Study on the Prediction of Imbalanced Bank Customer Churn Based on Generative Adversarial Network

Bo Li* and Jiuzuo Xie

College of Computer Science and Engineering, Chongqing University of Technology, Chongqing, China

*Corresponding author email: libo@cqut.edu.cn

Abstract. The imbalanced commercial bank customer data will lead to the unpredictability of the minority class. Therefore, this paper proposes an imbalanced data method based on generative adversarial network to deal the problem of poor prediction performance of traditional classifiers for minority class. This paper method is based on the generative adversarial network to generate minority class samples to improve imbalanced data. Finally, the classifier is used to train the balanced data to improve the prediction performance of minority class. In this experiment, the data of a commercial bank customer were measured with indicators such as F1, Precision, and compared with traditional data sampling methods such as SMOTE, BSSMOTE. This method is feasible and applicable to the classification of imbalanced data of banks by observing the experimental results, which has better application value.

1. Introduction

Imbalanced data classification is an extremely difficult field in binary classification problems [1]. The most valuable problem in this kind of problem is the identification of minority class sample. For example, customers churn in the field of commercial Banks, because the cost to retain existing customers is less than the cost to develop new customers [2]. If you cannot accurately know those churn customers, it will cause irreparable losses to the enterprise. But as a result of the imbalanced data classification features, there are a lot of the traditional classification method for minority prediction effect is very poor, such as Bayes and Support vector machine (SVM), most of these methods are designed in order to maximize the majority class sample accuracy, the prediction accuracy of majority class is higher than minority class [3]. From this point of view, the research and exploration of classification strategy is of great significance in dealing with imbalanced classification. For solving this problem, many scholars at home and abroad have put forward many methods in recent years [3]. One of the common methods is to process imbalanced data. The key idea is to reduce the difference between imbalanced classes by preprocessing the training set. In other words, by changing the distribution of minority class and majority class in the training set, a balance between different classes can be obtained. Data generation is a method to increase the proportion of minority class by copying and synthesizing minority samples. SMOTE[4] is a classic generation method. SMOTE is to generate minority class samples along the line segment, many variations of this algorithm have also been proposed such as Majumder, Bennin[5] proposed ADASYN and BSMOTE. Other methods focus on classification algorithms. Bagging and Boosting in Ensemble learning is a common method to improve the performance of classification. In addition, cost sensitive learning is also proved to be an better strategy to deal the imbalanced problem.
Generative Adversarial Network (GAN) is different from the traditional artificial neural network. It learns through a competitive process of two networks [8]. Generator (G) uses random variables as input learning generates false data that can fool the discriminator, and discriminator (D) tries to distinguish the real data from the generated data. If the training goes well, G will generate high imitation data that is very similar to the real data and fool D, since its inception, GAN has become a method widely used in many areas of artificial intelligence, including image segmentation and speech generation, etc. [9].

This paper proposes an imbalanced classification strategy based on GAN. The model is applied to the generation of minority samples. Machine learning is used to classify the original data and the processed data, and the validity of the model is verified by contrast experiments.

2. Materials and Related Methods

2.1. Common Imbalanced Data Classification Methods

Under normal circumstances, there are two methods to deal with unbalanced classification problem. One is based on data and the other is algorithm [3]. RUS removed random samples from majority, FUS removed majority that existed on the boundary between two classes. SMOTE is a more common method of data generation.

SMOTE generates minority classes by taking a sample of each minority class and giving a synthetic example along the line segment connecting the nearest neighbor of all selected N minority classes [12]. According to the required sample size, randomly select closest samples from N nearest neighbors. This process is described in SMOTE as shown in figure 1. First, for each observation x of the minority class, its N nearest neighbors are identified, as shown in the square sample in the figure. Then randomly select N neighbors. Finally, minority class sample is copied along the line that connects the real sample n with its nearby samples.

