of having a class j in node t. An estimate of this conditional probability is N_j/N. Where N is the total number of instances in the node and N_j is the number of class j instances in the node. The impurity of a node is a function of this conditional probability. CART uses Gini Index for defining the impurity function, its formula is:

$$I(t) = \sum_{i \neq j} p(i \mid t) p(j \mid t) \quad (2)$$

Like the entropy function, this measure will reach a maximum when all classes are evenly distributed in the node and it will be at a minimum if all instances in the node belong to one class. There are no issues with the bias discussed previously as the CART algorithm that uses the Gini index is a binary split algorithm and does not have to deal with highly branching splits.

Step 2: Stopping Rules and Building the Final Model [4] [5]: CART uses backward pruning algorithms. This means that they will grow a tree until it is not possible to grow it any further and thus the only stopping rule is when there are only 2 instances left in a node. When all nodes are like this, the tree growing process will end. Pruning will be necessary to build smaller tree models that perform better on new data and not just on the training data. The idea is to remove leafs that have a high error rate. CART uses Pruning in which each node in the tree model has a certain number of instances that are misclassified, say E out of a total of N instances in the node. The training error rate (we will call f) for each node is then simply E/N.

Step 3: Assigning Classes to Tree Nodes [4] [5]: Every node in a tree carries with it a particular classification. This classification is usually determined by a simple majority. In a given node, the class attached to it will be the class that is most well represented by the instances in the node. Leaf nodes are different in that their classification is final, and it is from them that a model's predictive performance is determined. Each node will have an error rate, say e, which is the proportion of misclassified instances in it. The probability that a particular classification will be correct is then simply 1-e. The probability of a correct prediction from the model is then the weighted average of these probabilities from each leaf. These estimates can be based on training data or on a separate and independent test data used to validate the model.

3. ADVANTAGE AND DISADVANTAGE OF CART

CART is nonparametric. it does not require variable to be selected in advance .its Results are invariant to monotone transformation of its independent variable. it Can Handle Outliers. it is Flexible and has an ability to adjust in time [6] CART Also have some disadvantages like it May have unstable decision trees. It splits only by one variable [6]

4. LITERATURE REVIEW: APPLICATION OF CART

Several Application in the literature have considered classification and regression tree. Some of the application are frequently used in numerous areas like financial, medical [3, 8, 9,], spatial data mining [12] and many more areas.

CART Algorithm of decision tree is very useful in predicting financial and risk management areas [3, 19]. CART Also Find its Application in Spatial Data Prediction [12] Weed Classification [20].

Heidi Boerstler and John M. de Figueiredo in [8] present CART as computerized recursive partitioning program to identify potential high users of services among low-income psychiatric outpatients. They present Sociodemographic variables, clinical variables source of referral and the most recent psychiatric treatment setting. They focus on admission to outpatient psychiatric treatment for predicting high use of outpatient psychiatric services

Garzotto Mark ,beer Tomasz et al in [9] present a tree analysis using CART Algorithm for Improved Detection of the Prostate Cancer, they Build decision tree for patients suspected of having prostate cancer using classification and regression tree (CART) analysis. They collected Patients and Methods Data of 1,433 referred men with serum prostate-specific antigen (PSA) levels of ≤ 10 ng/mL who underwent a prostate biopsy. The Factors analyzed included demographic, laboratory, and ultrasound data (ie, hypoechoic lesions and PSA density [PSAD]). Twenty percent of the data was randomly selected and reserved for study validation. In this paper CART analysis was performed in two steps, initially using PSA and digital rectal examination (DRE) alone and subsequently using the remaining variables.

Chee Jen Chanage in [3] Present CART Algorithm for Partitioning Groups in Biomedical. He Provide good information about the prediction of Blood Pressure with certain attribute of the patients.

W. Hannover, M.Richard et al in [14] proposed Classification and Regression tree model for decision making clinical practice. They develop a clinical application using CART. They present a model that identify and describe changes in a patient's risk of treatment failure that may be developed to support decisions during ongoing treatment. This paper conclude that CART has potential for making treatment decision support because CART present tree structure that identifies changes in risk status associated with available treatment option.

Srinivas Mukkamata, Qing Zhang Liu et al in [15] present Computational intelligent techniques that can be useful at the diagnosis stage to assist the Oncologist in identifying the malignancy of a tumor. In this paper they perform a t-test for significant gene expression analysis in different dimensions based on molecular profiles from micro array data, and compare several computational intelligent techniques for classification accuracy on selected datasets. For finding accuracy of classification Linear genetic Programs, Multivariate Regression Spines (MARS), Classification and Regression Tress (CART) and Random Forests are use.

