Over-sampling Algorithm Based on VAE in imbalanced Classification

Chunkai Zhang[1], Ying Zhou[2], Yingyang Chen[3], Yepeng Deng[4], Xuan Wang[5], Lifeng Dong[6] and Haoyu Wei[7]

[12345] Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
[6] Hamline University, 1536 Hewitt Avenue, St. Paul, the USA
[7] Sichuan University, Sichuan, China
ckzhang812@gmail.com

Abstract. The imbalanced classification problem is a problem that violates the assumption of uniform distribution of samples, classes differ in sample size, sample distribution and misclassification cost. The traditional classifiers tend to ignore the important minority samples because of their rarity. Oversampling, the algorithm uses various methods to increase the minority samples in the training set to increase the recognition rate of them. However, these over-sampling methods are too coarse to improve the classification effect of the minority samples, because they can't make full use of the information in the original samples, but increase the training time because of adding extra samples. In this paper, we propose to use the distribution information of the minority samples, use the variational auto-encoder to fit the probability distribution function of them without any prior assumption, and reasonably expand the minority class sample set. The experimental results prove the effectiveness of the proposed algorithm.

Keywords: Imbalanced Classification, Variational Auto-encoder, Over-sampling.

1 Introduction

The classification problem is a very important part of machine learning, and it is also the first step for artificial intelligence to understand human life. At present, most classifiers assume that the samples of different classes are evenly distributed, and the classification costs are the same. However, in reality, the data people are more concerned about is often scarce, such as the detection of credit card fraud and medical disease diagnosis. In the medical disease diagnosis, most of the results are normal while only a small proportion of the results are diagnosed as diseases, which indicates the different distribution in different classes samples. Second, if healthy people are misdiagnosed as diseases, they can be removed by other inspection methods, errors do not cause very serious accidents, but if the disease people are diagnosed as healthy, it may cause the patients to miss the best treatment time and cause serious consequences. This is the second feature of the imbalanced classification problems: different classes of misclassification costs are inconsistent. At the same time, if samples are classified as diseases

as much as possible because they are afraid to miss the dis-ease samples, it will cause a huge waste of medical resources and intensify conflicts between doctors and patients. Therefore, it is not feasible to determine all samples as disease, and the best way is to try to separate these two results as correct as possible. Due to the scarcity of the minority samples and the definition of global accuracy, the classifier pays less attention to the minority class, so the recognition performance is unsatisfying. Imbalanced classification problems arise in many fields, such as bioinformatics [1], [2], remote sensing image recognition [3], and privacy protection in cybersecurity [4]–[6]. The imbalanced problems cover widely and have a very important practical significance.

The traditional solutions to the imbalanced problems are divided into two parts: the algorithm-level methods and the data-level methods. The algorithm-level methods mainly focus on the different misclassification costs, such as improved neural network [7]: it uses the approximation of F1 value of the minority class as the cost function; the bagging algorithm [8] continues to enhance the misclassified the minority samples, and improve the recognition rate of the minority samples; structured SVM [9] uses the F1 value of the minority samples as the optimization function, and has a better performance in the classification of the minority samples.

The data-level methods focus on the imbalance of sample size, which mainly adjust the data sample size through resampling to reduce the impact on classification performance. The data-level methods can be divided into over-sampling, under-sampling and hybrid sampling. Over-sampling adds the minority samples in the training process, Over-sampling can effectively improve the classification performance of the minority class but it has no idea of the rationality. Under-sampling [10] removes the majority samples before training, which can quickly reach equilibrium, but may take a risk of losing valuable samples.

The oversampling method can be divided into random sampling and informed sampling. Random sampling means repeating the known samples, which includes simple repetition [11], linear interpolation [12], nonlinear interpolation [13], etc.; SMOTE [12], as a classic over-sampling algorithm, interpolates linearly in the minority samples, will increase the amount of information and rationality of synthesized samples in random oversampling, which improves the classification effect. Border-line-smote [14], to reduce the risk of overfitting, it selects the minority samples needing to be interpolated called boundary samples. The above oversampling methods only consider the influence of the sample size and the local sample distribution on the classification performance, ignoring the overall distribution of the sample, which is more informative for classification performance.

