Data Augment in Imbalanced Learning Based on Generative Adversarial Networks

Zhuocheng Zhou[1], Bofeng Zhang[1(✉)], Ying Lv[1], Tian Shi[1], and Furong Chang[1,2]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai, China
{bearing512,bfzhang,lvying2016}@shu.edu.cn
[2] School of Computer Science and Technology, Kashi University, Kashi, Xinjiang, China

Abstract. Imbalanced learning is a traditional problem in machine learning and widely occurs in many applications. Most of the methods apply simple geometric transformation for data augment to imbalanced datasets. Due to those methods learn from local information, they might generate noisy samples in the dataset with high dimension and special complexity. To solve the problem, we propose an improved Generative Adversarial Networks with modification function (GAN-MF) to approximate the true distribution of the minority class of the dataset. The model could generate data from an overall perspective to overcome the limitation of the simple geometric transformation. The performance of GAN-MF is compared against multiple standard oversampling algorithms on several imbalanced learning tasks. Experiments demonstrate that the model has an improvement in data augment for imbalanced learning.

Keywords: Imbalanced learning · Generative Adversarial Networks (GAN) · Data augment · Modification function

1 Introduction

Learning from the imbalanced dataset is challenging and meaningful in many common areas including fraud detection, healthcare and medical diagnosis and many other applications. The reason why the performance of classifier drops sharply when dataset is imbalanced is that most standard algorithms assume or expect that the class distribution is balanced. So that, features of the minority class might be missed or neglected.

Imbalanced learning refers to the dataset in which one or several classes are outnumbered than the others. The gap in number of instances among the classes is defined as the imbalanced ratio (IR) [7].

Supported by National Key R&D Program of China grant (NO. 2017YFC0907505) and the Xinjiang Natural Science Foundation grant (NO. 2016D01B010).

T. Gedeon et al. (Eds.): ICONIP 2019, CCIS 1142, pp. 21–30, 2019.
https://doi.org/10.1007/978-3-030-36808-1_3

Devoted to improving the performance of imbalanced learning, different methods have been proposed, those could be summarized into several categories. The first is from the data perspective, focusing on reinforcing the learning on the minority by the means of sampling and feature selection [2]. Besides, synthetic sampling for data augment is also widely used. The second is to encourage the classifiers to minimize the cost errors by introduced cost-sensitive, ensemble learning and kernel-based methods [3,9]. The third is to restructure the classifier to suit the task according to the background of the applications. For instance, the algorithm of transfer learning and genetic algorithm are integrated in imbalanced learning [1].

However, with the rise in data complexity, methods mentioned above might be insufficient, especially in terms of data augment. Inspired by the fact that Deep Generative Models (DGMs) can synthesize new samples based on the distribution captured from the overall class rather than local information [8]. We try to synthesize sampling based on Generative Adversarial Networks (GAN) for data augment to improve the imbalanced binary classification. Whereas, the vanilla GAN model is restricted to continuous derivable variables for the gradient policy and the instability in training for model collapse and vanishing gradient.

To settle the matters, we proposed a novel GAN model (GAN-MF) based on a modification function $f(\mathrm{x})$ to approximate the true data distribution of the minority class. With the help of the modification function, the numeric discrete detests in imbalanced learning are converted into datasets with approximate Gaussian distribution that could be accepted by the GAN model and be trained in a stable way. The performance of the model is compared against multiple standard over-sampling algorithms and another generative model of Variational Auto-Encoder (VAE) based on 6 classifies. Experiments show GAN-MF has improved the results in imbalanced learning tasks.

The sections in the paper are organized as follows. In Sect. 2, an overview of related previous works regarding to GAN models and imbalanced learning are described. In Sect. 3, the model of the GAN-MF and application to the imbalanced learning is stated. In Sect. 4, the experiments and the results are addressed in detail. Finally, conclusions are provided in Sect. 5.

