# 基于vae的过采样算法在不平衡分类中的应用

# Abstract

不平衡分类问题是指在分类问题中，不同类别的样本数量和分布、分类代价不一致，而在传统的分类算法中，却是假设样本均匀分布，从而产生的感兴趣的类别样本识别效果不佳，一直以来，学者们为了提高少数类的准确率，提出了许多不平衡分类算法，数据层面的算法作为预处理的可选部分，具有更广泛的应用，从而引起了非常多的关注，过采样算法中利用各种方法增加训练集中的少数类样本，以增加少数类的识别率。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，实验结果证明了本文算法的有效性。

关键字：不平衡分类，生成模型，变分自编码器

# Introduction

分类问题是机器学习中非常重要的一部分，也是人工智能理解人类生活的第一步。目前，大多数的分类器的设计中，都是假设不同类别的样本均匀分布的，分类代价也是一致的。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。在医疗数据疾病诊断中，大部分的诊断结果为正常，只有少部分的样本诊断为疾病，首先印证了数据分布的不一致；其次，如果将健康人群误诊为疾病，可以通过其他的检查方式去除这种错误，不会造成非常严重的事故，但如果将疾病诊断为健康，则可能会导致病人错过最佳治疗时间而造成严重后果，这就是不平衡分类的第二个特征：不同类别的分类代价不一致。同时，如果因为害怕错过疾病样本而将所有的样本都尽可能分类为疾病，则会造成大量的医疗资源的浪费，加剧医患矛盾等，因此将所有的样本全部判断为正类是不可行的，而只能尽量将两者分开，且减少疾病的误诊率。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。不平衡分类问题出现在许多领域中，例如生物信息学[1], [2]、遥感图像识别[3]、网络安全中的隐私保护[4]等。不平衡问题的覆盖范围广，具有十分重要的现实意义。

传统的不平衡问题通常的解决方案有两种思路：算法层面和数据层面。算法层面的主要是针对不平衡问题的第二个特征的研究工作，因为不同类别的分类代价不一致，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，改进的神经网络[5]：利用近似少数类F1值作为反向传播标准的代价函数；各种集成算法bagging[6]等，不断对错分的少数类进行加权，提高分类器对少数类的关注程度；结构化SVM[7]，使用少数类的F1值作为优化函数，因而在少数类的分类效果上较好的性能。

数据层面的研究工作是针对不平衡分类问题的第一个特征，主要是通过数据重采样调整数据的分布，以减轻不平衡对分类结果的影响。从采样的对象上，可以分为过采样、欠采样和混合采样等。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[8]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，则随机性过强且缺乏典型性。欠采样[9]是指删除混淆分类的多数类样本，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[10]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。

过采样方法从采样方式上，可以分为随机采样和有信息的采样：

随机采样是指利用样本直接进行采样，包括简单重复[11]、线性插值[12]、非线性插值[13]等；SMOTE[12]作为一种经典的过采样算法，为了增加在随机过采样中的合成样本的信息量和合理性，在少数类样本中进行线性插值，提高了少数类样本的分类效果。Border-line-smote[14]在其基础上，筛选了需要插值的少数类样本，主要针对边界样本做线性插值，降低了过拟合风险。以上的这些过采样方式均只考虑了样本数量对于分类效果的影响和局部的样本分布，而未能考虑样本的整体分布。

有信息的采样[15]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， Chen[16]等提出了基于分布的过采样算法，对少数类的概率分布函数进行建模，利用此模型进行过采样，实验结果证明，该算法的过采样效果比SMOTE和随机过采样的结果要好。不同的学者们提出了基于不同分布的过采样算法，例如论文中涉及到的高斯分布[16], [17]、威布尔分布[18]等，因为利用了样本分布信息，这些算法较随机插值方法有了比较大的进步。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，当样本分布符合为该假设分布时，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对不同的数据集的分类提升效果是不一致的。

数据层面的处理对于不平衡分类来说是非常重要的，不论实际问题中样本重要程度是否一致，数据层面的处理对最终的分类结果都会有比较正面的影响。由于影响数据集分类效果的因素不仅包含样本数量，样本分布也起到了非常重要的作用，而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，实验结果证明了算法的有效性。

本文所提出的算法利用vae对少数类样本分布进行建模，神经网络的输出维度不受限制，可以生成任意维度的数据，其强大的拟合能力，可以模拟任意的分布函数而不需要提前设定未知分布的概率分布函数，使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。

