Improved Decision Tree Classification (IDT) Algorithm For Social Media Data

Anu Sharma[1*]
[1]*er.anusharma18@gmail.com*
[1]Research Scholar,
Uttarakhand Technical University, Dehradun, India
Dr. M.K Sharma[2]
[2] Professor
Amrapali institute, Haldwani, India
Dr. R.K Dwivedi[3]
[3]Professor
Teerthanker Mahaveer University, Moradabad (U.P.), India.

Abstract

In this paper we used classification algorithms on social networking. We are proposing, a new classification algorithm called the improved Decision Tree (IDT). Our model provides better classification accuracy than the existing systems for classifying the social network data. Here we examined the performance of some familiar classification algorithms regarding their accuracy with our proposed algorithm. We used Support Vector Machines, Naïve Bayes, k-Nearest Neighbors, decision tree in our research and performed analyses on social media dataset. Matlab is used for performing experiments. The result shows that the proposed algorithm achieves the best results with an accuracy of 84.66%.

Keywords:
Classification; Data Mining; Decision Tree; KNN; Social Network.

1. Introduction

According to (Almasoud, et al, 2015) "Data mining refers to a collection of techniques used to extract hidden knowledge such as patterns, relationships, or rules from large data sets and to make appropriate predictions"[18].
The social media has now become gotten a focal point of data trade [4].Social network is a popular way to show the connection among the individuals in a gathering or network. It very well may be deciphered as a diagram in which a vertex relates to an individual in that gathering and an edge speaks to some type of relationship between the comparing people. Notwithstanding, informal organizations are extremely powerful articles, since new edges and vertices are added to the diagram over the time. Machine Learning is a method that can adapt consequently and foresee the outcomes dependent on the past perceptions. It gives an exact expectation [4]. In the field of machine learning and data mining, classification is a significant method of information investigation.. Classification is one of the supervised learning methods. Basic classification algorithms are: decision tree, association rules, Bayesian,neural network, genetic algorithm, KNN algorithm [1].

 The classification has two stages, specifically training as well as testing. Training implies making the machine to learn with the information and their associate class. Testing implies making the machine to check the data based on the past perception. In classification, the decision tree is generally mainstream and basic strategy to categorize the data into two or numerous classes. In this paper we used classification algorithms on social networking. Improved Decision tree algorithm has been implemented to show that it gives more accuracy than existing techniques.

2. Literature Review

M.Thangaraj et.al[7] examines various text classification techniques with their qualities, conceivable outcomes and shortcomings in extracting knowledge from data. Semi-supervised text classification is significant in text mining because to its classification efficiency.

Tanushree Dholpuria[10] discuss about the classification algorithms which are utilized for large data analysis that can be utilized to predict Box-Office performance dependent on twitter messages.

Deepa B[12] expressed that classification is done with the assistance of Logistic regression and decision tree classifier. In that decision tree create high exactness than the Logistic regression.

R. Nivedha and N. Sairam [13] characterize the health and non-health related information .The outcome is compare with the Naïve Bayesian classifier.

Mateusz Nurek et.al [15] expressed the aggregate arrangement classification algorithm obtained better outcomes on account of an unequal dataset.

Brijain R Patel[16] expressed that the performances of the algorithms are strongly depends on the "entropy, information gain and the features of the data sets".
Ihsan A. Kareem , Mehdi G. Duaimi[17] centers around C4.5 algorithm.

Bhaskar N. Patel. et.al[18] gather the information of 10 students and attempt to discover the knowledge with assistance of KNN and Decision tree. At last as far as accuracy and complexity decision trees are better than K-Means.

Yan-yan SONG et.al [19] present that decision tree method is a powerful statistical tool for "classification, prediction, interpretation, and data manipulation".

Satbir Kaur[20] introduced decision tree algorithm CART and C4.5. Presumed that C4.5 is more precise and consume less execution time to mine data with least error rate 0.04. Harvinder Chauhan, Anu Chauhan[21] concentrated on serial execution of decision tree algorithm which is memory resident, quick and simple to actualize.