2 Related Works

Because of the simplicity and effectiveness of the algorithm, algorithms based on data augment are most widely used [6]. They offer additional minority-class instances derived by applying simple geometric transformations for the training. As the most classic one, Synthetic Minority Over-sampling Technique (SMOTE) provides a mechanism in creating artificial data based on the feature space similarities among the existing minority in the d-dimension dataspace X. The new instance x_{new} is created by $(x_i + \lambda(x_j - x_i))$, where $x_{i,j}$ is the minority instance in X, and x_j is selected considering to the k-nearest neighbor for x_i. Therefore, x_{new} is created in the vector between $x_{i,j}$, located in a random percent of way from $x_{i,j}$ as $\lambda \in [0, 1]$.

However, SMOTE has a vague understanding of the boundary and might generate noisy samples. To modify the algorithm, several rules including Edited Nearest Neighbor, balanced and weight level have been introduced into the algorithm which is summarized in [5]. Since they learn from local information, they might be ineffective in dealing with data in high dimensions.

DMGs have been gradually introduced to data augment for the excellent capability to represent multidimensional and complex data. Neural augment is firstly proposed in imbalanced picture classification in [11]. After that, Balancing GAN (BAGAN) [10] goes further more by taking attention mechanism to the training. The method based on Conditional generative adversarial networks (CGAN) [4] has also been addressed in learning numeric imbalanced data where additional space Y, as the label of the instance, is introduced to extra valuable information from latent space.

Although many efforts have been made, little research has been conducted in using GANs in learning the numerical variables dataset, and there is hardly no evidence suggests whether it is effective for GANs to generate discrete skewed data in dealing with imbalance learning. Meanwhile, it is unknown whether GANs have a shortage of capacity and training time when compared with standard over-sampling methods.

3 GAN-MF Model for Imbalanced Learning

3.1 The GAN-MF Model

The aim of generative model is to learn the data probability distribution $p_{data}(x)$ over the real space R^d. Although GANs have shown excellent ability to capture the distribution in many applications, the vanilla GAN model has been proved to be unsuitable to deal with discrete data for the model has hardly no gradient in generation process [12]. In addition, the model has a problem in training for model collapse and gradient vanish.

Thus, we introduced a modification function to figure out the limitation of the GAN model. The modification function $f(x)$ serves the role to convert problems of discrete data into an approximate continuous variable one that can be served by the GAN model. In other words, the d-dimensional real space is R^d is mapped to a special vector space $R^{d'}$ where numerical differences in features are relatively smooth and representative features of the dataset are preserved.

As the result, the two-player minimax game between the discriminator D and the generator G is improved. As G acts the role of producing fake data with striking resemblance from the latent variable z, D tells the data from sampled from the true data distribution $p_{data}(f(x))$ apart from those forged by G, where z is defined on the latent space Z.

The value function of the GAN-MF model is described in (1), where $E()$ represents the calculated expectation. From the view of D, it will maximize the outs if given data from real data and minimize the output if given data from G. Thus, D is optimized followed as $\log(1 - D(G(z))$. At the same time, G tries the best to maximize the output of G when the fake is presented to D. G is

optimized by $\log D(f(x))$. Finally, the generator's distribution $p_g(x)$ approaches to $p_{data}(f(x))$. The distribution of discrete dataset is related to $G(z, \theta)$, where θ is tuning parameters of the G.

$$\min_G \max_D V = E_{x \sim p_{data}(f(x))}(\log D(f(x))) + E_{z \sim p_z(Z)}[\log(1 - D(G(z)))] \quad (1)$$

3.2 Modification Function

With the help of modification function, the model has the ability to approximate the data distribution of the minority and generate augmented datasets that can present characteristics in a much smaller size than the simple geometric transformation.