Informed sampling [15] uses the distribution information in the sample to fit its probability distribution function (PDF) and sample it according to the PDF. Chen [16] proposed a normal distribution based oversampling approach, and this approach assumes the minority class distribution as the Gaussian normal distribution, the parameters are calculated from the minority samples with EM algorithm, the experimental results are better than SMOTE and random oversampling. Different scholars have proposed oversampling algorithms based on various distributions, such as the Gaussian distribution [16], [17], Weibull distribution [18], etc. Due to the distribution information, these algorithms have made greater progress than random oversampling method. However, the

problems are also obvious: there is a prior assumption about the real distribution and all the features are dependent from each other. If the real distribution meets this hypothesis, it will get better results, otherwise, the improvement is limited, so it is inconsistent in their effect on different datasets.

Data level methods are of great matter in imbalanced classification, as it can be regarded as a step in data preprocessing, it will have a positive effect on the final classification results. Since the factors that affect the datasets classification include not only the sample size, but also the sample distribution, while the current over-sampling methods do not make full use of distribution information and cannot guarantee the rationality of the generated samples.

In this paper, we propose a oversampling method based on the variational auto-encoder [19] (VAE) model to generate the minority samples. The proposed method is motivated that the distribution information plays an important role in oversampling methods, and aims at the rationality of the generated samples, we use VAE to increase minority instances, to our knowledge, first, the output dimension of the neural network is not limited so it can generate data of any dimension; second, the strong fitting ability of the neural network can simulate any distribution function without any prior knowledge in advance. We use this model to model the distribution of minority samples and oversample according to the model, the proposed method shows the superiority that it doesn't need any prior distribution assumption nor the dependent features assumption, the experimental results prove the effectiveness of the algorithm.

We organize the paper as follows. Section 2 describes related work of this paper. Section 3 presents the proposed algorithm and analyze it. Section 4 shows the experimental results. Section 5 concludes the paper.

2 Related work

In 2013, KM [19] proposed VAE: add variational inference to auto-encoder and use parameterization trick to make the variational inference combined with stochastic gradient descent. The overall structure of vae network is shown in Fig. 1, while it assumes the hidden variables to be a Gaussian standard distribution, it is easy to sample and the final probability distribution function is uncertain, coincides with the characteristics of distribution-based oversampling.

In VAE, we assume the variables are determined by the hidden compression code z, the encoder can map z to X, which makes z obey a particular distribution (such as Gaussian distribution, etc.). Knowing the possibility distribution function and its mapping function, we can sample z and encode z, to get new x to generate infinite sample theoretically. The structure of vae as shown below:

Assume z is a latent variable, and its distribution function is p(z), use Bayesian conditional probability formula to calculate P(X):

$$p(X)=\int p(X|z)p(z)dz \tag{1}$$

However, in z's prior distribution, most of z cannot generate reliable samples, that is p(X|z) tends to 0, so p(X|z)p(z) tends to 0. To simplify the calculation, only $p(X \mid z)$

need to be calculated. Considering the z with larger P(X|z), which is represented by P(z|X) form the encoder, but only considering this part of z cannot generate samples that are not in original data, so we need to assume the distribution of P(z | X) and complete the error through the decoder.

Q(z) is the assumption of the real distribution, we use KL divergence to calculate the difference between the real distribution and the assumption:

$$D(p\|q)=\int p(x)\log\frac{p(x)}{q(x)}dx \tag{2}$$

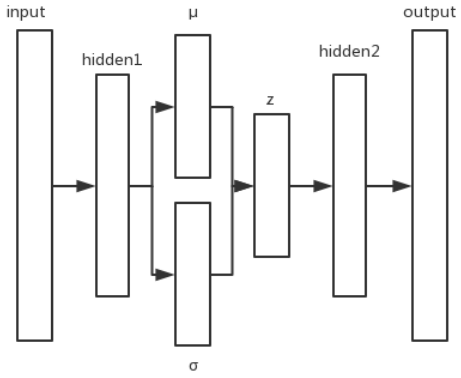

Fig. 1. Structure of variational auto-encoder.