3. Methodology

In the proposed algorithm, a decision tree algorithm is utilized. At that point, linear regression strategy is applied. Proposed improved decision tree algorithm embraces a post-pruning strategy for decision tree algorithm that estimates accuracy, prediction speed and training time. Figure 3.1 shows the architecture and Figure 3.2 shows the flowchart of Improved Decision tree algorithm.
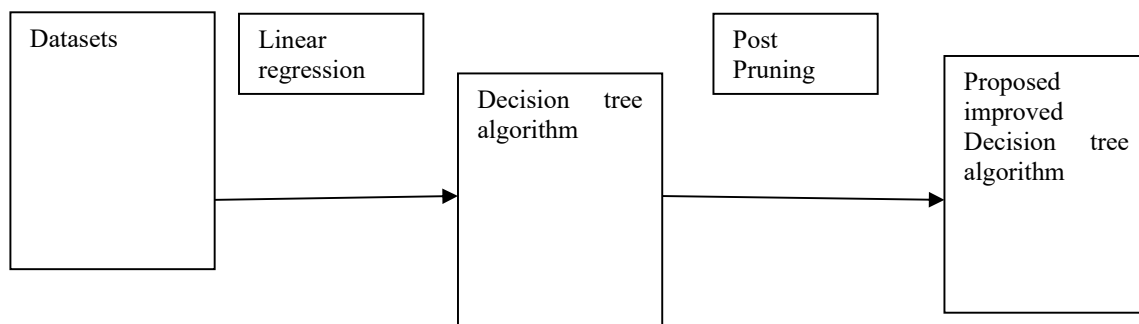
Fig 3.1 Architecture of Improved Decision tree algorithm
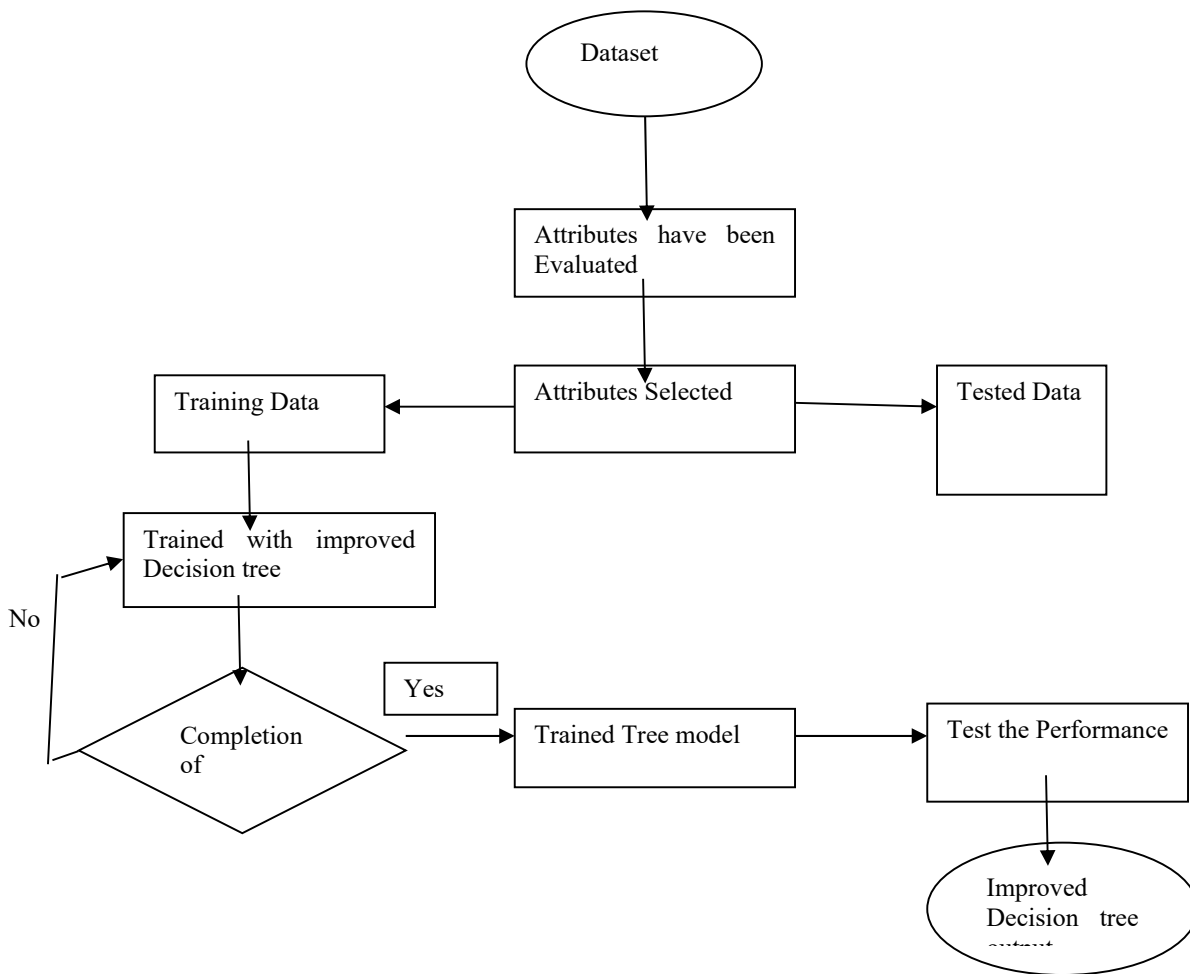
Fig 3.2 Improved Decision tree algorithm Flowchart