Jensen-Shannon divergence and Wasserstein distance are widely used as the way to measure the difference in data distribution and optimizer for the network. We defined a vector $x = (x_1, x_2, x_3, ...x_n)$ as a discrete multivariate random variable where values of x_i are from fractions and integers. When we try to evaluate the Wasserstein distance between two probability distributions P_a and P_b, where $P_{(a,b)}$ is over the set of values for x, we find that it is a Linear Program (LP) problem. Therefore, the runtime reflects exponential growth with the increase in dimensions of data and variety of variables.

$$W(P_a, P_b) = min_{\gamma \in \prod(P_a, P_b)} \sum_i \sum_j \gamma(x_i, x_j) d(x_i, x_j) \quad (2)$$

where $d(x_i, x_j)$ is the distance between x_i, x_j and $\prod(P_a, P_b)$ is defined as the set of joint probability distribution $\gamma(x_i, x_j)$ whose marginals are P_a and P_b.

The same problem also occurs in the JSD which is used in most GAN models. As a consequence, it is clear that learning directly from difference in discrete mathematical distribution is not easy. Since the fact that it is difficult to measure the difference in discrete data distribution, the modification $f(x)$ become significant to GAN-MF. The modification we proposed is shown in (3), where μ_i is the mean of the feature x_i is and σ_i is the standard deviation of x_i.

$$max(0, \frac{x_i - \mu_i}{\sigma_i}) \quad (3)$$

Suppose that the networks is defined as $U = Wx + b$, $Z = F(U)$, where $F()$ is the activation function and W, b is the vector of weights and bias. When the modification is worked to the algorithm, the networks is transformed into (4):

$$U(f(x)) = W[max(0, \frac{x_i - \mu_i}{\sigma_i})] + b \quad (4)$$

Therefore, if $x_i > \mu_i$, $U(f(x)) = W(\frac{x_i - \mu_i}{\sigma_i}) + b$. All the features has been transformed to an approximate Gaussian distribution $\mathbf{N}(0,\ 1)$ which could be accepted by the GAN model and positive to the convergence of the networks. If $x_i < \mu_i$, x_i would be 0 in x. The vector would become sparse and the features would be more independent.

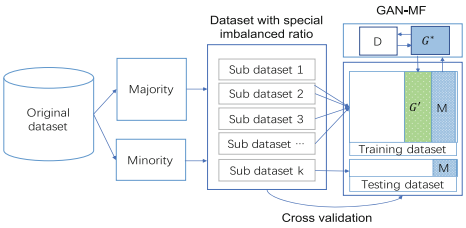

Fig. 1. GAN-MF for imbalanced learning. M refers to the minority of the dataset, $G^{'}$ is the augmented dataset generated by G^{*} for training. (Color figure online)

4 Experiments and Results

The framework of the GAN-MF Model in imbalanced learning we proposed is shown in Fig. 1.

(1) k-fold cross validation is applied with $k = 5$, the dataset is factitiously divided into 5 parts. Each part has approximately equal instances for both classes.
(2) All the hyperparameters of classifiers are performed with maximum accuracy under original dataset and used in subsequent experiments.
(3) The minority examples M colored in blue in Fig. 1 are isolated for training the GAN model G^{*} with tuning parameters.
(4) G^{*} as a generative model could generate artificial dataset $G^{'}$ by receiving random noise as input. Hence, the dataset used for training the classifies is composed of $G^{'}$ (colored green in Fig. 1) and sampled from real ones in M.

In this work, we rebalanced the dataset to the equal IR to the traditional methods. It made sure that classifies could learn unbrokenly. We doubled or tripled the number of minority classes in training for the methods based on deep generative models since they learn from an overall view.

Datasets. Several datasets from the Machine Learning Repository UCI and a credit card detection dataset were chosen for experiments. Aiming to objectively test the performance of the GAN-MF model, by the means of the under-sampling and random-sampling, the datasets from UCI were generated into additional dataset according to IR of 4, 10 on purpose. This procedure was applied only when the instances of the minority in the sub-dataset is no less than 5. Table 1 shows the datasets in detail. Values separated by comma in the table cells are related to the same dataset over original status and different IR in 4 and 10.

