ISSN No:-2456-2165

Grid Search Hyper-Parameter Tuning and K-Means Clustering to Improve the Decision Tree Accuracy

Shivam Kumar[1]*, Tushar Singh[2], Smita Singh[3], Shivam Singh[4]
B. Tech. CSE Department Students, IMS Engineering College,Ghaziabad, Uttar Pradesh, India
Mr. Naveen Kumar Rathore (Asst. Professor, IMS Engineering College, India)

Abstract:- Representation and quality of the instance data are the foremost factors that affects classification accuracy of the statistical - based method Decision tree algorithm which gives less accuracy for binary classification problems. Experiments shows that by using clustering and hyper-parameter tuning, the decision tree accuracy can be achieved above 95%, better than the 75% recognition using decision tree alone.

Keywords:- *Classification, Clustering, K-means, Decision Tree, Hyper-parameter Tuning, Grid Search, Customer Churn, Logistic Regression.*

I. INTRODUCTION

Predicting a correct decision on data set is very much important, it can be in any field either in business, engineering, civil planning, law or in other real-life areas. The base of decision making ishaving the correct data, a dataset is the collection of related discrete items of related data that may be accessed individually or in combination or managed as whole entity. The dataset which we are dealing with our research experiment is customer churn that is a binary ('Yes'/'No') classification problem, in this we are predicting the customer churn for the business. Here the churn prediction isabout detecting which customers are likely to leave the business. The algorithm that works best on such binary classification problems is logistic regression, which gives the decent accuracy around 85%, but this accuracy is not enough to make correct decisions in a business. Therefore, we are introducing a hybrid algorithm that uses Decision Tree as the classifier, in which clustering and hyper-parameter tuning are used as the supportive elements. Although the accuracy of decisiontree on this customer churn dataset is very less, near to 75%. It is because the Decision Tree algorithm have many limitations like the independency between samples which has drastic effects on its accuracy, along with this it is not suitable for big size datasets. Decision Tree causes overfitting problem for the model in case data does not fit well. For the setup of our hybrid algorithm the data filtration is used in the initial phase, it is the refining of data which are either null values or values that are not supported by Decision Tree algorithm. After the data filtration, clustering is used to group the data elements on the basis of similarity and dissimilarity and makes the data more structured which makes Decision Tree to perform better. After using clustering outcome with the Decision tree, it shows a little hike and the accuracy reached near to that of logistic regression, but there is still a large possibility for the improvement. The next step of our hybrid algorithm is hyper-parameter tuning which pull out the bad and irrelevant parameters forthe churn prediction. After using this on the outcome of clustering, the accuracy of Decision Tree gets improved and touches the 90% banner. As the customer churn dataset has only two possible outcomes, i.e. 'Yes'/'No'. So due to the nature of our dataset we didn't take much benefit of clustering but the results become more better.

This hybrid algorithm is applied on 20 datasets and there is a large benefit of clustering for the datasets which has several class values that are to be predicted. Our customer churn dataset has only two class values so we used hyper-parameter tuning on this dataset without other customization and the Decision Tree accuracy touched the 95% banner.

The hybrid algorithm has improved the accuracy of Decision Tree by more than 25%. So, this can be concluded that the achieved results of experiments are far greater than the best suited machine learning model (logistic regression) on this binary classification dataset.

