哈 尔 滨 工 业 大 学

硕 士 学 位 论 文 中 期 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院

学　　科　计算机技术

导　　师　张春慨

研　究　生　周 颖

学　　号　**16S051076**

中期报告日期　**2018 年 4 月 28 日**

目　录

1 课题主要研究内容及进度

1.1 课题主要研究内容

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，模型训练是基于样本分布均匀的假设，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，另一方面不同类别的数据分类难度也不相同，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，造成人们感兴趣的类别识别效果较差。在这类问题中，传统的机器学习方法是基于每类样本的分类代价相同的假设，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。例如在医疗健康体检数据中，绝大多数的体检报告结果是正常的，只有少部分的体检报告是真实需要治疗的，在这些少数类中，如果产生误诊断为正常，则会对病人的权益产生极大的损害，甚至错过最佳救治时间，但是如果为了保证较低的误诊率而将大多数体检报告诊断为有病，则会浪费很多医疗资源，甚至影响医患关系。在这一类的情况下，不同类别的分类代价不同，而我们感兴趣的类别识别难度较高，即构成不平衡分类问题。

一直以来，在不平衡分类问题中，为了提高少数类的准确率，学者们提出了许多具有代表性的算法：数据重采样，代价敏感学习，基于训练集划分的 bagging 方法和基于分类器集成的 boosting 方法等。因为数据层面的处理作为数据预处理的部分，是所有机器学习算法中的必要过程，因此这部分的方法可以结合到任意方法中，具有更广泛的应用场景。在这类算法中，主要思想是对已有的数据分布进行调整，使其符合样本均匀分布的假设，处理过的数据则采用一般的模式分类算法，以获得较好的分类结果。

本课题将从以下几个方面开展研究：

（1）一种新的数据集不平衡度衡量方法

传统的数据集不平衡度都是考虑样本数量不平衡，而忽略了样本分布对于其分类结果的影响，因此，本课题从样本分类难度的不一致考虑数据集的不平衡程度，改进传统的基于样本集合大小的不平衡率计算方法，提高不平衡率和最终分类结果的相关度。

（2）基于变分自编码器（VAE）的过采样方法

利用变分自编码器对少数类样本的分布进行建模，并对该模型随机采样，并利用提出的不平衡度量筛选随机采样的生成样本，生成更加合理的合成样本。

（3）基于条件变分自编码器-生成对抗模型（CVAE-GAN）的过采样方法

利用 CVAE-GAN 对训练集的全部样本进行建模，并对该模型随机采样少数类样本，充分利用已有样本信息，生成更加合理的合成样本。

1.2 进度介绍

目前已经完成了新的不平衡问题度量方法的设计和验证，并验证了基于变分自编码器（VAE）的过采样方法的可行性，并提出了整体的流程架构，经过实验验证，新的不

平衡问题度量方法与分类结果的相关性较传统的不平衡率有较大的提升，基于变分自编码器（VAE）的过采样方法也能较好地提高分类效果。

2 目前已完成的主要研究工作及结果

2.1 一种新的不平衡度衡量方法

传统的不平衡率都是计算少数类和多数类的样本数量的比值，而忽略了样本分布在分类结果中产生的重要影响，甚至产生了不平衡率虽低但是分类结果却非常不好的情况。本课题中提出的不平衡度衡量方法，不仅可以考虑到样本数量的因素，同时，更重要的样本分布也被考虑在其中，提高了不平衡度与分类结果间的相关性。

