*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727*
*PP 51-56*
*www.iosrjournals.org*

# A Pragmatic Analysis of Implementing Multivariate Decision Tree Algorithm for Supervised Classification of Online Customers

D.Kalaivani [1]* Dr.P.Sumathi [2],
[1]*Associate Professor & Head, Department of Computer Technology, Dr.SNS Rajalakshmi College of Arts and Science,Chinnavedampatty,Coimbatore,Tamilnadu,India*
[2]*AP, PG & Research Department of Computer Science,Govt. Arts College, Coimbatore, Tamilnadu, India*

***Abstract:*** *Data mining is used to transform the data available from the web servers to track the continually changing attitude of online customers. The Business Organizations which are involved in online trading must classify them accurately. Based on the preferences and expectations of various levels of online customers the buying behaviour and pattern also vary. This research paper discusses about the Multivariate Decision Tree Algorithm to perform the Supervised Classification of Online Customers. According to Market Basket Analysis Classification process is taken place. This Classification will be certainly helpful to the organizations to address the expectations of online buyers and lead them to successful sustainability in the online trading. Multivariate Decision Tree deals with attribute correlation. It uses linear machine co-relation of attributes that influences online purchase. Entropy is used to measure uncertainty of attribute occurrences.*
*The experiments and results shown in this research paper are about classification accuracy, efficiency and sensitivity. The results in this paper on this dataset is tested under WEKA tool. It is illustrated that the accuracy of supervised classification under linear machine Multivariate Decision Tree is better than Univariate Decision Tree Algorithm.*

***Keywords:*** *DataMining, Classification, Online Buyers, Linear Machine,Univariate Decision Tree,Multivariate Decision Tree,Market Basket Analysis.*

## I. Introduction

Data Mining is a technique used to analyse and transform the data into useful information which in turn gives knowledge. DataMining is non-trivial extraction of implicit,previously unknown,imaginaable and useful information from data.It is reasoning the data available in large databases.It is the use of software techniques for finding patterns and consistency in sets of data [2]. Data Mining is an interdisciplinary field involving: Databases, Statistics, and Machine Learning. There are various techniques available for data mining as given below:-

A. Association Rule Learning: - This is also called market basket analysis or dependency modelling. It is used to discover relationship and association rules among variables.

B. Clustering: - This technique creates and discovers group of similar data items. This is also called unsupervised classification.

C. Classification: - This can classify data according to their classes i.e. put data in single group that belongs to a common class. This is also called supervised classification.

D. Regression: - It tries to find a function that model the data with least errors.

E. Summarization: - It provides easy to understand and analysis facility through visualization, reports etc. [1].

It is possible to mine data with computer that automates this process. Various data mining tools are available in market some are:-

- Environment for DeveLoping KDD-Applications Supported by Index-Structures (ELKI)
- jHepWork
- Konstanz Information Miner (KNIME)
- Orange (software)
- RapidMiner
- Scriptella ETL — ETL (Extract-Transform-Load) and script execution tool
- WEKA
- MALLET
- KEEL
- CMST DataMiner

WEKA is the most sophisticated and popular Data Mining tool used in different applications used in different for visualization and algorithms for data analysis and predictive modeling. It is also used for data preprocessing, clustering, classification, regression, visualization and feature selection.

Classification is used to manage data,tree modelling of data is used to predict the information from data.A Decision Tree is a Decision Support System that uses tree-like structure or graph to promote decisions and their possible after-event results,resource costs and utility.This verfies problem known as Supervised Classification because the dependent attribute and the counting of classes or values are available.Tree Modelling of data helps to make prediction about new data.Tree Complexity has its effect on accuracy[7]

## II. Objective

The main objective of this researh paper is to evaluate the customer data under Linear Machine Multivariate Decision Tree Algorithm and successfully classify the online buyers so that any Business Organization implementing this algorithm shall successfully and accurately classify their target customers and adapt the changing Markting Merchandising trend.This leads the Business Organization to a successful and sustainable stan in the market of online trading. Decision trees are the most powerful approaches in knowledge discovery and data mining.[5] It includes the technology of research large and complex bulk of data in order to discover useful patterns. This idea is very important because it enables modelling and knowledge extraction from the bulk of data available. All theoreticians and specialist are continually searching for techniques to make the process more efficient, cost-effective and accurate.

