MACHINE LEARNING PROJECT REPORT

Github Repository Link : https://github.com/RachanaHS/Machine_learning

MEGHANA V NAYAK
PES1201701339
BTECH CSE V SEM
PES UNIVERSITY BANGALORE

SRUTHI MADINENI
PES1201700051
BTECH CSE V SEM
PES UNIVERSITY BANGALORE

RACHANA HS
PES1201701726
BTECH CSE V SEM
PES UNIVERSITY BANGAORE

***Abstract*— We have implemented k-Nearest Neighbor Algorithm to classify the stars and quessars given the catalog file.The data given was found to be imbalanced.So there was a need for balancing the data for better classification.We used K fold cross validation to test and train the model.Different K fold value gave different accuracy**

Keywords—Classification, majority class , minority class ,k-fold cross validation , unbalanced data,confusion matrix,individual class accuracy

I. INTRODUCTION

K Nearest Neighbor Algorithm is one of the machine learning model which is used for both classification and regression problems. This model takes into account the k - nearest neighbors based on the distance measures.The different distance measures are as follows-Euclidean Distance,Manhattan Distance etc.K value can be user defined or optimal value for a particular dataset.

II. PRE-PROCESSING

A. UNBALANCED DATA

Observing the data given and by plotting the simple bar chart we realized that the data given is highly imbalanced.There were around 3824 Quessars and 471 stars.Simply doing classification without balancing might give a better accuracy ,but the accuracy is only because of the majority class.Hence it will lead to overfitting.Hence we used a simple random upsampling to balance the data.

B. DROPPING OF COLUMNS

We dropped colums like pred,spectrometric_redshift and id to simplify our further classification process.

III. HOW DID WE ARRIVE AT WHICH MODEL TO USE

We were told to implement the model from scratch without the library functions.Hence we verified first by using the sklearn library function.We applied KNN,decision tree on our model.KNN gave a better accuracy compared to decision tree.Hence we went with KNN as our model.

A. k-Nearest neighbor model

- Calculate the Euclidean distance between the test and the train data.If the distance is small similar is the record and larger the difference dissimilar is the record.
- K-fold cross validation method to test and train the model.Depending on the K-fold value every iteration the test and the train data changes and each iteration gives a different accuracy.We have verified our model with different K-fold value and noted the accuracy.
- neighbors_get(train_set,test_row,n_neighbors) is the function defined to get the n_neighbor number of neighbors depending on the least Euclidean distance value.
- Each neighbour obtained from the training data has a class label which is used classify the test data.The predicted label of the test data depends on the label with maximum neighbors obtained.
- After computing the predicted value to compute the accuracy we check with the class label of the data.If the actual values is equal to the predicted values then the count of correctly classified data is incremented.Finally the accuracy is computed by comparing with the total actual values with which it is compared.

B. Equations

We have used only Euclidean distance for our kNN model.

Euclidean distance=sqrt(sum i to N (X1_i – X2_i)**2)

C. I. *ACCURACY*

CATALOG 1

SL.NO	ACCURACY CALCULATION		
	Fold value	***neighbors***	***Accuracy***
1.	4	4	96.633
2.	5	5	96.555
3.	5	10	93.950
4	10	5	96.891
5	10	10	93.950

CATALOG 3

SL.NO	ACCURACY CALCULATION		
	Fold value	***neighbors***	***Accuracy***
1.	4	4	95.973
2.	5	5	95.945
3.	5	10	92.256
4	10	5	96.492
5	10	10	92.919

The above tables shows the accuracy calculation for different Fold value and neighbors for 2 different catalogs.We have shown the comparison between catalog1 and catalog3. Similarly we have done for other catalogs also.

From the above table its clear that the accuracy is highest with value 96.891 and 96.492 with fold value being 10 and neighbors being 5 for catalog1 and catalog 3 respectively.

II.*INDIVIDUAL CLASS ACCURACY*

Catalog1 gave individual class0 accuracy as 100 % and class1 accuracy as 93.109% for k fold =5 and neighbors=5 with overall accuracy of 96.555%.

Catalog gave individual class0 accuracy as 97.986% and class1 accuracy of 85.83% for k fold=5 and neighbors =5 with overall accuracy of 91.913%.

IV. CONFUSION MATRIX

We have created a confusion matrix to describe the performance of our model. Here we have TP_count, TN_count, FN_count,FP_count.

For catalog1 we get precision=100 ,Recall=93.109 and

Specificity=100

REFERENCES

[1]Snehanshu Saha,Nithin Nagaraj,Archana Mathur,Rahul Yedida, “Evolution of novel activation functions in neural network training with applications to classification of exoplanets”

[2]https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/