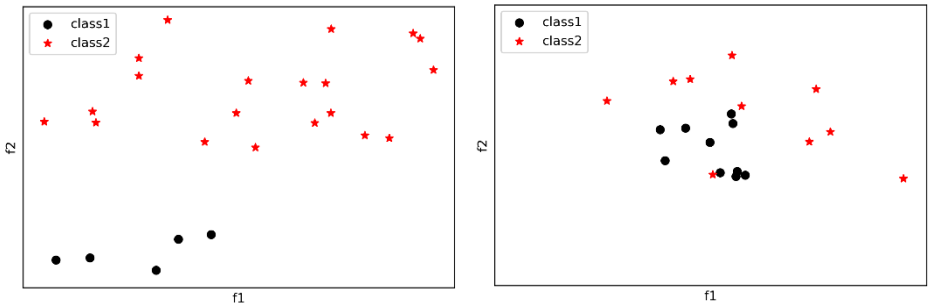
2.1.1 数据集不平衡度量

传统的不平衡数据集中，主要是指不同类别的样本数量差距悬殊，不平衡率（IR）的定义如公式(2-1)：

以二分类数据集为例，数据集为 X，其中包含 P 类样本（少数类样本）N_+个，N 类样本（多数类样本）N_-个，传统的不平衡率 IR 计算结果为

$$IR=\frac{N_-}{N_+} \tag{2-1}$$

这种定义下的不平衡率在样本均匀分布的情况下是合适的，因为基于边界划分的分类器在追求全局准确率的情况下，会倾向于将少数类样本分类为多数类样本，即在极端情况下，数据集中包含了 99%的多数类样本，将所有的样本全部分类为多数类样本的情况下，全局准确率依然高达 99%，但是这种分类是无意义的，因此，IR 在这类情况下确实能反映数据集的偏度。但是在如图 2-1 所示的情况下，IR 的定义却不足以准确反映数据集的分类难度，图 2-1 (a)中，IR 为 4，而 (b)中的 IR 却为 1，从 IR 角度上来说，应该是(a)中的数据集比(b)中更加难以分类，但是肉眼可见，(a)中两个类别有清晰的线性边界，但是在同样的线性模型下，(b)中的样本却无法得到 100%的分类结果。因此实际上(b)中的数据集比(a)中的更难分类，这与 IR 的比较结果相反，因此 IR 中无法体现数据集本身分布的复杂程度。



(a) (b)

图 2-1 不同平衡率下的数据集分布

在这个课题中，我们提出一种改进的不平衡率，每个样本的分类难度由其近邻的种类决定，即如果某个样本的近邻中包含越多的同类样本，则该样本的分类难度越低，相反，如果某个样本都是被其他类别的样本包围，则分类器很难分类正确；IGIR 定义如公式(2-7)：

$$T_+=\frac{1}{N_+}\sum_{x\in P}\frac{1}{k}\sum_{r=1}^{k}IR(x,X)=\frac{1}{N_+}\sum_{x\in P}t_k(x) \tag{2-2}$$

$$wei\text{-}T_+=\frac{1}{N_+}\sum_{x\in P}1/k\sum_{r=1}^{k}weight*IR(x,X)=\frac{1}{N_+}\sum_{x\in P}weight*t_k(x) \tag{2-3}$$

$$T_-=\frac{1}{N_-}\sum_{x\in N}t_k(x) \tag{2-4}$$

$$wei\text{-}T_-=\frac{1}{N_-}\sum_{x\in N}weight*t_k(x) \tag{2-5}$$

$$GIR=T_-\text{-}T_+ \tag{2-6}$$

$$IGIR=\sqrt{wei\text{-}T_-*wei\text{-}T_+} \tag{2-7}$$

$$f=argmax\frac{(\mu_{i1}-\mu_{i2})^2}{\sigma_{i1}^2+\sigma_{i2}^2} \tag{2-8}$$

$$\mathrm{CM}_{k(j)}=IR\left(\frac{Number\ of\ patterns\ j'\ in\ N_j\ with\ y_j'=y_j}{k}\leq 0.5\right) \tag{2-9}$$

$$\mathrm{CM_k}=\frac{1}{n}\sum_{j=1}^{n}CM_{k(j)} \tag{2-10}$$

其中IR(x, X)是一个符号函数，当 x 与其近邻的类标相同是为 1，不同时为 0。$\mu_{i1},\mu_{i2},\sigma_{i1},\sigma_{i2}$为第 i 个特征中的正类和负类子集的均值和方差。IGIR 对应的算法如下：