3.1 Proposed Algorithm

Input: Social Network Dataset
Output: Improved Decision Tree (IDT)

Method:

1. We have taken social network dataset.
2. Thereafter attributes have been evaluated and then selected.
3. Next we have trained improved decision tree algorithm.
4. On completion of training, we will test the performance.
5. If training is not complete then we goto step [3]

4. Experiment

In this research, we have taken Facebook metrics dataset. We explored with four distinctive classification algorithms. The algorithms that we utilized are KNN, SVM, Decision tree, and Naive Bayes. At that point we analyzed the performance of the above classifiers with improved decision tree (IDT). We observe that an improved Decision Tree (IDT) classification algorithm performs better accuracy at that point existing classifiers.
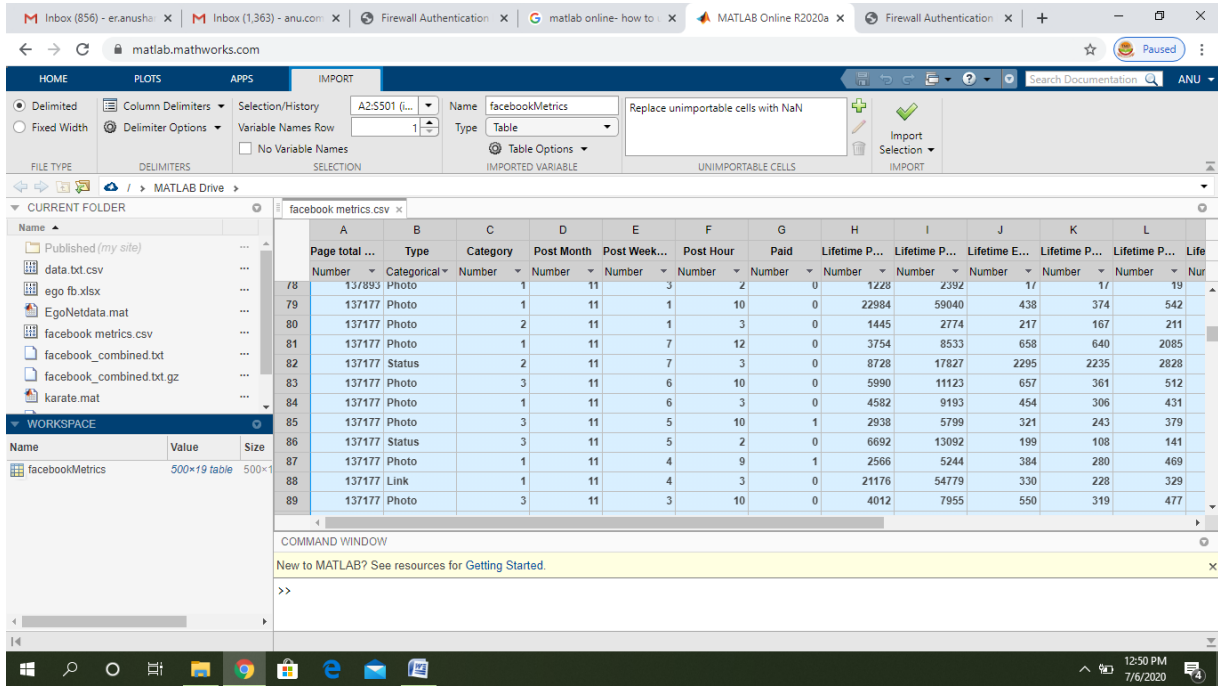
Fig 4.1: face book dataset