Architecture of the GAN-MF Model. In this work, both G and D used a module of multilayer perceptron with one single hidden layer. No convolution layers was need. Binary cross-entropy was served as the loss function. Relu was selected as the activation function in the output layer for G when Sigmoid was used in D. As Adam optimizer was used as the optimizer in G, Stochastic

Table 1. Description of the datasets in detail.

Dataset	Features	Majority instances	Minority instances	IR
Segment	16	1980,1000,1000	330,250,100	6,4,10
German	24	700,400,400	300,100,40	2.333,4,10
Pima	8	500,400,400	268,100,40	1.8656,4,10
Liver	10	416,400,400	165,100,40	2.491,4,10
Haberman	3	255,200,200	81,50,20	2.778,4,10
Ionosphere	34	255,200,200	126,50,20	1.786,4,10
Breastcancer	16	458,400,400	241,100,40	6,4,10
Credit card	29	284315	492	577.876

Gradient Descent(SGD) was chosen in D. Dropouts was used in G with a probability of 0.5. The input random noise followed a normal distribution. The other hyperparameters of the networks are described in Table 2. The optimal range for the numbers of epochs shoule be 5000–15000, much smaller than the one in the picture. The batch size should be set carefully to ensure that the final number of minority class instances is sufficient for the training. No dimensionality reduction methods were used. All samples with missing values were deleted.

Table 2. Parameters for GAN-MF model in detail. Including dimension d_z, number of hidden units for G and D, learning rate and batch size. The values in the same cell refers to the parameters under the IR of 4 and 10. N_G and N_D is defined as the number of units for hidden layer of G, D.

Dataset	d_z	N_G	N_D	Learning rate	Batch size
Segment	80,120	100,50	30,130	0.0005,0.0005	20,10
German	150,100	90,80	50,50	0.0005,0.0005	16,8
Pima	20,8	50,45	80,80	0.0005,0.0005	8,8
Liver	50,25	35,50	20,30	0.0001,0.0005	5,5
Haberman	10,10	20,20	10,15	0.0005,0.0005	10,8
Ionoshere	200,120	30,25	90,90	0.0005,0.0005	8,8
Breastcancer	70,70	90,90	30,30	0.0005,0.0005	10,10
Creditcard	200	36	100	0.0001,0.0001	10

Assessment Metric. F-measure, the geometric mean of specificity and sensitivity (G-mean) and Area Under the ROC Curve (AUC) were chosen as the assessment criteria. k-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Trees (DT), AdaBoosting classifier, Nave Bayes (NB) and an ensemble learning method based on the simple voting (Vote) method were chosen as classifies.

Furthermore, a ranking score and the Friedman test were given for more holistic evaluation of the results. The ranking score was applied to each data augment method for the experiments of 14 datasets under different assessment metrics and classifiers. In the rank, the best performing method ranks 1 and the worst one ranks 6. Besides, we defined the under-fitting as the situation that F-measure was under 50% and G-mean was under 40%. The under-fitting methods were set 6 in the rank. The Friedman test is a non-parametric statistical test, and widely used to detect the difference between treatments across multiple research attempts. The null hypothesis in the work is whether GAN-MF model is as effective as traditional over-sampling methods for data augment in imbalanced learning.

Results. The meaning ranking results are summarized in Fig. 2, where each plot is related to three assessment metrics and a classifier. Each mean rank is the result of 14 datasets based on the same classify. From a macro perspective, the model of GAN-MF has shown the improvement in most classifiers and datasets.

