Comparative Study of ML Algorithms on Real Datasets
By: Safiia Mohammed

1 Introduction

The purpose of this project is to study different Machine Learning algorithms and to compare their performance to find the best learning method to analyze and build a predictive model. In this study, three Algorithms (Logistic Regression, k-nearest neighbors and Support Vector Machines (SVM))are analyzed depends on their accuracy , training time and testing time. The shuttle dataset is used to perform the comparative analysis.[1]

2 Dataset Preparation

2.1 The Data

The dataset consists of 9 attributes with numerical data and the number of instances are 58000. The first attribute is the time. The last one is categorical with 7 classes :1 Rad Flow, 2 Fpv Close, 3 Fpv Open, 4 High, 5 Bypass,6 Bpv Close and 7 Bpv Open . [1]

2.2 Data Preprocessing

Before training the models on the data, the null values are checked, by performing basic functions on data , no Null values are found. in each algorithm, train data is divided into train and test data.The main issue with this data is the unbalance size of the classes, the majority of the data is classified in class 1.

2.3 Dataset Visualization and Basic Statistical Analysis

Data has been visualized using Seaborn Library[2]. data is visualized using number of plots like: countplot which counts the data in each class, pairplot is used to summarize the relationships between dataset attributes, it is also shows that, the data is tend to be non linear. Another plot is heat map, heatmap illustrates the correlation between attributes.
Some Basic statistical Analysis are perform to understand more about the dataset. describe() function calculates the basic statistical functions like, mean, standard deviation, min , max ...etc. /In addition, count() is used to count the number of non-null observations. plots also included like, pie chart and kurtosis to highlight information about data distribution.

3 Models

Three basic Machine Learning algorithms have been tried, Logistic Regression, KNN and SVM. for each models of these algorithms, the train data is divided into data - which contains all attributes except the category attribute- and class attribute which contains 7 classes. Before training the dataset, data has been

standardization / normalization, this is also done for test dataset.
in addition, the train time , test time and accuracy are calculated.(see Table 1). also after training the data, k-fold cross validation is performed to each model[3]. finally the prediction result is visualized using heatmap and different matrices.

3.1 Logistic Regression

To solve the problem of imbalance of the data, the class-weight parameter is fine tuning by assigning weight to each class depend on its size. [4]

3.2 k-nearest neighbors

KNN parameters are fine tune by trying different values for the K and the method that we use to calculate the distance. the KNN algorithm usually is not affected by imbalanced classes, so no type of data resampling is applied.[5]

3.3 Support Vector Machines

In SVM, the RBF kernel is used and number of parameters are fine tuning. For example, gamma and cost parameters are running with cross validation with different values to get the best value, GridSearchCV is used to perform the selection of the best parameters. in other hand, class-weight parameter is set to 'balanced' to make balance between different classes' size.[6]

3.4 Different classifiers for this dataset

In the case that there are other choices of classifier, we can think about which when is less sensitive to imbalanced data, Accually KNN and SVM are good choices, but we can also think about Deep learning which has different techniques like data augmentation to add more data.

4 Results and Findings

This section summarize the comparative analysis of the three models. The results show that the KNN performs well in the accuracy and it has the fastest training and testing time. That usually because the data has low dimensions, otherwise the KNN can not perform well with high dimensional data. SVM has also very good accuracy however it is slow and takes long time to fit, train and test the dataset. Logistic Regression is fast however the accuracy is less comparing with KNN and SVM, the Non linearity of the data usually affect the accuracy of the logistic regression.

Model	Accuracy	Train Time	Test Time
LR	93.1%	0.708	0.003
KNN	99.8%	0.124	0.002
SVM	99.1%	7.273	0.011

Table 1: Summary of used Models' Accuracy and train and test times.

5 Conclusion and Limitations

5.1 Limitations

The main limitation for this project that i faced is the time, many enhancement can take place and many additions can be added to improve the comparative study. the re-sampling methods and how we can tackle the imbalance data can be more exploring and different methods can be applied for the future work.

5.2 Conclusion

In conclusion, we observed that different algorithms are compared according to their accuracy and execution time (train and test). The imbalance of data can affect the prediction, however finetuning and different resampling techniques can improve the prediction accuracy.

References

[1] Newman, D.J., Asuncion, A., 2007. UCI Machine Learning Repository. University of California, Irvine, Dept. of Information and Computer Sciences .

[2] https://www.analyticsvidhya.com/blog/2019/09/comprehensive-data-visualization-guide-seaborn-python/ [Access: 9, May , 2020]

[3] https://scikit-learn.org/stable/modules/generated/sklearn.model$_{s}election.KFold.html[Access$ $10, May,$ 2020]

[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear$_{m}odel.LogisticRegression.html$” $[Access :$ 11, $May,$ 2020]

[5] Website: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html [Access: 12, May , 2020]

[6] Website: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html [Access: 12, May , 2020]