- *Flow Chart*

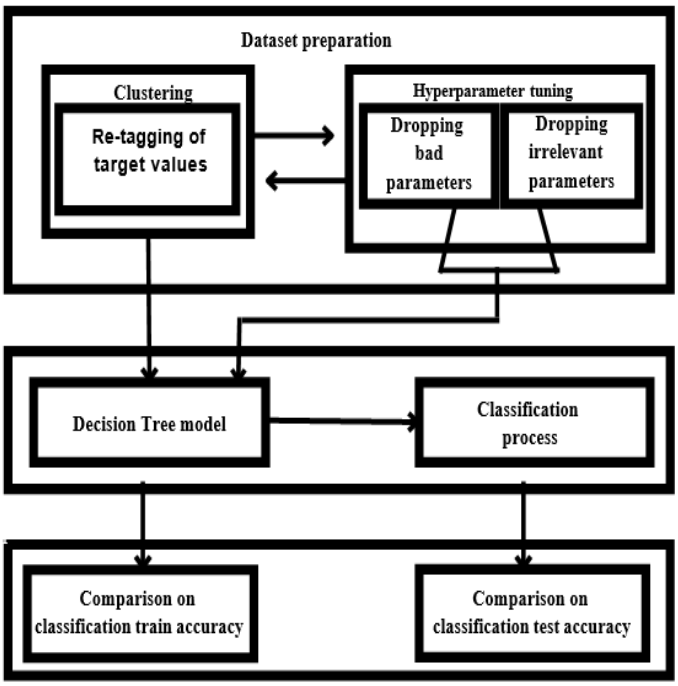

Fig 1:- Flow chart

ISSN No:-2456-2165

➢ *Data Filtration*

Data Filtration is the process of refining the data, for example if a dataset contains no value or null value then data filtration will remove that null valuefrom the dataset. Through data filtration we can have the set of all useful data.

➢ *Clustering*

Clustering is the process of dividing the data points into different groupssuch that same type of data is in one group and the data which are not similar are indifferent groups. We can say that, clustering is basically the collection of data points on the basis of similarity and dissimilarity of data.

➢ *Parameter Tuning*

Parameter Tuning is the process of removing the unwanted parameters from the dataset, it is very useful as it controls the overall behavior of machine learning model. In parameter tuning the parameters should be adjusted insuch a way that the best predictions can be achieved by the model. This is the technique of adjusting the elements which control the behavior of the model.

➢ *Classification Process*

Classification is the process of classifying or categorizing thedata points on the basis of the class or of same group. It can be performed on structured and unstructured data. In classification process, a program learns from given data or observation and then classifies new observation into a number of groups and classes.

II. K-MEANS CLUSTERING FOR PRE-PROCESSING

Normally we solve such binary classification problems using logistic regression because it performs best on binary class labels, being fast and relatively simple. Also, it can be applied to multi-class problems. The general accuracy of this classifier varies from 80% - 85%, but it is far from the best. Hence by making the data more structured with k-means clustering process our Decision Tree model makes it easier to train and perform better.

Clustering is an important tool for a wide variety of applications in data mining, statistical data analysis, data compression and vector quantization. The goal of clustering is to group data into clusters in such a way that the similarities between data members within the same clusters are maximum while the similarities between data members from different clusters are minimal.

K-means is well known prototype-based partitioning clustering technique that attempts to find a user-specified number of clusters(K), which are represented by their centroids.

The K-means algorithm is as follows:

➢ Choose starting centers of K clusters.
➢ Generate a separate division by putting data to its nearest cluster centers.
➢ Compute new cluster centers which will be used as centroids for clusters.
➢ Repeat steps 2 through 3 until the clusters are stable.

❖ *Pre-Processing Steps* -

A. Data cleaning

➢ Removing duplicates
➢ Removing irrelevant observations and errors Removing unnecessary columns
➢ Handling inconsistent data Handling outliers and noise

B. Handling missing data

C. Data Integration

D. Data Transformation

➢ Feature Construction
➢ Handling Skewness
➢ Data Scaling

E. Data Reduction

➢ Removing dependent (highly correlated) variables
➢ Feature selection

address	income	ed	employ	equip	callcard	wireless	longmon	...	pager	internet	callwait	confer	ebill	loglong	logtoll	lninc	custcat	churn
7.0	136.0	5.0	5.0	0.0	1.0	1.0	4.40	...	1.0	0.0	1.0	1.0	0.0	1.482	3.033	4.913	4.0	1.0
12.0	33.0	2.0	0.0	0.0	0.0	0.0	9.45	...	0.0	0.0	0.0	0.0	0.0	2.246	3.240	3.497	1.0	1.0
9.0	30.0	1.0	2.0	0.0	0.0	0.0	6.30	...	0.0	0.0	0.0	1.0	0.0	1.841	3.240	3.401	3.0	0.0
5.0	76.0	2.0	10.0	1.0	1.0	1.0	6.05	...	1.0	1.0	1.0	1.0	1.0	1.800	3.807	4.331	4.0	0.0
14.0	80.0	2.0	15.0	0.0	1.0	0.0	7.10	...	0.0	0.0	1.0	1.0	0.0	1.960	3.091	4.382	3.0	0.0

Table 1 :- Data

III. EXPERIMENTAL OPERATIONS

➢ *K-means for structuring data?*

The process of clustering groups similar data together, resulting in a more structured overall data.The cluster finding is done using K-means algorithm. In this we are creating cluster equal to our class (target parameter) label which will group the data with corresponding class value.

