Imbalanced classification based on VAE

ZhouYing
Shenzhen Graduate School
Harbin Institute of Technology
Shenzhen, China‘

***Abstract*—Classification problem is a very important part of machine learning. But in the unbalanced classification, the traditional machine learning method is easy to ignore the identification of the minority sample, resulting in a low recognition rate. In this paper, variational auto-encoder is used to expand the minority sample, and a reconstruction model using auto-encoder is proposed. The results show that the proposed algorithm can achieve better results on a specific set of data.**

Keywords—unbalanced classification; generation model; variational autoencoder; reconstruction model component

I. INTRODUCTION

The classification problem is a very important part in machine learning. At present, most of the classification problems are assuming that the number of samples of different classes are equal, but in reality, people are more concerned about the data is often scarce, for example, credit card fraud detection, medical diagnosis etc.

There are two ways to solve the traditional imbalance problem: the data level and the algorithm level. The data level is mainly sampling, including over sampling, under sampling and mixed sampling. Over sampling refers to the process of training model for minority class samples are easier to be ignored, increase the number of training process, common sampling method has simple duplicate sample resampling, the linear interpolation of the SMOTE for the sample, over sampling can effectively improve the classifier to the minority class attention, but a simple copy sample will not include additional information, and the SMOTE random interpolation method, is the lack of typical and has strong randomness. Undersampling is a method of removing large numbers of samples from a large number of samples. Undersampling can quickly reach equilibrium, but there is a risk of losing valuable samples. Mixed sampling is the unification of the above two sampling methods, and there is also the risk of losing valuable samples.

The algorithm level is mostly based on the idea of cost sensitive classification, namely for the minority class and majority class classification results of different typical punishment, the direct use of the F1 value as the cost function of the standard backpropagation, and integrated algorithm, heuristic algorithm, is more costly punishment for classification error the minority class.

Because of the resampling method in generating data is more direct copy or simple linear interpolation, did not make full use of information distribution between samples, this paper aiming at the imbalance of minority class samples in the classification of the problem is difficult to obtain, and the distribution relationship between samples, the minority class samples were expanded by VAE the model has very generation ability of neural network, and the classifier easy to ignore the minority class problem, put forward to different categories were trained in a reconstruction model, using the reconstruction error classifier, the results show that, in a particular data set, the model can effectively improve the classifier's F1 value.

II. RELATED WORK

A. *Variational autocoder*

Using neural networks to generate samples, it is assumed that the final form of the sample is determined by some potential variables, Z, and its prior distribution and posterior distribution are simulated by neural networks. The advantage of neural network model is that its output dimension is arbitrary, so it can generate data of arbitrary dimension.

According to the observation data generation model is that its distribution, and generate valid data did not appear in the observed data, a natural idea is to calculate the P in the original sample space (X) to estimate the probability of size, but because they do not know the distribution of data, so P (x) cannot be computed according to the Bayesian formula. Assuming that z is a potential variable

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

But for most of the Z, can generate reliable samples, namely P (X|z) to 0, P (X|z) P (z) to 0, in order to simplify the calculation, assuming that P (z) is the same distribution, you only need to calculate P (X / z), we used directly to P (X|z) Z large value calculation and sampling.

The coding part of the encoder can be obtained, and its P (z|X) distribution is Q (z), and the KL divergence is used to calculate the distribution fitting error.