## III. Literature Review

Customer who buy online shall be categorised as follows:

**Generator:** The individual who endeavors to induce others in the gathering concerning the result of the choice and regularly assemble data and endeavor to force their decision criteria on the choice.

**Dominator:** The person who attempts to persuade others in the group concerning the outcome of the decision and typically gather information and attempt to impose their choice criteria on the decision.

**Decider:** The person with the power or potentially budgetary expert to settle on a definitive decision in regards to whether to purchase, what to purchase, how to purchase, or where to purchase.

**Buyer and Users:** The individual who directs the transaction and makes the genuine buy and the users are the individual who expends or utilizes the item or service. In addition through showcasing course books and shopper analysts here and there utilize somewhat extraordinary terms a portion of the stages, thus the investigation of purchaser conduct concentrates predominantly on these seven phases and how a scope of components impact each phase of customers' choice

- Need recognition, issue mindfulness
- Search for information
- Pre-buy assessment of choices
- Purchase
- Consumption
- Post-Consumption Evaluation
- Divestment

**Predictive Modelling**

Regression technique can be adjusted for predication. Regression analysis can be utilized to demonstrate the connection between at least one autonomous factors and ward factors. In information mining free factors are traits definitely known and reaction factors are what need to anticipate. Lamentably, some true issues are not just forecast.[8] For example, deals volumes, stock costs, and item disappointment rates are all extremely hard to foresee in light of the fact that they may rely on upon complex collaborations of numerous indicator factors. In this way, more mind boggling strategies (e.g., calculated relapse, choice trees, or neural nets) might be important to gauge future esteems. A similar model sorts can frequently be utilized for both relapse and characterization. For instance, the CART (Classification and Regression Trees) choice tree calculation can be utilized to fabricate both characterization trees (to order straight out reaction factors) and relapse trees (to conjecture persistent reaction factors). Neural systems also can make both order and relapse models**.** [7]

Co-efficient Determination in Regression: One way to assess fit is to check the coefficient of determination, which can be computed from the following formula.

$R^2 = \{ ( 1 / N ) * \Sigma [ (x_i - x) * (y_i - y) ] / (\sigma_x * \sigma_y ) \}^2$

where N is the number of observations used to fit the model, $\Sigma$ is the summation symbol, $x_i$ is the x value for observation i, x is the mean x value, $y_i$ is the y value for observation i, y is the mean y value, $\sigma_x$ is the standard

deviation of x, and $\sigma_y$ is the standard deviation of y. Computations for the sample problem of this lesson are shown below.

| | |
|---|---|
| $\sigma_x = \text{sqrt} [ \Sigma ( x_i - x )^2 / N ]$<br>$\sigma_x = \text{sqrt}( 730/5 ) = \text{sqrt}(146) = 12.083$ | $\sigma_y = \text{sqrt} [ \Sigma ( y_i - y )^2 / N ]$<br>$\sigma_y = \text{sqrt}( 630/5 ) = \text{sqrt}(126) = 11.225$ |
| $R^2 = \{ ( 1 / N ) * \Sigma [ (x_i - x) * (y_i - y) ] / (\sigma_x * \sigma_y ) \}^2$<br>$R^2 = [ ( 1/5 ) * 470 / ( 12.083 * 11.225 ) ]^2 = ( 94 / 135.632 )^2 = ( 0.693 )^2 = 0.48$ | |

A coefficient of determination equal to 0.48 indicates that about 48% of the variation in statistics grades (the dependent variable) can be explained by the relationship to math aptitude scores (the independent variable). This would be considered a good fit to the data, in the sense that it would substantially improve an educator's ability to predict student performance in statistics class.

## IV. Decision Trees

Decision trees are highly effective tools in many areas such as data and text mining, information extraction, machine learning, and pattern recognition[16] Decision tree offers many benefits to data mining, some are as follows:- • It is easy to understand by the end user. • It can handle a variety of input data: Nominal, Numeric and Textual • Able to process erroneous datasets or missing values • High performance with small number of efforts •This can be implemented data mining packages over a variety of platforms [3]. A tree includes: - A root node, leaf nodes that represent any classes, internal nodes that represent test conditions (applied on attributes).

