# 基于vae的不平衡分类

# Abstract

分类问题是机器学习中非常重要的一部分，在传统的分类问题中，我们都假设不同类别的样本数量是相同的，因而每个样本的分类代价是一致的。但是在现实的数据集中，一方面我们对不同类别的感兴趣程度不一样，另一方面数据的分布也不相同，因而出现了不平衡分类问题：即在分类问题中，不同类别的样本数量和分布不同。在这类问题中，传统的机器学习方法中，每类样本的分类代价是一样的，由于少数类数量过少，为了追求全局准确率，传统分类器容易忽视少数类样本的识别，造成少数类识别率很低。一直以来，学者们为了提高少数类的准确率，采用对少数类进行过采样的方式增加数据集中少数类的数量，但是这些过采样方法都太过粗糙以至于不能很好地提高少数类的分类效果，反而因为增加了样本数量而增加训练时间。在本文中，我们提出利用少数类的分布信息，使用变分自编码器对少数类样本的概率分布函数进行拟合，合理扩充少数类样本集合，实验结果表明，本文算法是有效的。

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，由于数据集中需要关注的类别中样本数量很稀少，导致分类器对少数类关注度不够，因此识别效果不佳。

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，从采样的对象上，可以分为过采样、欠采样和混合采样等等。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。

从采样方式上，可以分为有信息的采样和随机采样，随机采样是指利用样本直接进行采样，包括简单重复、线性插值、非线性插值等；有信息的采样指的是利用样本中的分布信息，对其分布进行拟合，并根据建模内容进行采样，例如论文中涉及到的高斯分布、威布尔分布等，这些算法较随机插值方法有了比较大的进步，因为利用了样本分布信息，但是在这些方案中存在的问题同样也很明显：算法假设样本服从某种先验分布，例如正态分布等，然后利用EM算法对分布中的参数进行计算，当样本分布确实为该假设分布时，这种算法通常能获得比较好的效果，但是如果样本先验分布不服从假设分布时，这种过采样算法能够获得的提升效果往往有限，而在数据类型的数据集中，我们难以获得样本的真实分布情况，因而这种基于分布的算法效果不是很稳定。

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，例如结构化SVM，就是可以使用少数类的F1值对分类面进行改进，因而在少数类的分类效果上较好的性能。

由于目前的过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，也无法保证插值样本的合理性，本文中针对不平衡分类中的少数类样本难以识别的问题，根据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，并针对插值样本的合理性问题，提出了利用独立的第三方分类器，对生成样本进行评价，实验结果证明该模型可以有效提高分类器的F1值。

本文的贡献如下：

1. 利用vae对少数类样本分布进行建模，并利用该模型对少数类样本过采样，解决分类问题中的不平衡问题。
2. 对vae中采样生成的少数类，本文中提出了有效的机制进行评价，以避免加入不好的样本进入到原始样本集中，提高了分类性能，并减少过拟合的风险。

本文的结构如下：2中介绍了本文的相关工作，主要是vae等；3中介绍了本文方法及对该方法的分析；4中介绍了基本数据集和实验结果等；5中为论文结论和将来的扩展。

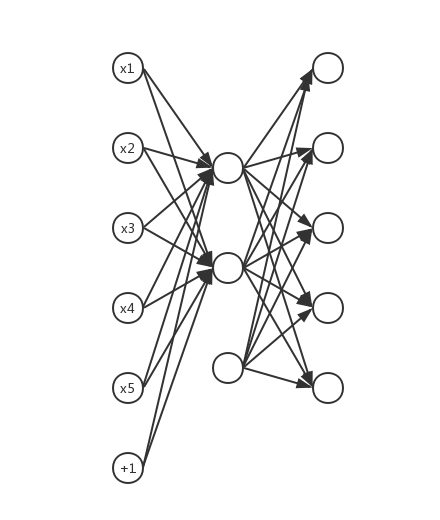
# Related work

GAN和vae作为两种最流行的生成方式，GAN将生成样本和真实样本做对比，将分类器的性能表现作为生成样本质量的衡量指标，得到了广泛好评，但由于GAN中真实样本时作为参考数据直接输入到网络中的，因而无法对GAN的generator本身性能有比较好的衡量。

即无法直接计算与某个真实样本对应的伪样本，而这在无法可视化的数值型数据库是比较受影响的，而vae是利用ae结构解决了这一问题，可以直接生成原始样本，因此，本文中采用vae结构对少数类样本分布进行拟合，并对模型进行采样，以解决少数类的过采样问题。

## Vae

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



自编码器由两部分组成，编码部分和解码部分，假设编码部分构成函数f(x)，而解码部分构成函数g(x)，隐层空间为z，则自编码器的目标函数为

自编码器中的隐层变量z可以认为是原始空间中变量的压缩编码，由于实数的无限性，理论上该类形式的编码可以存储无限信息。但这些信息在原始的自编码器中，都是由原始的X决定的，无法随意生成，为了能够生成原始样本中没有出现的样本，2013年KM对其进行改进，将变分推断加入到自编码器中，并提出了重参数化的机制，使得变分推断可以和随机梯度下降相结合，因而产生了vae，在vae中，我们假设变量是由隐层的压缩编码z决定的，而该压缩编码z服从某种特定的分布（如高斯分布等），因此在已知该特定分布的情况下，可以根据其CDF进行采样，理论上可以产生无限样本。

