GAN-Based Semi-supervised For Imbalanced Data Classification

Tingting Zhou
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China
Dongguan Data Science software Technology Co., Ltd
Dongguan, China
e-mail: Zttingop@163.com

Wei Liu*
Chengdu Vocational and Technical College of Industry
Chengdu, China
Dongguan Data Science software Technology Co., Ltd
Dongguan, China
e-mail: weiliu0405@gmail.com

Congyu Zhou
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China
Dongguan Data Science software Technology Co., Ltd
Dongguan, China
e-mail: Lfe007@163.com

Leiting Chen
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China
Institute of Electronic and Information Engineering
UESTC in Guangdong
Dongguan, China
e-mail: richardchen@uestc.edu.cn

***Abstract*—Most of the traditional classification algorithms are based on the premise that the datasets are uniformly distributed or roughly equivalent. Once the sample dataset is not balanced , the classification performance drops sharply. To efficiently deal with the imbalance of data, an improved generative adversarial network (GAN) algorithm is proposed in this work . Firstly, we construct artificial samples so that more minority-class's data can be obtained via optimizing GAN loss function. Secondly, we build a fully-connected network for structured data classification. Finally, experimental evaluations are conducted on two open structured-datasets and the results of the proposed algorithm demonstrate a good applicability for the classification of structured data.**

Keywords-component; GAN; semi-supervised learning; imbalanced data; classification

I. INTRODUCTION

Imbalanced data is a challenging field in classification task. An imbalanced dataset consists of majority-class and minority-class, and the number of samples in minority-class is much smaller than that of majority-class. Samples of the minority-class usually account for 0.1 -5% of the total samples. Credit card fraud data, cancer incidence data and medical insurance fraud data are typically imbalanced data[1].

Most of traditional classification algorithms are based on the premise that the datasets are uniformly distributed or roughly equivalent. Once the sample dataset is not balanced , the classification performance drops sharply.

Recently, there have been considerable research efforts devoted to imbalanced data classification. Xinmin Tao et al.[2] propose an imbalanced data support vector machines (SVM) classification algorithm. However, the selectionof parameters is empirical . MinHan et al.[3] propose a mixed classification algorithm consisting of radial basis function neural network and random forest. This algorithm first makes samples equilibrium, and then trains and tests the neural network of minority-class.But it improves the classification performance of minority-class,at the expense of memory and time. Wang et al.[4] combine a direct-push SVM and an edit-neighbor rule to solve imbalanced data.

In the light of the idea[8] that combines GAN model with semi-supervised learning, we propose an imbalanced data classification method in this paper, where we reduce the imbalance of data via improving GAN to construct more artificial samples.

II. METHOD

A. GAN Model

GAN model[5][6]is a learning generative model based on game theory，which consists of a generator network and a discriminator network. The generator network G(z) produces fake samples that imitate the real samples according to the discriminator network D(x)[7]. z is a random noise and x is the real sample.

The optimization of GAN model is a minimal - maximization problem. The cost function of GAN model can be described as follows:

*Corresponding author

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

where $p(x)$,is a real sample distribution. $p(z)$ is a noise distribution and E (•) represents the calculated expectation. In (1), GAN model includes optimization process of discriminator (2) and optimization process of generator (3).

$$\max_D V(D,G) = E_{X\sim p(x)}\left[\log(D(x))\right] + E_{Z\sim p(z)}\left[\log(1-D(G(z)))\right] \quad (2)$$

