哈 尔 滨 工 业 大 学

硕士学位论文开题报告

题 目：生成式模型的改进及其在不平衡分类中的应用

院　　（系）　计算机科学与技术

学 科 /专 业　计算机科学与技术

导　　师　张春慨

研　究　生　周　颖

学　　号　**16S051076**

开题报告日期　**2017 年 9 月 19 日**

深圳研究生院制

二〇一七年九月

目　录

1 课题的来源及研究的目的和意义

1.1 课题来源

通过查阅与专业相关的资料及文献，根据与导师讨论和自己所学的专业知识确定论文题目。

1.2 研究的目的和意义

不平衡数据集通常是指样本数据集的类别分布表现出不平衡的问题，具体表现为某些类的样本数量远远少于其他类，在这类问题中，具有较多样本的类别称为多数类，样本数量较少的类别则称为少数类，物以稀为贵，稀有的信息往往能获得人们更多的关注，而在现实问题中，也存在大量这样需要关注却又出现频率低的事务，它们虽然很重要，但利用传统的分类器却难以被正确分类。数据集的不平衡问题会影响分类器的性能：当一些传统的方法面对不平衡问题时，往往不能取得很好的效果[1]，因为现有的成熟的分类器设计都是基于类分布大致均衡这一假设，用于训练的数据集是大致平衡的，以提高数据集的总体分类准确率为目标，忽略了少数类样本的重要性。所以当这些分类器应用到不平衡数据集的分类问题时，对于多数类的有较高的识别率，但少数类的识别率却很低。实际问题中存在大量不平衡样本集的分类问题，有些问题是原始数据的分布就存在不平衡，如信用卡欺诈行为检测，网络入侵检测，医学疾病诊断等，这些问题都是以少数类的信息为关注的重点，例如在疾病监测过程中，绝大多数的结果都指向为正常，但是少量的患病结果才是值得重视的，且如果忽视患病结果而将其判定为正常，会使病人和医院遭受巨大损失，因此在这些问题上，少数类的识别效果应该受到更多关注。实际上不仅是类别数目的不平衡影响分类算法的结果，不同类的数据间的重合程度也会影响分类结果。而且数据集本身的概念复杂性、类重叠与类的不平衡分布的问题，使得不平衡数据的学习受到了广泛的关注和研究

2 国内外研究现状及分析

2.1 国外内研究现状

2.1.1 不平衡样本分类

不平衡样本分类主要有以下几个原因导致传统分类器少数类识别效果不佳：1.样本稀缺问题，这个问题中包含样本绝对稀缺和相对稀缺问题，绝对稀缺问题是指样本数量真实较少，导致传统分类器少数类识别效果不佳，相对稀缺问题是指少数类的样本数量较多，只是相对多数类的比例较低，Japkowicz 和 Stephen 的研究结果表明，这种情况下只要数据样本足够，并不一定会引起分类器性能下降，但绝对稀缺导致的少数类样本分布不集中、数量过少从而引起分类器性能下降[1]；2.噪声问题，噪声数据对稀有类将产生更大的影响[2]；3.决策面偏移问题，基于决策面的分类器如 SVM 等，因为追求全局准确率而导致决策面偏向多数类，而基于概率的分类器如朴素贝叶斯等则会因为少数类的先验概率较低而降低将样本判定为少数类的概率；4.评测指标问题，分类器评测指标的科学性直接影响着分类器的性能，因为分类器训练的目标是实现最高的评测指标，传统的模式分类方法一般以准确率作为分类器评测指标，但是以准确率为评测指标的分类器倾向于降低稀有类的分类效果[3]。

针对以上问题，目前学者们提出了很多的解决方案，由于绝对稀有的少数类样本会导致传统分类器性能下降，因此可以通过重新构建平衡数据集来解决该问题，数据层面的解决方案试图通过“采样”对不平衡样本集的样本分布进行重构，以使其从不平衡状态转化为均衡分布。由于不平衡样本存在正负类样本边界重叠、噪音数据和类内不平衡问题，所以从数据层面解决，主要是解决数据的分布问题，强调边界样本研究的重要性，寻求方法解决类内不平衡问题。边界重合对不平衡样本的分类结果影响很大，有研究表明当正负类样本边界重合时，少数类样本的比例和在边界的分布对分类的正确率影响很大，所以考虑对数据集的重采样对少数类在边界的分布调整使得分类效果最佳；另外当数据