![Figure 1. Smote algorithm diagram (Square is minority class, round is majority class)](image)

2.2. Generative Adversarial Network

In 2014, Goodfellow and other experts proposed Generative Adversarial Network (GAN)[6]. G(z) is the simulated real sample generated by D(x) where z is a random noise and x is a real sample. Optimization of GAN can be regarded as how to minimize and maximize [8]. The loss function can be expressed as:

\[
\min_G \max_D V(D, G) = E_{X \sim p(x)}[\log D(x)] + E_{Z \sim p(z)}[\log(1 - D(G(z)))]
\]  

(1)

\[p(x)\] is the true sample distribution. \[p(z)\] is the noise distribution, \(E(*)\) is the expectation. In (1), GAN model includes both a discriminator optimization process (2) and a generator optimization process (3).

\[
\max_D V(D, G) = E_{X \sim p(x)}[\log D(x)] + E_{Z \sim p(z)}[\log(1 - D(G(z)))]
\]  

(2)

\[
\min_G V(D, G) = E_{Z \sim p(z)}[\log(1 - D(G(z)))]
\]  

(3)

The Generative Adversarial Network is a kind of network structure which can obtain new samples by training samples [8]. GAN's main task is to estimate the sample distribution of the training set, and
then use the sample distribution to generate another sample similar to the training set. Figure 2 below is a flowchart:

Figure 2. The GAN flowchart

At present, this model is widely used in image vision, anomaly detection and Credit card fraud. Compared with the traditional generation model [9], the GAN model doesn’t need to generate synthetic data based on real data to approximate real data [8].

3. Construction of Data Generation Model Based on GAN

3.1. Feature Engineering

Feature Engineering can screen out attributes with better effects for training, so that the training effect of the model can be improved. Feature processing, feature analysis and feature selection are the three main steps of feature engineering. Feature selection is the key process. Random Forest is an integrated learning algorithm. It has an excellent performance in classification problems with low overhead. Feature selection based on RF can not only solve the over-fitting problem caused by too many features, but also improve the training speed and model effect by improving features.

The idea of feature selection based on RF is looking at the contribution of individual variables on each tree. Gini index (GI) and out-of-pocket data error rate are usually used to measure. This paper used GI for evaluation. It uses FIS to represent feature scores. If there are n features, $X_1, X_2, X_3, \ldots, X_n$, To calculate the GI index $G_{IS}^{(GI)}$ of $X_i$, The following formula (4) is the calculation formula of GI:

$$G_{IS}^{(GI)} = G_{ni} - G_{lj} - G_{ij}$$

In formula (4), K means there are K categories, $p_{nk}$ refers to the ratio of category k in node n. In an objective sense, It means that two samples are randomly selected from any node n, and they mark different probabilities. The change of GI index before and after the branch of node n is:

$$G_{IS}^{(GI)} = G_{ni} - G_{lj} - G_{ij}$$

Finally, all the obtained importance scores were centralized and normalized.
3.2. GAN Model Network Structure in this Paper
Before training with GAN, first use random forest for feature selection, then use PCA algorithm for dimensionality reduction, and retain 25 features for training. The generated network in this paper uses four hidden layers. The input of the generated network is 200-dimensional random noise, and the number of hidden layer nodes is 64, 128, 256, and 512. The activation functions are all LeakyReLU. LeakyReLU is a special variant of ReLU function [9]. Although the output of half the range of the ReLU domain is 0, it can make the network structure more sparse and alleviate overfitting [9]. However, the ReLU function has a "fragile" flaw during training, that is, if the weight is less than 0 when the network is training for the first time, the subsequent training will always be 0. LeakyReLU function can solve this phenomenon. LeakyReLU will still have a small gradient of non-zero value output when the neuron is inactive, thus avoiding the possible "evolution" of the neuron. It can be seen from the formulas (9) and (10) of both ReLU and LeakyReLU. ReLU refreshes negative values to zero, and LeakyReLU refreshes negative values to non-zero slopes. a_i is a fixed value in the interval (1, +∞). Based on LeakyReLU’s angle, if the slope is small, it is very different from ReLU, otherwise, the effect is better. Finally, the number of output layer nodes of the generated network is 25, and the activation function of output layer is tanh.

\[
\text{ReLU}(x) = \begin{cases} x_i & \text{if } x \geq 0 \\ \frac{x_i}{a_i} & \text{if } x < 0 \end{cases}
\]