本文的结构如下：2中介绍了本文的相关工作；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。

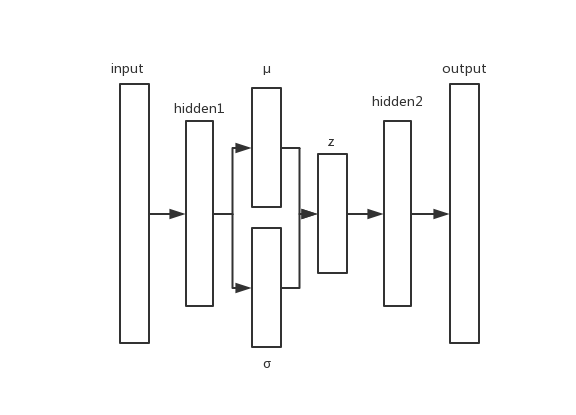
# Related work

本文中提出利用神经网络对少数类样本的概率分布函数进行拟合，由于神经网络的拟合能力很强，因而可以避免假设分布和样本的原始分布不一致的情况，提高生成样本的质量，并提高最终的分类效果。

Vae(variational auto-encoder)

2013年KM[19]提出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，的网络的整体结构是编码器-解码器形式的，而在中间的隐藏变量中，则假设其为正态分布，这一特点恰好符合基于分布的过采样的特点：易于采样而最终的概率分布函数形式不定。在vae中，假设变量是由隐层的压缩编码z决定的，跟在其后的encoder可以将z映射成最终的可见形式，而该压缩编码z服从某种特定的分布（如高斯分布等），在已知该特定分布的情况下，可以根据其CDF进行采样并通过encoder部分，理论上可以产生无限样本。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的结构如图所示：



假设z为潜在变量，其概率分布函数为，利用贝叶斯条件概率公式计算得：

但在z的先验分布中，大多数的z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，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，在自编码器的encoder中，映射出，这部分的z是能够直接再通过decoder映射回X的，但是如果只是计算这部分的z，则无法生成原始数据中没有的样本，所以需要假设的分布形式，其误差则通过decoder部分补全。

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。Vae网络模型的目标函数为

Apply Bayes rule to, we can get both and

Apply the into it

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 does depend on X, and in particular, one which makes small:

因为是固定的，所以最小化转变为最大化等式右边的值，其中为z通过decoder后生成的X的概率，计算为同原始样本的交叉熵或者均方误差，后者可以理解为的假设先验分布同encoder中映射出的z的分布的误差。

# Method

本文中针对不平衡分类问题，在基于分布的过采样的思路的基础上，提出利用vae这类的生成模型：将vae作为过采样方法，并对隐层空间z进行采样，生成最终的少数类合成样本，以提高少数类的分类效果。本文中的网络结构如图1所示，由于原始的vae是应用于图像生成中，因此合成结果可以天然可视化；但本文中直接使用了序列数据，因而需要一定程度的改进：

1. 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；
2. 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，因此本文中假设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。

方法的数学描述如下所示：

给定训练数据集，其中为d维数据样本，为相应的类标，分别表示负类和正类，我们分别用P和N来表示正类样本子集和负类样本子集，其中P中包含个正类样本，N中包含个负类样本，并且。

在训练vae模型的过程中，首先对训练集中的第维特征出现的特征数进行统计，排除离散特征。计算公式如下：

，

如果，则第j列特征为离散特征，反之则为连续特征，将数据集中的特征按照顺序分成连续特征和离散特征，并取出连续特征作为最终的训练集合。则

用训练一个vae模型并随机采样，设合成样本为，则

是最终的合成样本，则最终的训练集为。

算法架构：

|  |
| --- |
| Algorithm 1：VAEOS：VAE-based oversampling approach for imbalanced learning. |
| Input: : dataset with N samples, consisting of positive samples in P and negative samples in N  Output: Classifier H trained with dataset after the oversampling algorithm  Procedure:   1. Divide X into training dataset and testing dataset 2. Data preprocess according to formula () 3. Compute for each feature in , decide each discrete feature and continuous feature 4. Decide according to formula 5. Use to train a model and randomly sample with the corresponding model 6. Synthesize the according to formula 7. Train a classifier H with |

