**哈尔滨工业大学**

**硕士学位论文开题报告**

**题 目：生成式模型在不平衡样本分类方法中的应用**

**院 （系） 计算机科学与技术**

**学 科 /专 业 计算机技术**

**导 师 张春慨**

**研 究 生 周 颖**

**学 号 16S051076**

**开题报告日期 2017年9月22日**

**深圳研究生院制**

**二〇一六年九月**

目 录

[1 课题的来源及研究的目的和意义 1](#_Toc493407097)

[1.1课题来源 1](#_Toc493407098)

[1.2研究的目的和意义 1](#_Toc493407099)

[2 国内外研究现状及分析 2](#_Toc493407100)

[2.1国外内研究现状 2](#_Toc493407101)

[2.1.1生成模型 2](#_Toc493407102)

[2.1.2不平衡样本分类的数据层面方法 3](#_Toc493407103)

[2.1.3不平衡样本分类的算法层面方法 4](#_Toc493407104)

[2.2国内外文献综述及简析 5](#_Toc493407105)

[3 主要研究内容及研究方案 6](#_Toc493407106)

[3.1研究内容 6](#_Toc493407107)

[3.1.1数据层面解决不平衡样本问题 6](#_Toc493407108)

[3.1.2算法层面解决不平衡样本问题 6](#_Toc493407109)

[3.2研究方案 7](#_Toc493407110)

[4 预期目标 7](#_Toc493407111)

[5 已完成的研究工作及进度安排 7](#_Toc493407112)

[5.1已完成的研究工作 7](#_Toc493407113)

[5.2进度安排 7](#_Toc493407114)

[6 已具备的研究条件和所需条件及经费 8](#_Toc493407115)

[6.1实验室条件和经费保障 8](#_Toc493407116)

[6.2所需条件及经费 8](#_Toc493407117)

[7 预计困难及解决方案 8](#_Toc493407118)

[7.1预计困难与技术难点 8](#_Toc493407119)

[7.2解决方案 8](#_Toc493407120)

[参考文献 9](#_Toc493407121)

# 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

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，或者是根据现有样本，合成一些新的样本。

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[1]，传统的过采样[2]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。

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，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[3]，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。不平衡样本是生物信息数据的一个常见的问题，实际问题中也存在大量不平衡样本集的分类问题:信用卡欺诈行为检测，网络入侵检测，医学疾病诊断等。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生成模型

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[4]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[5]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，文献[6]考虑条件概率分布，文献[7]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，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，文献[8]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。文献[9]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。

生成式对抗网络(Generative Adversarial Networks)，利用神经网络和另一个分类器，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[10]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[11]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[12]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[13]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在文献[14]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（vae）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，不同的簇会有不同的优势。

### 2.1.2不平衡样本分类的数据层面方法

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。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。

SMOTE[2]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。文献[15]中采用了SMOTE和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。杨泽平[16]提出对多数类进行主动欠采样，即消除远离决策面的多数类样本，去掉离群点可能会有助于提高分类器的准确度，并且减少不同类别在数量上对分类器的影响，但对提高正类样本准确度没有直接帮助。

Sáez J A等人在2014年提出了SMOTE–IPF[17]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。然而，SMOTE仿造生成样本的有效性与可信度一直被部分研究者所诟病。石凤兴[18]在2015年对类内不平衡问题进行研究，提出了改进的基于密度的聚类方法先对少数类聚类，在根据不同的过采样率对少数类簇过采样，从而有效的解决了类内不平衡的问题。

### 2.1.3不平衡样本分类的算法层面方法

算法方面地研究大致可以分为代价敏感学习、集成学习以及其他的一些学习方法。

文献[19]将代价敏感思想与Adaboost结合，在权重更新时，对误分的正负类样本采用不同的权重更新策略。针对自编码器模型的任务相关性不强的问题，文献[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，而本文[20]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。

许多研究使用采样方法与集成学习方法结合，例如SMOTEBoost[21]，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[22]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，将这些多数类与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。

由于神经网络强大的非线性拟合能力，使其在分类问题中常常取得不错的效果，在不同类别样本数量不同的情况下，文献[23]对每个类别的图形都进行建模，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。

## 2.2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。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。但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。

# 3 主要研究内容及研究方案

## 3.1研究内容

### 3.1.1数据层面解决不平衡样本问题

数据层面研究，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。

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，这就是为什么分类和生成模型都在图像领域首先获得应用的原因之一。因此在该问题的研究中，我们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，以达到对少数类过采样的目的，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。