算法 2-1 IGIR 计算步骤

Algorithm 1: Computing the IGIR

Input: A dataSet X,label Y,number of nearest neighbors k in k-NN
Output: IGIR

for x *in* X *with label* y_x **do**
  $M \leftarrow$ the k nearest neighbors of x
  $t_k(x) \leftarrow \frac{1}{k}\sum weight * IR(x, M)$
end
$wei-T_- \leftarrow \frac{1}{N_-}\sum t_k(x) * sgn(y_x == 0)$
$wei-T_+ \leftarrow \frac{1}{N_+}\sum t_k(x) * sgn(y_x == 1)$
IGIR $\leftarrow \sqrt{wei-T_- * wei-T_+}$
return IGIR

2.1.2 不同指标与分类结果的关系

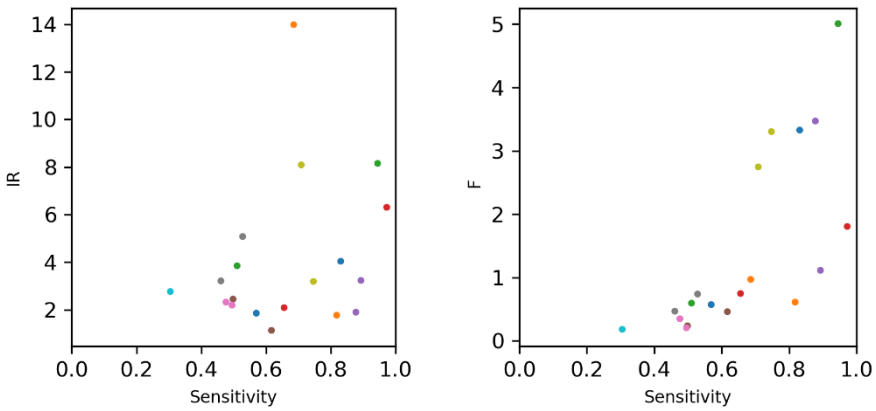
传统的不平衡率指标只考虑了样本集合大小同分类结果间的影响，而未考虑样本分布对其分类结果的影响，从不同衡量指标同分类结果的相关系数中，可以看出这种影响。

表 2-1 中为各个数据集的不同指标的计算结果，其中 GIR 为泛化的不平衡率，IGIR 为改进的泛化不平衡率。分类器为 C4.5，结果采用 10 次 10 折交叉验证，以尽量避免随机性。

表 2-1 不同指标及分类结果

	IR	GIR	CM	F	IGIR	F1_min	Gmean	Sensitivity
breasttissue	4.05	0.30	0.24	3.33	0.46	0.78	0.83	0.83
breastw	1.90	0.05	0.05	3.47	0.57	0.91	0.92	0.88
diabetes	1.87	0.22	0.49	0.58	0.37	0.57	0.66	0.57
german	2.33	0.37	0.52	0.35	0.33	0.47	0.60	0.48
glass	3.20	0.19	0.12	3.31	0.53	0.75	0.80	0.75
haberman	2.78	0.43	0.46	0.18	0.32	0.24	0.35	0.30
ionosphere	1.79	0.39	0.21	0.61	0.46	0.84	0.87	0.82
movement	14.00	0.44	0.06	0.98	0.47	0.59	0.77	0.69
satimage	8.15	0.04	0.01	5.01	0.59	0.95	0.97	0.94
Segment*	6.32	0.04	0.03	1.81	0.58	0.97	0.98	0.97
sonar	1.14	0.16	0.34	0.46	0.47	0.59	0.58	0.62
spect	3.85	0.61	0.33	0.60	0.32	0.50	0.65	0.51
vehicle	3.25	0.10	0.12	1.12	0.54	0.88	0.92	0.89
vertebral	2.10	0.14	0.31	0.75	0.47	0.67	0.71	0.66
wpbc	3.21	0.39	0.43	0.47	0.32	0.42	0.56	0.46
yeast0	5.08	0.38	0.21	0.74	0.41	0.49	0.68	0.53
yeast1	2.46	0.34	0.42	0.24	0.37	0.48	0.62	0.50
yeast2	2.21	0.26	0.48	0.21	0.37	0.49	0.61	0.49
yeast6	8.10	0.33	0.08	2.75	0.47	0.69	0.82	0.71

