Turkish Journal of Computer and Mathematics Education Vol.12 No. 7 (2021), 1465-1470

Research Article

Some Novelties in Classification and Analysis of Facebook data – A Decision Tree based Approach

[1]Prashant Bhat , [2]Pradnya Malaganve
[1]Dept. Computational Sciences and IT
(Asst. Professor)
Garden City University, Bengaluru
Bengaluru, India
prashantrcu@gmail.com
[2]Dept. Computational Sciences and IT
(Research Scholar)
Garden City University, Bengaluru
Bengaluru, India
pradnyamalaganve@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

ABSTRACT: As of 2020 statistics, Facebook is the huge social media platform universally having 2.6 billion monthly active users all over and generates data in petabytes every day. Hence Knowledge Discovery from such a huge data is very essential. At present days, Knowledge Discovery is a significant research area. To get the ultimate answers for many research questions in data mining, the final hope is knowledge that can be achieved from different forms of data. If the data has known associations or the data is labelled, supervised approach i.e., Classification method can be used. To accomplish this task, we propose a novel approach to classify the Facebook data at most accuracy.

Keywords— Knowledge Discovery; Facebook; Classification; Accuracy

I. INTRODUCTION

Facebook users are billions in number. Hence Facebook social media generates 4PB(Petabytes) of data per day. That means 40Lacks Gigabytes every day. Thus, the data is huge, stored in Hive which has maximum data storage capacity. Online business on social media platforms is gaining lot of popularity by attracting the customers or users towards their product and brands. And Many social media platforms like Instagram, Facebook, LinkedIn are witness for this. These platforms have most of the active users every day hence it is easy to gain user's attention towards the products by playing attractive advertisements visions. And will definitely help in leading good business strategies. And Facebook allows to upload videos, images, text and links regarding the post, one wishes to explore about any business. This is a great way to communicate with customers which promotes the online business directly [20].

If one is thinking to create a Facebook business page and doing online business, it is worth spending the time to engage with it. According to web statistics 2020, 74% high income earners are the users of Facebook which is the second top most social media and by having 83% of high-income earner, YouTube holds first place. Starting from small business to huge companies are promoting themselves on social medias. In the present work, cosmetic company's Facebook page data is used and it contains 19 different attributes and 500 rows. Attributes such as Type, Life time post total reach, Life time post consumptions, etc. shows the user's reaction on the company's page as well as the reaction of users on each post uploaded. Thus, we have proposed a novel approach to classify this Facebook data. Supervised method [1] is used to label the instances of Type attribute as, video, photo, link and status. And have produced the confusion matrix that determines the proposed algorithm's efficiency as, correctly classified and misclassified instances by the novel approach. And calculated the accuracy of Classification report [2]. As a result, the proposed algorithm has produced the best Accuracy rate for the Facebook data. Hence this approach can help in classifying Facebook data efficiently and can be used to make future predictions on Facebook data which is related to online business. The paper continues with attributes description of the Facebook dataset, proposed algorithm for classifying Facebook data, Experiments and Results and Conclusion.

II. ATTRIBUTES DESCRIPTION

A. Dataset

In the present work, used a cosmetic company's Facebook dataset which includes 19 different attributes and 500 instances. And the dataset carried six missing instances hence filled those missing values by zero (0). Among

19 attributes, two attributes such as type and paid are holding nominal values which are non-numeric and remaining 17 attributes are holding numeric values.

B. **Attributes of the dataset**

Page total likes: Number of users have liked the cosmetic company's Facebook page.
Type: Type of the content i.e. link, video, photo or status. Category: Characterization of the content.
Post month: Post published month.
Post week: Post published week.
Post hour: Post published time.
Paid: Contains yes/no values which indicates whether the cosmetic company has paid to the Facebook for advertising its products.
Life time post total reach: Number of unique users who viewed the page post.
Life time post total impression: Number of times the post from company's page is appeared, even though it is clicked or not. Life time engaged users: Number of unique users clicked anywhere on the post.
Life time post consumers: Total number of users clicked on the page.
Life time post consumptions: Total number of clicks anywhere on the post
Lifetime Post Impressions by people who have liked your Page: Number of impressions from users who liked the page.
Lifetime Post reach by people who like your Page: Total number of unique users viewed a page post just for liked it.
Lifetime People who have liked your Page and engaged with your post: Number of unique users who liked a page and also clicked on the post.
Comment: Number of comments for the post.
Like: Number of likes for the post.
Share: Number of shares for the post.
Total Interactions: Total of comments, likes and shares.