(9)

\[
\text{LeakyReLU}(x_i) = \begin{cases} x_i & \text{if } x_i \geq 0 \\ \frac{x_i}{a_i} & \text{if } x_i < 0 \end{cases}
\]

(10)

The discriminant network also uses four hidden layers. Part of the input data comes from real data, and the rest comes from the generator's generated data. The number of discriminator input layer nodes is 25, the number of nodes in the 4 hidden layers is 512, 256, 128, 64, respectively, and the activation function is the LeakyReLU function. The number of nodes in the output layer is 2, the activation function is Sigmoid, and the loss function is binary cross entropy. The GAN model is shown in Figure 3.

4. Experiment and Result

4.1. Evaluation Metric
Accuracy is often used as the criterion to evaluate the classification effect in classification problems, but for imbalanced data problems, it is not enough to use Accuracy as the measurement index. Because in unbalanced data, there will be most classes with very high prediction accuracy, while a few classes with very low prediction accuracy. In this case, F-mean and g-mean as evaluation will be used in this paper. These two indexes can not only objectively evaluate the performance of the data but also give consideration to the classification accuracy of various categories. The confusion matrix is shown in Table 1. Where FN and FP refer to the number of positive and negative samples wrongly classified, while TP and TN refer to the number of positive and negative samples correctly classified.
Table 1. Confusion matrix

|                  | Predicted Minority class | Predicted Majority class |
|------------------|--------------------------|--------------------------|
| Real Minority class | TP                       | FN                       |
| Real Majority class | FP                       | TN                       |

According to the table 1, the following evaluation indexes are obtained, where the precision and recall formulas are shown in (11) and (12).

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (11)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (12)
\]

Under normal circumstances, the relative importance of Precision and Recall is tempered by \( \beta \), usually 1. If the Precision and Recall of a few classes are both relatively large, the f-mean value will be larger, which can accurately reflect the classification performance of a few classes, as shown in the following formula (13).

\[
\text{F-mean} = \frac{(1+\beta^2) \times \text{Precision}}{\beta^2 \times \text{Precision} + \text{Recall}} \quad (13)
\]

4.2. Dataset and Experiment Design

The dataset is the customer data set of a commercial bank. The data imbalance ratio reaches 20:1, and the number of samples is 80,000. Data pre-processing is also required before training, but this paper does not repeat detailed elaboration. The number of GAN iterations was set as 5000, with 30% of the test data and 70% of the train data. In order to verify the effect of the model, this experiment will be compared with the usual methods of unbalanced treatment of data, such as RUS and SMOTE. The experimental process is shown in figure 4.

![Figure 4. Experimental flow chart](image)

4.3. Experiment Result and Analysis

This experiment uses decision tree (DT) and random forest (RF) to verify the problem of unbalanced classification. Experimental comparison of GAN with BS, AS, ROS, SMO, RUS. The experimental results are shown in Figures 5 and 6. In Figures 5 and 6, we use random forest and DT algorithm to test the overall accuracy of each sampling algorithm and GAN on the test set. It can be seen from Figures 5 and 6 that the overall accuracy of the data processed by the algorithm in this paper is superior to other algorithms.
The above experiments prove that the proposed algorithm improves the overall accuracy of the sample, but it is still not comprehensive enough to reflect the improvement of the imbalanced data classification effect. Therefore, we also need to use the Precision and F-mean metrics mentioned above. And compared with various algorithms, the results are shown in Table 2 and Table 3, respectively.

**Table 2.** Precision value comparison

| Algorithm | None | AS  | BS  | ROS | RUS | SMO | GAN |
|-----------|------|-----|-----|-----|-----|-----|-----|
| RF        | 0.27 | 0.21| 0.21| 0.24| 0.12| 0.21| 0.97|
| DT        | 0.11 | 0.2 | 0.21| 0.23| 0.11| 0.2 | 0.99|

**Table 3.** F-mean value comparison

| Algorithm | None | AS  | BS  | ROS | RUS | SMO | GAN |
|-----------|------|-----|-----|-----|-----|-----|-----|
| RF        | 0.09 | 0.35| 0.35| 0.39| 0.2 | 0.35| 0.66|
| DT        | 0.1  | 0.34| 0.34| 0.37| 0.19| 0.33| 0.76|