从图 2-2 中可以看出，数据集的不同指标同少数类的分类结果间的相关性是不同的，为了能够更好预测少数类分类结果，应当采用相关性更高的衡量指标。其中 IR 和 F 值并未表现出很强的线性相关性，而 CM 和 IGIR 则呈现出比较明显的线性相关性，同时，IGIR 的分布更为扁平，即同一分类难度下的数据集的 IGIR 呈现出更少的变化趋势，具有更强的指示效果。



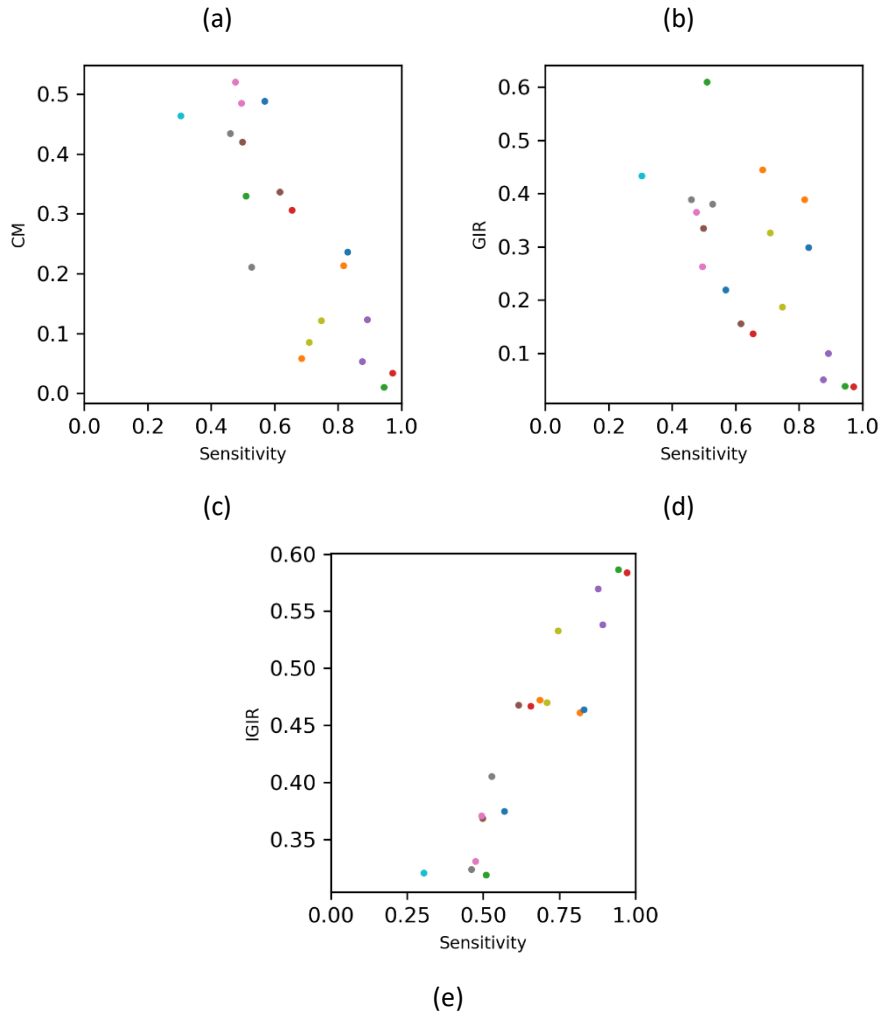


图 2-2 不同指标同分类结果的关系

为了定量分析不同指标同分类结果间的关系，本文使用可决系数R^2进行进一步分析，R^2表明了因变量的变化同自变量变化间的关系。从表 2-2 中可知，IGIR 同数据分类结果的相关系数较以前的 IR 和 GIR 等指标有较大的提升。此外，不同的分类器产生的分类结果与同一个指标也会有不同的相关度，这是因为不同的分类器在进行分类时，自身的分类假设与数据集特性间的契合度不同造成的。而在划分边界的分类 SGDclassifier 中，IGIR 反映了不同类别样本的分类边界的清晰度，IGIR 越高，则分类边界越清晰，基于边界的分类器的分类结果越好。