Tree Size Basically, decision makers prefer a decision tree because it is not complex as well as easy to understand. Tree complexity has its effect on its accuracy. Usually the tree complexity can be measured by a metrics that contains: the total number of nodes, total number of leaves, depth of tree and number of attributes used in tree construction. Tree size should be relatively small that can be controlled by using a technique called pruning [4].

Rule Induction in trees Decision tree induction is closely related with the rule induction. Each path that starts from the root of a decision tree and ends at one of its leave represents a rule. These rules can be generated very easily.There are two types of approaches in Decision Tree Algorithms (i)Univariate Decision Tree Algorithm (ii)Multi-variate Decision Tree Algorithm.

(i)Univariate Decision Tree Algorithm : In this technique, splitting is performed by using one attribute at internal nodes. For ex. X<2, y>=10 etc. There are many algorithms for creating such tree as ID3, c4.5 (j48 in weka) etc. This algorithm is an extension of ID3 algorithm and possibly creates a small tree. It uses a divide and conquers approach to growing decision trees that was leaded by Hunt and his co-workers (Hunt, Marin and Stone, 1966) [5].

(ii) Multivariate decision tree Algorithm:In this Algorithm Multivariate Decision Tree is able to generalize well when dealing with attribute correlation and its result is also easy to humans. When working with Univariate DT's, they test single attribute more than once that may result in inefficient tree in some situations. Multivariate DT performs different tests with the data by using more than one attributes in test leaves. Test condition in these trees may be as x + y>10. This technique is a non linear combination of attributes at every test nodes [4].

***Entrophy:*** "Entropy" is used in this process. Entropy is a measure of disorder of data. Entropy is measured in bits, nats or bans. This is also called measurement of uncertainty in any random variable. Just suppose that there is a fair coin, if single toss is performed on that coin than its entropy will be one bit. A series of two fair coins tosses will have entropy of two bits. Now if coin is not fair than there is uncertainty and this provides lower entropy rate.

Entropy for any P can be calculated as :

$$Entropy(p) = -\sum_{j=1}^{n} \frac{|pj|}{|p|} \log \frac{|pj|}{|p|}$$

The conditional Entrophy is:

$$Entropy(j|p) = \frac{|pj|}{|p|} \log \frac{|pj|}{|p|}$$

If base is 2 for logarithm than entropy measurement unit will be in bits, if base is 10 than unit is dits. Information Gain is used for measuring association between inputs and outputs. It is a state to state change in information entropy. Finally information gain can be calculated as:-

$$Gain(p,j) = Entropy(p - Entropy(j|p))$$

LM (Linear Machine) A linear machine is a set of R linear discriminant functions that are used collectively to assign an instance to one of the R classes.[14] Here p is an instance description that consists 1 and the n features that describe the instance. Discriminat function gi(p) for multivariate

***[4]. Each test node will follow the form:-***

$$\sum_{i=1}^{n+1} w_i y_i > 0$$

Where wi are real-valued coefficients, yi are attributes and n is total no of attributes in an instance. Figure 3 and 4 shows the difference between univariate and multivariate space partitioning and also represents that how multivariable test conditions are placed on internal nodes [6].

C. Pruning Pruning is very important technique to be used in tree creation because of outliers. It also addresses overfitting. Datasets may contain little subsets of instances that are not well defined. To classify them correctly, pruning can be used. There are two types of pruning:- 1. Post pruning (performed after creation of tree) 2. Online pruning (performed during creation of tree) [8]. Pruning algorithms Separate and Conquer rule learning algorithm is basis to prune any tree. There are various rule learning schemes. All separate-and-conquer rule learning schemes are modifications of the same simple algorithm that starts with an empty set of rules and the full set of training instances. This simple Separate and Conquer algorithm is given below:-

Procedure SeparateAndConquer (D: instances)

R := empty set of rules • while not D is empty • r := best single rule for D • R := add r to R • remove those instances from D that are covered by r • return R Some Separate and Conquer rule learning schemes are:- • Reduced-error pruning for rules • Incremental reduced-error pruning • Incremental tree-based pruning [3].