KL divergence definition:

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

It can be seen from the equation that the divergence of KL tends to be 0. if the PQ is closer

Compute D[Q(z)||P(z|X)]

The relationship between $E_{z\sim Q}P(X|z)$ and P(X) is one of the corner-stones of variational Bayesian methods. We begin with the definition of Kullback-Leibler divergence (KL divergence or D) between P(z|X) and Q(z), for some arbitrary Q (which may or may not depend on X):

$$\mathrm{D[Q(z)||P(z|X)]} = E_{z\sim Q}[logQ(z) - \log \mathrm{P(z|X)}](3)$$

We can get both P(X) and $P(X|z)$ into this equation by applying Bayes rule to P(z|X):

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

Here, $\log P(X)$ comes out of expectation because it does not depend on z. Negating both sides, rearranging, and contracting part of $E_{z\sim Q}$ into a KL-divergence terms yields:

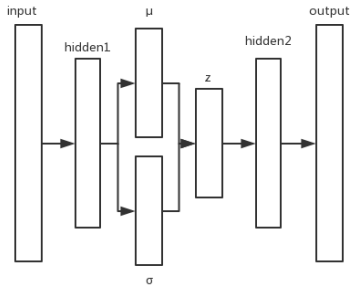
$$\log P(X) - D[Q(z)||P(z|X)] = E_{z\sim Q}\left[logP(X|z) - D[Q(z)||P(z)]\right] \quad (5)$$

Note that X is fixed, and Q can be any distribution, not just a distribution which does a good job mapping X to the z's that can produce X. Since we're interested in inferring P(X), it makes sense to construct a Q which dose depend on X, and in particular, one which makes $D[Q(z)||P(x||X)]$ small:

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

From equation (6) can be seen, as long as the $D[Q(z)||P(z|X)]$ tends to 0, it reached us maximum (x) to P. [4]

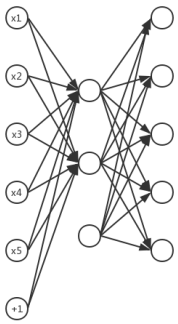
But we are on the form and distribution of the latent variables are not clear, need to use the fitting ability of neural network to the real distribution fitting, so the network structure of VAE to fit to the real hidden layer, the distribution of Z.



In the calculations we assume that Z obeys normal distribution N(μ,σ), we should start from the N(μ,σ) sampling, but the sampling operation of μ and σ is not differentiable, the conventional gradient descent method by error back-propagation (GD) cannot be used. Through reparemerization, we first from N (0,1) on sampling, then $z=\sigma\cdot\epsilon+\mu$. In this way, $z\sim N(\mu,\sigma)$, and the output from the encoder to the Z, involves only linear operation, (of neural network is constant), therefore, can be used to optimize [4] GD.

B. Auto-encoder

Since the encoder was proposed by LeCun in 1987, it was first used for dimensionality reduction or feature learning, and the network architecture is shown in fig.:



Since the encoder error loss function network input and output, the network can be regarded as composed of two parts, from the input to the hidden layer to the output of the encoder and decoder by hidden layer mapping, in addition to the development of many varieties, such as the network weights are penalized sparse since [5] encoder, request the network weights to 0 this network structure, a small range of input changes is more robust. And the noise from the encoder [6] increase in the input, adding noise in the input, since the encoder of the noise of input reconstruction, error is still the original input and output error, the network structure to the noise pollution input reconstruction, thus better network robustness.

Since the use of encoders is divided into supervised and unsupervised type, since the encoder itself is an unsupervised network, but due to the unsupervised learning result is not very good at present, so the auto encoder in the use of supervised mechanism still joined the most, there are a lot of unbalanced classification using self-made device classification most of the first layer using the method of training in pre training from the encoder, after using the mechanism of supervised fine-tuning, which add a softmax layer in the hidden layer output from the encoder, according to its output prediction standard tuning [7].

This paper adopts [7] SMOTE+ supervised self encoder architecture, the first to adopt SMOTE to carry on the data sampling data set to balance, after the use of AE reconstruction of the data, finally adding softmax layer classification. The results show that the model has some advantages in improving the effect of imbalanced classification, and the AUC values on UCI data sets have obtained relatively high test results.

In view of the task dependency model is not strong, this paper proposed [8] training at AE loss with supervised components that add a softmax layer in the model, and in training according to the classification results, adjust the parameters from the encoder. The traditional training process is layer by layer training, and finally add a softmax layer, and this article [8] in the training process directly added the softmax layer, and its classification results are directly added to the loss function.

