St. Joseph's Journal of Humanities and Science
ISSN: 2347 - 5331
http://sjctnc.edu.in/6107-2/

Efficiency of Cart and Regression Trees in Predicting Student's Absenteeism in an Academic Year

- G. Suresh MCA., M.Phil.*
- K. Arunmozhi Arasan M.Sc., M.Phil.**
- S. Muthukumaran MCA., M.Phil.***

Abstract

Educational data mining is being used to study the data available in the educational field and bring out the hidden knowledge from it. Classification methods like decision trees, rule mining can be applied on the educational data for predicting the students behavior. This paper focuses on the reason for the leave taken by the student in an academic year. The first step of the study is to gather student’s data by using a questionnaire. We collected data from 123 students who were under graduates from a private college which is situated in a semi-rural area. The second step is to clean the data which is appropriate for mining purpose and choose the relevant attributes and the classification is done using the gender attribute. Decision tree was constructed using CART Algorithm by using the Gini index as the splitting criterion. This knowledge is used to identify the reason for the leave taken by the student and help to improve the quality of the environment and also to improve the performance of the student.

Keywords: Data Mining, Decision Trees, CART, Regression Trees.

Introduction

Currently many educational institutions, especially small-medium educational institutions are facing problems with the lack of attendance among the students. The universities will allow the students to write the semester if they have attendance above 80%, student having attendance percentage below 80% will lack attendance and are not permitted to write the semester exam. All educational institutions are facing this problem and so this research aims to find the reason for students absence in the college and take immediate actions to overcome this problem.

Literature Survey

Catherine Butchart, Firat Ismailoglu, presented their paper in 2012 in that, they studied about the increasing number of oldest old people in the future. They identify the potential predictors of inpatient mortality in patients

* Assistant Professor, PG and Research, Department of Computer Applications, St. Joseph's College of Arts and Science (Autonomous), Cuddalore, Tamil Nadu, India. E-mail: sureshg2233@yahoo.co.in
** HOD and Assistant Professor, Department of Computer Science and Applications, Siga College of Management and Computer Science, Villupuram, Tamil Nadu, India. E-mail: arunlucks@yahoo.co.in
*** Assistant Professor, Department of Computer Science and Applications, Siga College of Management and Computer Science, Villupuram, Tamil Nadu, India. E-mail: muthulecturer@rediffmail.com, muthumphil11@gmail.com

over 90 years old admitted acutely to the hospital due to various medical emergencies in two UK centers[1].

Chong Yau Fu presented his paper in combining loglinear model with CART in application to birth data in 2003. In that CART method is based on variation (impurity) reduction for binary splitting; a tree structure was produced as a result of the splitting process[2].

Sunita B. Aher and Lobo, L.M.R.J. presented their paper on the comparison of the five classification algorithm to choose the best classification algorithm for Course Recommendation system. These five classification algorithms are ADTree, Simple Cart, J48, ZeroR & Naive Bays Classification Algorithm. They compare these six algorithms using open source data mining tool Weka & present the result.

F. Questier, R. Put presented their paper on CART approaches for both supervised and unsupervised feature selection in 2004. They describe feature selection by modelling one response variable (y) by some explanatory variable (x).

Background Knowledge

Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that are used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large. Classification and prediction have numerous applications, including fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis. Classification is used to find the class label for the data and prediction is used find the value in the class label[5]

Classification By Decision Tree Induction

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flow chart tree structure, in which 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 topmost node in a tree is the root node[6].

Decision Tree Induction Algorithm

During the late 1970s and early 1980s J.Ross Quinlan a researcher in machine learning developed a decision tree algorithm known as ID3(Iterative Dichotomiser). ID3 adopt a greedy (i.e. nonbacktracking) approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. Most algorithms for decision tree induction also follow such a top-down approach, which starts with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built[7]. A basic decision tree algorithm is summarized below.

Algorithm: Generate Decision Tree

Generate a decision tree from the training tuples of data partition D.

Input:

- Data partition, D, which is a set of training tuples and their associated class lables.
- Attribute_list, the set of candidate attributes.
- Attribute_selection_method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting_ attribute and possibly either a split point or splitting subset.