## V. Experiment And Results

*Result of Univariate decision tree approach*

Steps to create tree in WEKA

1 Create datasets in MS Excel, MS Access or any other & save in .CSV format.

2 Start the WEKA Explorer.

3 Open .CSV file & save in .ARFF format.

4 Click on classify tab & select J48 from choose button.

5 Select any appropriate test option.

6 Click on Start button & result will be displayed.

Result is displayed in Classifier output window. Result can be viewed in separate window. To do this, right click on result list. To view tree in graphical form, click on "visualize tree" option in pop-up menu

## VI. Conclusion

This paper presents discussion about Decision Trees with the Univariate and the Multivariate approaches. Weka is used as data mining tool that provides various algorithms to be applied on data sets. The J48 algorithm is used to implement Univariate Decision Tree approach, while its results are discussed. The Multivariate approach is introduced as the Linear Machine approach that makes the use of the Absolute Error Correction and also the Thermal Perceptron Rules. Decision Tree is a popular technique for supervised classification, especially when the results are interpreted by human. Multivariate Decision Tree uses the concept of attributes correlation and provides the best way to perform conditional tests as compare to Univariate approach. WEKA provides an algorithm, called M5P that is used to create classification and regression tree with a multivariate linear regression model where p stands for prime. This algorithm provides linear model as classes with some percent of approximated errors. [10] Applying this algorithm on dataset is just same as applying J48 algorithm The research study concludes that Multivariate approach is far better than Univariate approach while it allow us dealing with large amount of data.