Formula (2) shows that if two distribution is close, KL divergence will tend to 0. And the loss function of VAE model is

$$argmin\ D(Q(z)\|P(z|X)) \tag{3}$$

Apply the formula (2) to the formula (3)

$$D[Q(z)||P(z|X)]=E_{z\sim Q}[\log Q(z)-\log P(z|X)] \tag{4}$$

Apply Bayes rule to P(z|X), we can get both P(X) and P(X|z)

$$D(Q(z)||P(z|X))=E_{z\sim Q}[\log Q(z)-\log P(z)]+\log P(X) \tag{5}$$

Apply the $D[Q(z)||P(z|X)]$ into it, note that X is fixed, and Q can be any distribution, not just a distribution which does a good job at mapping X to the z's to produce X. since we're interested in inferring P(X), it makes sense to construct a Q which does depend on X, and in particular, one which makes $D(Q(z)||P(z|X))$ small: Because P(X) is fixed, the minimum $D(Q(z)||P(z|X))$ will transform to maximize the value of the right side of the equation, and logP(X|z) is the probability of X decoded by z. It is calculated as the cross-entropy or mean-squared error of the original sample. The latter can be regarded as the difference between the assumption and the distribution of z in the encoder.

$$\log P(X)-D[Q(z)||P(z|X)]=E_{z\sim Q}[\log P(X|z)]-D[Q[z]||P(z)] \quad (6)$$

3 The proposed method

In this paper, an oversampling method based on VAE is proposed, motivated by the idea that the distribution information is important in oversampling method. Without any prior assumption of the real PDF of the minority samples nor the independent assumption in the features, the proposed method can automatically model the PDF with the oral data. However, there is also a trick in the proposed, there might have discrete features in the data, while the features generated by the stochastic gradient descent must be continuously differentiable, so this part of the features must be selected before vae training using formula (9), and after generating the continuous features, use 1-NN to classify the generated continuous and combine the continuous features with the discrete features of the nearest original sample into a new composite sample.

We don't have enough information about whether a feature is discrete or not, so we assume that it is a discrete feature if there are no more than 2 distinct values in all the feature values. In fact, it is useless in classification if there is only one distinct value among the whole dataset.

Given training dataset $X = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$, $x_i \in R^d$ is the sample of d dimension, $y_i \in \{0,1\}$ is the labels represent negative and positive. We use P and N to represents a positive class sample subset and a negative class sample subset, where P contains N_+ positive samples, N contains N_- negative samples, and $N_+ + N_- = N$.

During the training of the VAE model, $nelements_j$ represents the number of distinct feature values in j_{th} dimension in the positive subset, the formula is shown as (7):

$$nelements_j = \sum_i^{N_+} distinct\{x_{ij}\}, 1 \le j \le d \quad (7)$$

$$x_i = \{x_{i1}, x_{i2}, \cdots, x_{ik}\} \cup \{x_{i(k+1)}, \cdots, x_{id}\} \quad (8)$$

$$s.t. \begin{cases} nelements_j > 2, & 1 \le j \le k \\ nelements_j \le 2, & k+1 \le j \le d \end{cases} \quad (9)$$

If $nelements_j$ is no more than 2, the feature j is discrete, otherwise, the feature is continuous. Divide the features in the positive subset into continuous and discrete features in order and the continuous features are used as the final training set.

$$Xtrainvae = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{N_+1} & \cdots & x_{N_+k} \end{bmatrix} \quad (10)$$

Train a VAE model with $Xtrain$ and randomly sample it, assume $Xnew$ is synthetic a sample:

$$\begin{cases} Xfinal_{ij} = Xnew_{ij} \cup X_{lm}, k+1 \le m \le d \\ s.t.\ agrmin \sum \left| |Xnew_{ij} - X_{lj}| \right|^2, 1 \le j \le k \end{cases} \quad (11)$$

X_{final} is the final synthetic sample, and $X \cup X_{final}$ is the final training set called X_{ov}.