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\overline{y})^2} \tag{2-11}$$

表 2-2 可决系数

	F1_min	Gmean	sensitivity
IGIR	**0.92**	**0.88**	**0.93**
IR	0.18	0.34	0.28
GIR	-0.70	-0.58	-0.67
CM	-0.80	-0.85	-0.84
F	0.70	0.70	0.71

2.1.3 基于 IGIR 的样本筛选

IGIR 除了能用于衡量数据集的不平衡度外，还能指导不平衡分类中的采样方法。采样方法指的是通过对数据样本的处理，使其分布发生变化，以减轻不平衡度对分类结果的影响。以过采样为例，传统的过采样方法是以如何生成新的样本为研究重点，即假设加入新样本就可以增加少数类的信息量并提高最终的分类结果，但是在实践中，加入不当的新样本甚至会降低少数类的识别效果，这就是产生样本的质量问题。这个问题可以通过不断训练分类器，考察分类器对未知样本的识别效果来解决，但是这样会浪费大量的时间来不断生成新样本和训练新模型，而不平衡率又不足以对生成样本的质量做出评价，因此，一个能够评价样本生成质量的衡量指标是非常重要的，而这个评价指标应当尽可能能够指示样本加入训练集后对分类结果的影响，即指标同分类结果应当紧密相关。

在产生新的样本后，根据 IGIR 的计算公式，对每个新样本计算其在原始训练集中的近邻的类标，如果该新样本的单一 IGIR 超过了原始训练集中正类子集的均值，说明该样本更靠近正类样本端，反之则更靠近负类样本。过于靠近正类样本端的新样本加入对训练分类器并无太多优势，更重要的是那些在正类和负类边界的样本，这些样本有利于分类器划分更清晰的正负类边界，即有小部分的近邻是正类样本而大部分近邻是负类样本，但是不能全是负类样本，这样容易牺牲负类样本的识别效果。

2.2 基于分布的过采样方法

因为数据分布对数据集的分类结果产生了更大的影响，因此基于分布的过采样算法产生的合成样本应当会对分类结果有更好的提高效果，而传统的基于分布的过采样算法均是假设了少数类样本的先验分布，利用 EM 算法计算分布参数，之后采样得到合成样本，这种方法难以保证样本数据与假设间的一致性。本文中提出利用 VAE 对少数类样本进行建模，利用 VAE 的变分特性，不需要预先少数类样本分布形式而能够对其分布进行建模，采样简单，样本合理。

2.2.1 基于 VAE 的过采样方法

2013 年 KM 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VAE，VAE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 2-3 所示：

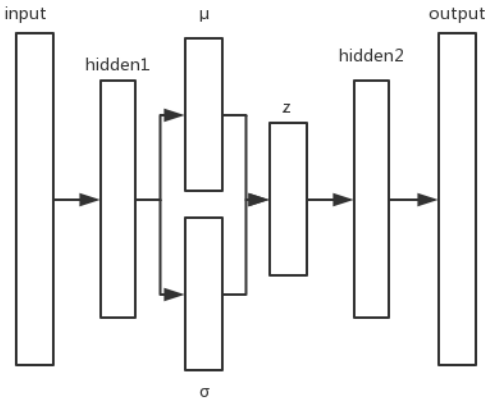


图 2-3 变分自编码器网络结构

假设 z 为潜在变量，其概率分布函数为p(z)，利用贝叶斯条件概率公式计算得：

$$p(X)=\int p(X|z)p(z)dz \tag{2-12}$$

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$p(X|z)$趋于 0，则$p(X|z)p(z)$也趋于 0，为了简化计算，则只需要计算$p(X|z)$，我们采用直接对$p(X|z)$较大值的 z 进行计算和采样，在自编码器的 encoder 中，映射出P(z|X)，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P(z|X)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P(z|X)分布为Q(z)，采用 KL 散度来计算分布拟合误差：

