# 基于vae的不平衡分类

# 摘要 简介 实验和结论

# 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 autoencoder is used to expand the minority sample, and a reconstruction model using autoencoder 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

分类问题是机器学习中非常重要的一部分，而在不平衡分类问题中，传统的机器学习方法容易忽视少数类样本的识别，造成少数类识别率很低，在本文中，利用变分自编码器对少数类样本进行扩充，并提出了采用自编码器的重建模型进行分类，结果表明，本文算法在特定的数据集上可以取得目前比较好的效果。

关键字：不平衡分类，生成模型，变分自编码器，重建模型

# Introduction

分类问题是机器学习中非常重要的一部分，目前，大多数的分类问题都是假设不同类别的样本数目是相等的，但是现实问题中，人们比较关心的数据往往是比较稀少的，例如，信用卡欺诈行为检测，医学疾病诊断等，不平衡问题有两种：一是数据真实稀有，而另一种则是数据相对稀有。在数据相对稀有的情况下，即少数类的总体样本数量较大，只是相对于多数类的样本数量来说比较稀缺，我们采用对多数类欠采样即可以获得比较好的分类效果，因为我们不缺乏训练一个好的分类器需要的信息量，而在数据真实稀有的情况下，目前的采样算法则难以发挥其优势，因为传统的采样算法并不会从现有的数据中获取额外的信息，即使是SMOTE方法，也只是针对现有数据的随机插值，并不会针对现有数据的隐含分布来进行信息扩充等。

由于深度生成模型可以在未知分布上获得质量好的样本，因此本文中针对数据真实稀有的情况下，利用生成模型生成更多的少数类样本，从而提高分类器在少数类上的分类性能。

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.

传统的不平衡问题通常的解决方案有两种思路：数据层面和算法层面，数据层面的主要是采样，包括过采样、欠采样、混合采样等。过采样是指在模型训练过程中针对比较容易被忽视的少数类样本，增加其训练次数的过程，常见的过采样方法有简单复制样本的重采样、针对样本进行线性插值的SMOTE[1]等，过采样可以有效提高分类器对少数类的关注度，但是简单复制样本并不会加入额外信息，而SMOTE这种随机插值的方法，则缺乏典型性且拥有较强的随机性。欠采样[2]是指针对大量的多数类样本，对其进行删除的方法，欠采样可以很快达到平衡状态，但可能有丢失有价值样本的风险。混合采样[3]则是对以上两种采样方式的统一，也存在着丢失珍贵样本的风险。

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. Under sampling is a method of removing large numbers of samples from a large number of samples. Under sampling 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.

算法层面的是大多是基于代价敏感的思路，即针对少数类和多数类的分类结果进行不同的分类惩罚，比较典型的有直接利用F1值作为反向传播标准的代价函数，还有各种集成算法、启发式算法等，多是针对少数类的分类错误进行更大的代价惩罚。

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.

由于目前的重采样方法在生成数据时多是采用直接复制或者简单的线性插值的方法，没有充分利用样本之间的分布信息等，本文中针对不平衡分类中的少数类样本难以获取的问题，以及样本之间的分布关系，利用神经网络中非常具有生成能力的vae模型对少数类样本进行扩充，并针对分类器容易忽视少数类的问题，提出了对不同类别分别训练一个重建模型，利用重建误差对其进行分类的分类器，结果证明，在特定的数据集上，该模型可以有效提高分类器的F1值。

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.

# Related work

## Vae

（变分自编码器）（对其进行梳理并整理成文字形式）

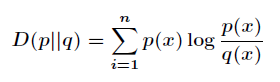
利用神经网络生成样本，假设样本的最后表现形式由某些潜在的变量决定，设其为z，并对其先验分布和后验分布用神经网络进行模拟。神经网络的生成模型优势在于其输出维度任意，因此几乎可以生成任意维度的数据。

生成模型是根据观测数据得出其分布，并生成观测数据中未曾出现的有效数据，一个自然的想法是在原始样本空间中计算P（X）以估计其出现的概率大小，但是因为不知道数据分布，所以P（x）无法计算，根据贝叶斯公式，假设z为潜在变量，