Algorithm 1: Oversampling using VAE

Input: A dataSet X,sampling rate k
Output: x_{ov} after oversampling

$X \leftarrow \frac{X-\overline{X}}{s}$
$X_{train}, X_{test} \leftarrow \text{divide}(X)$
for *each feature* j *in* X **do**
    $nelements_j \leftarrow \sum_1^{N_+} distinctX_j$
end
Decide $X_{trainvae}$ with formula (10)
$vae \leftarrow \text{trainvae}(X_{trainvae})$
$X_{new} \leftarrow \text{sample}(vae)$
Synthesize X_{final} with formula (11)
$X_{ov} \leftarrow X_{final} \bigcup X_{train}$
return X_{ov}

The whole process is described as Algorithm 1, firstly, normalize the dataset to scale the range of data, and divide(X) is a function which can split the dataset as training set and testing set, in imbalanced classification, to keep the distribution unchanged in these subsets, the positive and negative samples are split separately. Secondly, choose the features with over two distinct values and use them as the $X_{trainvae}$. Thirdly, train a VAE model and sample from the trained model, suppose the generated samples as X_{new}. Finally, add the discrete features for the generated samples using their nearest neighbors' discrete features, and these are X_{final}.

4 Experiment

4.1 Dataset and Evaluation

In this paper, all datasets are from UCI [20] Machine Learning Repository, and some of them are multi-label datasets, so we select one class as the minority class and the remaining samples as majority class. The missing values are supplemented by the most frequent value. After that we use normalization, the formula is shown in (12):

Table 1. Dataset.

Index	Dataset	Samples	Attributes	Minority	Imbalance ratio
1	breast-w	699	9	241	1.90
2	vehicle	846	18	199	3.25
3	segment-challenge	1500	19	205	6.32
4	Diabetes	768	8	268	1.87
5	Ionosphere	351	34	126	1.79

$$x_{inew}=\frac{x_i-\bar{x}}{s} \tag{12}$$

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i, s=\sqrt{\left(\frac{1}{n-1}\sum_{i}^{n}(x_i-\bar{x}\right)^2} \tag{13}$$

In traditional classification method, global accuracy is used as the evaluation, but in the imbalanced problem, this evaluation will mask the classification performance of the minority class. In extreme conditions, assume the dataset only contain 1% minority class, if the classifier decides all samples as majority class, the accuracy still can reach 99%, and however, the recognition rate of the minority samples is 0. In binary imbalanced classification, the confusion matrix in Table 2. Confusion metrics. is often used to evaluate the performance of the classifier.

Table 2. Confusion metrics.

	Positive prediction	Negative prediction
Positive class	True positive(TP)	False negative(FN)
Negative class	False positive(FP)	True negative(TN)

Among them, FN is the number of the positive samples misclassified as negative, and FP is the number of the negative samples misclassified as positive. There are some new evaluation metrics based on confusion matrix to calculate the accuracy and recall of imbalance data such as F-value, G-mean [21].

$$precision=\frac{TP}{TP+FP} \tag{14}$$

$$recall=\frac{TP}{TP+FN} \tag{15}$$

$$F\text{-}value=\frac{(1+\beta^2)\times recall\times precision}{\beta^2\times recall+precision} \tag{16}$$

Where $\beta\in[0,+\infty]$.

$$Gmean=\sqrt{\frac{\mathrm{TP}}{\mathrm{TP+FN}}\times\frac{\mathrm{TN}}{\mathrm{TN+FP}}} \tag{17}$$

In this experiment, we choose $\beta=1$ for F-value, it is the average between recall and precision. Gmean is the geometric mean of the classification accuracy of the minority class and majority class. Only when the precision of minority class and precision of majority are high at the same time, gmean will be maximum.