Output: A decision tree

Method:

```
1. Create a node N:
2. If tuples in D are all of the same class C,
3. Return N as a leaf node labeled with the class C
4. If attribute_list is empty then
5. Return N as a leaf node labeled with the majority class in D;//majority voting
6. Apply Attribute_selection_method (D, attribute_list) to find the "best" splitting_ criterion";
7. Label node N with splitting_criterion;
8. If splitting_attribute is discrete-valued and multiway splits allowed then//not restricted to binary trees
9. Attribute_list→attribute_list-splitting attribute;//remove splitting_attribute
```

```
10. For each outcome j of splitting_criterion // partition the tuples and grow subtrees for each partition.
11. Let Dj be the set of data tuples in D satisfying outcome j;//a partition
12. If Dj is empty then
13. Attach a leaf labeled with the majority class in D to node N;
14. Else attach the node returned by Generate_decision_tree(Dj,attribute_list)to node N; endfor
15. Return N;
```

Classification and Regression Tree (Cart)

Classification and Regression Trees (CART) was introduced by Breiman et al. It is a statistical technique that can select from a large number of explanatory variables(x) those that are most important in determining the response variable (y) to be explained. The CART steps can be summarized as follows[8].

1. Assign all objects to root node.
2. Split each explanatory variable at all its possible split points (that it is in between all the values observed for that variable in the considered node.)
3. For each split point, split the parent node into two child nodes by separating the objects with values lower and higher than the split point for the considered explanatory variable.
4. Select the variable and split point with the highest reduction of impurity.
5. Perform the split of the parent node into two child nodes according to the selected split point.
6. Repeat steps 2-5, using each node as a new parent node, until the tree has maximum size.
7. Prune the tree back using cross-validation to select the optimal sized tree.

For regression trees with numerical response variable the impurity calculated at step 4 can be defined as the total sum of squares of the response values around the mean of each node. For a node with n objects the impurity is then defined as:

Impurity = $\sum_{i=1}^{n}(yi - \bar{y})^2$

For classification trees with a categorical response variable the impurity is defined with i.e. the Gini index of diversity. The Gini index of a node with n objects and c possible classes is defined as:

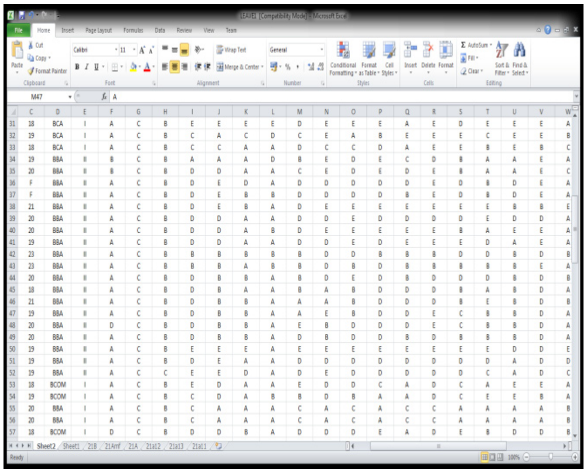
Gini = 1- $\sum_{j=1}^{c}\left(\frac{nj}{n}\right)^2$

Where n_j is the number of objects from class j present in the node.

Data Collection

The data are collected from a private college at Ulundurpet in Villupuram district. There were 123 records collected from the students who are doing under graduate course who belong to the age group 18 to 23. Among the 123 students 85 were male and 38 were female candidates. The data are stored in Microsoft Excel 2010. A questionnaire was prepared and given to the students of the department of BCA, BBA, BCOM. The questionnaire contains 30 questions and five point scale was used. The analysis is done by using the gender attribute.

The data used for data mining contains 123 records and have 30 dimensional attribute namely name, gender, age, department, year, mode of transport, college location, home location, test, cinema, festival, sick, miss bus, friend leave, subject boring, staff question, exam study, result, occasionally, institution work, part time job, assignment, pay fees, native, accident, dress code, commitment friends, college care, impress, problem in college. For our study name is not necessary; so we omitted the attribute and took the 29 attribute for classification.



