ISSN: 2393-8528

IJICSE
Open Access

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 5 Issue 2; March-April 2018; Page No. 01-05

Implementation of Novel DTA (Novel Decision Tree Data Mining Algorithm)on University Students Behaviour in Sharing Information on Facebook using Data Mining

Er. Navneet Kaur[1], Er. Jasdeep Singh Mann[2]

M.Tech Scholar[1], BMSCE, Sri Muktsar Sahib, Punjab, India.

Assistant Professor[2] (Dept. of CSE), BMSCE, Sri Muktsar Sahib, Punjab, India

ABSTRACT

In this research paper, we have shown Social networking sites have gained massive eminence because of the opportunities they give people to connect to each other in an easy and timely manner. Evidently, the fastest growing ecumenical social network during the past few years is Facebook. Although its popularity is declining in Europe and America, number of Facebook users in Thailand, especially in Bangkok is still growing as reported by the Electronics Transaction Development Agency (Thailand), and Siam News-Network that

Bangkok achieves has the highest rank in number of Facebook users in the world where as the age of majority users is the youth. Data mining is used for a variety of purposes in both the publicprivate and sectors. Industries such as sharing information, banking, insurance, medicine, and retailing commonly use data mining to enhance research,reduce costs, and increase sales. For example, the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment.

Keywords: NovelDTA, CART, IDTA, C4.5.

INTRODUCTION

Decision tree learning is a method widely used in data mining. The goal is to create a model that predicts the value of a target variable based on various input variables. An example is shown on the right. Each internal node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf node represents a value of the target variable given the values of the input variables represented by the path from the root node to the leaf node.

The Objectives of this Research Work are:

- To combine the features of Classification algorithms C4.5 and CART Classification algorithm.
- To find Classifications in large dataset efficiently.
- To reduce the sum of square error and achieve accuracy.
- Compare the sum of square error of proposed algorithm with the existing Classification algorithms.

CART and C4.5 classification algorithms

CART and C4.5 are developed by Quinlan for applying *Classification Models*, also called *Decision Trees,* from data. We are given a set of accounts. Each record has the same construction, consisting of a number of quality/value pairs. One of these attributes represents the *group* of the record. The problem is to decide a decision tree that on the basis of answers to questions about the non-category attributes predicts the correctly value of the category attribute. Usually the category attribute takes only the values {true, false}, or {success, failure}, or something equivalent. In any case, one of its values will mean failure.

The basic ideas behind CART are that:

In the decision tree each node corresponds to a non-categorical attribute and each path to a possible value of that attribute. A leaf of the tree specifies the expected value of the definite attribute for the records described by the path from the origin to that leaf. [This defines what a Decision Tree actually is.]

In the decision tree at every node must be related the non-categorical attribute which is most useful among the attributes not so far measured in the path from the root. [This defines what a “Good” decision tree is.]

Entropy is used to predict how informative a node is. [This tells what we mean by "Good". By the way, this notion was used by Claude Shannon in Information Theory.]

**Corresponding author: Er. Navneet Kaur*

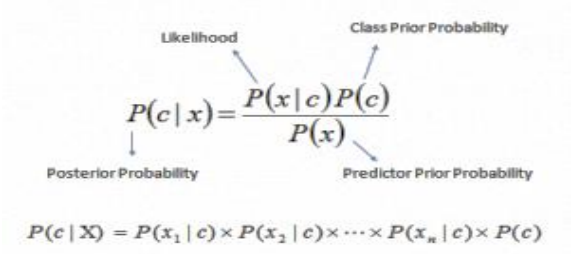
C4.5

C4.5 builds decision trees from a set of training data in the same way as the CART, using the concept of information entropy. The training data set is a set $S = s_1, s_2, \cdots$ of already classified samples. Each sample $s_i = x_1, x_2, \cdots$ is a vector where $x_1, x_2, \cdots$ represent attributes or features of the sample. The training data is augmented with a vector $C = c_1, c_2, \cdots$ where $c_1, c_2, \cdots$ represent the class to which each sample belongs.

At each node of the decision tree, C4.5 chooses one attribute of the data set that most efficiently splits its set of samples into subsets endowed in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data set.The attribute with the largest normalized information gain is chosen to make the decision. The C4.5 algorithm then executes a procedure on the smaller sublists.
This algorithm has a few base cases.