Vae的结构如图所示：

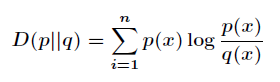
利用神经网络生成样本，假设样本的最后表现形式由某些潜在的变量决定，设其为z，并对其先验分布和后验分布用神经网络进行模拟。神经网络的生成模型优势在于其输出维度任意，因此几乎可以生成任意维度的数据。

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，

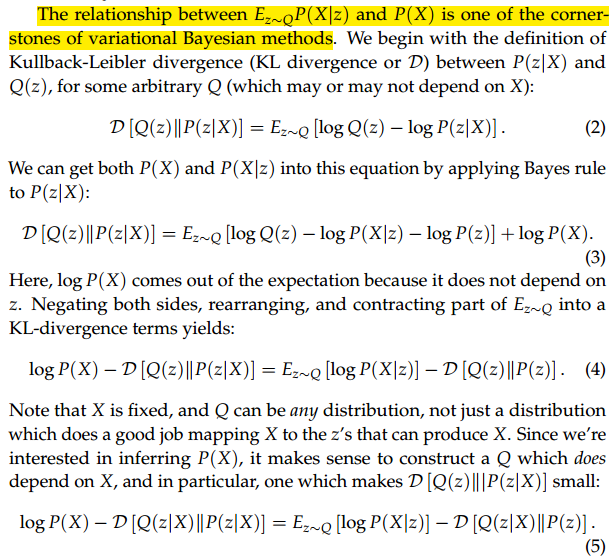
但对于大多数的z来说，都无法生成可靠的样本，即趋于0，则也趋于0，为了简化计算，假设是相同分布的，则只需要计算，我们采用直接对较大值的z进行计算和采样。

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：

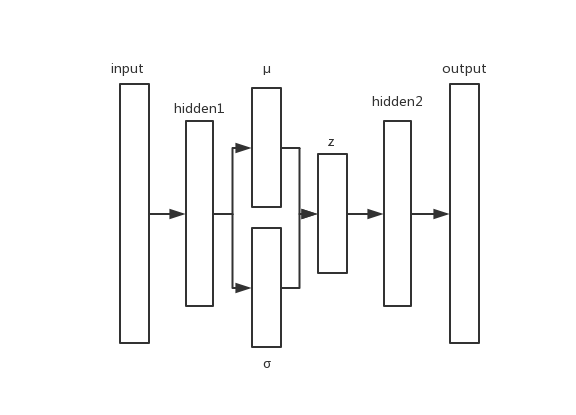
从等式中可以看出，如果pq比较接近的话，KL散度会趋于0.

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，产生一定数量的正类生成样本，并根据选择机制将生成的样本筛选，得出信息量较大的样本。

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。

# Method

由于原始的vae是应用于图像生成中，因此可以天然可视化；但本文中直接使用了序列数据，因而需要对其进行比较多的改进：

首先，由于序列数据中可能存在着离散特征，而vae中使用的随机梯度下降使得所生成的特征必定是连续可微分的，因此在vae开始训练前，需要将这部分特征进行筛选，并在生成后对生成的样本进行1NN分类，将其最靠近的原始样本的离散特征部分和生成的连续特征组合成新的合成样本；在机器学习中，因为我们是假设所有的样本都是独立同分布的，因此在样本生成后，由于无法保证这些随机采样的样本均是正类样本，所以我们需要对这些生成样本再进行筛选，筛选机制是利用无类标的迁移学习，将那些能够提高分类性能的合成样本筛选出来，加入到最后的生成样本中。该算法的流程图如下：

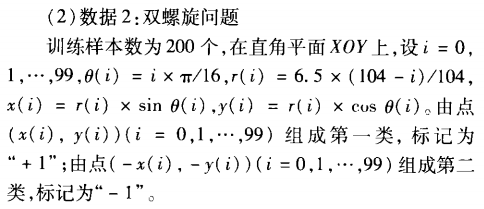
由于样本数量稀少，无法可靠判断某个特征是否为离散特征，因此本文中认定已经出现的特征值少于2个的，均认定其为离散变量，实际上，如果某个特征在各类样本中只有1个值的话，该类特征对于分类是没有作用的。

方法验证

从生成数据集中选取二维数据，进行可视化验证，以证明生成的样本确实有可取之处。

例如那个螺旋数据集等。

训练样本数为200个，在平面直角XOY上，设i=0,1···，99，（容后在写）



# Experiment

数据集描述

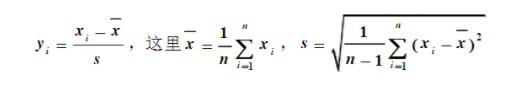
文中实验数据来自于UCI机器学习数据库，都是比较常用的二分类不平衡数据集，具体描述如表中所示