With fewer training data for data augment, we observe that the GAN-MF outperforms all other data augment methods when the voting algorithm

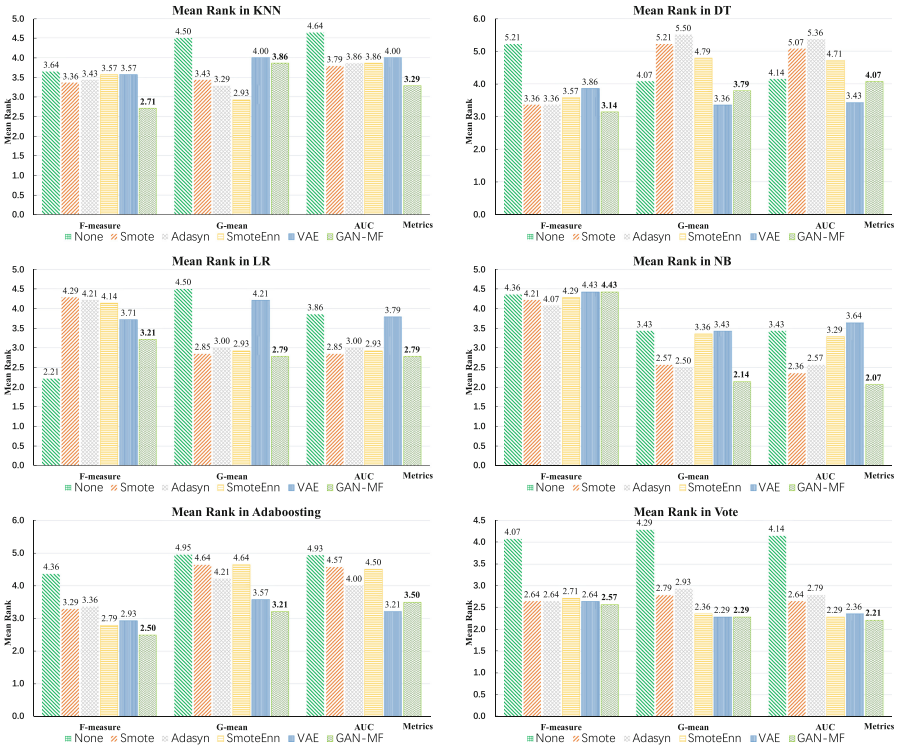

Fig. 2. Result for mean ranking of various data augment methods.

Table 3. Result of Friedman test. If $p < \alpha$, reject the hypothesis.

$\alpha = 0.05$	KNN	DT	LR	NB	Adaboost	Vote
X_γ^2	6.81	3.857	5.857	4.333	9.95	4.904
p	0.235	0.570	0.320	0.502	0.077	0.427

is selected as the classifier. It is also clear that the GAN-MF has an advantage to the metric of G-mean and AUC in more than four-fifths of cases.

The result of the Friedman test is shown in Table 3. All the p-values are more than the given standard value and the hypothesis are all not rejected where $\alpha = 0.05$. It means that the performance of the classifies show no bias in different methods and GAN-MF model is superior to traditional methods in data augment for imbalanced learning.

In the terms of the vibration in the mean rank for the GAN-MF, it should be noted that F-measure might be sick since the classifies would be favor to the majority and mark a high score for original imbalanced data. Both GAN and VAE have done a bad performance especially in the dataset with fewer features and instances which result in the drop in meaning rank.

As it can be seen from Table 4, G-mean and AUC have improved appreciably and F-measure holds the line when augmented data synthesized by GAN-MF is used in training. Each result is the average of the cross validation. The instance of

Table 4. Results of credit card fraud detection.