$$\min_G V(D,G) = E_{Z\sim p(z)}\left[\log(1-D(G(z)))\right] \quad (3)$$

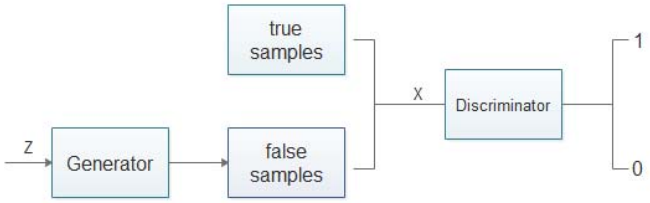

Figure 1. Operation principle of GAN

B. GAN-based Semi-supervised Learning

The semi-supervised learning method is used to train a classifier when most of the training samples are not labeled. Although unlabeled samples do not directly contain label information, they are useful for modeling if they are sampled independently and identically from the same data source with the labeled samples. Im Salimans et al.[8] propose SSGAN algorithm, which combines GAN model with semi-supervised learning method for unstructured data such as images. When GAN model is used in semi-supervised classification[9], the output layer of the discriminator network is converted to a softmax classifier. Assuming that there are K classes in the training dataset. When training GAN model, the samples from generator can be classified as the (K + 1)th class. Meanwhile, softmax classifier also needs to add an output neuron which provides a probability. The probability represents that the input of discriminator comes from generator. The loss function can be described as follows:

$$L = L_{supervised} + L_{unsupervised} = E_{X,Y\sim P(x,y)}\left[\log p(y\mid x)\right] + E_{Z\sim P(z)}\left[\log p(y=K+1\mid x)\right] \quad (4)$$

where y is the label of samples and the total loss is disassembled into supervised loss $L_{supervised}$ and unsupervised loss $L_{supervised}$:

$$L_{supervised} = E_{X,Y\sim P(x,y)}\left[\log p(y\mid x, y<K+1)\right] \quad (5)$$

$$L_{unsupervised} = E_{X,Y\sim p(x,y)}\left[1-\log p(y=K+1\mid x)\right] + E_{Z\sim p(z)}\left[\log p(y=K+1\mid x)\right] \quad (6)$$

C. Improved SSGAN

In SSGAN model, the discriminator is converted to a classifier, and the samples in the classifier includes the real samples with label and the fake samples without label. The fake samples are generated by the generator. According to $L_{unsupervised}$ in (6), there will be generator loss $\min(E_{Z\sim p(z)}\log[p(y=K+1\mid x)])$, when the classifier classifies the fake samples into one of *K* possible classes , and the loss is passed to the generator. We hope the discriminator puts the fake samples into the (K+1)th class as little as possible. However, the generator generates both minority-class and majority-class samples, so that, the classes of generated samples are uncertain, In the results, imbalanced datasets remain a problem.

To solve this problem, we enforce to generate minority-class samples by constructing artificial samples[10]. The generator network of GAN model is used to generate minority-class samples without producing samples of majority-class by modifying the generator loss function as follows:

$$L_{generator} = \max(E_{Z\sim p(z)}\log[p(y=m\mid x)]) \quad (7)$$

where m represents the label of minority-class. The discriminator loss function according to (5) and (6) can be obtained as follows:

$$L_{discriminator} = \max\{E_{X,Y\sim p(x,y)}\log[p(y\mid x, y<K+1)] + E_{Z\sim p(z)}\log[p(y=K+1\mid x)]\} \quad (8)$$

Supposing that the label of positive samples is 0 majority-class and the label of negative samples is 1 (minority-class)According to (7) the generator loss can be obtained by

$$L_{generator} = \max(E_{Z\sim p(z)}\log[p(y=1\mid x)]) \quad (9)$$

That is, the generator expects that the fake samples will be classified by the discriminator as 1, instead of 0. In (9), since the fake samples are enforced to label "1", the generated samples must belong to the minority-class

III. EXPERIMENTS

In this section we evaluate our proposed algorithm by comparing it with SSGAN and traditional random forest algorithm(RF). In this paper, we only focus on structured data. The experiments are conducted on two public datasets: Medical Insurance Fraud Dataset from Alibaba cloud[11],

short for Alibaba-MIFD dataset and Risk Prediction of Login Information dataset from JD[12], short for JD-RPLI dataset.

The Alibaba-MIFD dataset with 16000 samples consists of 5% negative samples and 95% positive samples, and the JD-RPLI dataset contains 132718samples and consists of 2.7% negative samples and 97.3% positive samples. Obviously, the two datasets are imbalanced datasets.

