CS4044 Pattern Recognition Assignment

Submitted by

Abhiram Haridas B150177CS
Abin P B150453CS
Arun Joseph B150102CS
Nandu B C B150144CS

Department of Computer Science and Engineering
National Institute of Technology Calicut
Calicut, Kerala, India - 673 601

March 29, 2019

Bank Marketing

Abhiram Haridas Abin P Arun Joseph Nandu B C

The assignment aimed at analyzing the data related with direct marketing campaigns of a Portuguese banking institution to predict whether a client would subscribe a term deposit.

The Bank Marketing data was available in two datasets:- one with 45211 tuples, 16 attributes, 5289 positive samples and 39922 negative samples, and an additional dataset with 41188 tuples, 19 attributes, 4640 positive samples and 36548 negative samples. Both the datasets are skewed, as they contain more negative samples than positive samples. Also, some attributes had unknown values like, *job*, *education* and *contact*. A large number of tuples have *pdays* as -1 or 999 and *poutcome* as unknown or nonexistent, indicating clients that have not been contacted in the past.

Further actions which were performed on both the datasets includes adding dummy attributes for categorical features. The target output, *y*, was factorized to obtain integer values. The *pdays* attribute values were modified by replacing -1 with 999. The attribute *duration* was dropped entirely, since it has direct correlation with the final output. Unknown values are taken as a distinct class for each category. Then, PCA was performed on both the datasets.

After performing Principal Component Analysis(PCA), it was found that in the first dataset, 80% of the variance was captured by 2 attributes, while 95% of the variance was captured by 5 attributes. In the additional dataset, 80% of the variance was captured by 6 attributes, while 95% of the variance was captured by 10 attributes. Before training, oversampling was performed to compensate for the skewed dataset by increasing the size of the positive class 16 times. Finally, bias was added to the dataset. The dataset was split into 70:30 as training and test sets.

The above processed dataset was trained using different models. This includes Logistic Regression with polynomial features with degree 2, *lbfgs* optimiser and 500 iterations, which gave an accuracy of 65% and F1 score of 45%. Using Support Vector Machines(SVM) with polynomial features with degree 3, scaled kernel coefficient and a penalty parameter of 1, an accuracy of 60% and F1 score of 45% was attained. Using a Random Forest Classifier with 5 trees, an accuracy of 73% and F1 score of 15% was attained. For Artificial Neural Network, a Multi-layer Perceptron Classifier(MLPC) with 3 hidden layers, Rectified Linear Unit activation function, regularization parameter of 0.005 and a learning rate of 0.001 was used, which gave an accuracy of 65% and F1 score of 35%.

It is observed that SVM classifier gave a slightly better result when compared to Logistic Regression. But the F1 score for both the models were only about 45%. The low F1 score might be due to lack of good attributes and not having enough data for training the positive class.

The code is available at: https://github.com/arun-Joseph/Bank-Marketing.