It can be seen from Table 2: In the indicator Precision, the combination effect of DT, RF and GAN in this paper works best, and the results obtained by GAN in this paper are the highest compared to AS, BS, ROS and other methods. Table 2 is the experimental analysis of the accurate performance of predicting minority class samples, but this does not reflect the classification of the overall minority class samples by GAN in this paper. Therefore, we also need to use the F1 index to evaluate the overall effect evaluation of the prediction of the small sample. The experimental results are shown in Table 3 above. The F1 index of the algorithm in this paper on the DT and RF classifiers has obtained better results. In summary, the method in this paper has achieved better results than ROS, SMOTE and other algorithms in accuracy, F1, and Precision.
5. Conclusion
In the field of classification problems, the problem of imbalanced data classification is particularly important. This paper has implemented the application of generative adversarial network commonly used in the field of image processing to the problem of bank customer churn. In this paper, a model of sample generation based on the GAN is proposed, which solves the problem of minority generation and classification in the bank customer churn. In addition to the accuracy is used, the performance of this method is evaluated by f-mean, and compared with ROS and SMOTE. The experimental results confirmed that the strategy avoids the blindness of the synthetic sample in SMOTE, improve the quality of the synthetic sample, and make the data generation more targeted. Furthermore, the classification results have better accuracy, f-mean values. It is feasible and applicable to apply it to the imbalanced classification of bank customers.

Acknowledgments
This work is supported by the Natural Science Foundation Project of Chongqing Science and Technology Commission. (Grant No. cstc2015jcyjBX0113), Chongqing higher education teaching reform research project(173155), Chongqing graduate education and teaching reform research project(yjg183114).

References
[1] Lincy Mathews, Seetha hari. Learning from imbalanced data[M]// Encyclopedia of Information Science and Technology. 2018.
[2] Ding jun-mei, liu gui-quan, li hui. Application of improved random forest algorithm in customer loss prediction in telecom industry [J]. Pattern recognition and artificial intelligence,2015,28(11):1041-1049.
[3] Li yanxia, chai yi, hu youqiang, Yin hongpeng. Summary of unbalanced data classification methods [J]. Control and decision,2019,34(04):673-688.
[4] Georgios D, Fernando B, Felix L. Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-Means and SMOTE[J]. Information Sciences, 2018:S0020025518304997-.
[5] Anima Majumder, Laxmidher Behera, Venkatesh K. Subramanian. GMR based pain intensity recognition using imbalanced data handling techniques[C]// International Conference on Signal & Information Processing. IEEE, 2017.
[6] Goodfellow I J, Pouget-Abadie J, Mirza M,et al.Generative adversarial nets[C]/Proc of International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 2672-2680.
[7] Lu qian-wen, tao qingchuan, zhao ya-jing, liu mi-xiao. Simplification of cartoon draft drawing based on generative antagonism network [J]. Chinese journal of automation,2018,44(05):840-854.
[8] Cheng xianyi, xie lu, zhu jianxin, hu bin, shi shuan. GAN overview of generative antagonist network [J]. Computer science,2019,46(03):74-81.
[9] A. Sethia, R. Patel and P. Raut, "Data Augmentation using Generative models for Credit Card Fraud Detection," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-6.
[10] Gong J, Kim H. RHSBoost: Improving classification performance in imbalance data[J]. Computational Statistics & Data Analysis, 2017, 111:1-13.
[11] Braytee, Ali, W. Liu, and P. Kennedy. "A Cost-Sensitive Learning Strategy for Feature Extraction from Imbalanced Data." International Conference on Neural Information Processing Springer International Publishing. 2016.
[12] Chawla, N. V, et al. "SMOTE: Synthetic Minority Over-sampling Technique." Journal of Artificial Intelligence Research 16.1(2001):321-357.
[13] Chen shengling, shen siqi, li dongsheng. Unbalanced data integration learning method based on sample weight updating [J]. Computer science,2018,45(07):31-37.