In imbalanced data sets, it is very effective to classify the samples by using the information in the class. The document [9] uses RBM to reconstruct the images and calculate their reconstruction errors, and the accuracy is very high. In recent years, since the encoder and the latent variable model theory. From the encoder with the front edge of generative modeling, because the structure of the auto encoder is simple and easy to realize, and can add a different loss function, and denoising since

encoder has strong robustness, can have a good fitting effect of scattered distribution the data in the training process of this paper we use denoising from encoder form reconstruction model, and use it to classify.

The traditional self coder minimizes the following objectives

$$L(x, g(f(x))) \quad (7)$$

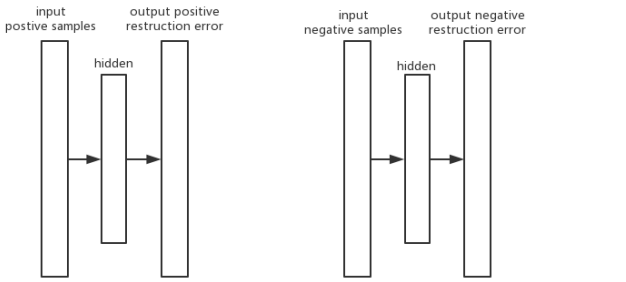
Among them, L is a loss function that punishes the difference between G (f (x)) and X, commonly used in the L2 norm. If the model is given too much capacity, L simply makes g (f (x)) an identity function;

In contrast, denoising minimizes the self coder

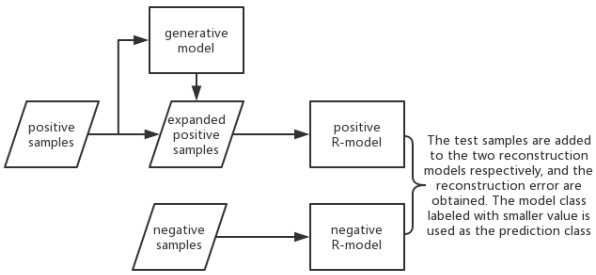
$$L(x, g(f(\tilde{x}))) \quad (8)$$

The X is a x noise pollution, therefore, denoising since the encoder must restore the original x, which is not a simple copy input.

The structure is as follows:



In this article, we mainly analyze the imbalance in the two categories. Therefore, we have trained two models, which are the positive sample model and the negative class model. The classification process is as follows:



Algorithm description:

The original data is divided, and the training set and test set are segmented

Separate training positive samples and negative samples, then input the positive samples into the generative model, generate the same number of positive classes, generate samples, and add them to the original positive samples

The positive samples and the negative samples are input into the reconstruction model, and the corresponding class reconstruction model is generated

The test samples are input into the two models respectively, and their reconstruction errors are calculated respectively

In the voting process, weighted voting mechanism can be adopted to obtain better minority class recognition performance

In the generative model, the traditional VAE is used as the generation network, and the network structure is shown in Figure 1

Since the objective function tends to minimizing the overall error in the classification algorithm in the traditional rate, but the imbalance in the sample classification, samples of minority class mistakes cost more, in the literature [10] in unbalanced classification, the threshold is set to 0.5 is not reasonable, and put forward an optimum classification threshold the architecture of novel, at the same time, the paper also presents in the extreme case, the reliability of the AUC value, so in the analysis of experimental results, the prediction index, using only the F-value and gmean to analyze the overall classification at the same time, the results in this paper taking a weighted voting mechanism, and compared in different threshold, the performance of the algorithm.

III. Experiment and Analysis

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Data set description

The experimental data from the UCI machine learning database are two commonly used imbalanced data sets, which are described in detail in the table.