类内不平衡问题时少数类由于存在数据碎片，导致分类器难以找到少数类的全部边界。对于这些问题，应该考虑研究合适的采样方法以期真实的还原少数类样本的分布情况。采样方法，可以被归纳为“过采样方法”和“欠采样方法”两大类。过采样方法的思想是通过增加少数类样本的数量，欠采样的方法移除一些多数类样本，使多数类与少数类达到平衡。

最原始的过采样算法是复制少数类样本，但是这样容易导致过拟合，并且这种方法对提高少数类的识别率没有很大帮助。SMOTE[4]算法为每个稀有类样本随机选出几个邻近样本，并且在该样本与这些邻近的样本的连线上随机取点，生成无重复的新的稀有类样本。在少数类样本中，不同的样本对分类器的最终决策影响程度不同，但SMOTE算法对所有少数类样本都进行插值产生新样本，造成少数类边界扩大影响决策面，降低少数类的识别效果，因此Han等[5]认为远离边界的样本点对分类决策面影响较小，提出了只针对边界少数类进行过采样，这种方式能强化边界点，因此在基于决策面的分类器结果中取得了不错的识别效果。Sáez J A等人在2014年提出了SMOTE－IPF[6]来解决噪声数据和边界样本问题，是基于SMOTE方法的一种改进算法，先使用SMOTE增加负类样本的数量，在迭代的删除噪声点和边界点，SMOTE-IPF在许多情况下能取得较好的结果。但是SMOTE只是在原始空间中随机插值，可能会造成增加后的样本与原始样本分布不一致的问题，因此于化龙[7]等在2012年提出了生成均值和方差与原始分布一致的少数类样本，保证采样前后分布的一致性，提高分类器的性能。Sanabila[8]等在2017年提出GenOMe算法，直接利用分布函数对少数类建模并重采样，以保证采样前后分布的一致性。

分类器算法方面的研究大致可以分为代价敏感学习、集成学习以及一类学习（one-class learning）等其他的学习方法。代价敏感学习主要考虑训练分类器时不同的分类错误会有不同的惩罚力度。例如在医疗中，“将病人误诊为健康人的代价”与“将健康人误诊为病人的代价”不同；在金融信用卡盗用检测中，“将盗用误认为正常使用的代价”与将“正常使用误认为盗用的代价”也不同。许多研究使用采样方法与集成学习方法结合，例如 SMOTEBoost[9]，

SMOTEBoost 是将 SMOTE 和 BOOST 相结合，每次迭代过程中，都是用 SMOTE 方法合成一些少数类，使用加入了合成少数类的训练集训练基分类器。EasyBOOST[10]与 SMOTEBoost 相反，将 Adaboost 和欠采样技术结合起来，在 Adaboost 算法迭代的每一轮，都对多数类样本进行随机欠采样，抽取的多数类与少数类相同，将这些多数类样本与所有少数类样本一起组成平衡的数据集训练基分类器。

一类学习是指当样本数量不平衡时，并且当特征空间中混杂有大量噪音特征时，基于学习单一稀有类样本的产生式模型，相比于学习两类问题的判别式模型具有更好的性能[11]。Hayat 等[12]对每个类别的图像都进行建模，使样本数量对模型的影响降到最低，利用重建误差进行分类，实验结果表明：该重建模型能够取得当时最好的分类效果，同时 ROC 曲线结果表明，该算法在不平衡数据集上的效果也比较理想。

2.1.2 生成式模型

生成模型是目前机器学习中研究比较多的一个领域，在机器学习中，我们得到的数据通常假设其为独立同分布数据，生成样本则是指通过对其概率密度分布进行建模，并在该分布上进行采样的结果。但是由于数据的高维分布，使得我们难以通过建模方式来获取其分布，生成模型则是在不直接对数据分布进行建模的前提下，对数据样本进行扩充。

生成式模型不仅在人工智能领域占有重要地位,生成方法本身也具有很大的研究价值。生成方法和判别方法是机器学习中监督学习方法的两个分支。生成式模型是生成方法学习得到的模型，生成方法涉及对数据的分布假设和分布参数学习，并能够根据学习而来的模型采样出新的样本[13]，传统的过采样[4]方法只是针对数据进行插值或者是直接对数据进行重采样，没有有效利用样本中蕴藏的分布信息，因此在因为分布而导致的分类效果不好的不平衡数据集上，通常难以取得令人满意的过采样效果。