III. PROPOSED ALGORITHM

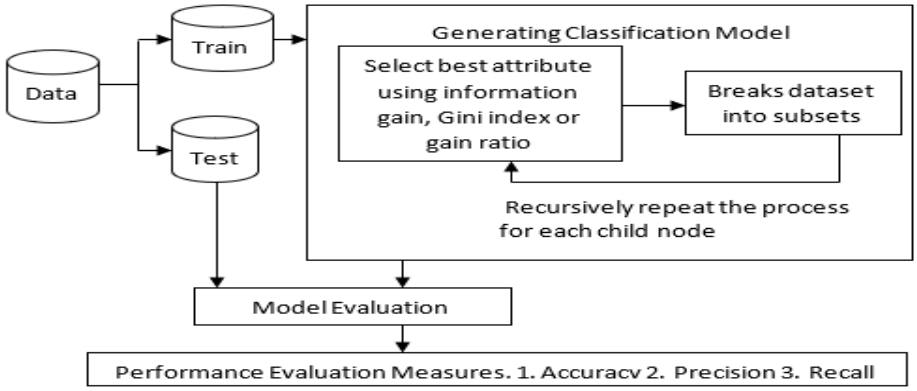

Fig. 2. Flow diagram for the proposed algorithm

In the initial step of the algorithm shown in Fig.1, imported the required python libraries to read the dataset and for further calculations. After the Facebook dataset is read, checked the missing values from the dataset and found 6 missing instances so to fill the missing values, used python pre-processing method and filled it with zero (0) by considering it as a global variable. The code for proposed algorithm is written using python platform and as it reads only numerical values, converted four instances from nominal to numeric data type. i.e., attribute "Type" contains, four nominal values hence considered number 1 for photo instances, 2 for status instances, 3 for video instances and 4 for link instances. And also attribute "paid" contains two nominal values such as yes and no hence denoted those values by numerals 1 and 0 respectively then proceeded the further implementation.
Create a root node and assign all training instances to it.

In the next step, calculate the entropy of the entire dataset.

```
Import pandas as pd
Import numpy as np

Step1: input dataset
        Dataframe= pd.read_csv("Path of the dataset")
Step2: #Seperating the attributes which has particular value
        # at_array denotes n*1 array
        Def separate (at_array, value)
        List1 = []
        For i=1 to length(at_array):
                If(at_array[i] == value):
                        List1=[list1,i]
        Return list1
Step3: #consider that the labels are having the values from 1 to m
        Def calc_Entropy(labels):
        Entropy=0
        For i=1 to m:
        Prob_i = length(separate(labels, i)) / length(labels)
        Entropy - = prob_i * log (prob_i)
        Return entropy
        # used to find the impurity or randomness of the dataset
Step4: #Calculate the frequent values
        Def calc_frequent(labels):
        Connt = -info
        Id1 = non
        For I = 1 to m:
                Count – I = length(separate(labels,i))
                If (count – I >count):
                        Count = count – i
                        Id1 = i
        Return id1
Step5: #Used to find out the best feature of root node
        # Calculate the information gain
        Def info_gain (I1,r,current)
         # I1 = left, r = right, current denotes current uncertainty
        P = float (length(I1)) / (len(I1) + len(r))
        Return current – p * gini(I1) – (I - P) * gini(r)
Step6: #Calculate the gini index
        Def gini_i (specificvalue, k)
        Gini = 1 – (prob(specificvalue = k))²
        # A feature which has lowest gini index is considered for splitting
        Return gini
Step7: # Calculate the accuracy for correctly classified instances
        Accuracy = (TP+TN) / (TP+TN+FP+FN)
```

