Machine Learning Report

Chitra Periya
College of Engineering
Northeastern University
Toronto, ON
periya.c@northeastern.edu

Abstract

In this report, I explored the performance of the supervised learning algorithm Artificial Neural Networks on two interesting datasets. Breast Cancer Dataset and FedEx datasets. These dataset are different in terms of the data formats, the size, the number of features they contain,etc. This exploration for data analysis and building models using ANN helped in understanding the tasks involved in exploratory data analysis and how the classification models behave. The learning curve, model complexity and training time of each algorithm on both datasets have been explored and analyzed.

1 Datasets

I used the datasets - Dataset 1: Breast Cancer Dataset ; Dataset 2: FedEx Dataset

Table 1: The basic feature of both datasets.

	Data Set Characteristics	Attribute Characteristics	Associated Tasks	Number of Instances	Number of Attributes
Dataset 1	Multivariate	Real	Classification	6259	32
Dataset 2	Multivariate	Real	Classification	3604175	15

1.1 Data characteristics

Both the imported datasets were analysed to identify the number of rows and columns/features present, the data types associated with the columns, presence of missing values,etc.

Dataset 1 has numeric data and didn't have any missing values. The relationship between features were identified using the correlation matrix and outliers were identified using box and scatter plots and removed. The imbalanced data 'Diagnosis' that was present was identified and handled using SMOTE and transformed using LabelEncoding,
The skewness was identified and removed. The features used in the training and testing data were standardized using StandardScalar before passing to ANN for training and testing.

Dataset 2 is a multi-class dataset with categorical and numerical data and had values as NA in the Delivery status and 2 other fields. 2 types of analysis was done.One by removing the rows with NA values and the other by replacing with mean, median. Both had the same performance in terms of accuracy in ANN. As we are focusing on Delivery status as the target variable, the imbalance in the data was analysed and handled using SMOTE function.
To resolve the imbalance, data transformation was done. Label Encoding and one hot encoding were used to find how it varies.
Histograms and box plots were used to identify the spread and skewness. Heatmap was used to show the correlation between numerical variables.

Outliers were identified using the InterQuartile range and removed.
The data was split into training and test in 70-30 ratio and standardized using the StandardScalar function.

The histograms of classes from both datasets are shown in Figure 1. Class unbalance is evident in both datasets and must be accounted for in calculating the accuracy scoring function.

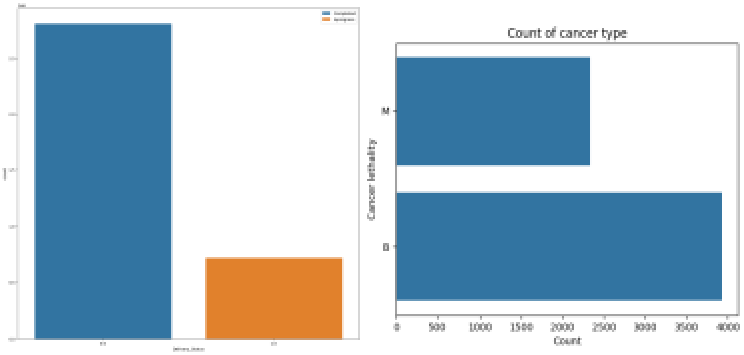

Figure 1:

Imbalance in Delivery status Imbalance in Cancer Lethality

1.2 Why are these interesting datasets?

Both datasets are interesting for their practical applications. The datasets are multivariate, they are not similar. Datset 1 involves binary classification while Dataset 2 has multi-class classification. 1t is interesting as it gives a better understanding on the preprocessing and data cleansing needed before an algorithm is applied to it. Ihe helps understand relationship between the features and how it impacts the classification. It helps in understanding how ANN trains and classifies the test data and creates a confusion matrix giving the actual results and false positive and false negative counts in the data tested. We can get a better idea on which algorithm suits the dataset better.

Though the datasets are used for classification, they can be used for regression as well.

2 Artificial Neural Network

I chose to train the data by controlling the epochs and batch size. The loss and accuracy seemed to change based on the epochs and batch size.

When the epoch for set more than 26, the accuracy reduced and then again increased. So, set the epoch as 26 which returned an accuracy of 96% for test data.

For Breast Cancer dataset:

Accuracy: 0.9824561403508771

Recall: 0.9767441860465116

Precision: 0.9767441860465116

f1: 0.9767441860465116

Confusion matrix:
Positive cases are 777 for M and 749 for B. The False positive and Negative cases are: 12 and 33.

For FedEx dataset, we get an 99% accuracy for the same epoch of 26. When the batch size was set from 50000 to 30000, the false negatives and false positives. But, when the batch size was reduced to 512, the false positive and false negatives increased.

The performance of FedEx dataset test data prediction and confusion matrix:

Accuracy: 0.9987876768917587

Recall: 0.9950732974577844

Precision: 0.9989800528137036 f1: 0.9970228480722884

confusion Matrix: [[840870   219]

[1062 214498]]

score is: 0.9987876768917587

Based on the confusion matrix we find the positive cases for Completed and Inprogress as: 840870 and 214498 . The false negative and false positive as : 219 and 1062. This indicates that 219 scenarios were wrongly identified as negative and 1062 cases were wrongly identified as positive.

7 Conclusions

In this report, I have analyzed the performance of Artificial Neural Network supervised learning algorithm on two interesting datasets.Based on the type of dataset, preprocessing, removal of outliers, and transformations that are done differ, and they give different accuracies on training and testing indicating that ANN is better on which type of dataset. We have to look into the features used to make sure that it isn't the reason for precision of more than 95%.

Acknowledgments

The learning code is adapted from:

https://www.kaggle.com/code/itsmohammadshahid/eda-breast-cancer-prediction/notebook

https://www.kaggle.com/code/iftikar24/fedex-project-one

and references therein.

References

[1] Detect and Remove Outliers using Python | Aman Kharwal (thecleverprogrammer.com)

[2] https://www.geeksforgeeks.org/steps-for-mastering-exploratory-data-analysis-eda-steps/