生成式模型通过观测数据学习样本与标签的联合概率分布 P(X, Y)，训练好的模型能够生成符合样本分布的新数据，它可以用于有监督学习和无监督学习。

在有监督学习任务中，根据贝叶斯公式由联合概率分布 P(X,Y)求出条件概率分布 P(Y|X)，从而得到预测的模型，典型的模型有朴素贝叶斯、混合高斯模型和隐马尔科夫模型等。无监督生成模型通过学习真实数据的本质特征，从而刻画出样本数据的分布特征，生成与训练样本相似的新数据。生成式模型的参数远远小于训练数据的量，因此模型能够发现并有效内化数据的本质，从而可以生成这些数据。生成式模型在无监督深度学习方面占据主要位置，可以用于在没有目标类标签信息的情况下捕捉观测到或可见数据的高阶相关性。

目前生成模型通常有三种：自回归模型，定义数据的分布并对分布参数进行回归模拟，以获得最小误差或者是最大似然概率等，这种生成方法主要包括最大似然估计法、近似法[14]等。以真实样本进行最大似然估计，参数更新直接来源于已知样本，导致训练出的模型结果受限。而利用近似法学习到的生成式模型由于目标函数难以直接优化，而通常转向在学习过程中逼近目标函数的下界，这样自然会导致模型的生成效果不如意，而马尔科夫链的最大劣势则在于计算复杂度较高。

变分自编码器（Variational Auto-Encoder），对隐藏层变量的分布进行假设和采样，并利用神经网络映射到样本空间，生成新样本。Kingma[15]等首次在 2013 年提出了变分自编码器的概念，即将贝叶斯模型和自编码器进行结合，提出在隐层空间中进行近似的观点，对隐层分布进行采样，并最小化该假设分布和真实分布的KL散度，与神经网络中常采用的平方误差或熵的目标函数，以获得比较有效的生成样本。在此基础上，针对贝叶斯模型中的不同条件概率的形式，生成了诸多变分自编码器的变种，比如加入训练数据的类标信息，并以不同的概率分布形式表现，Kingma 等[16]考虑条件概率分布$p_\theta(x|y)$，在生成样本的过程中加入类标信息，产生质量更加好的样本。Sohn 等[17]假定额外信息 y 与隐变量 z 没有直接的关系，因此条件概率$p_\theta(z|y) = p_\theta(z)$，从而产生另一种条件变分自编码器。由于原始的变分自编码器需要对隐层空间的数据分布进行假设模拟，因此总体来说还是一种近似的思路，Chen 等[18]在原始变分自编码器的基础上，提出模糊的隐层分布而不是通常使用的高斯分布等。Louizos 等[19]则是在原始

变分自编码器的基础上进行了改进，即对优化中的目标函数进行了改进，因为生成样本后续常常会加入其它任务，因此在加入分类任务中，根据特征与类标之间的相关性，对目标函数进行加权，以达到更重要的特征具有更好的生成质量的目的。针对自编码器模型的任务相关性不强的问题，Sun 等[20]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，而该文献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。

生成式对抗网络(Generative Adversarial Networks)，如图2.1所示，利用生成器和一个判别器，将噪声z直接映射到样本空间，并根据分类器的结果调整该映射函数，以生成有效样本。在前两种生成方式中，由于需要对真实世界进行建模，则需要对模型进行一定的假设估计，而建模的好坏则直接影响最终的生成样本的质量，另一个困难则是真实数据通常比较复杂，从而导致建模的计算量很大。Goodfellow等[21]提出了生成式对抗网络，利用了博弈论中的纳什均衡理论，即该生成模型中包含一个生成器和一个分类器，在模型趋于稳定的状态下，生成器生成质量好的样本以至于分类器难以区分生成样本和真实样本。CGAN[22]则是在原始的GAN基础上加入了更多的先验知识，比如类标等信息或者别的多模态信息，比如对图像的描述语言等。DCGAN[23]中总结了许多对于GAN这的网络结构设计和针对CNN种网络的训练经验。比如，他们用strided convolutional networks替代传统CNN中的pooling层，从而将 GAN中的生成模型（G）变成了完全可微分的，结果使得GAN的训练更加稳定和可控。InfoGAN[24]则是在输入信息z中加入了更直观、可解释的变量c，希望因此而产生更可靠的输入信息，利用互信息的建模方式，将c定义为与生成结果相关度大的信息变量。结果显示c对于GAN的训练确实有帮助，该模型下的生成结果更出色，其次，利用c的天然特性，控制c的维度，使得infoGAN能控制生成图片在某个特定语义维度的变化。Goodfellow等[23]在文献中分析了各种生成式模型的优缺点，并介绍了不同的GAN模型架构和具有不同的损失函数的GAN模型，并指出GAN的生成