KL 散度定义：

$$D(p||q)=\int p(x)log\frac{p(x)}{q(x)}\mathrm{dx} \tag{2-13}$$

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

$$argmin\ \mathrm{D(Q(z)||P(z|X)} \tag{2-14}$$

$$D[Q(z)||P(z|X)]=E_{z\sim Q}[\log Q(z)-logP(z|X)] \tag{2-15}$$

在 P(z|X)上应用贝叶斯公式，可以同时得到P(X)和P(X|z)。

$$D(Q(z)||P(z|X))=E_{z\sim Q}[\log Q(z)-logP(z)]+logP(X) \tag{2-16}$$

将D[Q(z)||P(z|X)]应用到公式（2-14）中

$$logP(X)-D[Q(z)||P(z|X)]=E_{z\sim Q}[logP(X|z)]-D[Q[z]||P(z)] \tag{2-17}$$

其中 X 是固定的，Q 为任意分布而不是某个特定的能将 X 映射为可以生成 X 的 z 的分布，采用更具有可能性的 z 来推断 X，将 X 能够映射的 z 应用到公式中，得到

$$logP(X)-D[Q(z|X)||P(z|X)]=E_{z\sim Q}[logP(X|z)]-D[Q(z|X)||P(z)] \tag{2-18}$$

因为$P(X)$是固定的，所以最小化$D(Q(z)||P(z|X))$转变为最大化等式右边的值，其中$\log P(X|z)$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z的假设先验分布同 encoder 中映射出的 z 的分布的误差。

本文中针对不平衡分类问题，在基于分布的过采样的基础上，提出利用 VAE 这类的生成模型作为过采样方法，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 2-3 所示，因为数据集中存在着连续特征和离散特征，而传统的过采样算法中，对这些特征没有区分度，导致生成的样本并不存在实际意义，且 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，只适合用于连续特征的生成，因此将这部分的过采样分为两个阶段：连续特征生成和离散特征生成。

（1）连续特征生成

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。

给定训练数据集$X=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中$x_i\in R^d$为 d 维数据样本，$y_i\in\{0,1\}$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N_+个正类样本，N 中包含N_-个负类样本，并且$N_+ + N_- = N$。

在训练 VAE 模型的过程中，首先对训练集中的第 j 维特征出现的特征数$nelements_j$进行统计，排除离散特征。计算公式如(2-19)(2-20)：

$$nelements_j = \sum_{i}^{N_+} distinct\{x_{ij}\}, 1\le j\le d \tag{2-19}$$

$$x_i = \{x_{i1}, x_{i2}, \cdots, x_{ik}\} \cup \{x_{i(k+1)}, \cdots, x_{id}\}$$
$$\text{s.t.} \begin{cases} nelements_j > 2, & 1\le j\le k \\ nelements_j \le 2, & k+1\le j\le d \end{cases} \tag{2-20}$$

如果$nelements_j\le 2$，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

$$XtrainVAE=\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{N_+1} & \cdots & x_{N_+k} \end{bmatrix} \tag{2-21}$$

用Xtrain训练一个 VAE 模型并随机采样，设合成样本为Xnew，采样数为 M，则

$$Xnew=\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{M1} & \cdots & x_{Mk} \end{bmatrix} \tag{2-22}$$

（2）离散特征生成

$$\begin{cases} Xfinal_{ij}=Xnew_{ij} \cup X_{lm}, k+1 \le m \le d \\ s.t.\ agrmin \sum \left|\left|Xnew_{ij}-X_{lj}\right|\right|^2, 1 \le j \le k \end{cases} \tag{2-23}$$