Brain Canada, Georgia Thomas et al in [16] present CART for doing Automatic segmentation and classification of Zebrafish phenotyping. They present a prototype for automated segmentation and classification of histology image in animal models. Oliver Wirjadi, Thomas M.Beeeuel in [17] presents a supervised learning method for image classification. they use CART Decision tree algorithm to choose the features that are most relevant for a given application and apply to evaluate our system on the classification task of meningioma cells. Huajin in [18] uses CART for predicting hip Fracture Recursive Partitioning Methods. Dennis White Et Al in [21] present a Mapping of multivariate spatial Relationships from Regression Tree by Partitions of color Visual Variable.

5. INTRODUCTION TO MULTIVARIATE DATASET: IRIS FLOWER DATASET

Multivariate dataset play very important role in Multivariate statistics, which is a form of statistics Analysis. It encompassing the simultaneous observation and analysis of more than one statistical variable. One of the applications of multivariate statistics is multivariate analysis.

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by Sir Ronald Aylmer Fisher (1936) as an example of discriminant analysis. [1] It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the geographic variation of Iris flowers in the Gaspé Peninsula. [2] The Dataset consists of 50 samples from each of three species of Iris flowers. These Species are Iris setosa, Iris virginica and Iris versicolor. Four features were measured from each sample; they are the Sepal Length, Sepal Width, Petal Length and Petal Width with centimeter (cm) as their units Based on the combination of the four features.

In dataset a term sepal is derived from Latin word separatus "separate" + petalum "petal" which is a part of the flower of angiosperms, flowering plants. Colour of sepal in most flowers is green and lies under the more conspicuous petals. Collection of sepal is called calyx, whereas the

collection of petals is called the corolla. Together, these two structures are known as the perianth of the flower. [1]

The term tepal is usually applied when the petals and sepals share the same color, or the petals are absent and the sepals are colorful. When the flower is in bud, the sepals enclose and protect the more delicate floral parts within. Morphologically they are modified leaves. The calyx or the sepals and the corolla or the petals are the outer sterile whorls of the flower, which together form what is known as the perianth [1] Fisher developed a linear discriminant model to determine which species they are. It contain a iris flower table as shown below only some sample rows are shown and dotted symbol in a row means it consists more statistics values for the same attribute and Species.

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
		……..		
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
		……..		
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
		……..		

Fig. 3: Data Set Table

6. EXPERIMENTAL RESULTS

During Classification of Multivariate dataset, one of the complexity issues is to find Best Split Attribute. In CART for finding best split attribute Gini Index is used. It is also called Impurity Function. After Growing Full Length Tree, next task is find optimal tree using Pruning Technique. In Pruning, Cross-validation and cost of misclassification is used, which is shown in the graph. Following graph shows a plot between Misclassification error and number of terminal nodes.