- All the samples in the data list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

Naive Bayes model is quite easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is also known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



Proposed Algorithm NovelDTA

The NovelDTA algorithm combines the features of C4.5 Classification algorithm whose feature of insertion and splitting is same as B-Tree algorithm and Partitioning Classification algorithm CART algorithm. The algorithm is applied on Facebook dataset which is collected from a social networking site Facebook. The NovelDTA algorithm first make call to tree algorithm which is named as DTA algorithm that build a tree containing more than 1500 Classifications on Facebook dataset. The insertion and splitting of this tree algorithm is same as C4.5 algorithm but in this algorithm each node of the tree stores the node or tree label, the Classification number and the number of instances in that Classification. These large numbers of Classifications are difficult to predict and understand. After that the algorithm makes call to CART algorithm Classification algorithm which Classifications the leaf nodes of the CART Classification algorithm. In C4.5 we have to prior define the number of Classifications. In this paper the comparison is done among proposed algorithm, C 4.5 and CART algorithm by changing the number of Classifications.
Steps for Novel DTA Algorithm are:-

Novel DTA-Proposed Algorithm

Input: Training set T, Attribute set S.

Output: Decision tree Tree.

1. Start

2. Compute class frequency(T)

3. Set Tree={}

4. Choose one attribute as class attribute(a€S) and computeInformation gain(I(a,T))=p/p+nlog2(p/p+n)−n/p+nlog2(n/p+n)

5. Foreach attributes b€S**do**

6. Compute Information gain(b,T)

Information gain(I(b,T))= −pi/pi+nlog2(pi/pi+ni)−ni/pi+nilog2(ni/pi+ni)

7. Forv€values(b,T)**do**

8. Set Tb,v as the subset of T with attribute b=v

9. Compute Entropy(b,T)

Entropy(b,T)==$\sum_{i=1}^{v}(pi + ni$/p+n) I(pi, ni)

10. End For

11. Compute Gain(b,T)

Gain(b,T)=I(a,T)-Entropy(b,T)

12. End For

13. Set abest=max{Gain(b,T)}

14. Attach abest into Tree

15. For v€values(abest,T)**do**

16. End For

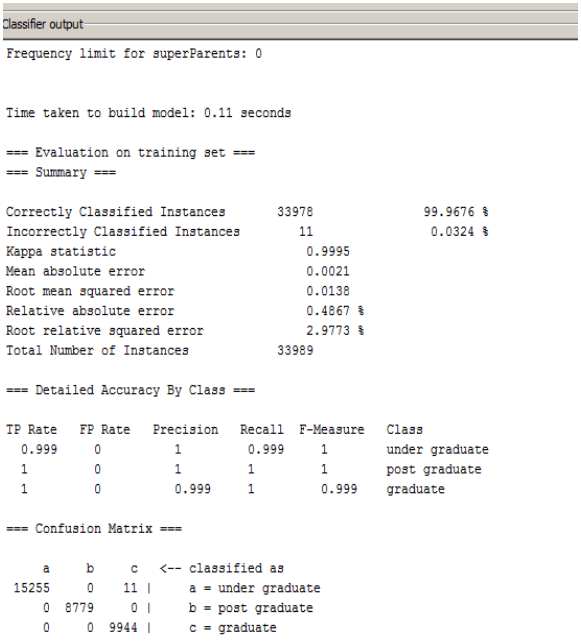
17. Return Tree

18. End

Table: 1 Attributes of Facebook dataset

Gender	**Numeric**
Age	**Numeric**
Area of Education	**Numeric**
Information Shared	**Numeric**
Product	**Numeric**
No. of Facebook Friend	**Numeric**
No. of Hours Used in a Day	**Numeric**
No. of Group Joined	**Numeric**
No. of social networking sites joined	**Numeric**
Education	**Nominal**

IMPLEMENTATION AND RESULT ANALSIS



```
Classifier output

Frequency limit for superParents: 0


Time taken to build model: 0.11 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       33978               99.9676 %
Incorrectly Classified Instances        11                0.0324 %
Kappa statistic                          0.9995
Mean absolute error                      0.0021
Root mean squared error                  0.0138
Relative absolute error                  0.4867 %
Root relative squared error              2.9773 %
Total Number of Instances            33989

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.999     0          1         0.999     1        under graduate
  1         0          1         1         1        post graduate
  1         0          0.999     1         0.999    graduate

=== Confusion Matrix ===

     a     b     c   <-- classified as
 15255     0    11 |     a = under graduate
     0  8779     0 |     b = post graduate
     0     0  9944 |     c = graduate
```

Figure 1.1 shows the result of Novel DTA algorithm when applied on the processed dataset.

