

Summary 12

Shaun Pritchard

Florida Atlantic University

CAP 6778

October -28 0-2021

M. Khoshgoftaar

## **Feature Extraction for Class Imbalance Using a Convolutional Autoencoder and Data Sampling**

The purpose of this study was to address the problem of high-dimensional and highly imbalanced data. The researchers applied feature extraction techniques to credit card fraud data focusing on binary class imbalances. There were 248,807 instances and 30 independent variables in the fraud detection dataset. The binary dataset's dependent variable was set one for fraudulent transactions and two for non-fraudulent transactions.

Data mining involves extracting features from original data to create new subsets of more accurate or richer features for analysis. Feature extraction techniques used in this experiment were Principal Component Analysis (PCA) and Convolutional Autoencoder (CAE). By implementing linear transformations, PCA is used to reduce the dimensionality of large sets of data by transforming a large set of variables into a smaller one that still contains most of the information. In CAE models, nonlinear functions are calculated using the convolution neural network operator, which exploits the observation to encode and reconstruct the input.

They also evaluate the effects of using these two feature extraction methods while using Synthetic Minority Oversampling Technique (SMOTE) data sampling method and ANalysis Of VAriance (ANOVA) metrics To understand the statistical significance of the inferences . The researchers implemented feature extraction and data sampling techniques using the random forest classifier for the inferences of all experiments in this study.

To conduct this experiment, the researchers used SMOTE oversampling and feature extraction in multiple variations with and without it to augment the data. The researchers evaluated the use of SMOTE only, feature extraction only, both SMOTE and feature extraction,

and a baseline without using SMOTE or feature extraction. The results showed that CAE used with SMOTE provided the best results of the experiment and highest performance measure for class imbalance.