Table 1 data set description

Dataset	Total samples	Total attributes	Minority class	Majority class	Imbalance rate
Ionosphere	351	34	126	225	1:1.8
german	1000	24	300	700	1:2.3
wpbc	198	33	47	151	1:3.2

B. Evaluation index based on confusion matrix.

In the two category, confusion matrices are often used to evaluate the performance of classifiers, which are defined as follows.

Table 2 confusion matrices for two classification problems

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

Among them, TP represents the positive class are correctly predicted, judged to be positive, TN is negative is correctly predicted the number of samples is negative, FN said it is wrong decision is negative, while FP negative is the wrong judgment is negative type of situation, the traditional classification methods. Usually used to overall accuracy rate as the evaluation index, and the imbalance problem, because of the small number of positive class, the overall accuracy rate as the evaluation index

will cause the classifier is not sensitive to the minority class, in extreme cases, if the data set contains only 1% of the minority class, if the classifier will determine all samples for the majority of all class, the overall accuracy rate can still reach 99%, but this is very detrimental to the minority class we care about, so the traditional classification algorithms will lead to the minority class easily into the majority class, leading to the recognition rate of the minority class At present, there are some new classification indexes for imbalanced data. According to the confusion matrix, the accuracy and recall rate can be calculated, such as AUC, F-value and G-mean.[11]

F-value is a classification evaluation index to measure the accuracy and recall rate. It is more biased to evaluate the classification performance of minority groups, as defined below:

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

Among them, the accuracy rate $\mathrm{precision} = \frac{TP}{TP+FP}$, recall rate $\mathrm{recall} = \frac{TP}{TP+FN}$, β value as [0,+∞].In this experiment, β =1, The F-value at this point represents the average of the weights between recall and accuracy.

G-mean represents the geometric average of the minority class classification accuracy and the majority class classification accuracy, which is used to evaluate the overall classification performance of the classifier:

$$\mathrm{G-mean} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$$

G-mean indicates that only when a few classes and most class classifications have high accuracies, G-mean is the most valuable at this point.

C. *Experimental results and analysis*

a.

The contrast algorithm of this paper is the SMOTE-SDAE algorithm implemented in this paper, and the traditional naive Bayes algorithm, the algorithm is set as follows:

This paper uses the 10- fold cross validation, the minority class and majority class and division, to ensure the data distribution is consistent with the original distribution, this paper uses the algorithm of VAE is the three layer of hidden layer settings, including the hidden layer potential variable Z, and the reconstruction model is a layer of hidden layer DAE classification using the reconstruction error, but also the reconstruction error of the minority class and majority class with weighted voting mechanism.

Comparison of F1 values of different algorithms %

	ionosphere	german	wpbc
SMOTE-SDAE	78.42	**81.92**	83.71
VAE-R	**92.29**	47.59	**86.98**
Naïve Bayes	90.99	58.07	76.38

Comparison of gmean values of different algorithms %

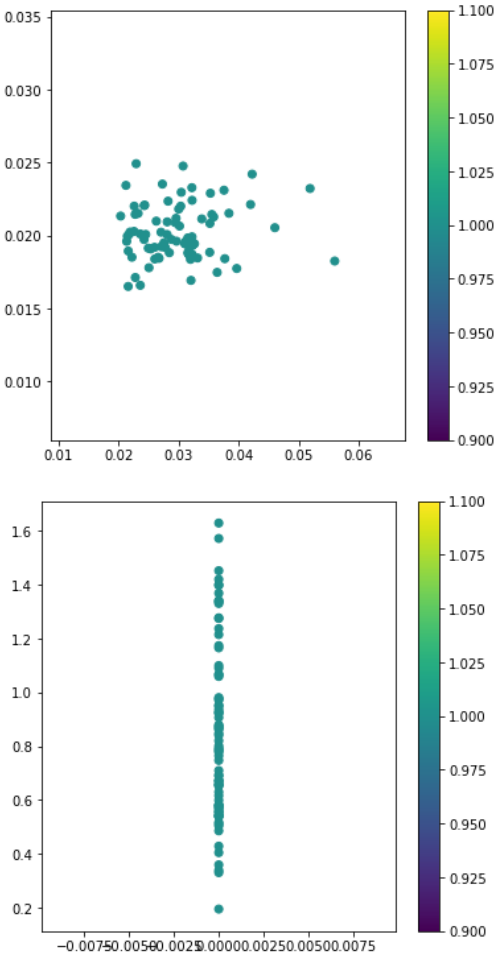
	ionosphere	german	wpbc
VAE-R	**86.08**	35.34	9.1
SMOTE-SDAE	76.15	**84.50**	**65.24**