Fig. 1. Proposed algorithm to calculate the classification accuracy of the Facebook dataset

Entropy: Decision Tree partitions the dataset into subsets starting from the root node and the subset will be homogeneous (similar value). Entropy is used calculate the similarity or homogeneity of a sample of a dataset. If it is totally homogeneous, then the entropy is zero and if the sample is equally portioned then it has entropy of one [3].

$$\text{Entropy} = -P \log_2 P - q \log_2 q \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Partition all the instances and calculate the information gain of every single feature and get that feature with highest information gain [13].

Information gain: It is calculated on decrease in entropy when dataset is split on an attribute. Decision Tree is formed by finding the attribute which gives highest information gain [8]. Steps to calculate the information gain is shown below:

Step 1: Calculate entropy

Step 2: Dataset is divided on different attributes. For every branch, the entropy is calculated and added proportionally to get total entropy. The result is subtracted from the entropy before spit. Now we get the information gain.

Step 3: Select attribute with highest information gain and consider that as the decision node. Split the dataset by its branches and repeat for all the branches.

Step 4: If entropy is zero, that branch is a leaf node and if entropy greater than zero needs to split further.

Step 5: Till the dataset is fully classified, algorithm runs recursively.

Recognize the feature which gives highest information gain. Set that particular feature as splitting criterion at current node. If information gain is 0(zero), then set current node as leaf node and return. Elaborate for every feature value an outgoing branch and consider unlabeled nodes at the end [4].

Partition the dataset along with the values of highest information gain feature and discard this feature from the dataset. Consider every partition as child node of the current node [6]. Calculate the Gini index and a feature which has the lowest Gini index is taken for splitting the dataset [11].

For every child node, if child node has instances from single class, then assign it as leaf node else, Repeat the steps starting from calculating entropy till partitioning the dataset using Gini index until the final criteria is satisfied.

Finally calculate the Classification accuracy for the Facebook dataset [19].

$$\textbf{Accuracy} = (TP+TN)/(TP+TP+FP+FN) \quad \ldots\ldots\ldots\ldots (2)$$

Accuracy is the percentage of correctly classified instances. Where, TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative

Decision Tree helps building regression and Classification models in tree structured format. It breaks the dataset into several smaller subsets and at the same time it enhances tree development. At the end, outcome will be a tree with leaf nodes and decision nodes.

IV. EXPERIMENTS AND RESULTS

```
[[160   1   3   3]
 [  3  15   0   2]
 [  3   1   0   0]
 [  4   0   0   5]]
              precision    recall  f1-score   support

           1       0.94      0.96      0.95       167
           2       0.88      0.75      0.81        20
           3       0.00      0.00      0.00         4
           4       0.50      0.56      0.53         9

    accuracy                           0.90       200
   macro avg       0.58      0.57      0.57       200
weighted avg       0.90      0.90      0.90       200
```

Fig. 3. Generated confusion matrix and calculated classification Accuracy using proposed algorithm

```
[[164   5   0   1]
 [  2  16   0   0]
 [  2   1   0   0]
 [  3   2   1   3]]
              precision    recall  f1-score   support

           1       0.96      0.96      0.96       170
           2       0.67      0.89      0.76        18
           3       0.00      0.00      0.00         3
           4       0.75      0.33      0.46         9

    accuracy                           0.92       200
   macro avg       0.59      0.55      0.55       200
weighted avg       0.91      0.92      0.91       200
```

Fig 4. Generated confusion matrix and calculated classification Accuracy after changing the Maximum depth value