## IV. Figures And Tables

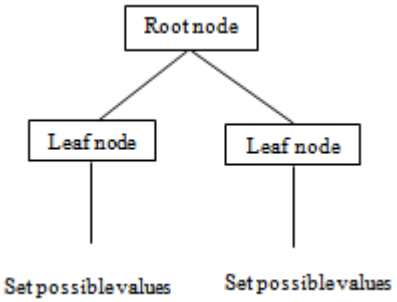


**Figure.1** Decision Tree Model

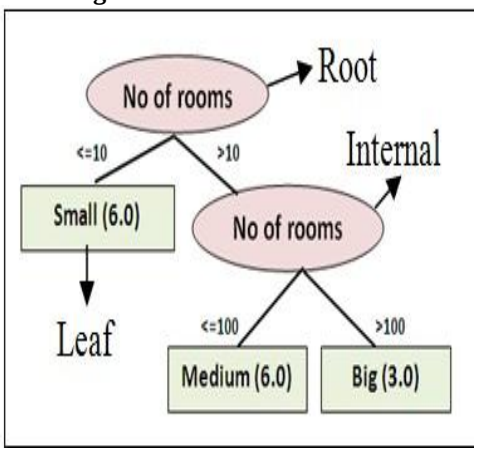


**Figure.2** A Sample Decision Tree

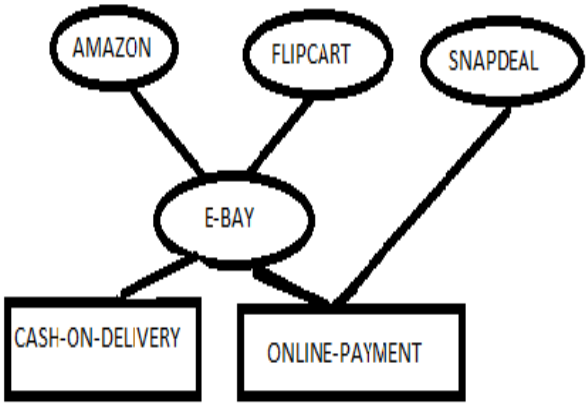


**Figure.2** Decision Tree Visualization

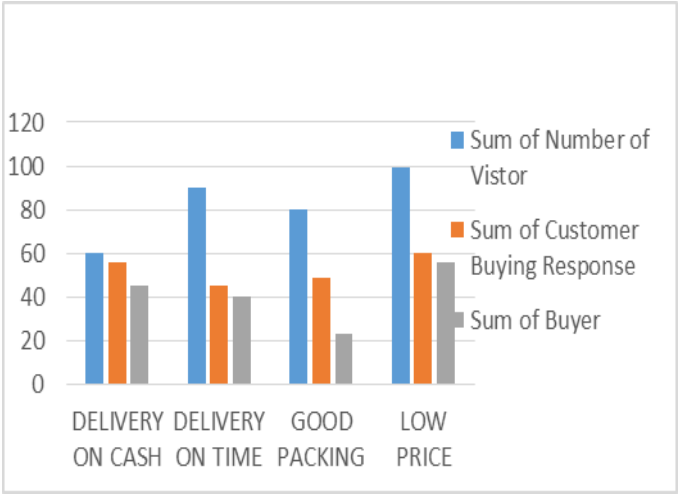


**Figure.3:** Customer Buying Response

**Table.1** Website Visitors & Buyer

| Number of Visitor | Service by Vendor | Customer Buying Response | Actual Buyer |
|---|---|---|---|
| 60 | DELIVERY ON CASH | 56 | 45 |
| 80 | GOOD PACKING | 49 | 23 |
| 90 | DELIVERY ON TIME | 45 | 40 |
| 99 | LOW PRICE | 60 | 56 |

## Acknowledgments

I would like to thank my colleagues, research supervisor and my family members for their continuous support rendered to me while doing to this work.

## References

[1] "Data Mining" from Wikipedia the free Encyclopedia. Web. <http://en.wikipedia.org/wiki/Data_mining>.
[2] Berzal, Fernando, Juan-Carlos Cubero, and Nicol as . "Building multi-way decision trees with numerical attributes." 31. Web. 5 Apr. 2013.
[3] Rokach, Lior, and Oded Maimon. "DECISION TREES." 28. Web. 1 Feb. 2013.
[4] Frank, Eibe. "Pruning Decision Trees and Lists." (2000): 218. Web. 5 Apr. 2013.
[5] Quinlan, J. R. "Improved Use of Continuous Attributes in C4.5." 14. Web. 11 Jan. 2013.
[6] Korting, Thales S. "C4.5 algorithm and Multivariate Decision Trees." 5. Web. 2 Feb. 2013.
[7] Gholap, Jay. "PERFORMANCE TUNING OF J48 ALGORITHM FOR PREDICTION OF SOIL FERTILITY." Web. 2 May 2013.
[8] Moertini, Veronica S. "TOWARDS THE USE OF C4.5 ALGORITHM FOR CLASSIFYING BANKING DATASET." Vol. 8 No. 2, October 2003 (2003): 12. Web. 24 Jan. 2013.
[9] Ozer, Patrick. "Data Mining Algorithms for Classification." (January 2008): 27. Web. 5 May 2013.
[10] Dolado, J. J., D. Rodríguez, and J. Riquelme. "A Two Stage Zone Regression Method for Global Characterization of a Project Database." (2007): 13. Web. 5 Apr. 2013.
[11] Rokach, Lior. "Data Mining with Decision Trees: Theory and Applications." 69 (2008): Web. 3 Feb. 2013.
[12] Ga·sperin, Matej. "Case Study on the use of Data Minig Techniques in Food Science using Honey Samples." (February 2007): 18. Web. 8 May 2013.
[13] Utgoff, Paul E. "Linear Machine Decision Tree." (1991): 15. Web. 6Feb.2013.
[14] JUNEJA, DEEPTI, et al. "A novel approach to construct decision tree using quick C4.5 algorithm." Oriental Journal of Computer Science & Technology Vol. 3(2), 305-310 (2010) (2010): 6. Web. 18 Feb. 2013
[15] Hong Hu, Jiuyong Li, Ashley Plank, "AComparative Study of Classification Methods for Microarray Data Analysis", published in CRPIT, Vol.61, 2006.
[16] Xiang yang Li, Nong Ye, "A Supervised Clustering and Classification Algorithm for Mining Data With Mixed Variables" ,IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, Vol. 6, No. 2, 2006, pp. 396-406 ]http://stattrek.com/regression/regression-example.aspx