15th International Conference on Innovative Computing, Information and Control (ICICIC2021)
Online, September 15–16, 2021

A Mixture-Model-Based Framework for Fraud Detection

Chang-Jung Wu and Chuan-Wei Ting*

Information and Communications Research Laboratories, Industrial Technology Research Institute
No. 195, Sec. 4, Chung-Hsing Rd., Chutung, Hsinchu 31040, Taiwan
*Corresponding author: cwting@itri.org.tw

Introduction

In a fraud detection system (FDS), the class distribution of legitimate and fraudulent transactions is usually unbalanced, which may cause a problem for an algorithm's performance.

This paper presents a new framework for FDS based on the concept of the mixture model to handle imbalanced data processing and the classification algorithm.

Methodologies

In this paper, the Gaussian mixture model (GMM) is used for modeling original data, whether it is fraudulent (minority) or legitimate (majority) data, and then generate data for over-sampling or under-sampling. Meanwhile, the GMM is also applied to being a classification model for determining fraudulent and legitimate transactions.

To penalize the overfitted model, we used the Bayesian information criterion (BIC), a model complexity selection criterion, to estimate the number of mixture component ***M*** in GMM.

Mathematical Formulas

- For a d-dimensional feature y, the mixture of Gaussian densities for class c is defined as:

$$p_{GMM}(y|\lambda_c) = \sum_{j=1}^{M} w_{c,j} \cdot (2\pi)^{-\frac{d}{2}} |\Sigma_{c,j}|^{\frac{1}{2}} \times exp\left\{-\frac{1}{2}(y-\mu_{c,j})^T \Sigma_{c,j}^{-1}(y-\mu_{c,j})\right\}$$

Model Parameter Set | Mixture Weight | Covariance Matrices | Mean Vector

- Maximum Likelihood (ML) estimation:

$$\lambda_{ML} = arg\max_{\lambda} p(Y|\lambda)$$

- Transaction classification can also be performed through finding the most likely class $\hat{c}$ from input transactional data $\tilde{y}$ using ML estimates of each class c, i.e. $\hat{c} = arg\max_{c} p(\tilde{y}|\lambda_c)$

$$BIC\left(\hat{\lambda}(M), M\right) = -\log p\left(Y|\hat{\lambda}(M)\right) + \frac{\rho}{2}\#(M)\log N$$

Dataset Information

- **Name:** BankSim (A synthetic dataset of transactional data.)
- **Records:** 594,643 records in total. Only 7,200 **(1.21%)** records are fraudulent transactions.
- **Time Coverage:** 6 months of transactions.
- **Columns:** 8 columns. (2 were selected for the experiment.)

Figures

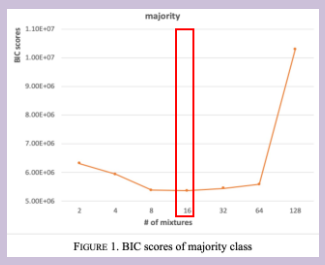

FIGURE 1. BIC scores of majority class

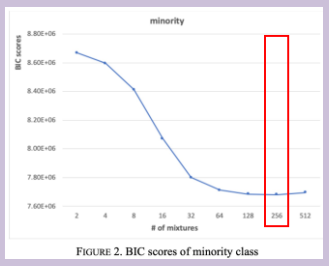

FIGURE 2. BIC scores of minority class

Table

TABLE 1. Fraud detection results

Classifier	Preprocessing	Accuracy	Precision	Recall
LR	- -	99.543%	89.061%	70.944%
	RUS	95.875%	22.545%	98.806%
	GMMUS	96.148%	23.779%	98.861%
	ROS	96.536%	25.755%	98.819%
	SMOTE	98.148%	39.262%	96.597%
	GMMOS	96.500%	25.579%	98.986%
NB	- -	93.908%	16.578%	99.944%
	RUS	95.836%	22.386%	98.764%
	GMMUS	95.883%	22.867%	98.500%
	ROS	95.872%	22.549%	98.875%
	SMOTE	95.627%	21.545%	98.792%
	GMMOS	95.961%	23.657%	98.597%
GMM	- -	96.374%	25.341%	97.542%
	RUS	97.375%	31.047%	91.056%
	GMMUS	97.371%	31.922%	93.875%
	ROS	96.555%	26.271%	96.208%
	SMOTE	98.659%	48.005%	83.722%
	GMMOS	98.668%	48.178%	84.056%

Conclusions

We have presented a GMM framework for both imbalanced data processing and classification. Importantly, we determined the number of mixture components via the BIC model selection criterion. Experiments on the BankSim dataset confirmed the superior performance of the GMM-based sampling method for imbalanced data processing and GMM classifier.