Fig. 3. represents the output received for confusion matrix and classification Accuracy score of Facebook dataset. And also generated the interpretation of performance measures such as precision, recall, f1 score and support. Where in precision is the ratio of correctly predicted positive instances to the total number of positively predicted instances. Recall is the ratio of correctly predicted positive instances to the whole number of instances present in the actual class [17]. F1 score is the average of precision calculated and recall. And Support indicates the number of times one particular instance occurs [18]. Hence the proposed algorithm is able to classify the Facebook dataset as follows; As we have used train test split method to divide the dataset in two different partitions, the algorithm uses 60% of data for training and remaining 40% of data for testing. Hence, total number of instances considered for testing are 200 out of 500 instances and remaining 300 instances come under training set. As a result, the algorithm has correctly classified 177 instances out of 200 testing instances and 23 instances are misclassified. The confusion matrix represents that the algorithm has correctly classified 160 instances as photos, 15 instances as status, 0 i.e., no instances are classifying under video type and 5 instances are correctly classified as link with the good Accuracy of 90%. And also, the values of precision, recall, f1 score and support for all four types of instances are generated in fig 3.

Out[80]:

	max_depth	train_acc	valid_acc
0	1	0.890000	0.850
1	2	0.900000	0.865
2	3	0.916667	0.875
3	4	0.953333	0.910
4	5	0.956667	0.890

Fig. 4. Accuracy score of first five rows of the dataset

To check the variation in the Accuracy level and in order to get better Accuracy, we have changed the maximum depth in the proposed algorithm hence taken the maximum depth from 1 to 10, it trains the model 10 times and changes the variable maximum depth and stores the training and testing(validation) Accuracy for every model. Fig 5. Shows the first five rows of the dataset. It has training and testing (validation) Accuracy score corresponding to a set of range of maximum depth. And the values are plotted in the form of graph shown in fig 6. In fig 6, blue colored line (upper one) denotes training Accuracy and Orange line (below one) denotes testing(validation) Accuracy.

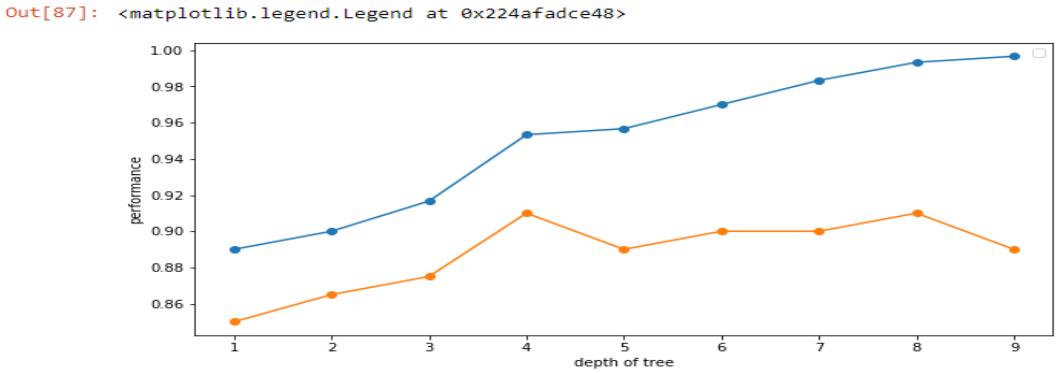

Fig 6. Graph to represent the performance measure

After plotting the values on the graph in fig6., it is visible that when maximum depth value is 1, training and testing Accuracy are low hence assigning lower value to the maximum depth does not allow the model to learn different patterns and can be considered as underfitting. When the maximum depth value goes on increasing, both training and testing Accuracy are increasing. The magnitude of increase in the Accuracy level of training data is higher than the magnitude of increase in the Accuracy level of testing data hence in the proposed algorithm, set the maximum depth value as 4 and kept the random state to 10 and assigned the leaf node value to 25. Thus, the variation in the confusion matrix as well as in the classification Accuracy is achieved by proving higher Accuracy than which is shown in fig 3. i.e., previous Accuracy was 90% and after the changes made in the present algorithm the new Accuracy is 92% as shown in fig 4.

Conclusion

As studied and went through number of literatures, found that, no work is done on classifying the Facebook in order to do the Knowledge Discovery as well as to give the high Accuracy score of classifying Facebook data. Hence the proposed algorithm has done the work by classifying the Facebook data in an efficient manner with 90% Accuracy. And also proved that the variations made in different variables can give even better Accuracy and in present algorithm we have achieved 92% of Classification Accuracy for the Facebook dataset. And we like to conclude that the proposed algorithm is the best suitable method in classifying the Facebook data.