Metric	Methods	None	Smote	Adasyn	SmoteEnn	VAE	GAN-MF
F-measure	KNN	0.99962	0.99839	0.99839	0.99822	0.99962	**0.99964**
	DT	0.80092	0.85970	0.85956	**0.85972**	0.81792	0.83056
	LR	**0.99957**	0.98131	0.93163	0.98128	0.81752	0.86709
	NB	**0.98873**	0.98766	0.97642	0.98759	0.98723	0.98765
	Adaboosting	0.85081	0.83962	0.83865	0.83865	0.98773	**0.99887**
	Vote	0.99778	0.99353	0.98388	0.99331	0.99947	**0.99965**
G-mean	KNN	0.84812	0.89138	0.89013	0.89686	0.84812	**0.90812**
	DT	0.67981	0.57068	0.58051	0.60438	0.71749	**0.71979**
	LR	0.77057	0.92087	0.90380	0.93396	0.71803	0.83957
	NB	0.89860	0.91339	0.91993	0.91332	0.91576	**0.92349**
	Adaboosting	0.66077	0.77537	0.76366	0.76366	0.79003	**0.84511**
	Vote	0.87027	0.91506	0.91865	0.91598	0.87436	**0.91948**
AUC	KNN	0.86145	0.89785	0.89682	0.90276	0.86145	**0.91389**
	DT	0.76956	0.70109	0.70920	0.71636	0.76339	**0.77974**
	LR	0.80050	0.92197	0.90644	**0.93503**	0.78298	0.85602
	NB	0.90253	0.91570	0.92111	0.91563	0.91782	**0.92469**
	Adaboosting	0.73837	0.81944	0.80377	0.80377	0.81732	**0.85964**
	Vote	0.87997	0.91850	**0.92157**	0.91930	0.88362	0.90901

the augmented data generated by GAN-MF for training is about 1100, about one hundredth of the ones based on simple geometric transformation. Since GAN-MF learn the screwed dataset from an overall way, the classify can capture the representative feature in a more effective way. In the terms of G-mean and AUC, the GAN-MF model has outperformed in two-thirds of classifies. The classify of Adaboosting and KNN have been obviously improved by the GAN-MF.

5 Conclusion

In this work, we proposed a GAN-MF model for data augment to improve the imbalanced learning. Since the model learns the dataset from an overall view, it can generate data for augmentation based on the learned distribution. Modification function is employed to converts the numeric discrete detests into the one that could be train in a stable way. The model has been evaluated on several datasets, with much fewer augmented data, the model has done a good performance for most classifies, especially in dataset with high dimension.

More work should be taken to overcome the limitation of the model in stabilization, capacity and training time. The whole model still suffers from collapse problem. Our future work will try more different networks as well as take more other deep generative models into practice.

References

1. Al-Stouhi, S., Reddy, C.K.: Transfer learning for class imbalance problems with inadequate data. Knowl. Inf. Syst. **48**(1), 201–228 (2016)
2. Chen, H., Li, T., Fan, X., Luo, C.: Feature selection for imbalanced data based on neighborhood rough sets. Inf. Sci. **483**, 1–20 (2019)
3. Ding, S., et al.: Kernel based online learning for imbalance multiclass classification. Neurocomputing **277**, 139–148 (2018)
4. Douzas, G., Bacao, F.: Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Syst. Appl. **91**, 464–471 (2018)
5. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **42**(4), 463–484 (2011)
6. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: review of methods and applications. Expert Syst. Appl. **73**, 220–239 (2017)
7. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2008)
8. Hong, Y., Hwang, U., Yoo, J., Yoon, S.: How generative adversarial networks and their variants work: an overview. ACM Comput. Surv. (CSUR) **52**(1), 10 (2019)
9. Lu, H., Yang, L., Yan, K., Xue, Y., Gao, Z.: A cost-sensitive rotation forest algorithm for gene expression data classification. Neurocomputing **228**, 270–276 (2017)
10. Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., Malossi, C.: BAGAN: data augmentation with balancing GAN. arXiv preprint arXiv:1803.09655 (2018)

11. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621 (2017)
12. Srivastava, A., Valkov, L., Russell, C., Gutmann, M.U., Sutton, C.: VEEGAN: reducing mode collapse in GANs using implicit variational learning. In: Advances in Neural Information Processing Systems, pp. 3308–3318 (2017)