A. Structured Data Preprocessing in GAN Model

As a typically structured data, the two datasets above are stored in a relational database using a two-dimensional table Its' data structure can be regarded as the vector, where the dimension of the vector corresponds to the number of features.Each sample in Alibaba-MIFD dataset has 69 features, such as the cost of medicine, time of hospital stay, the amount of the declaration, JD-RPLI `dataset` has 15 features, such as login time, login-city, user account.

During preprocessing, first, extract high-dimentional descriptors from original features[13][14];then, normalize the data, and take normalized data as one-dimensional tensor to send into the GAN model respectively, moreover, build a fully connected GAN model to be more suitable for structured data. In practice, discriminator and generator contains three fully-connected layers in training process.

B. Comparison Results and Analysis

We randomly choose 12000 samples from the Alibaba-MIFD dataset as the training set, and the remaining 4000 samples as the test set, and 721-dimentional descriptors are extracted from original features. In the JD-RPLI dataset, 99538 samples are chosen as training set and 33180 samples as the test set , and 683-dimentional descriptors are extracted from original features.

In table I and II, We select 1:1,3:1,7:1,10:1 and19:1 as the ratios of positive samples to negative samples in Alibaba-MIFD, while 1:1,5:1,10:1,20:1,27:1 and 35:1 in the JD-RPLI dataset.

It can be found in table I that the accuracy of our model is always higher than SSGAN and RF. The classification accuracy of the proposed method is 23.88% and 5.92% higher than the resultof SSGAN algorithm and RF algorithm when the ratio is 1:1, respectively. With the increase of the ratio of positive samples, the accuracy of the three models is increasing, and the advantage of the proposed method is most obvious. When the ratio risesto 19: 1, the proposed method improves the performance by 42.71% over SSGAN and 1.99% over RF..

As shown in table II, the results of JD-RPLI dataset indicate that our method is significantly better than the SSGAN. The classification accuracy of the our method is 6.63% higher than the result of SSGAN on the ratio of 1:1.And when the ratio rises to 35:1,the accuracy is improved from 58.89% to 96.92%.

Due to the sparsity of JD-RPLI data features, the accuracy of the proposed is lower than the RF. But the performance of ours approximates those of RF as the number of positive samples increases.

TABLE I. PERFORMANCE COMPARISON ON ALIBABA-MIFD DATASET

Ratio \ Model	*SSGAN*	*RF*	*Ours*
1:1	51.79%	69.75%	75.67%
3:1	52.08%	82.54%	85.55%
7:1	52.78%	88.79%	91.30%
10:1	53.81%	91.75%	92.97%
19:1	54.81%	95.53%	97.52%

TABLE II. PERFORMANCE COMPARISON ON JD-RPLI DATASET

Ratio \ Model	*SSGAN*	*RF*	*Ours*
11	52.26%	86.59%	58.89%
5:1	55.38%	95.24%	81.26%
10:1	57.41%	96.88%	88.45%
20:1	57.90%	97.93%	92.28%
27:1	58.20%	98.25%	93.50%
35:1	58.89%	99.14%	96.92%

For a more intuitive observation, experimental results on two types of datasets are shown in Figure 2 and 3 , respectively. The value of ratio v is defined by

$$v = \frac{the\ number\ of\ \ positive\ samples}{the\ number\ of\ \ negative\ samples} \tag{10}$$

With the increase of the value of ratio, the accuracy of SSGAN in training process has been hovering around 50%. The accuracy of our method could be gradually increased to 90% .