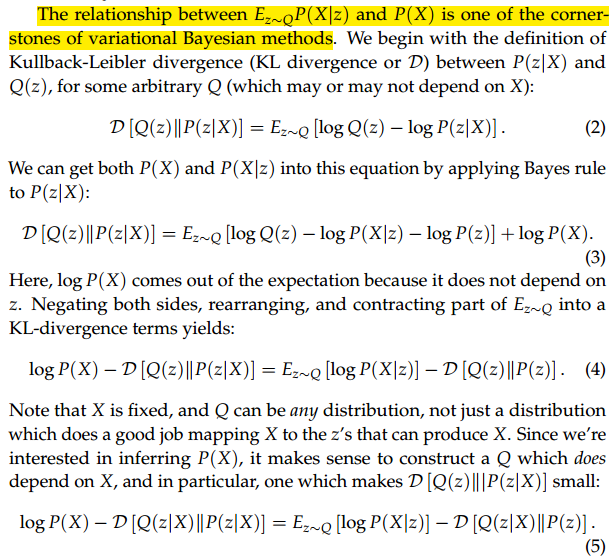
但对于大多数的z来说，都无法生成可靠的样本，即趋于0，则也趋于0，为了简化计算，假设是相同分布的，则只需要计算，我们采用直接对较大值的z进行计算和采样。

由自编码器的编码部分可得，设其产生的分布为，采用KL散度来计算分布拟合误差

KL散度定义：

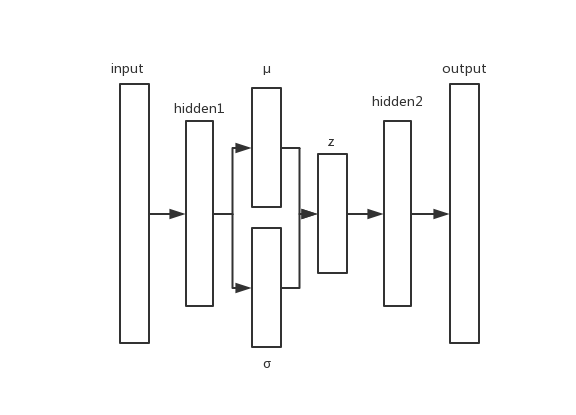
从等式中可以看出，如果pq比较接近的话，KL散度会趋于0.

计算



从等式（5）中可以看到，只要趋于0，则达到了我们最大化P（x）的目的。

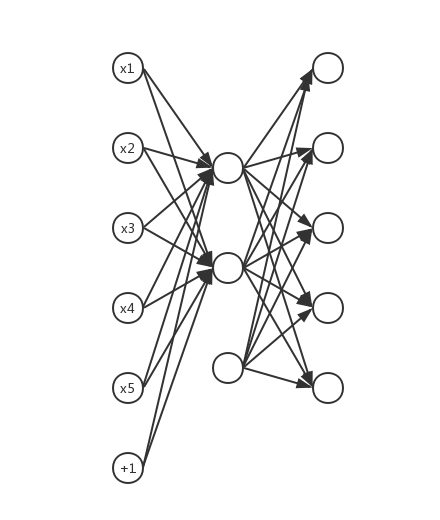
但是我们对潜在变量的形式和分布都不清楚，需要利用神经网络的拟合能力对其真正的分布进行拟合，因此vae的网络结构需要多隐藏层，以对z的真实分布进行拟合。



在计算中我们假设z服从正态分布N(μ,σ)，我们应该从 N(μ,σ) 采样，但这个采样操作对 μ 和 σ 是不可导的，导致常规的通过误差反传的梯度下降法（GD）不能使用。通过 reparemerization，我们首先从 N(0,1) 上采样 ϵ，然后，z=σ⋅ϵ+μ。这样，z∼N(μ,σ)，而且，从 encoder 输出到 z，只涉及线性操作，（ϵ 对神经网络而言只是常数），因此，可以正常使用 GD 进行优化[4]。

## 自编码器

自编码器是由LeCun在1987年提出的，最开始用于对数据的降维或特征学习，网络架构如图:



自编码器的损失函数为网络输出与输入的误差，该网络可以看成是由两部分组成，由输入到隐藏层的编码器和由隐藏层映射到输出的解码器，此外还发展出许多变种，例如对网络权重进行惩罚的稀疏自编码器[5]，要求网络权重趋于0，这样的网络结构对输入的小范围变化更具有鲁棒性。还有在输入中增加噪声的去噪自编码器[6]，在输入中加入噪声，要求自编码器对被噪声污染过的输入进行重建，误差仍为输出与原始输入的误差，该网络结构能够对噪声污染的输入进行重建，也因此网络的鲁棒性较好。

自编码器的使用分为有监督类和无监督类型，自编码器本身是无监督网络，但是由于目前的无监督学习发展结果并不很好，所以大部分的自编码器在使用时仍然加入了有监督的机制，有很多的不平衡分类采用了自编器进行分类，多数是采用先利用逐层训练的方法对自编码器进行预训练，之后采用有监督微调的机制，即在自编码器的隐层输出上加上一层softmax层，根据其输出的预测类标进行微调[7]。

本文[7]中采用了SMOTE+有监督自编码器的架构，首先对数据采取SMOTE方式进行过采样至数据集平衡，后采用AE对数据进行重建，最后加入softmax层进行分类。结果显示该模型对提高不平衡分类效果有一定的好处，并在UCI数据集上的auc值取得了比较高的测试结果。

针对模型的任务相关性不强的问题，本文[8]提出在AE的训练loss中加入有监督分量，即在模型中添加一个softmax层，并在训练中根据其分类结果，调整自编码器的参数。传统的训练过程是逐层训练，最后添加一个softmax层，而本文[8]则在训练过程中直接添加了softmax层，并且将其分类结果直接添加入loss函数。

在不平衡数据集中，利用类内的信息对样本进行分类是非常有效的，文献[9]利用RBM对图像进行重建并计算其重建误差，精度很高。近年来，自编码器与潜变量模型理论的联系将自编码器带到了生成式建模的前沿，由于自编码器的结构简单易于实现，并且可以加入不同的loss函数，且去噪自编码器具有比较强的鲁棒性，能够对分布较散的数据具有比较好的拟合效果，因此在本文的训练过程中我们利用去噪自编码器构成重建模型，并利用其进行分类。