4.2 Experiment results

In this paper, we compare different oversampling algorithms such as NDO-sampling [17] and random interpolation algorithm SMOTE [12] (SMO). The classifier is naïve Bayes to reduce the impact of classifier's parameters on classification performance. To reduce the randomness in the final results, each algorithm calculates the average of 10 times with 10-fold cross-validation. The results of NDO are from the corresponding

paper, and k=5 in SMOTE, the structure of the proposed method is shown in Fig. 1, we use the random sample in generating new samples.

The result shown in the Table 3 indicate that vae performs better in generating samples than NDO and SMOTE when the number of oversampling is the same, as the VAE can generate more reasonable samples with more information. With the growth of oversampling rate, all the sampling methods can help to improve the classification performance, which indicate that the original minority samples don't contain enough information for a classifier to recognize them correctly from the negative samples.

In the meanwhile, from the result in Table 4, compared with the traditional oversampling algorithms which sacrifice some majority samples to ensure the classification performance of minority, the proposed method can guarantee the rational distribution of synthetic samples and improve the classification performance of the majority samples, which indicates a stronger classifier.

The proposed method can produce more reasonable samples, which can be concluded form the result shown in Table 5, as the classifier trained with the samples generated by the proposed method can get a better overall classification performance, as the $Gmean$ is the geometric mean of the accuracy of the minority samples and the majority samples, and with a higher oversampling rate, the classifier gets a best result with the proposed oversampling method.

The experimental results also show that for all oversampling methods, a higher oversampling rate can lead to a better classification performance, when the minority samples after oversampling are equal to the majority ones in size, the best classification performance is reached, this indicates that the size has limited effect on the classification performance, more informative samples and stronger classifier play a bigger role.

Table 3. F1-min of different algorithms and oversampling rate.

	100%			200%			300%		
	VAE	NDO	SMO	VAE	NDO	SMO	VAE	NDO	SMO
1	**94.65**	94.38	94.38	**94.63**	94.38	94.38	**94.67**	94.38	94.38
2	**58.22**	55.66	56.26	**58.12**	56.63	56.45	**58.78**	56.27	56.45
3	**66.88**	65.47	62.44	**69.45**	66.55	61.15	**71.21**	61.90	61.15
4	65.51	65.93	**66.27**	**66.96**	66.74	66.33	**66.34**	65.59	66.33
5	**87.00**	82.34	80.54	**85.89**	82.63	82.71	**84.26**	81.44	82.71

Table 4. F1-maj of different algorithms and oversampling rate.

	100%			200%			300%		
	VAE	NDO	SMO	VAE	NDO	SMO	VAE	NDO	SMO
1	**96.95**	96.89	96.89	**96.94**	96.89	96.89	**96.96**	96.89	96.89
2	**74.50**	72.06	72.35	**75.90**	71.98	71.87	**77.82**	71.73	71.90
3	91.68	**91.79**	89.69	**92.78**	92.22	89.41	**93.43**	92.47	89.03
4	79.78	79.73	**80.25**	77.44	**77.78**	76.71	**76.29**	74.95	74.48

5	**93.70**	89.62	87.79	**93.42**	89.84	88.56	**92.82**	90.07	89.45

Table 5. Gmean of different algorithms and oversampling rate.

	100%			200%			300%		
	VAE	NDO	SMO	VAE	NDO	SMO	VAE	NDO	SMO
1	**96. 50**	96. 35	96. 35	**96. 50**	96. 35	96. 35	**96. 55**	96. 35	96. 35
2	**75.14**	72.71	73.27	**75.19**	73.50	73.18	**75.79**	73.30	73.34
3	**90. 34**	89. 55	88. 91	**91. 40**	89. 23	89. 36	**91. 88**	89. 41	89. 01
4	73.15	73.57	**73.84**	74.04	**74.07**	73.01	**73.35**	73.24	73.03
5	**88.54**	86.47	85.32	**87.46**	86.66	85.99	86.07	86.86	**86.98**

5 Conclusion

In this paper, we propose an oversampling algorithm based on VAE, in order to make full use of the distribution information in the dataset, it can generate more reasonable samples with no prior assumption of the real distribution nor the assumption that the features are independent, what's more, we separate the features into discrete and continuous ones, use the nearest discrete features as the features of the generated samples, to generate samples with real meaning as can as possible. The experiment results prove the effectiveness of the proposed method, it can improve the overall performance rather than the minority samples. The sampling is still too rough to guarantee the generated samples' impact on the classifiers, and the future work is to overcome this drawback.

