Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (SBS)

Hadeel Ahmad, Bassam Kasasbeh, Balqees Aldabaybah, Enas Rawashdeh, Deval Shaileshkumar Mali

October 23, 2022

1 The main idea

Credit card fraud is a growing problem nowadays and it has escalated during COVID-19 due to the authorities in many countries requiring people to use cashless transactions. Every year, billions of Euros are lost due to credit card fraud transactions, therefore, fraud detection systems are essential for financial institutions. The authors mainly focus on processing unbalanced data by using an under-sampling technique to get more accurate and better results with different machine learning algorithms. They propose a framework that is based on clustering the dataset using fuzzy C-means and selecting similar fraud and normal instances that have the same features, which guarantees the integrity between the data features.

2 The methodology

This case requires to develop a customer segmentation to define marketing strategy. The sample dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables. The dataset contains real transactions collected from European cardholders in September 2013.

In the dataset, there are 31 features in total. According to the relation between the features and to ensure that SBS is performed on features that do really affect the classification process, some features have been removed from the model. 'Time', 'Class' and 'Amount' have been removed because they do not want these features to affect the classification process. The number of features is reduced to 28 features and the dataset is ready for the clustering phase.

After the clustering phase is done, in which similar data instances are grouped in one cluster. Then they have get the balanced dataset from the clusters, to be used as an input for the different classification models. This step describes in detail the proposed SBS technique that is based on instances similarity selection. In this paper, the data instances are distributed in three portions: 50:50, 34:66, 25:75 as (fraud: normal). Once, they get three balanced portions of the dataset with different ratios (50:50, 34:66, 25:75). The balanced dataset is divided into two portions: 70% for the training set, and 30% for the testing set, and then the ML algorithms namely: ANN, LR, KNN, NB are used to train the models.

3 The results

The experimental results have been implemented using Python. There has been four Machine Learning algorithms are tested on the original dataset, and on the balanced datasets obtained by our SBS technique. To benchmark SBS performance, researchers have compared the obtained results among RUS and CBUFN. To ensure the fairness of comparison, the dataset has been distributed in three proportions, which are taken as follows: Class A: (50% fraud, 50% normal), Class B:(34% fraud, 66% normal), and Class C:(25% fraud, 75% normal) (to ease the comparison with [6]). In this paper, the performance measure of our method (SBS) is investigated on six evaluations metric including: Accuracy (ACC), Precision (P), F-Measure (F), Sensitivity (SEN), Specificity (SPE), and Area Under the ROC curve (AUC). The performance measure of our method (SBS) is investigated on six evaluations metric including: Accuracy (ACC), Precision (P), F-Measure (F), Sensitivity (SEN), Specificity (SPE), and Area Under the ROC curve (AUC).

Then they have investigated the problem of imbalanced credit card dataset, the study work was carried out with the purpose of finding the best under-sampling technique that gets rid of the RUS problem and guarantees better results. They propose a framework (SBS) to solve the problem of imbalanced class distribution by using Fuzzy C-means. The performance of the experiment is compared with other methods to emphasize the power of SBS technique and to prove it's superiority. SBS aims to solve the problem of RUS and guarantees the similarity and integrity of the instances' features. In the future work, they will continue to research on the basis of enhancing the proposed framework to achieve better and optimal results, which can give better performance in dealing with imbalanced credit card dataset.

4 Recommendation

They can also use multi-classifier framework and an ensemble model with multiple machine learning classification algorithms can be designed, in which the Behavior-Knowledge Space (BKS) is leveraged to combine the predictions from multiple classifiers.