# Experiment

数据集描述

文中实验数据来自于UCI机器学习数据库，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。

表1 数据集描述

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 索引 | 数据集 | 样本总数 | 属性数 | 少数类 | 多数类 | 不平衡率 |
| 1 | breast-w | 699 | 9 | 241 | 485 | 1.90 |
| 2 | vehicle | 846 | 18 | 199 | 647 | 3.25 |
| 3 | segment-challenge | 1500 | 19 | 205 | 1295 | 6.32 |
| 4 | Diabetes | 768 | 8 | 268 | 500 | 1.87 |
| 5 | Ionosphere | 351 | 34 | 126 | 225 | 1.79 |
| 6 | Sonar | 208 | 60 | 97 | 111 | 1.14 |

数据预处理

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：

其中

基于混淆矩阵的评估指标

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，

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为正类的样本个数，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的样本数，而FP则为负类被错误判定为负类的样本数，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、AUC值等方法[20]。在文献[21]中说明在不平衡分类的极端情况，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。

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

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：

只有在少数类和多数类分类精度同时都高的情况下，此时，的值最大。

对比算法及参数

本文中对比了不同的过采样算法，基于分布的NDO算法 [7]，以及经典的随机插值算法SMOTE算法，分类器统一采用了朴素贝叶斯，以减少分类参数对分类效果的影响。在交叉验证中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。为了减少随机性对对比结果的影响，每个算法都计算了10次10折交叉验证的平均分类结果。

|  |  |  |  |
| --- | --- | --- | --- |
| 算法 | 本文方法 | NDO | SMOTE |
| 参数 | Vae：5层  Hidden1:[]  Hidden2:[]  Hidden3:[]  Encoder和decoder为对称结构 | Same with paper [7] | k=5 |

实验结果及分析

不同算法的对比

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 100% | | | 200% | | | 300% | | |
|  | VAE | NDO | SMOTE | VAE | NDO | SMOTE | VAE | NDO | SMOTE |
| 1 | 94.22 | **94.38** | **94.38** | **94.99** | 94.38 | 94.38 | **94.99** | 94.38 | 94.38 |
| 2 | 58.07 | 55.66 | 56.26 | 58.75 | 56.63 | 56.45 | 59.36 | 56.27 | 56.45 |
| 3 | 66.49 | 65.47 | 62.44 | 69.56 | 66.55 | 61.15 | 70.86 | 61.90 | 61.15 |
| 4 | 65.88 | 65.93 | 66.27 | 65.50 | 66.74 | 66.33 | 64.98 | 65.59 | 66.33 |
| 5 | 87.02 | 82.34 | 80.54 | 87.62 | 82.63 | 82.71 | 86.20 | 81.44 | 82.71 |
| 6 | 58.34 | 71.85 | 70.40 | 43.66 | 73.90 | 70.41 | 32.21 | 73.89 | 70.43 |

不同算法的对比

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 100% | | | 200% | | | 300% | | |
|  | VAE | NDO | SMOTE | VAE | NDO | SMOTE | VAE | NDO | SMOTE |
| 1 | 96.77 | **96.89** | **96.89** | 97.22 | 96.89 | 96.89 | 97.21 | 96.89 | 96.89 |
| 2 | 74.40 | 72.06 | 72.35 | 76.43 | 71.98 | 71.87 | 78.22 | 71.73 | 71.90 |
| 3 | 91.54 | 91.79 | 89.69 | 92.81 | 92.22 | 89.41 | 93.30 | 92.47 | 89.03 |
| 4 | 79.85 | 79.73 | 80.25 | 78.19 | 77.78 | 76.71 | 77.86 | 74.95 | 74.48 |
| 5 | 93.53 | 89.62 | 87.79 | 93.98 | 89.84 | 88.56 | 93.49 | 90.07 | 89.45 |
| 6 | 61.33 | 63.77 | 64.67 | 62.05 | 65.13 | 65.03 | 66.13 | 65.02 | 65.39 |

不同算法的对比

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 100% | | | 200% | | | 300% | | |
|  | VAE | NDO | SMOTE | VAE | NDO | SMOTE | VAE | NDO | SMOTE |
| 1 | 96.04 | 96.35 | 96.35 | 96.47 | 96.35 | 96.35 | 96.45 | 96.35 | 96.35 |
| 2 | 75.00 | 72.71 | 73.27 | 75.79 | 73.50 | 73.18 | 76.31 | 73.30 | 73.34 |
| 3 | 90.60 | 89.55 | 88.91 | 91.47 | 89.23 | 89.36 | 91.78 | 89.41 | 89.01 |
| 4 | 73.43 | 73.57 | 73.84 | 72.92 | 74.07 | 73.01 | 72.48 | 73.24 | 73.03 |
| 5 | 88.83 | 86.47 | 85.32 | 89.08 | 86.66 | 85.99 | 87.81 | 86.86 | 86.98 |
| 6 | 58.48 | 67.31 | 67.37 | 49.97 | 68.81 | 67.59 | 40.20 | 68.73 | 67.81 |