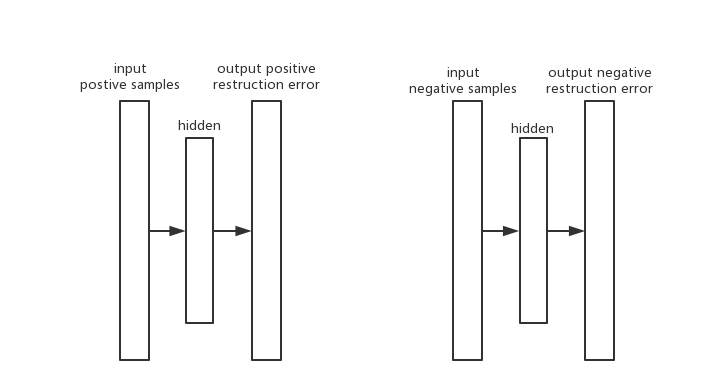
传统的自编码器最小化以下目标

其中L是一个损失函数，惩罚与的差异，常用的有范数等。如果模型被赋予过大的容量，仅仅使得学成一个恒等函数；

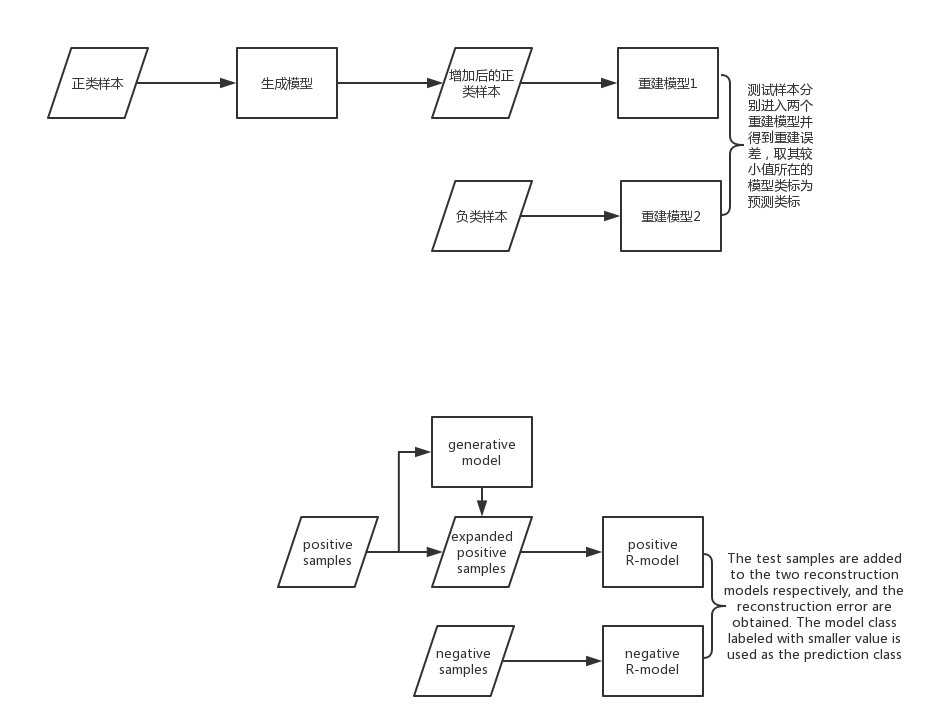
相反，去噪自编码器最小化

其中是被某种噪声污染的,因此，去噪自编码器必须还原原始的x，而不是简单的复制输入。

结构如下：



在这篇文章中，主要是分析二分类中的不平衡问题，因此本文训练了两个模型分别为正类样本的重建模型和负类样本的重建模型，分类流程如下：



算法描述：

对原始数据进行划分，分割出训练集和测试集

分开训练集中的正类样本和负类样本，将正类样本输入到生成模型中，产生相同数量的正类生成样本，并加入到原来的正类样本中

将正类样本和负类样本分别输入到重建模型中训练，分别生成类对应的重建模型

将测试样本分别输入到两个模型中，分别计算其重建误差，进行投票

投票过程中，可以采用加权投票机制，以获得较好的少数类识别性能

在生成模型中，本文采用了传统的vae作为生成网络，网络结构如图1所示

在传统的分类算法中由于目标函数倾向于最小化整体的错误率，但在不平衡样本分类中，少数类样本的错误代价更大，在文献[10]中说明在不平衡分类中，阈值设为0.5是不合理的，并提出了一个寻找最佳分类阈值的novel架构，同时，该文献还提出了在极端情况下，AUC值的不可靠性，因此在实验结果分析中，只采用了F-value和gmean等能够分析整体分类情况的预测指标，同时，本文结果采取了加权投票机制，并对比在不同的阈值下，本文算法的性能。

# Experiment

数据集描述

Data set description

文中实验数据来自于UCI机器学习数据库，都是比较常用的二分类不平衡数据集，具体描述如表中所示

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

表1 数据集描述

Table 1 data set description

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集 | 样本总数 | 属性数 | 少数类 | 多数类 | 不平衡率 |
| Ionosphere | 351 | 34 | 126 | 225 | 1:1.8 |
| german | 1000 | 24 | 300 | 700 | 1:2.3 |
| wpbc | 198 | 33 | 47 | 151 | 1:3.2 |

基于混淆矩阵的评估指标

Evaluation index based on confusion matrix.

在二分类中，常常使用混淆矩阵来评估分类器的性能，其定义如下

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

表2 二分类问题的混淆矩阵

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) |

其中，TP代表正类被正确预测，判定为正类，TN则是负类被正确预测为负类的样本数，FN则表示正类被错误判定为负类，而FP则为负类被错误判定为负类的情况，在传统的分类方法中，通常采用以整体准确率作为评价指标，而在不平衡问题中，由于正类的数量较少，则采用整体准确率为评价指标会导致分类器对少数类不敏感，极端情况下，如果数据集中仅仅包含1%的少数类，如果分类器将所有样本全部判定为多数类时，整体准确率仍然可以达到99%，但是这对我们关心的少数类是非常不利的，因此传统的分类算法会导致少数类容易被分成多数类，导致少数类的识别率较低，目前出现了一些新的不平衡数据的分类评价指标，根据混淆矩阵中可以计算准确率和召回率等，例如 AUC、 F-value 和 G-mean 等方法[11]。

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.