表1 数据集描述

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |

数据预处理

在UCI数据集中，有些是多类标数据集，为了追求高的不平衡率，我们选中其中某一个类别为少数类，其余的样本均归为多数类，同时，数据集中存在着值缺失的情况，为了保证数据集的完整性，我们使用最频繁出现的属性值作为该缺失属性的补充结果。针对属性值未在某个范围内的数据集，我们采用了归一化，将其进行缩放，公式如下所示：



基于混淆矩阵的评估指标

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

表2 二分类问题的混淆矩阵

|  |  |  |
| --- | --- | --- |
|  | 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 F-value 和 G-mean 等方法[11]。

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：

其中，准确率，召回率,β取值为[0,+∞]。本实验中取β=1，此时的F-value表示召回率和准确率之间权重的平均。

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表示只有在少数类和多数类分类精度同时都高的情况下，此时，G-mean的值最大。

实验结果及分析

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。

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。

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。

生成样本质量对分类结果的影响

该实验基于ionosphere数据集，由于vae中假设是高斯分布，因此在符合高斯分布的情况下，能够获得比较好的结果，在F1值和G-mean都能取得比较高的效果。

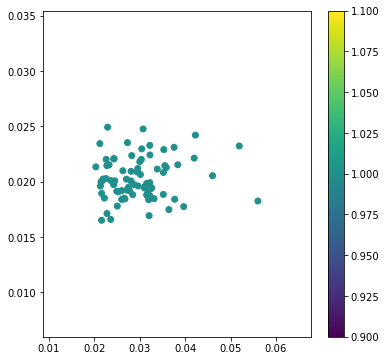
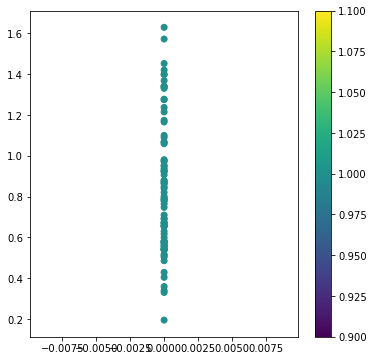
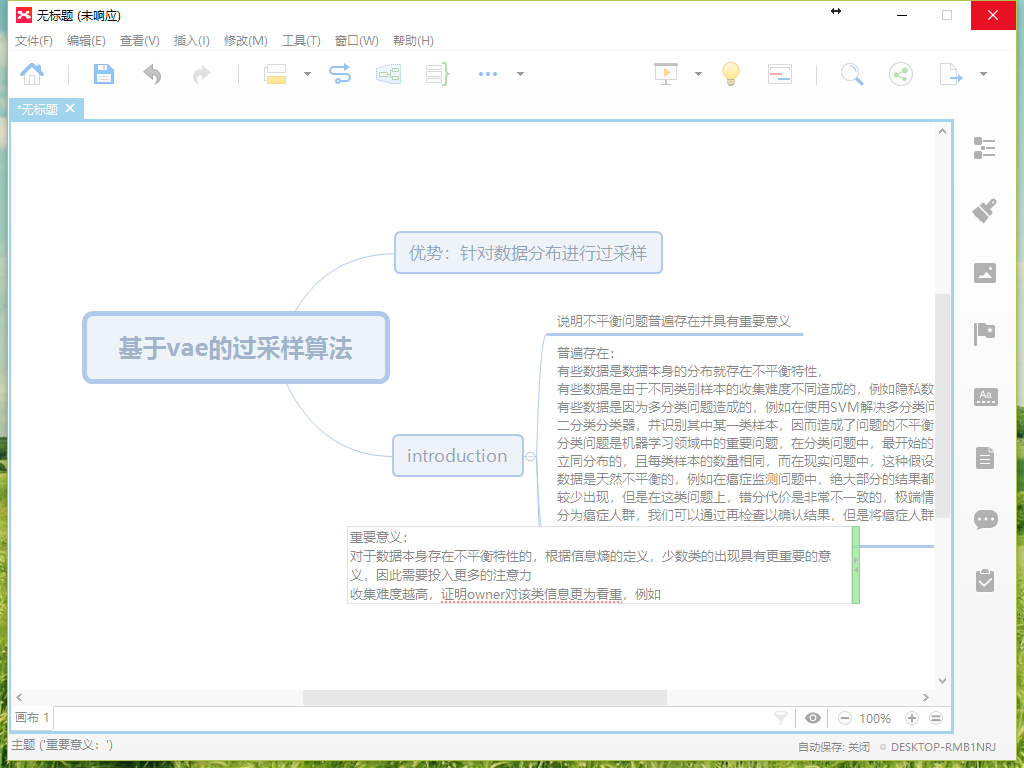
 

图1中的测试结果为F1=0.84，图2结果为F1=0.79，从中可以看出，虽然在假设中认为无论的分布形式如何，都能通过神经网络将其映射成真实的潜变量分布，但是实际上由于神经网络的容量问题，且神经网络容易陷入局部极小值，不一定能将完全无偏差的映射到真实的潜变量分布，也因此会造成生成样本的质量不佳问题。



# 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。

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.