Table 2 Comparison among NovelDTA, IDTA, CART & C4.5 algorithms with Correctly Classified Instances on Facebook Dataset

	CART	C45	IDTA	NOVELDTA
Correctly classified instance	44.91	57.34	94.69	99.76

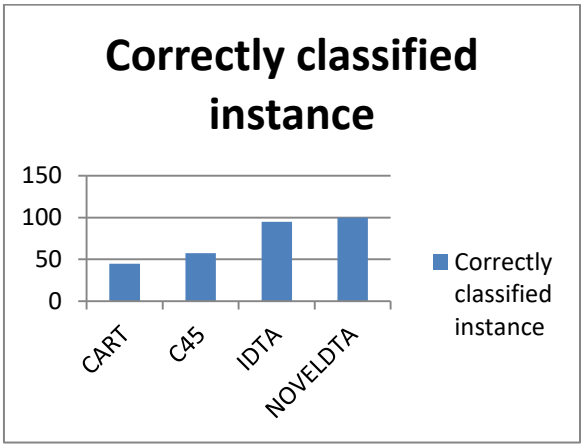


Figure 1.2 Graphical Representations of Correctly Classified Instances

Table 3 Comparison among NovelDTA, IDTA, CART & C4.5 algorithms with Incorrect Classified Instances on Facebook Dataset

	CART	C45	IDTA	NOVELDTA
Incorrect classified instance	55.08	42.65	5.3	0.03

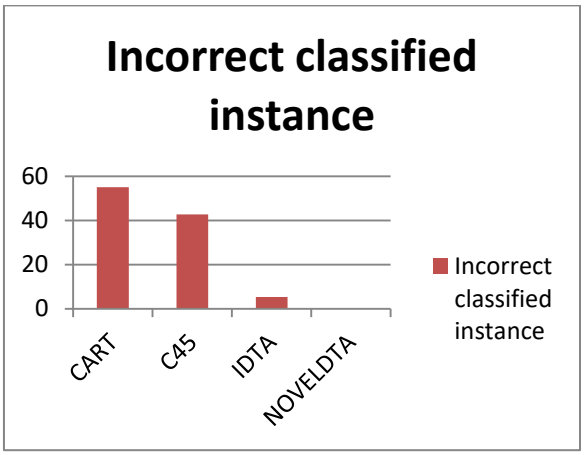


Figure 1.3 Graphical Representations of Incorrectly Classified Instances

Table 4 Comparison among NovelDTA, IDTA, CART & C4.5 algorithms with Error Rate on Facebook Dataset

	CART	C45	IDTA	NOVELDTA
Error Rate	100	91.71	32.29	2.97

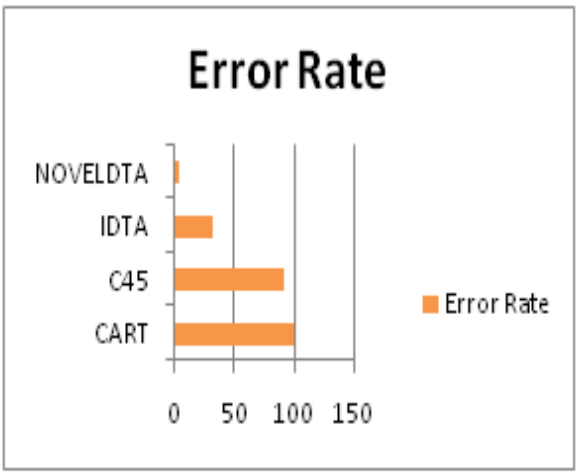


Figure 1.4 Graphical Representations of Error Rate

Table 5 Comparison among Novel DTA, IDTA, CART & C4.5 algorithms with CCI, ICI, Error Rate on Facebook Dataset