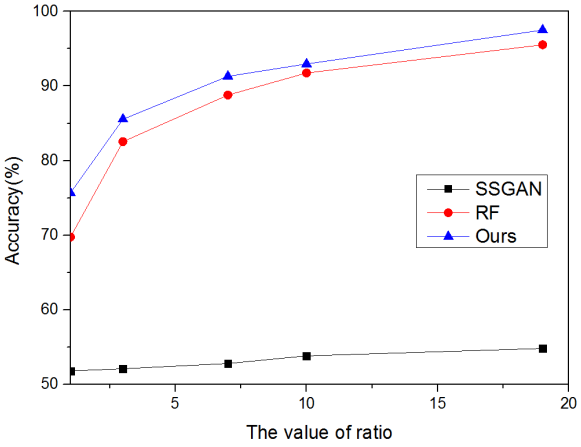

Figure 2. Performance on Alibaba-MIFD dataset

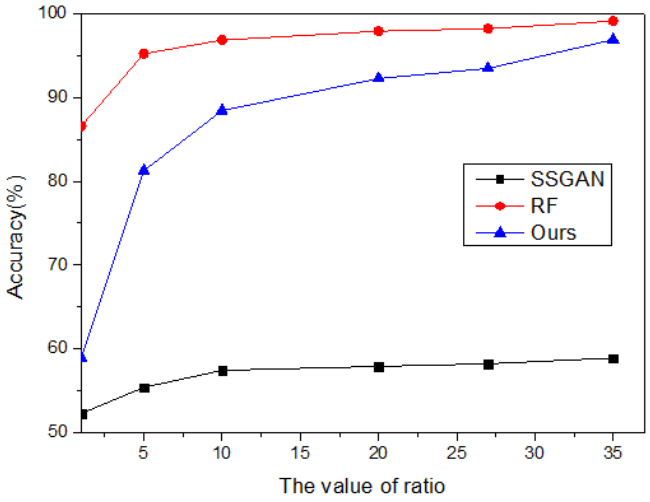

Figure 3. Performance on JD-RPLI dataset

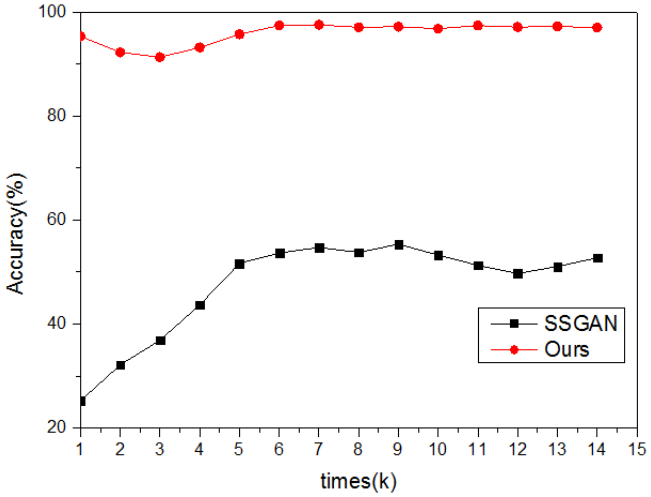

Figure 4. Optimization process on Alibaba-MIFD dataset

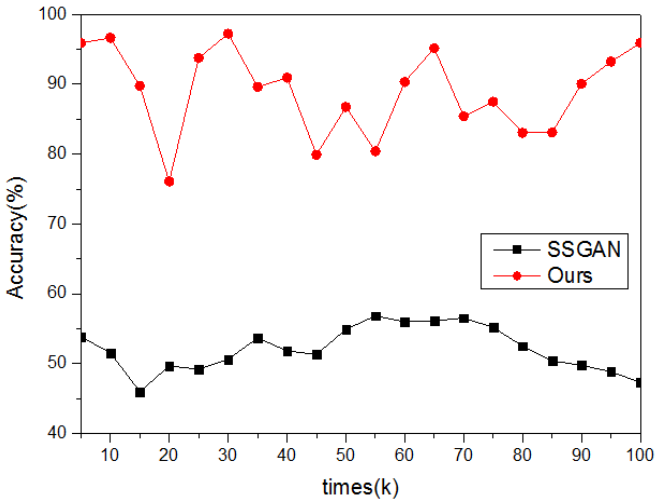

Figure 5. Optimization process on JD-RPLI datase

Figure 4 and 5 demonstrate optimization process on training sets, from Alibaba-MIFD dataset and JDRPLI dataset, respectively. As training times increasing, the accuracy of our methodcan be kept higher than the SSGANall the time. It can be seen that the accuracy of our method is kept above 90% while the accuracy of SSGAN is less than 60% on the Alibaba-MIFD training set. In the JDRPLI case, the fluctuation of the accuracy on SSGAN is in the range of 45% ~ 57%, that on our method is 75% ~96%.