F-value 是衡量准确率和召回率的分类评价指标， 比较偏向对少数类的分类性能评价，定义如下：

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:

Among them, the accuracy rate ，recall rate ,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表示少数类分类精度和多数类分类精度的几何平均值，用来评价分类器的整体的分类性能，其定义如下：

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:

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.

实验结果及分析

Experimental results and analysis

本文的对比算法为论文中实现的SMOTE-SDAE算法[7]，以及传统的朴素贝叶斯算法，算法设置如下：

本文中采用10-折交叉验证，对少数类和多数类同时进行分割，以保证数据分布同原始分布较一致，本文的算法vae中采用的是三层隐藏层的设置，其中包括产生潜在变量z的隐藏层，而重建模型则是含有一层隐藏层的DAE，利用重建误差来进行分类，同时还对少数类和多数类的重建误差有加权的投票机制。

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.

不同算法的F1值比较 %

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 |

不同算法的gmean值比较 %

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** |

从实验数据中可以看到，在加入比较有效的生成样本的前提下，本文提出的分类器可以获得比较好的结果，但是该分类器倾向于将样本划分成少数类，因此在样本数量较少的情况下，多数类容易全部被误分，gmean值很容易为0。

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.

本文提出的算法在样本总体数量较少的情况下，能够得到比较好的效果，即少数类是绝对稀缺的情况，因此在ionosphere和wpbc数据集上的F1值比较高，能达到目前state-of-art水平，但是在german这种样本数量较多的数据集中，算法的重建模型容量不足以拟合整个数据集，会造成随机分类的情况。

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.

生成样本质量对分类结果的影响

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.

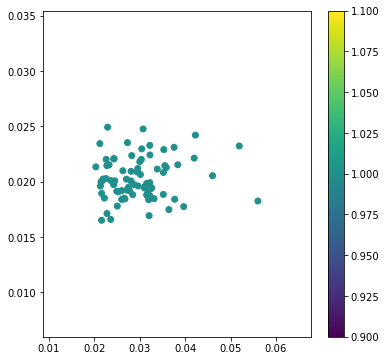
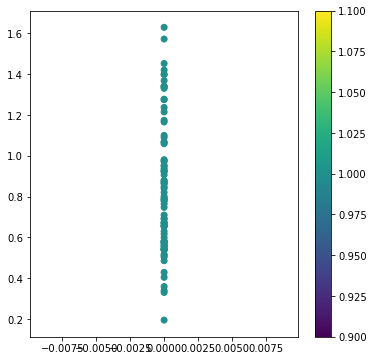
 

图1中的测试结果为F1=0.84，图2结果为F1=0.79，从中可以看出，虽然在假设中认为无论的分布形式如何，都能通过神经网络将其映射成真实的潜变量分布，但是实际上由于神经网络的容量问题，且神经网络容易陷入局部极小值，不一定能将完全无偏差的映射到真实的潜变量分布，也因此会造成生成样本的质量不佳问题。

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.

# Conclusion

在这篇论文中，我们利用生成模型来代替传统的重采样机制，这样可以充分利用数据集中的信息，同时，利用自编码器对不同类别的样本的特征获取，根据其重建误差来对样本进行分类，这样，分类器是有两个相对独立的自编码器组成，能够有效防止分类器对少数类的忽视，但是本文算法中，投票机制中的权重会在最终结果上有非常大的影响，同时，在不加权时，本文中的算法会倾向于将多数类误分成少数类，也会造成分类器的性能下降，在生成样本的质量衡量标准上仍然有很多不足，目前只能依靠分类器的结果来衡量生成样本质量，在以后的研究中应当会改进。

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.

Reference

[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.

[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.