M47 fx A

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
31	18	BCA	I	A	C	B	E	E	E	E	D	E	E	E	A	E	D	E	E	E	A
32	19	BCA	I	A	C	B	C	A	C	D	C	E	A	B	E	E	E	C	E	E	B
33	18	BCA	I	A	C	B	C	C	A	A	D	C	C	D	A	E	E	B	E	B	C
34	19	BBA	II	B	C	B	A	A	A	D	B	E	D	E	C	D	B	A	A	E	A
35	20	BBA	II	B	C	B	D	D	A	A	C	E	D	E	D	E	B	A	A	E	C
36	F	BBA	II	A	C	B	D	E	D	A	D	D	D	D	D	E	D	B	D	E	A
37	F	BBA	II	A	C	B	D	E	B	B	D	D	D	D	B	E	D	B	D	E	A
38	21	BBA	II	A	C	B	D	E	B	A	D	E	E	E	E	E	E	E	B	B	E
39	20	BBA	II	A	C	B	D	D	A	A	D	D	E	D	D	D	D	E	D	D	A
40	20	BBA	II	A	C	B	D	D	A	B	D	E	E	E	E	E	B	A	E	E	A
41	19	BBA	II	A	C	B	D	D	A	A	D	D	E	D	E	E	E	D	A	E	A
42	23	BBA	II	A	C	B	B	B	B	B	B	D	D	B	B	B	D	D	B	D	B
43	23	BBA	II	A	C	B	B	B	A	B	B	D	B	D	B	B	B	B	B	E	A
44	20	BBA	II	A	C	B	D	B	B	A	B	D	E	D	B	D	D	D	B	D	B
45	18	BBA	II	A	C	B	D	B	A	A	B	A	B	D	D	D	B	A	B	D	A
46	21	BBA	II	A	C	B	D	B	B	A	A	A	B	D	D	D	B	E	B	D	B
47	19	BBA	II	A	C	B	D	B	B	A	A	E	B	D	D	E	C	B	B	D	A
48	20	BBA	II	D	C	B	D	B	B	A	E	B	D	D	D	E	C	B	B	D	A
49	20	BBA	II	A	C	B	D	B	B	A	D	B	D	D	B	D	B	B	B	D	A
50	19	BBA	II	A	C	B	E	E	E	A	E	E	E	E	E	E	E	E	D	D	E
51	19	BBA	II	A	C	B	D	E	A	A	D	D	D	D	D	D	D	D	A	D	D
52	19	BBA	II	A	C	C	E	E	D	A	D	E	D	D	D	D	D	C	A	D	C
53	18	BCOM	I	A	C	B	E	D	A	A	E	D	D	C	A	D	C	A	E	E	A
54	19	BCOM	I	A	C	B	C	D	A	B	B	D	B	A	A	D	C	E	E	B	A
55	20	BBA	I	A	C	B	C	A	A	A	C	A	C	A	C	C	A	A	A	A	B
56	20	BBA	I	A	C	B	C	A	A	A	C	A	C	A	C	C	A	A	A	A	B
57	18	BCOM	I	D	C	B	D	D	B	A	D	D	D	E	A	D	E	B	D	D	B

Figure 1: Data Used For Mining

System Framework

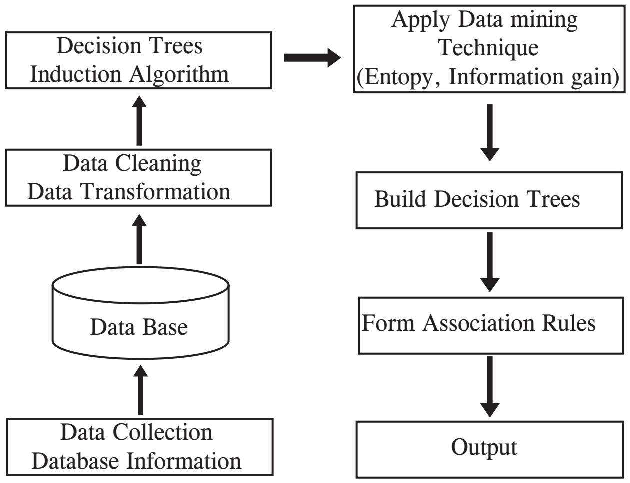


Figure 2: System Frame Work

Experimental Setup

Tanagra is a open source software used for data mining. It supports all the basic type of data formats like *.xls, *.txt etc. and it is very user friendly. The data set is implemented in Tanagra by the following method.

- Open the Tanagra software and go to the file menu and click open then insert the data you want to evaluate.
- Go to Data Visualization and select view data set and drag in to the data set.
- Select the view data set and right click and click execute and then click view then the data is displayed on the right side screen.
- Drag the Define status from the icon and give the target attribute as gender and in input select all the attributes.
- Go to spv learning and select C-RT and drag into the Define status and right click it and click execute and then click view; then the results are displayed in the right side screen.