From the experimental data we can see that in the premise of adding more effective generating sample under the proposed classifier can get good results, but the classifier tends to divide samples into the minority class, so in the case of small sample size, most easily all misclassified, gmean value is very easy for 0.

The proposed algorithm in this paper is the total number of samples under the condition of less, can get good results, that minority class is relatively scarce, so the ionosphere and wpbc data set the F1 value is relatively high, can reach the current level of state-of-art, but this kind of concentrated in German the large number of data, the capacity of the reconstruction model the algorithm is not enough to fit the entire data set, will cause the random classification.

D. *Influence of sample quality on classification results*

The experiment is based on the ionosphere data set, since Q (z) is assumed to be Gauss distribution in VAE, P (z|X) can obtain better results in the case of Gauss distribution, and can achieve relatively high results both in F1 and G-mean.



In Figure 1 the test results for F1=0.84, figure 2 Results for F1=0.79, as can be seen, although whether Q in hypothesis (z) to the distribution form, can through the neural network map it into the latent variable distribution of real, but actually because

of the capacity of neural network, and neural network is easy to fall into the local minimum value, not necessarily will be Q (z) mapping without bias to the latent variable distribution of real, thus will cause the generated sample of poor quality problems.

IV. CONCLUSION

In this paper, we use the generative model to replace the traditional resampling mechanism, which can make full use of the data centralized information, at the same time, the characteristics of different categories of samples from the encoder to obtain, according to the reconstruction error of classification, the sample in this way, the classifier is two relatively independent from the encoder and can effectively prevent the classifier to the minority class ignored, but this algorithm, weighted voting mechanism in the impact will be very large, in the final results at the same time, the weighted, the algorithm will tend to be most types of error into the minority class, performance decline will result in quality of classifier. Generate sample criteria, there are still many problems, can only rely on the classifier to measure the results of generating sample quality in future research should be changed in.

REFERENCES

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, no. 1, pp. 321–357, 2011.G.

[2] D. L. Donoho and J. Tanner, "Precise Undersampling Theorems," Proceedings of the IEEE, vol. 98, no. 6, pp. 913–924, 2010.

[3] I. Bilinskis, Hybrid Sampling. John Wiley & Sons, Ltd, 2007.

[4] C. Doersch, "Tutorial on Variational Autoencoders," 2016.

[5] A. Ng, "Sparse autoencoder," 2011.

[6] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in International Conference on Machine Learning, 2008, pp. 1096–1103.

[7] 张成刚, 宋佳智, 姜静清, and 裴志利, "一种改进的降噪自编码神经网络不平衡数据分类算法," 计算机应用研究, no. 5, pp. 1329-1332, 2017.

[8] Y. Sun, H. Mao, Y. Sang, and Z. Yi, "Explicit guiding auto-encoders for learning meaningful representation," Neural Computing & Applications, vol. 28, no. 3, pp. 429–436, 2017.

[9] M. Hayat, M. Bennamoun, and S. An, "Learning Non-linear Reconstruction Models for Image Set Classification," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 37, no. 4, pp. 713–727, 2015.

[10] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, "Finding the Best Classification Threshold in Imbalanced Classification ☆," Big Data Research, vol. 5, pp. 2-8, 2016.

[11] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," Data Mining & Knowledge Discovery, vol. 28, no. 1, pp. 92–122, 2014.