Running a decision tree algorithm on more structured data gives more accurate results. So, the training of decision tree algorithm will be better and somehow, we can also bypass the overfitting problem of the model. The goal of performing cluster analysis is to sort different objects or data points into groups in such a way that the degree of association between two objects if they belongto the same group, and lower if they belong to different groups. The allure of decision trees is in their ease and interpretability.

ISSN No:-2456-2165

➢ *Grid Search for hyperparameter tuning:*

Grid Search is an optimization tool used for hyper parameter tuning. We define the grid of parameters we want to search, and we choose the best combination of parameters for our data. Ourgoal of using grid search is to find a specific combination of parameters.

After checking all the grid points, we know which parameter combination is best for our prediction. In machine learning we compare different models with each other and try to find the model that works best. Grid search allows us to find the best parameters for a data set.

➢ *Classification Process:*

Classification is the process of classifying or classifying data points on thebasis of class or same group. This can be done on structured and unstructured data. In the classification process, a program learns from a given data or observation and then classifies the new observation.

IV. RESULTS

Before interpreting the results obtained for the experiment, there is some basic information that should be conveyed about the test procedure. The test had to be reliable. Therefore, all the results obtained (in terms of accuracy, standard deviation of accuracy, standard error of misclassification rate) were averaged over a thousand iterations.

However, the dataset had to be tested manually with the k-means and grid search algorithm. The dataset results are explained one by one in the previous sections. In this section, all the analyzed results are combined to draw some logical conclusion.

```
: #by hyperparameter tuning parameters
  MyHybridAlgo().doClassification()

  Optimal parameters combination: {'max_depth': 3, 'min_samples_leaf': 6, 'min_samples_split': 2}

  Training accuracy hybrid algorithm : 70.55555555555554 %
  Test accuracy of hybrid algorithm : 95.0 %
  <built-in method upper of str object at 0x000001D1B3F2D3A0>  24330.72 ms
```

V. CONCLUSIONS

This paper generally focuses on the improvement in the accuracy of decision tree algorithm using grid search hyper-parameter tuning and k-means clustering, we have taken the reference data of Customer Churn for the analysis of accuracy. Prediction of customer churn is very important in e-commerce and other businesses to maintain the competitiveness in the market. Decision based prediction using machine learning in customer relationship management to predict potential losses and formulate new marketing strategies and customer retention measures according to the results of prediction and this prediction will be efficient and accurate.

The primary and first objective was to study the effectiveness of customer segmentation and the longitudinal timeliness of customer purchase behavior and the predictive effect of models before and after customer segmentation according to multivariate variables. The experimental results proved that there is a significant improvement in the customer segmentation of each prediction index.

The results of this study also have some limitations. We used a data set consisting of data of about customers and the selection of data is limited to a certain extent. The outcome of customer segmentation greatly affects the forecasting performance of the model.

REFERENCES

[1]. Y. Achoura, Hamid R., Pourghasem 2020. How do machine learning techniques help in increasing accuracy of landslide susceptibility maps?. Geosc. Front. II. Pages 871-883. Xiancheng Xiahou and Yoshio Harada. B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. Journal of Theoretical and applied Electronic Commerce Research. Ritu Rani and Rahul Sahu, Nov. 2017 BOOK RECOMMENDATION USING K-MEAN CLUSTERING AND COLLABORATIVE FILTERING. IJESRT. , 6(11), 145-150

[2]. Ali Mustapha, S.A. Ali, N. Sulaiman, and N. Mustapha. Information Technology Journal 8 (8):1256-1262, 2009. ISSN 1812-5638. K-Means Clustering to Improve the Accuracy ofDecision Tree Response Classification. Academia. Asian Network for Scientific Information. Swati patel. ISBN-13: 978-6138838197. july2019 .K-means Clustering Algorithm: Implementation and Critical Analysis.