Evaluation

The analysis of the data is done on the basis of gender i.e. the reason why a male student takes leave and the reason why a female student takes leave from the college by applying Entropy and Information gain

Table 1: Total Records

MALE	FEMALE	TOTAL
85	38	123

$$\text{Gini} = 1-\sum_{j=1}^{c}\left(\frac{nj}{n}\right)^2$$

$$\text{Gini (D)} = 1-\left(\frac{85}{123}\right)^2-\left(\frac{38}{123}\right)^2 = 0.428$$

Table 2: Total records for the attribute Job

JOB	MALE	FEMALE	TOTAL
Strong agree	25	1	26
Agree	28	2	30
Neutral	6	8	14
Disagree	14	13	27
Strong Disagree	12	14	26
TOTAL	**85**	**38**	**123**

The Gini for the attribute part time job is

$$=\left(\frac{26}{123}\right)\left[1-\left(\frac{25}{26}\right)^2-\left(\frac{1}{26}\right)^2\right]+$$

$$\left(\frac{30}{123}\right)\left[1-\left(\frac{28}{30}\right)^2-\left(\frac{2}{30}\right)^2\right]+$$

$$\left(\frac{14}{123}\right)\left[1-\left(\frac{6}{14}\right)^2-\left(\frac{8}{14}\right)^2\right]+$$

$$\left(\frac{27}{123}\right)\left[1-\left(\frac{14}{27}\right)^2-\left(\frac{13}{27}\right)^2\right]+$$

$$\left(\frac{26}{123}\right)\left[1-\left(\frac{12}{26}\right)^2-\left(\frac{14}{26}\right)^2\right]$$

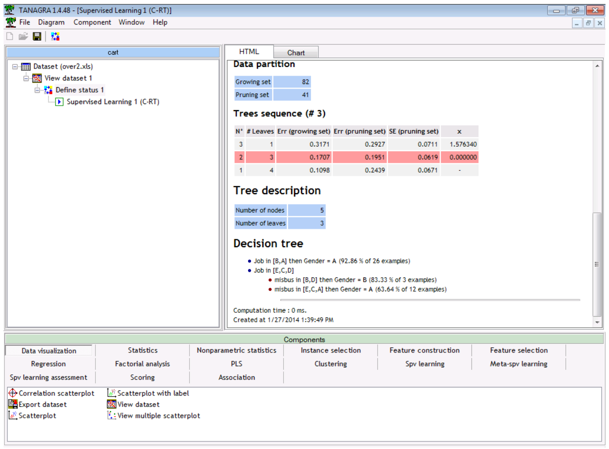
$$= 0.0156 + 0.0302 + 0.0553 + 0.1096 + 0.1050$$
$$= 0.3157$$

$$\text{Gini (job)} = 0.428 - 0.3157$$
$$= 0.1123$$

The information gain for the attribute job has the highest value in the database and the other values in the table are listed below.

The job attribute is taken as the root node of the tree and the 123 records are split by the branches of the job as strongly agree 26 records and Agree 30 records and Neutral 14 records and Disagree 27 records and Strongly Disagree 26 records and the process is applied for the each table and decision tree is formed as below.

Result for Our Data Set



Data partition

Growing set	82
Pruning set	41

Trees sequence (# 3)

N°	# Leaves	Err (growing set)	Err (pruning set)	SE (pruning set)	x
3	1	0.3171	0.2927	0.0711	1.576340
2	3	0.1707	0.1951	0.0619	0.000000
1	4	0.1098	0.2439	0.0671	-

Tree description

Number of nodes	5
Number of leaves	3

Decision tree

- Job in [B,A] then Gender = A (92.86 % of 26 examples)
- Job in [E,C,D]
  - misbus in [B,D] then Gender = B (83.33 % of 3 examples)
  - misbus in [E,C,A] then Gender = A (63.64 % of 12 examples)

Computation time : 0 ms.
Created at 1/27/2014 1:39:49 PM

- Job in [B,A] then Gender = **A** (92.86 % of 26 examples)
- Job in [E,C,D]
  - misbus in [B,D] then Gender = **B** (83.33 % of 3 examples)
  - misbus in [E,C,A] then Gender = **A** (63.64 % of 12 examples)

Figure 3: Decision Tree in Tanagra

The results obtained from Tanagra is shown in the figure above. Here job attribute is taken as the root node because it has the highest information gain; so all job is the root node and it has five discrete attributes as its values called A) Strongly Agree, B) Agree, C) Neutral, D) Disagree and E) Strongly Disagree. All the five are the branches of the node job and all the 123 records are split according to the values present in the each node and the process is repeated at each node until the leaf node comes or until there is no root node. Here in Tanagra the root is dark circle and next child node is blank circle using this dots we can identify the nodes in Tanagra.