效果好于变分自编码器（variational auto-encoder）原因可能并不在于其优化了不同的损失函数，而是在于其模型本身涉及到的对抗过程，他还评论对于精确定义的概率密度函数的模型中，虽然模型允许直接优化训练集上的似然函数，但是该种模型簇的训练是非常有限的，不同的函数簇会有不同的优势。

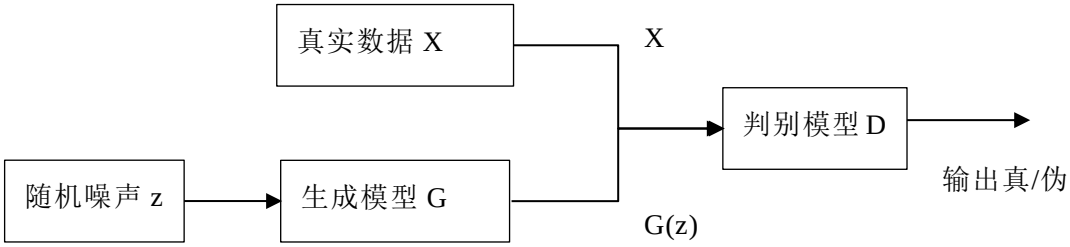

图 2.1 生成式对抗网络架构

2.2 国内外文献综述及简析

不平衡分类问题中，由于数据的数量和分布不均衡，分类器对所有数据的优化权重相同的问题，导致了感兴趣的少数类识别效果不理想，传统的解决方案都是针对少数类的数量和分布的问题进行解决，或者是在分类器的决策面上采用加权的形式，使得分类器更关注少数类的准确率。

目前的不平衡分类方法都是为了解决以上问题提出的，为了缓解少数类数量过少而导致的分类结果不理想的问题，学者们提出了对少数类过采样和对多数类欠采样的方法，伴生着混合方法等。根据采样方法侧重点的不同，又可以分为相同权重和不同权重的采样，相同权重是指同一类别中的样本，在采样中是平等对待的；而不同权重的采样，则是一种嵌入式的采样方法，因为样本在划分分类面的时候所起的作用不同。不同权重的采样指该算法侧重采样边界样本，主要是以 SVM 分类器的结果为准，因为 SVM 的特性，使得算法很容易区分边界样本，还有利用少数类和多数类的距离来确定边界样本等。同时，学者们还提出了融合样本采样和特征选择的算法，选择样本和特征时，又融合了很多种的启发式搜索算法，使得分类器进一步提高对少数类的识别率。

由于不同的数据集的分布不同，且样本数量和属性维数相差很大，因此分

类算法众多，但是普遍适用的算法数量却非常少，且算法架构都是基于集成或者是代价敏感提出的，集成方法中融合了不同的样本选择，不同的集成框架等，还有因此而提出的结构化 SVM；代价敏感类的则多是通过样本和属性权重来实现，并且伴随着不同的不平衡分类评价指标产生的模型，比如直接优化分类结果 F1 值的神经网络分类架构等。

生成式模型利用已有的数据样本，对其分布进行拟合和采样，以生成未知新样本，这样可以缓解不平衡分类中的少数类数量过少的问题。如表 2.1 所示，生成式模型从最开始的直接对样本分布进行建模，无法对生成样本直接进行评价，到现在的利用自编码器架构形成有监督式的生成结构，还有 GAN 中直接利用分类器来评价生成样本的质量，这都提示了一个问题：在表示方法上，评价样本质量是件非常艰难的事情，因此目前的生成式模型最开始多是应用于图像、视频等领域，因为样本可以直接可视化且样本数量众多，卷积神经网络的强效特征提取能力，在该领域的生成模型成果众多，但是在数据样本的生成领域中则很少见到有成果，一方面是因为数据样本难以可视化，主观很难判断模型质量，数据样本目前的可视化方法都是需要降维的，而在降维期间产生的误差难以估计，此外，生成模型在图像领域中不是应用于分类，而是更高层次的应用，比如由文字生成图像，自动驾驶等场景。

