BYTE PANACHE COMPETITION

FOUNDATIONS OF MODERN MACHINE LEARNING

OBJECTIVES

To build a support vector machine classifier to classify a pulsar star.

1. perform exploratory data analysis
2. compare the accuracy with various kernels such as rbf and linear
3. analyse model using confusion matrix to select the best model
4. evaluate the performance of the model using various classification metrics.

INTRODUCTION

Support Vector Machines (SVMs in short) are supervised machine learning algorithms that are used for classification and regression purposes. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

MATERIALS & METHODS

Predicting a pulsar star data set was used in the project. Pulsars are a rare type of Neutron star that produce radio emission detectable here on Earth. The data set contains 16,259 spurious examples caused by RFI/noise, and 1,639 real pulsar examples. The class labels used are 0 (negative) and 1 (positive).

Attribute Information:
Each candidate is described by 8 continuous variables, and a single class variable. The first four are simple statistics obtained from the integrated pulse profile. The remaining four variables are similarly obtained from the DM-SNR curve . These are summarised below:

- Mean of the integrated profile.
- Standard deviation of the integrated profile.
- Excess kurtosis of the integrated profile.
- Skewness of the integrated profile.
- Mean of the DM-SNR curve.
- Standard deviation of the DM-SNR curve.
- Excess kurtosis of the DM-SNR curve.
- Skewness of the DM-SNR curve.
- Class

The following module was used to build the model:

SVC imported from sklearn.svm

Signature: class sklearn.svm.SVC(*, C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache$_size$ = 200, $class_weight$ = $None$, $verbose$ = $False$, max_iter = -1, $decision_function_shape$ = $'ovr'$, $break_ties$ = $False$, $random_state$ = $None$)

RESULTS

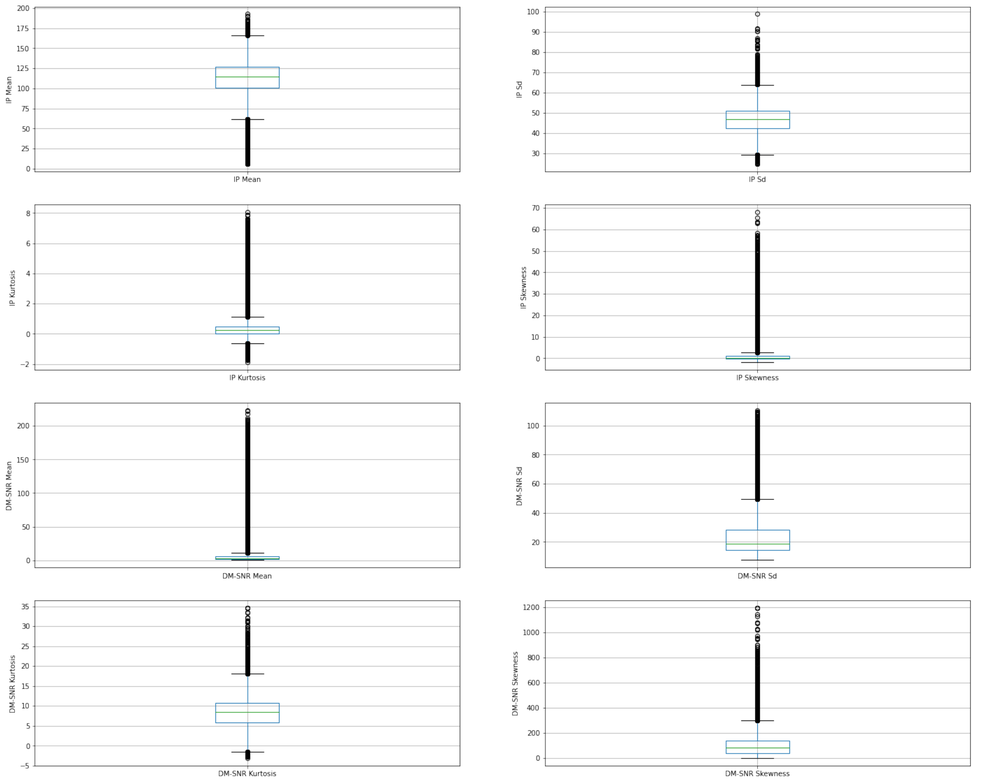

Figure 1: Outliers in the dataset

There are outliers in our dataset. So, as we increase the value of C to limit fewer outliers, the accuracy increased. This is true with different kinds of kernels.

CONCLUSION

We get maximum accuracy with rbf and linear kernel with C=100.0 and the accuracy is 0.9832. So, we can conclude that our model is doing a very good job in terms of predicting the class labels. But, this is not true. Here, we have an imbalanced dataset. Accuracy is an inadequate measure for quantifying predictive performance in the imbalanced dataset problem. So, we must explore confusion matrix that provide better guidance in selecting models.

REFERENCES

[1] Aurelien Geron. *Hands on Machine Learning with Scikit-Learn and Tensorflow*. Publisher, 2nd edition, 2017.

[2] https://en.wikipedia.org/wiki/support-vector$_machine$.

FUTURE RESEARCH

We would like to explore the confusion matrix that provide better guidance in selecting the modals. We would also like to build a K-Nearest Neighbor(KNN) classifier to classify the pulsar star and compare the accuracy to find out which method is suitable to better classify the pulsar star.

CONTACT INFORMATION

GROUP K

Pranamya P Bhat, Anushka Modi, Lalitha Evani, Jampala Srilasya, Sireesha Ramaiahgari