REFERENCES

1. Himani Sharma1, Sunil Kumar, “A Survey on Decision Tree Algorithms of Classification in Data Mining,” International Journal of Science and Research, ISSN: 2319-7064, Value: 6.14 Impact Factor: 6.391, Y:2015
2. Harsh H. Patel, Purvi Prajapati, “Study and Analysis of Decision Tree Based Classification Algorithms,” International Journal of Computer Sciences and Engineering, Vol.-6, Issue-10, E-ISSN: 2347-2693, Oct. 2018
3. Mr. Brijain R Patel, Mr. Kushik K Rana,“A Survey on Decision Tree Algorithm For Classification,” International Journal of Engineering Development and Research, Volume 2, Issue 1, ISSN: 2321-9939, Y: 2014
4. Gregor Stiglic1, Simon Kocbek, Igor Pernek, Peter Kokol, “Comprehensive Decision Tree Models in Bioinformatics,” Volume 7, Issue 3, e33812, March 2012
5. Chaitanya Manapragada, Geoffrey I. Webb, Mahsa Salehi, ,“Extremely Fast Decision Tree,” KDD’18, August 2018, London, United Kingdom
6. Badr Hssina, Abdelkarim Merbouha,Hanane Ezzikouri,Mohammed Erritali, “A comparative study of decision tree ID3 and C4.5,” International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications.
7. Siva S. Sivatha Sindhu, S. Geetha, A. Kannan, “Decision tree based light weight intrusion detection using a wrapper approach,” Expert Systems with Applications 39, (2012) 129–141

8. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye , Tie-Yan Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, NIPS 2017
9. W. Nor Haizan W. Mohamed, Mohd Najib Mohd Salleh, Abdul Halim Omar, "A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms," IEEE International Conference on Control System, Computing and Engineering, 23 - 25, Penang, Malaysia, Nov. 2012
10. A. S. Galathiya, A. P. Ganatra and C. K. Bhensdadia, "Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning," International Journal of Computer Science and Information Technologies, Vol. 3 (2) „3427-3431, Y:2012
11. Marina Milanović, Milan Stamenković, "Chaid Decision Tree: Methodological Frame And Application," Economic Themes 54(4): 563-586, Published 2017
12. Saman Rizvi , Bart Rienties , Shakeel Ahmed Khoja, "The role of demographics in online learning; A decision tree based approach," Computers & Education 137 (2019) 32–47, Y:2019
13. Mrinal Pandey, Vivek Kumar Sharma, "A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction," International Journal of Computer Applications (0975 – 8887) Volume 61– No.13, January 2013
14. Leszek Rutkowski, Maciej Jaworski, Lena Pietruczuk , Piotr Duda, "The CART decision tree for mining data streams," Information Sciences 266, 1–15, 2014
15. D.Lavanya and Dr.K.Usha Rani, "Ensemble Decision Tree Classifier For Breast Cancer Data," IJITCS Vol.2, No.1, February 2012
16. Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, Shrikanth Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," Speech Communication 53, 1162–1171, Y:2011
17. Dewan Md. Farid, Li Zhang , Chowdhury Mofizur Rahman, M.A. Hossain, Rebecca Strachan, "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks," Expert Systems with Applications 41, 1937–1946, Y: 2014
18. T.Miranda Lakshmi, A.Martin, R.Mumtaj Begum, Dr.V.Prasanna Venkatesan, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data," I.J.Modern Education and Computer Science, 5, 18-27, Y:2013
19. Prashant Bhat, Pradnya Malaganve and Prajna Hegade," A New Framework for Social Media Content Mining and Knowledge Discovery," IJCA (0975 – 8887), Volume 182 – No. 36, January, Y: 2019
20. Prashant Bhat and Pradnya Malaganve, "Review on Social Media Content Mining and KDD," IJRAR, Volume 5, Issue 3, M:July, Y:2018