表 2.1 各类生成方法对比

表示方法	优化目标	评测方法
自回归模型	最小误差或最大似然	优化后是否接近目标函数
变分自编码器	变分下界	优化后是否接近目标函数
对抗生成网络	生成器和判别器各自的目标函数	判别器是否能区分生成样本和真实样本

3 主要研究内容及研究方案

3.1 研究内容

在一些不平衡分类问题中，少数类样本只是相对稀有，并不一定会导致分类器性能下降，而少数类绝对稀有情况下容易导致分类器在少数类上的识别率降低，因此本文中主要是为了解决少数类绝对稀有的不平衡分类问题，为了提高分类器性能，我们利用过采样算法，增加少数类样本数量，从而提高分类器性能[1]，传统的 SMOTE 算法无法有效利用数据集中的样本分布信息，且其产生的样本的有效性和可信度一直为研究者所诟病，而对直接对少数类样本的概率分布函数进行建模则会受到建模函数的限制，比如当使用正态分布拟合其概率分布函数时，如果数据的先验分布不是正态分布，则过采样产生的样本效果也不会很好。神经网络的拟合能力强大，输出维度不受限制，因此本文中提出利用神经网络模拟其概率分布函数，以保证样本前后概率分布的一致性，提高生成样本的可信度和分类器的性能。

向量样本的生成不同于图片的生成，图片可以很容易地用肉眼看出其质量是否合格，但向量样本则难以做到这一点，可视化过程中的损失无法控制，且肉眼无法评价其好坏。因此在该问题的研究中，我们拟解决在过采样过程中如何评价生成样本的质量，并选择好的生成模型，因此需要找到衡量生成样本质量的标准，并在该标准下找到一个好的生成模型，设计一个框架，并同其他过采样算法对比，验证其效果。同时，由于深层生成模型是应用在数量较少的少数类上，因此需要考虑当训练样本数量过少的情况下，神经网络模型可能难以训练到收敛的问题；或者是利用标签信息，利用数据库中所有的样本以合成样本空间的整体分布，并在生成时采用少数类样本的标签，以生成新样本，该方案中需要解决的是模型因为训练数据的不平衡，而导致模型偏向多数类的问题。

3.2 研究方案

目前已经获得公开的不平衡数据集 UCI 数据集中某些子类，都是出现在其他文献中的数据集，其中包括原始未经处理的数据集，以及为了获得更高的不

平衡率而采用的将其他类混合成多数类的数据集。

本文的研究内容框架如图 3.1 所示：主要是针对生成式模型进行研究，本文中利用少数类构建其概率分布模型p(X)，并对生成模型进行采样，以达到过采样的目的，该做法既保证了概率分布的前后一致性，又增加了生成样本的随机性，使生成的样本集合更加合理。

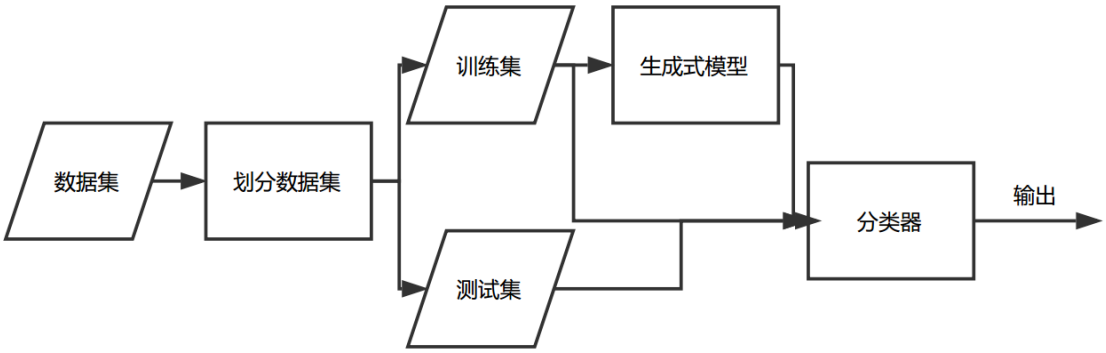

图 3.1 模型框架图

本文选用变分自编码器作为生成式模型，如图 3.2 所示，变分自编码器假设隐层z决定 X 的分布，因此我们从 z 的分布中采样，并对从 z 到 X 的映射过程进行建模，保证由 z 映射生成 X 的生成样本的质量。这部分由自编码器中的解码器部分构成，但是由于 z 的分布未知，需要从 X 中构建函数p(z|X)，以了解隐层 z 的条件分布， 因此需要利用函数将已知 X 的 z 的分布P(z|X)进行建模，因此采用自编码器中的编码器构成。由编码器产生的 z 的分布难以计算，在采样时则是以一个简单高斯分布拟合该未知分布P(z|X)，因此在自编码器的目标函数中，还加入了原始 z 分布和该简单分布的 KL 散度，以便在解码器中缩小该差异。该模型的自编码器结构使得生成样本的的过程成为一个有监督过程，能保证生成样本的合理性，而简单高斯分布则保证训练好的模型后的采样过程的可行性。