Acknowledge

This study was supported by the Shenzhen Research Council (Grant No. JSGG20170822160842949, JCYJ20170307151518535, JCYJ20160226201453085, JCYJ20170307151831260).

References

1. Wang, Y., Li, X., Tao, B.: Improving classification of mature microRNA by solving class imbalance problem. Scientific Reports. (2016).
2. Stegmayer, G., Yones, C., Kamenetzky, L., Milone, DH.: High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 1–1 (2016).
3. Leichtle, T., Geiß, C., Lakes, T., Taubenböck, H.: Class imbalance in unsupervised change detection – A diagnostic analysis from urban remote sensing. International Journal of Applied Earth Observation & Geoinformation. 60, 83–98 (2017).
4. Li, C., Liu, S.: A comparative study of the class imbalance problem in Twitter spam detection. Concurrency & Computation Practice & Experience. 30(4), (2018).

5. Singh, S., Liu, Y., Ding, W., Li, Z.: Empirical Evaluation of Big Data Analytics using Design of Experiment: Case Studies on Telecommunication Data. (2016).
6. Hale, M.L., Walter, C., Lin, J., Gamble, R.F.: A Priori Prediction of Phishing Victimization Based on Structural Content Factors. (2017).
7. Zhang, C., Wang, G., Zhou, Y., Jiang, J.: A New Approach for Imbalanced Data Classification Based on Minimize Loss Learning. In: IEEE Second International Conference on Data Science in Cyberspace. pp. 82–87 (2017).
8. Provost, F.: Machine Learning from Imbalanced Data Sets 101 (Extended Abstract). In: Soft Computing and Pattern Recognition (SoCPaR), 2011 International Conference of. pp. 435–439 (2008)
9. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: International Conference on Machine Learning. pp 104. (2004).
10. Donoho, DL., Tanner, J.: Precise Undersampling Theorems. Proceedings of the IEEE. 98 (6), 913–924 (2010).
11. Olken, F., Rotem, D.: Random sampling from databases: a survey. Statistics & Computing. 5 (1), 25–42 (1995).
12. Chawla, NV., Bowyer, KW., Hall, LO., Kegelmeyer, WP.: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 16(1), 321–357 (2002).
13. Zhang, C., Guo, J., Lu, J.: Research on Classification Method of High-Dimensional Class-Imbalanced Data Sets Based on SVM. In: IEEE Second International Conference on Data Science in Cyberspace. pp. 60–67. (2017)
14. Han, H., Wang, W.Y., Mao, BH.: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Lecture Notes in Computer Science. 3644 (5), 878–887 (2005).
15. Gao, M., Hong, X., Chen, S., Harris, CJ.: Probability density function estimation based over-sampling for imbalanced two-class problems. In: International Joint Conference on Neural Networks. pp. 1–8. (2012)
16. Chen, S.: A generalized Gaussian distribution based uncertainty sampling approach and its application in actual evapotranspiration assimilation. Journal of Hydrology. 552 (2017).
17. Zhang, H., Wang, Z.: A normal distribution-based over-sampling approach to imbalanced data classification. In: International Conference on Advanced Data Mining and Applications. pp. 83–96. (2011)
18. Li, DC., Hu, SC., Lin, LS., Yeh, CW.: Detecting representative data and generating synthetic samples to improve learning accuracy with imbalanced data sets. Plos One. 12(8), (2017).
19. Diederik P, K., Max, W.: Auto-Encoding Variational Bayes.
20. Amini MR, Usunier N, Goutte C, http://archive.ics.uci.edu/ml/datasets.html, last accessed 2018/03/22.
21. Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data Mining & Knowledge Discovery 28(1), 92–122 (2014).