Table 4.1 Performance of various classification algorithms

| Classification algorithms	Accuracy	Total misclassification cost	Prediction speed	Training time
DECISION TREE	81.80%	91	~4800 obs/sec	1.0035 sec
LINEAR SVM	84.20%	79	~2700 obs/sec	3.6376 sec
Guassian Naïve Bayes	56.60%	217	~4000 obs/sec	1.4255 sec
FINE KNN	78.40%	108	~2700 obs/sec	3.9467 sec
Weighted KNN	82.40%	88	~4100 obs/sec	2.3982 sec
IMPROVED DECISION TREE	**84.60%**	**77**	**~4500 obs/sec**	**1.2639 sec**
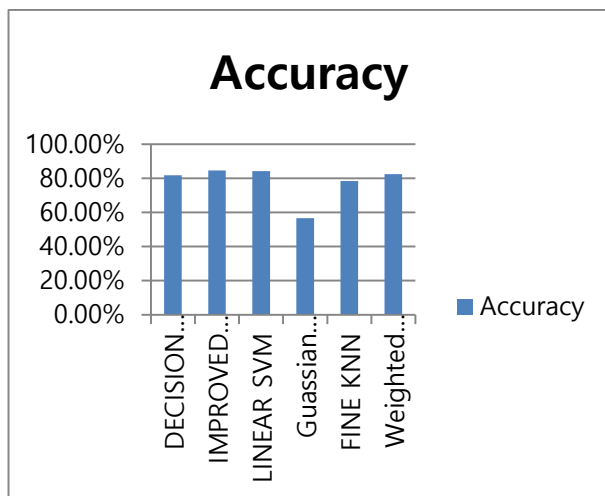
Fig 4.2 Accuracy Graph
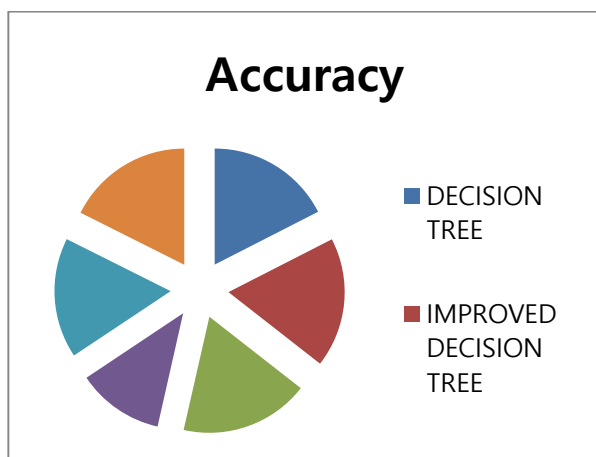
Fig 4.3 Accuracy Chart

5. RESULTS AND DISCUSSION

This paper proposed an improved decision tree classification on facebook data. The outcome is contrasted with the Naïve Bayesian classifier, SVM, Decision tree and KNN. The proposed improved decision tree gives better classification accuracy and performs well with the enormous data set. The resultant information is useful for social network analysis for further mining. Decision trees have consistently been an issue with huge dimensional data. To tackle this issue, improved decision trees are recommended.

References:

[1] S GANAPATHY, "An intelligent temporal pattern classification system using fuzzy temporal rules and particle swarm optimization, "S̄adhan̄a Vol. 39, Part 2, April 2014, pp. 283–302. c Indian Academy of Sciences".