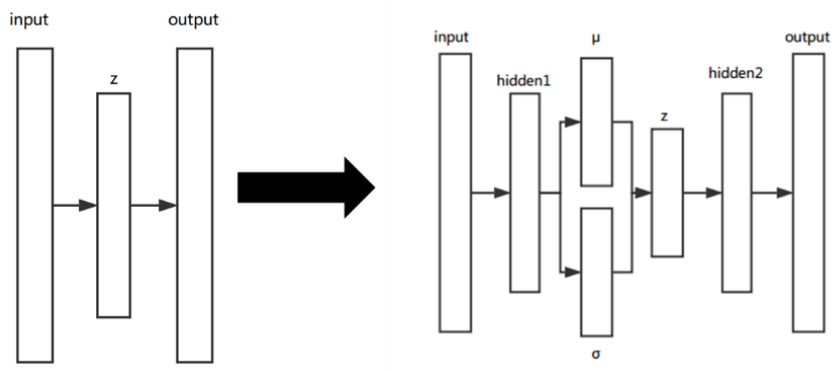

图 3.2 变分自编码器

4 预期目标

实现一个在传统的不平衡样本分类算法基础上改进的方法。使用了新的处理方法之后，使不平衡分类效果较使用之前有所改善。

其中，数据层面采用变分自编码器对少数类进行过采样，该做法既保证了概率分布的前后一致性，又增加了生成样本的随机性，使生成的样本集合更加合理，并提出模型质量的衡量标准，并在此标准下对生成数据结果进行分析。

5 已完成的研究工作及进度安排

5.1 已完成的研究工作

已经完成了有关生成式模型相关论文的阅读，基本掌握了相关背景知识，从相关数据库中下载了不平衡数据集，阅读了一些不平衡样本相关论文，基本了解了不平衡样本的问题，了解相关的分类器评价，下载相关的 UCI 数据集等。

表 5.1 UCI 数据集

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

对生成模型中的变分自编码器和生成式对抗网络有了基本了解，并实现了能够生成样本的基础变分自编码器，对算法的可行性做了实验验证，实验结果如表 5.2 所示，从结果中可以看到，本文提出的方案在一些 UCI 数据集上确实能取得比较好的效果，但是在某些 UCI 数据集上的效果还有待改进。

提出两种衡量生成样本的方法：向量可视化后的分布情况，和加入生成样本后分类器的分类结果。并对向量可视化方法进行了研究和实验结果对比，并选取了 PCA 方法进行数据可视化。

表 5.2 十折交叉验证结果

F1-value	Ionosphere%	German%	Wpbc%
SMOTE-SDAE	78.42	**81.92**	83.71
本文方案	**92.29**	47.59	**86.98**
Naïve Bayes	90.99	58.07	76.38

5.2 进度安排

2017 年 08 月-2017 年 09 月，阅读大量文献资料，对将要研究的项目做整体初步的了解。

2017 年 10 月-2018 年 01 月，对一些基本的生成与分类方法进行分析与记录，建立基本的生成式模型。

2018 年 02 月-2018 年 03 月，提出数据处理的方案与分类算法。

2018 年 04 月-2018 年 06 月，完成所设计的算法框架。

2018 年 07 月-2018 年 08 月，对各个生成模型用不同的数据进行实验，对比实验效果。

2018 年 09 月-2018 年 10 月，撰写、修改论文。

2018 年 11 月-2018 年 12 月，研究总结，完成论文答辩。

6 已具备的研究条件和所需条件及经费

6.1 实验室条件和经费保障

实验室拥有足够的工作站，GPU 服务器等。

6.2 所需条件及经费

打印参考资料和购买相关参考书籍可能需要经费。

7 预计困难及解决方案

7.1 预计困难与技术难点

课题中设计神经网络生成式模型，由于目前的生成式模型多是针对图像或者视频等可视化数据，在向量生成样本上存在着难以衡量样本质量和可视化的问题。在少数类样本真实稀少的情况下，神经网络的结构和收敛性难以保证，目前的结果是在重建模型上得到的，而在普通的分类器上会降低其性能，因此还需要提高生成样本的质量。