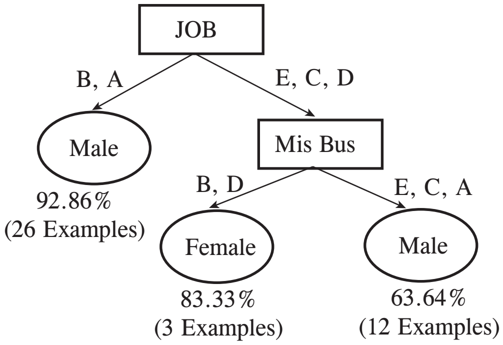


Figure 4: CART Tree for the Job Attribute

The CART Tree for our Dataset obtained from Tanagra is shown in the diagram above. The CART Algorithm used in this model for constructing the CART Tree uses the following parameters such as size before split 10, pruning set size 33%, x-SE rule 1.00, Random generator 1, Show all tree sequence(even if>15) 0.

Discussion

The following are the factors associated with student's absenteeism

- The root node split is considered to indicate the most important single variable, and in this study the job attribute was found to be the factor most strongly linked with student's taking leave.
- The next most important variable related to student's leave is the misbus attribute.

From the total of 123 examples containing 85 male and 38 female the following are the report.

- 92.86% of male students who belong to strongly agree and agree category apply leave to college due to going to job for earning money.
- 83.33% of female apply leave to college when they go to job for earning money and also when they miss the regular bus.
- 63.64% of male apply leave to college when they go to job for earning money and also when they miss the regular bus.

The remaining 27 attributes in the dataset are not considered as much important for the reason of student's absenteeism.

Recommendations

From the above experimental results we know that the job attribute plays a key role on job going students to earn money. To improve good learning environment and the quality of education in the rural and semi-rural areas, our suggestion is that change the college timings such as morning and evening sessions to avoid the students absenteeism for the classes. It was found that students apply leave due to the purpose of studying for the examinations; so if we grant enough study holiday, we can avoid students applying leave to college.

Conclusion

This research aims to study the pattern of students who put leave to the college frequently and the reason behind the students to apply leave. In this research, the decision tree techniques have been used because it is easy to interpret and contributing to the improved results to be compacted. The smaller models are easier to use and general users can understand more easily. We can obtain the best results in predicting the students' behaviour using Classification and Regression Trees (CART) algorithm.

References

1. Catherine Butchart, Firat Ismailoglu, "Identitficaation of possible determinants of inpatient mortality using classification and RegressionTree(CART)analysisinhospitalized oldest old patients",ELSEVIER,2012,
2. Chong Yau Fu, "Combining loglinear model with classification and regression tree(CART):an application to birth data", ELSEIVER, 2003.
3. Sunita B. Aher and Lobo L.M.R.J." Comparative Study of Classification Algorithms", International Journal of Information Technology and Knowledge Management July-December 2012, Volume 5, No. 2, pp. 239-243.
4. F. Questier, R. Put, "The use of CART and multivariate regression trees for supervised and unsupervised feature selection", ELSEVIER, 2004.
5. Michelian kumber and Jai weihan,"Data Mining Concepts and Techniques".
6. S.Anupama Kumar, Dr.Vijayalakshmi M.N, "Efficiency of Decision Trees in Predicting Students Academic Performance", D.C. Wyld,et al. (Eds):CCSEA 2011.
7. Sunitha B.Aher, Lobo L.M.R.J "Comparitive study of Classification Algorithms", InternationaljournalofInformationTechnology and Knowledge Management, 2012.
8. F. Questier, R. Put, D. Coomans "The use of CART and multivariate regression trees for supervised and unsupervised feature selection", ELSEVIER. 2004.