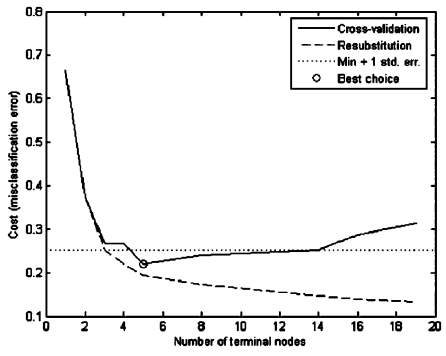


Fig. 4: Graph for Finding Best Splits during Splitting of the Tree

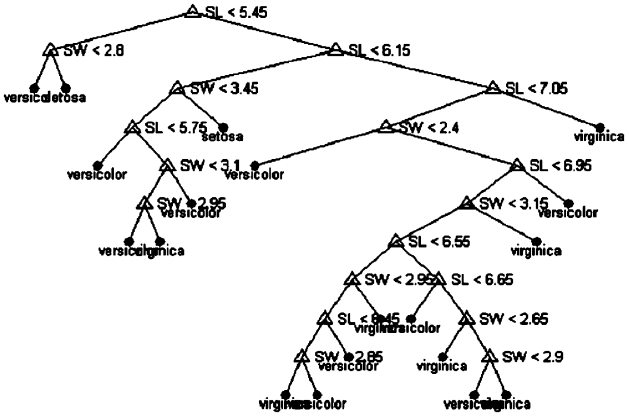


Fig. 5: Classification of Multivariate Dataset by CART Algorithm

7. CONCLUSION

In this paper Implementation of Statistic Application is shown using data mining decision Tree Algorithm, CART. Concept of Impurity function, Cross validation, Resubstitution method and Pruning is implemented for finding optimal decision Tree using CART. Multivariate dataset is applied on CART Algorithm. In Literature Review multiple aspects of CART Application is discussed. In Future it would be interesting to implement this algorithm for other areas like Aviation marketing, Stock Marketing, Market Trading, Bone Cancer Detection and also in Banking Sector.

REFERENCES

[1] Fisher, R.A. "The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics, 7:179–188. 1936.

[2] Edgar Anderson. "The Irises of the Gaspé Peninsula". Bulletin of the American Iris Society, 59:2–5, 1935.

[3] Chee Jen Chang, "Partitioning Groups using Classification and Regression Tree in Biomedical Research", 1/11/2002.

[4] http://wekadocs.com/node/2.

[5] Top 10 Algorithm in Data Mining According to the Survey Paper of Xindong Wu et al know (inf syst (2008) @ springer Verlog London Limited 2007, 14:1-37, 4th Dec'2007

[6] Roman Timofeev, "Classification and Regression Tree, Theory and Application", a Master Thesis Presented by at Center of Applied Statistics and Economics Humboldt University, Berlin December 2004.

[7] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques". Informatica, 31, 249-268, 2007.

[8] Heidi Boerstler and Jhon M.de Fiqueiredo, "Prediction of use of Psychiatric Service:Application of CART Algorithm.college of Business and Administrative and School of Nursing , University of Colorado, USA.

[9] Garzotto Mark ,beer Tomasz, M.Hudson R, "Improved Detection of Prostate Cancer using Classification and

Regression Tree Analysis", Journal of Clinical Oncology, July 2005.

[10] Anton Andriyashin, Financial Application of Classification and Regression Tree Center of Applied Statics and Economics Humboldt University, Berlin 2005.

[11] Anna Jurek and Danuta Zakrzewska, "Improving Na¨ýve Bayes Models of Insurance Risk by Unsupervised Classification", Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 137–144.

[12] L.Bel ,D.Allard, J.M. Laurent, R.Cheddadi and A.Bar-Hen, CART Algorithm for Spatial Data: Application to Environmental and Ecologogical Data, 2008.

[13] Elia Georgiana Petre, A Decision Tree for Weather Prediction PP:77-82, LXI, No 1/2009.

[14] W. Hannover, M.Richard, N.B. Hansen, Z. Martinovich. H.Kordy, A Classification Tree Model for Decision Making in Clinical Practice : An Application based on Data of the Germa Multicenter Study on Eating Disorder. Project TR_EAT Germany, Dec 2002.

[15] Srinivas Mukkamata,Qing Zhang Liu, Rajeev Verraghattam, Andrew , H. Sung , "Computattional Intelligent Techniques for Tumor Classification (Ubibs Microarray Gene Expression Data)" Dept of Comp.Sc, New MAxico Tech, Socorro NM , USA 2002.

[16] Brain Canada, Georgia Thomas, Keith Cheng, James Z.Wang, Automated Segmentation and Classification of Zebrafish Histology Image for High Throughput Phenotyping, 2008.

[17] Oliver Wirjadi, Thomas M.Beeeuel, Wolfgang Feiden and Yoo-jin Kim, Automated Feature Selection for the Classification of Meningioma Cell Nuclei, Institute of Neuropathology, 2006.

[18] Huajin, "Classification Algorithm for Hip Fracture Prediction Based on Recursive Partitioning Methods", Dept of Radiology, University of California, San Friancisco, South Chaina, 2004.

[19] Uzeyir Gurbanli, "Application of Analytical Methods in Risk Management", Institute of Information Technologies, Baku, Azerbaijan.

[20] Juraiza Ishak, Asnor, Md Tahir, Nooritawati Hussan, Aini Mustafa, Mohd Marzuki, "Weed Classification using Classification using Decision Tree", International Symposium , 2, Issue 26-28, Page 1-5, Aug 2008.

[21] Dennis White NAd Jean C. Sifneos Mapping Multivariate Spatial Relationships from Regression trees by Partitionins of Color Visula Variables.