[2] Mohammad Al Hasan Vineet Chaoji Saeed Salem Mohammed Zaki†, Link Prediction using Supervised Learning _Rensselaer Polytechnic Institute, Troy, New York 12180, {alhasan, chaojv, salems, zaki}@cs.rpi.edu

[3] Deepa B, JeenMarseline K.S, Social Media Data using Various Classification Algorithms in Datamaning, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-12S, October 2019

[4] R. Nivedha* and N. Sairam, "A Machine Learning based Classification for Social Media Messages", Indian Journal of Science and Technology, Vol 8(16), 63640, July 2015

[5] Muhammad Azam1 , Dr. Tanvir Ahmed2 , Dr. M. Usman Hashmi3 , Rehan Ahmad4 , Abdul Manan5 , Muhammad Adrees6 , Fahad Sabah7, "Improvement in Classification Algorithms through Model Stacking with the Consideration of their Correlation", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 3, 2019

[6] Zhen-Wu Wang1 , Si-Kai Wang1 , Ben-Ting Wan2 and William Wei Song2,3, "A novel multi-label classification algorithm based on K-nearest neighbor and random walk" ,International Journal of Distributed Sensor Networks 2020, Vol. 16(3) The Author(s) 2020 DOI: 10.1177/1550147720911892 journals.sagepub.com/home/dsn

[7] M. Thangaraj , M. Sivakami * ,
" TEXT CLASSIFICATION TECHNIQUES:
A LITERATURE REVIEW",
Interdisciplinary Journal of Information, Knowledge, and Management, Volume 13, 2018

[8] Zhiwei Cheng1, a, Caisen Chen1, b, Xuehuan Qiu1, c and Huan Xie1, d, "An Improved KNN Classification Algorithm based on Sampling", *Advances in Engineering Research, volume 114*

[9] https://www.guru99.com/data-mining-tutorial.html

[10] Tanushree Dholpuria1Dr. Yuvraj Rana2, "Literature Survey on Various Classification Algorithms in Machine Learning" , International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887 Volume 6 Issue II, February 2018.

[11] 1Umair Ahmed and 2T. Manoranjitham, "Literature Survey on Comparison of Supervised Learning

Classification Algorithms", International Journal of Pure and Applied Mathematics, Volume 118 No. 20 2018, 815-823 ISSN: 1314-3395.

[12] Deepa B, JeenMarseline K.S, "Social Media Data using Various Classification Algorithms in Datamaning", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-12S, October 2019

[13] R. Nivedha* and N. Sairam, "A Machine Learning based Classification for Social Media Messages", Indian Journal of Science and Technology, Vol 8(16), 63640, July 2015, ISSN (Print) : 0974-6846 ISSN (Online) : 0974-5645

[14] Ibomoiye Domor Mienyea, Yanxia Suna,*, Zenghui Wangb, "Prediction performance of improved decision tree-based algorithms: a review", 2nd International Conference on Sustainable Materials Processing and Manufacturing (SMPM 2019)

[15] Mateusz Nurek and Radosław Michalski , "Combining Machine Learning and Social Network Analysis to Reveal the Organizational Structures", Appl. Sci. 2020, 10, 1699; doi:10.3390/app10051699

[16] 1Mr. Brijain R Patel, 2Mr. Kushik K Rana, "A Survey on Decision Tree Algorithm For Classification", 2014 IJEDR | Volume 2, Issue 1 | ISSN: 2321-9939.

[17] Ihsan A. Kareem 1 , Mehdi G. Duaimi 2, "Improved Accuracy for Decision Tree Algorithm Based on Unsupervised Discretization", IJCSMC, Vol. 3, Issue. 6, June 2014, pg.176 – 183

[18] Bhaskar N. Patel, Satish G. Prajapati and Dr. Kamaljit I. Lakhtaria, "Efficient Classification of Data Using Decision Tree", Bonfring International Journal of Data Mining, Vol. 2, No. 1, March 2012

[19] Yan-yan SONG[1,2,*,*] and Ying LU[2,3,] "Decision Tree Methods: applications for classification and prediction", Shanghai Arch Psychiatry. 2015 Apr 25; 27(2)

[20] Satbir Kaur and Harjit Kaur , "Review of Decision Tree Data mining Algorithms: CART and C4.5",International Journal of Advanced Research in Computer Science, Volume 8, No. 4, May 2017 (Special Issue)

[21] Harvinder Chauhan, Anu Chauhan, "Implementation of decision tree algorithm c4.5", International Journal of Scientific and Research Publications, Volume 3, Issue 10, October 2013 1 ISSN 2250-3153.