### 3.1.2算法层面解决不平衡样本问题

算法层面研究，主要从对不平衡样本的分类结果上出发，由于对不平衡样本分类结果好坏的衡量标准不再是准确率，而是综合考虑了正负样本的分类错误的重要性，本质是对不同样本的分类错误的重视程度不一样，代价敏感型分类算法是对正负样本分类采取不同的惩罚，从而使算法更加注重少数类的分类正确性，代价敏感型的分类算法主要是研究代价矩阵的取值。集成算法可以把几个弱分类器集成起来，成为一个强分类器，其中adaboost也是强调误分样本的重要性，对错分样本的权重进行调整，所以对集成分类器的研究结合少数类样本的特点提高少数类的分类正确率，提高其泛化能力。Smote-Boost结合SMOTE和集成学习算法，集成分类算法的研究应该注重和结合采样技术，实际上是改变样本的分布并且集成。

本文中希望研究一个重建模型，将多数类和少数类样本分开训练，并根据相同类别具有更小的距离这一准则，利用重建误差进行投票分类。在这个方案中，涉及到的问题包括重建模型的基本结构，训练方式，目标函数构建，以及同样存在于数据层面研究中的模型在少数类中的收敛问题，由于最后的分类结果是根据重建误差计算得来，所以要求不同类别的重建模型，在训练时的误差应该尽量一致，否则会出现某个类别的重建模型容量过大，导致分类无效。

## 3.2研究方案

分析生成式模型特点，并针对不同的数据分布，生成比较好的样本。在不平衡的情况下，提出一种分类框架，达到比较好的少数类识别效果。

# 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，数据层面采用生成模型对少数类进行过采样，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；算法层面，提出新的重建模型框架，在该框架下，能够达到相当或者更好的分类效果，且该框架具有一定的普适性。

# 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。

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。

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。

## 5.2进度安排

2017年08月-2016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

实验室有多台工作站、CPU和个人电脑；拥有多种编程语言开发环境。实验室经费充足。

## 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。在数据量真实稀少的情况下，神经网络的结构和收敛性难以保证。

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

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] 王坤峰, 苟超, 段艳杰, 林懿伦, 郑心湖, and 王飞跃, “生成式对抗网络GAN的研究进展与展望,” 自动化学报, vol. 43, no. 3, 2017.

[2] 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.

[3] N. Japkowicz and S. Stephen, *The class imbalance problem: A systematic study*. IOS Press, 2002.

[4] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic Backpropagation and Approximate Inference in Deep Generative Models,” *Eprint Arxiv*, pp. 1278–1286, 2014.

[5] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” 2013.

[6] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, “Semi-Supervised Learning with Deep Generative Models,” *Advances in Neural Information Processing Systems*, vol. 4, pp. 3581–3589, 2014.

[7] K. Sohn, X. Yan, and H. Lee, “Learning structured output representation using deep conditional generative models,” in *International Conference on Neural Information Processing Systems*, 2015, pp. 3483–3491.

[8] X. Chen *et al.*, “Variational Lossy Autoencoder,” 2016.

[9] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, “The Variational Fair Autoencoder,” *Computer Science*, 2016.

[10] I. Goodfellow *et al.*, “Generative Adversarial Nets,” *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[11] M. Mirza and S. Osindero, “Conditional Generative Adversarial Nets,” *Computer Science*, pp. 2672–2680, 2014.

[12] S. Koo, “Automatic Colorization with Deep Convolutional Generative Adversarial Networks.”

[13] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets,” 2016.

[14] I. Goodfellow, “NIPS 2016 Tutorial: Generative Adversarial Networks,” 2016.

[15] 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.

[16] 杨泽平, “基于神经网络的不平衡数据分类方法研究,” 华东理工大学, 2015.

[17] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, “SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering,” *Information Sciences*, vol. 291, no. 5, pp. 184–203, 2015.

[18] 石凤兴, “针对类内不平衡样本分类方法的研究,” 哈尔滨工业大学, 2016.

[19] 陶新民, 刘福荣, and 杜宝祥, *不均衡数据SVM分类算法及其应用*. 黑龙江科学技术出版社, 2011.

[20] 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.

[21] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, “SMOTEBoost: Improving Prediction of the Minority Class in Boosting,” in *Knowledge Discovery in Databases: Pkdd 2003, European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003, Proceedings*, 2003, pp. 107–119.

[22] X. Y. Liu, J. Wu, and Z. H. Zhou, “Exploratory Undersampling for Class-Imbalance Learning,” *IEEE Transactions on Systems Man & Cybernetics Part B*, vol. 39, no. 2, pp. 539–550, 2009.

[23] 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.

**导师意见：**

**导师签字：**

**签字日期：**