不平衡分类中有各种分类框架，不同的数据集在不同的框架上表现往往差别巨大，选择一个具有普适性的框架则是非常难的。

在实现算法时算法细节描述不清晰，或者技术困难导致无法复现等问题。

7.2 解决方案

参考序列建模和生成式对抗网络的架构和相关的定量的生成样本评估标准。

我会广泛查阅资料、全面深入地了解相关知识，设计比较完善的实验，在老师的指导、同学的帮助下，也可以询问算法的原作者等，来克服困难。

参考文献

[1] JAPKOWICZ N, STEPHEN S. The class imbalance problem: A systematic study[M]. IOS Press, 2002.

[2] Weiss G M. Mining with rarity: a unifying framework[J]. Acm Sigkdd Explorations Newsletter, 2004, 6(1):7-19.

[3] Drummond C, Holte R C. Explicitly representing expected cost: an alternative to ROC representation[C]// ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2000:198-207.

[4] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research，2002，16(1)：321-357.

[5] Han H, Wang W Y, Mao B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]// International Conference on Intelligent Computing. Springer, Berlin, Heidelberg, 2005: 878-887.

[6] Sáez J A, Luengo J, Stefanowski J, et al. SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering[J]. Information Sciences, 2015, 291(5): 184-203.

[7] 于化龙,高尚,赵婧,等.基于过采样技术和随机森林的不平衡微阵列数据分类方法研究[J].计算机科学, 2012, 39(5): 190-194.

[8] Sanabila H R, Kusuma I, Jatmiko W. Generative oversampling method (GenOMe) for imbalanced data on apnea detection using ECG data[C]// International Conference on Advanced Computer Science and Information Systems. IEEE, 2017: 572-579.

[9] Chawla N, Lazarevic A, Hall L, et al. SMOTEBoost: Improving prediction of the minority class in boosting[J]. Knowledge Discovery in Databases: PKDD 2003, 2003: 107-119.

[10] Liu X Y, Wu J, Zhou Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, 2009, 39(2): 539.

[11] Raskutti B, Kowalczyk A. Extreme re-balancing for SVMs: a case study[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1): 60-69.

[12] Hayat M, Bennamoun M, An S. Deep Reconstruction Models for Image Set Classification[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(4): 713-27.

[13] 王坤峰等. 生成式对抗网络 GAN 的研究进展与展望[J]. 自动化学报, 2017, 43(3): 321-332

[14] Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models[J]. arXiv preprint arXiv:1401.4082, 2014.

[15] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[16] Kingma D P, Mohamed S, Rezende D J, et al. Semi-supervised learning with deep generative models[C]//Advances in Neural Information Processing Systems. 2014: 3581-3589.

[17] Sohn K, Yan X, Lee H. Learning structured output representation using deep conditional generative models[C]// International Conference on Neural Information Processing Systems. MIT Press, 2015: 3483-3491.

[18] Chen X, Kingma D P, Salimans T, et al. Variational lossy autoencoder[J]. arXiv preprint arXiv:1611.02731, 2016.

[19] Louizos C, Swersky K, Li Y, et al. The variational fair autoencoder[J]. arXiv preprint arXiv:1511.00830, 2015.

[20] Sun Y, Mao H, Sang Y, et al. Explicit guiding auto-encoders for learning meaningful representation[J]. Neural Computing & Applications, 2017, 28(3): 429-436.

[21] Goodfellow I, Pougetabadie J, Mirza M, et al. Generative Adversarial Nets[J]. Advances in Neural Information Processing Systems, 2014: 2672-2680.

[22] Mirza M, Osindero S. Conditional Generative Adversarial Nets[J]. Computer Science, 2014: 2672-2680.

[23] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[24] Chen X, Duan Y, Houthooft R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets[C]//Advances in Neural Information Processing Systems. 2016: 2172-2180.

[25] Goodfellow I. NIPS 2016 tutorial: Generative adversarial networks[J]. arXiv preprint arXiv:1701.00160, 2016.

导师意见：

报告合格/不合格

导师签字：

签字日期：