	CART	C45	IDTA	NOVELDTA
Correctly classified instance	44.91	57.34	94.69	99.76
Incorrect classified instance	55.08	42.65	5.3	0.03
Error Rate	100	91.71	32.29	2.97

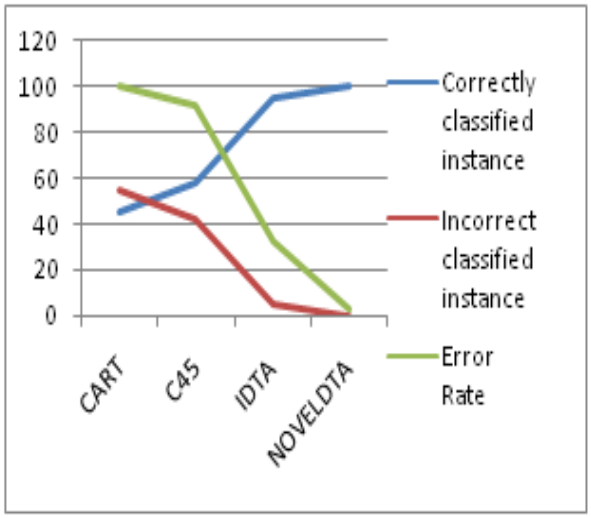


Figure 1.5Graphical Representations of CCI, ICI and Error Rate

CONCLUSION

In this research,study is being done on NovelDTA, IDTA, CART, C4.5 classification algorithms. The features of traditional CART, C4.5, IDTA algorithms are combined and a new algorithm NovelDTA is proposed. The comparison of proposed algorithm is done with the existing algorithm CART, C4.5, IDTA onFacebookdataset using WEKA data mining tool. The results by changing the CCI, ICI,

Error Rates value specifies that the proposed method gives better performance than CART, C4.5, IDTA by reducing the sum of square error which signifies that NovelDTA have high intra classification similarity and is more accurate. Also the proposed algorithm can handle large datasets more effectively.

REFERENCES

1. Shi Na , Liu Xumin, Guan yong , “Research on k-means Classification Algorithm An Improved k-means Classification Algorithm”, 2010 IEEE Third International Symposium on Intelligent Information Technology and Security Informatics.
2. ShuhuaRen, Alin Fan, “*K*-means Classification Algorithm Based on Coefficient of Variation”, 2011 IEEE 4th International Congress on Image and Signal Processing.
3. SaurabhShah,Manmohan Singh, “Comparison of A Time Efficient ModifiedK-mean Algorithm with K-Mean and K-Medoid algorithm”, 2012 IEEE International Conference on Communication Systems and Network Technologies.
4. ShaloveAgarwal, ShashankYadav, Kanchan Singh, “K-means versus K-means ++ Classification Technique”, 2012 IEEE Second International Workshop on Education Technology and Computer Science.
5. Y. Ramamohan, K. Vasantharao , C. KalyanaChakravarti , A.S.K.Ratnam, “A Study of Data Mining Tools in Knowledge Discovery Process”, International Journal of Soft Computing and Engineering (IJSCE)
6. DanJi, JianlinQiu(2010),”A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree”, IEEE International Conference on Computer and Information Technology.
7. Ismael Nidal, Alzaalan Mahmoud and Wesa mAshour (2014),“better Multi Threshold Birch Clustering Algorithm”, Worldwide Journal of Artificial Intelligence and request for Smart Devices.
8. Liu Yanxi (2013),"Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction." Journal of medical systems Vol.10,No.1.
9. ZhouDu Hai, BinLi Yong (2010),“An enhanced BIRC H Clustering Algorithm and request in Thermal Power”,IEEE Web information scheme and Mining.

10. D.Napoleon,P. G.Laxmi(2011),"An competent K-Means Clustering Algorithm for plummeting Time difficulty using consistent Distribution statistics Points", IEEE Trends in order science and computer.

11. A. Moore (2000),"X-means: Extending K-means with competent Estimation of The Number of Clusters",IEEE International Conference on mechanism knowledge.