过采样率同结果的关系

从实验数据中可以看到，在过采样数量相同的情况下，vae产生的样本比NDO和SMOTE生成的样本更能提高分类器的分类效果，同时，比起传统的过采样算法中，为了保证少数类的分类效果而牺牲掉部分多数类，本文中所提出利用vae进行过采样，保证了合成样本的分布，并且能够提高整体的分类效果。

从不同的过采样算法的对比结果来看，本文中所提出的过采样算法能够产生更加合理的样本，和以往的过采样算法会牺牲多数类的识别率不同，本文所提出的算法能够帮助分类器产生更合理的分类结果，提高整体的识别效果。

# Conclusion

在这篇论文中，我们利用生成模型来代替传统的过采样机制，以便充分利用数据集中的信息， 实验结果证明了该种思路的有效性，但是在采样中，仍然是粗糙的随机过采样的过程而缺少对生成样本的筛选机制，这是接下来的拓展工作。

# Reference

[1] Y. Wang, X. Li, and B. Tao, “Improving classification of mature microRNA by solving class imbalance problem,” *Scientific Reports*, vol. 6, p. 25941, May 2016.

[2] G. Stegmayer, C. Yones, L. Kamenetzky, and D. H. Milone, “High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2016.

[3] T. Leichtle, C. Geiß, T. Lakes, and H. Taubenböck, “Class imbalance in unsupervised change detection – A diagnostic analysis from urban remote sensing,” *International Journal of Applied Earth Observation & Geoinformation*, vol. 60, pp. 83–98, 2017.

[4] C. Li and S. Liu, “A comparative study of the class imbalance problem in Twitter spam detection,” *Concurrency & Computation Practice & Experience*, vol. 30, no. 4, 2018.

[5] C. Zhang, G. Wang, Y. Zhou, and J. Jiang, “A New Approach for Imbalanced Data Classification Based on Minimize Loss Learning,” in *IEEE Second International Conference on Data Science in Cyberspace*, 2017, pp. 82–87.

[6] F. Provost, “Machine Learning from Imbalanced Data Sets 101 (Extended Abstract),” in *Soft Computing and Pattern Recognition (SoCPaR), 2011 International Conference of*, 2008, pp. 435–439.

[7] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, “Support vector machine learning for interdependent and structured output spaces,” in *International Conference on Machine Learning*, 2004, p. 104.

[8] 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.

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

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

[11] F. Olken and D. Rotem, “Random sampling from databases: a survey,” *Statistics & Computing*, vol. 5, no. 1, pp. 25–42, 1995.

[12] 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, 2002.

[13] C. Zhang, J. Guo, and J. Lu, “Research on Classification Method of High-Dimensional Class-Imbalanced Data Sets Based on SVM,” in *IEEE Second International Conference on Data Science in Cyberspace*, 2017, pp. 60–67.

[14] H. Han, W. Y. Wang, and B. H. Mao, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning,” *Lecture Notes in Computer Science*, vol. 3644, no. 5, pp. 878–887, 2005.

[15] M. Gao, X. Hong, S. Chen, and C. J. Harris, “Probability density function estimation based over-sampling for imbalanced two-class problems,” in *International Joint Conference on Neural Networks*, 2012, pp. 1–8.

[16] S. Chen, “A generalized Gaussian distribution based uncertainty sampling approach and its application in actual evapotranspiration assimilation,” *Journal of Hydrology*, vol. 552, 2017.

[17] H. Zhang and Z. Wang, “A normal distribution-based over-sampling approach to imbalanced data classification,” in *International Conference on Advanced Data Mining and Applications*, 2011, pp. 83–96.

[18] D. C. Li, S. C. Hu, L. S. Lin, and C. W. Yeh, “Detecting representative data and generating synthetic samples to improve learning accuracy with imbalanced data sets,” *Plos One*, vol. 12, no. 8, p. e0181853, 2017.

[19] K. Diederik P and W. Max, “Auto-Encoding Variational Bayes.”

[20] 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.

[21] 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.