IV. CONCLUSIONS

In this paper ,we propose an improved method based on SSGAN model for imbalanced data classification. There are two main contributions, one is that only minority-class samples are generated in training process, and this will obtain a relative balance between minority-class and majority-classes. Another contribution is that GAN model is applied on structured data. Experimental results show that the proposed method achieve satisfactory performances.

In the future, we plan to study feature extraction from structured data by deep networks[15] and the sparsity of features by rough set approach[16].

ACKNOWLEDGMENT

This research is supported by the National High Technology Research and Development Program ("863"program) of China(NO.2015AA016010), Application Science and Technology Planning Project of Guangdong Province (NO.2015B010131002), and Major Science and Technology Projects of Dongguan (NO.2015215102).

REFERENCES

[1] Zhang C, Gao W, Song J, et al. An unbalanced data classification algorithm of improved auto encoder neural network. Eighth International Conference on Advanced Computational Intelligence. 2016. 95–99

[2] TAO Xin-min, TONG Zhi-jing, LIU Yu, FU Dan-dan. SVM classifier for unbalanced data based on combination of ODR and BSMOTE. Control and Decision. 2011,26(10):1535-1541.

[3] HAN Min,ZHU Xin-rong. Hybird algorithm for classification of unbalanced datasets. Control Theory and Applications. 2011,28(10):1485-1489.

[4] Wang A, Liu L, Jin X, et al. Adapting TSVM for fault diagnosis with imbalance class data. Control and Decision Conference (CCDC). China. IEEE. 2016. 2919–2923

[5] WANG Kun-Feng, GOU Chao, DUAN Yan-Jie,et al. Generative Adversarial Networks: The State of the Art and Beyond. Acta Automatic Sinica.2017,43(3)

[6] Poole B, Alemi A A, Sohldickstein J, et al. Improved generator objectives for GANs[J]. 2016.

[7] Larsen A B L, Sønderby S K, Larochelle H, et al. Autoencoding beyond pixels using a learned similarity metric[J]. 2015:1558-1566.

[8] Salimans T, Goodfellow I, Zaremba W, et al. Improved Techniques for Training GANs[J]. 2016.

[9] Kingma D P, Rezende D J, Mohamed S, et al. Semi-Supervised Learning with Deep Generative Models[J]. Advances in Neural Information Processing Systems, 2014, 4:3581-3589.

[10] Sun Z, Song Q, Zhu X, et al. A novel ensemble method for classifyin-g imbalanced data[J]. Pattern Recognition, 2015, 48(5):1623-1637

[11] https://tianchi.aliyun.com/

[12] http://jddjr.jd.com/item/1

[13] Liu W, Chen L T, Cai H B, et al. A canonical form-based approach to affine registration of DTI[J]. Multimedia Tools and Applications, 2017: 1-22.

[14] Dai Jianhua, Xu Qing. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. Applied Soft Computing, vol.13, no.1, pp. 211-221, 2013

[15] Li X, Yang F, Cheng H, et al. Multi-Scale Cascade Network for Salient Object Detection[C]// ACM, 2017:439-447.

[16] Jianhua Dai, Hu Hu, Wei-zhi Wu, Yuhua Qian, Debiao Huang. Maximal Discernibility Pairs based Approach to Attribute Reduction in Fuzzy Rough Sets, IEEE Transactions on Fuzzy Systems, DOI: 10.1109/TFUZZ.2017.2768044, 2017