Applying KCV-SMOTE and KFS to SVM in credit card fraud prediction

STATS 771 - Project Proposal

Pao Zhu Vivian Hsu (Student Number: 400547994)

2023-10-31

Introduction

Credit cards are an electronic form of payment that are popular for the convenience and protection they offer on everyday purchases. Due to its popularity, detecting credit card fraud is crucial for financial companies to maintain customer satisfaction and minimize losses. Machine learning techniques have become the predominant method to predict credit card fraud over the years and companies are constantly looking for ways to improve their predictions. In this paper, we propose a study that uses the support vector machine (SVM) algorithm with Kang and Zhang's method for online credit anti-fraud, a method that combines the K-fold cross-validation and synthetic minority oversampling technique (KCV-SMOTE) with key feature scanning (KFS) on a classifier (2022).

Literature Review

To this date, a variety of machine learning techniques are being used to predict credit card fraud. This section provides a summary of some recent techniques found in literature.

Itoo et al. (Itoo et al., 2021) built a logistic regression, K-nearest neighbour, and Naïve Bayes model to predict credit card fraud. Their logistic regression model was the strongest with an accuracy of 95%.

Kang and Zhang (2022) introduced the KCV-SMOTE and KFS method to improve poor model performance caused by the imbalance of fraudulent and non-fraudulent data. They combined this method with a linear regression approach using publicly available data from online sources. Their final model

No additional work has been done to explore the KCV-SMOTE and KFS method on other classifiers since their paper was published. This, our proposed study will apply their method on a different classifier, namely SVM.

Data and Methods

The proposed study will use a public data set on credit card fraud collected by Worldline and the Machine Learning Group of ULB, the Université Libre de Bruxelles (Worldline & Machine Learning Group - ULB, 2017). This data set is frequently used in credit card fraud research including the work done by Kang and Zhang Itoo et al. (2021).

We will build two models in this study: an SVM model that utilizes Kang and Zhang's KCV-SMOTE and KFS methodology (2022) and a basic SVM model. The KCV-SMOTE and KFS methodology removes features that negatively influence classification accuracy and speed (Kang & Zhang, 2022). Thus, we have decided to study this method on an SVM classifier since SVM is a strong classifier for classification problems and performs well for data with less features (Sulaiman et al., 2022). Here are details on our study's steps based on Kang and Zhang's work (2022):

1) Divide the data into a training set and a test set.
2) Use KCV-SMOTE on the training set to obtain a synthetic training set.
3) Perform key feature scanning on the synthetic training set to obtain a sub-training set.
4) Train the SVM classifier for each sub-training set separately and get a set of AUROC values. Sort the AUROC models and select the sub-training sets with the highest scores.
5) Train the intersection of the sub-training sets from the previous step to obtain the best model, and input the test set into the best model to obtain the final prediction result.

Once the models are built, we will compute their accuracy, precision, recall, F1-score, and area under the precision-recall curve (AUROC). These metrics will be used to compare models within the study and with Kang and Zhang's logistic regression models (2022). All analyses will be performed using Python.

Expected Results

This study will provide greater insight into the effectiveness of Kang and Zhang's method (2022) with an SVM classifier. After applying their method, we expect the final model to perform better than a basic SVM and logistic regression model. Results of our study will inform further research into the method and guide the industry in building more accurate credit card fraud models.

Timelines

There are two main components to this project, a paper and a presentation. By November 30, 2023, we will finish data collection, preliminary visualization, and data splitting. By December 31, 2023, we will complete the modelling. By January 31, 2024, we will compute performance metrics and compare the results between models. By February 28, 2024, we will complete a draft of the paper.

By March 31, 2024, we will finish the presentation slides. By April 20, 2024, we will finalize the paper and presentation. Finally, during the last week of April, the presentation will occur.

References

Itoo, F., Meenakshi, & Singh, S. (2021). Comparison and analysis of logistic regression, naïve bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, *13*, 1503–1511. https://doi.org/10.1007/s41870-020-00430-y

Kang, H., & Zhang, H. (2022). A new improved method for online credit anti-fraud. *Automatic Control and Computer Sciences*, *56*, 347–355. https://doi.org/10.3103/S0146411622040046

Sulaiman, R. B., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. *Human Centric Intelligent Systems*, *2*, 55–68. https://doi.org/10.1007/s44230-022-00004-0

Worldline, & Machine Learning Group - ULB. (2017). *Credit card fraud detection*. https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data/data