计算 VAE 中生成的合成样本的连续特征同原始的连续特征的欧式距离，并将最近邻中的离散特征补全为新样本中的离散特征，并按照原始的特征排序生成Xfinal，为最终的合成样本，则最终的训练集为X ∪ Xfinal。

算法 2-2 基于变分自编码器的过采样方法

Algorithm 1：VAEOS：VAE-based oversampling approach to imbalanced learning.
Input: X: dataset $X = P \cup N$ with N samples, consisting of N_+positive samples in P and N_- negative samples in N Output: Classifier H trained with dataset after the oversampling algorithm Procedure: 1. Divide X into training dataset Xtrain and testing dataset Xtest 2. Data preprocessing according to formula (2-41) 3. Compute $nelements_j$ for each feature j in Xtrain, decide each discrete feature and continuous feature 4. Decide XtrainVAE according to formula (2-21) 5. Use XtrainVAE to train a VAE model and randomly sample with the corresponding model 6. Synthesize the Xfinal according to formula (2-23) 7. Train a classifier H with Xtrain ∪ Xfinal

2.2.2 基于 CVAE-GAN 的过采样方法

在上一小节中，本文提出利用 VAE 对少数类样本的概率分布函数进行建模，并在分类前利用该模型过采样，以增加分类器的少数类信息，提高分类准确率。但是在该模型的训练过程中，只采用了占比很少的少数类样本，而丢弃了大量的多数类样本信息;同时，VAE 在生成过程中采用的是原始样本同生成样本的平方差，这造成了生成样本的平均化，生成图像时十分模糊；而在另一类生成模型 GAN 中，分辨器返回的生成信息又非常有限，容易造成生成器模型坍塌，即每次只能生成相似样本以通过分辨器，VAE 则可以缓解这一缺陷，因为 vae 要求每个原始样本和对应的生成样本是相似的。

在机器学习中，有一个非常重要的假设：样本是独立同分布的，这意味着多数类样本同样可以提供分布信息，对整体的样本分布函数进行建模，而在生成过程中，对生成样本的类标加以限制，即可以获得更加合理的生成样本。针对 VAE 的生成样本模糊的问题，GAN 则利用分辨器对生成模糊、不够真实的样本进行惩罚，提高样本生成的合理性。

2017 年，JBao 等提出将 CVAE 和 GAN 结合起来，不仅利用 vae 的架构生成与原始样本中相似的样本，且利用 GAN 中分辨器对原始样本和生成样本进行分类，即最终生成的样本，不仅需要拟合原始样本的分布，还需要足够逼真，这样既能够减缓 GAN 中模式坍塌问题，同时也能增加 VAE 生成样本中的细节清晰度。同时，由于在训练时

采用了全部的训练样本，在整个模型中增加了分类器，生成器需要生成不同类别的真实样本。

模型架构如图 2-4 所示：

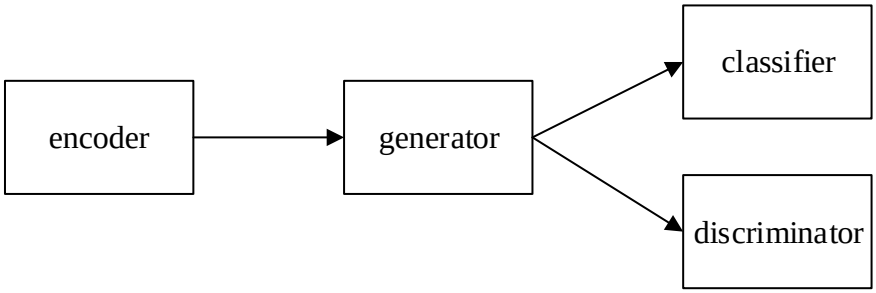


图 2-4 CVAE-GAN 的模型架构

由于在不平衡数据集中，不同类别的样本数量不同，而这会对最终的模型参数产生影响，即在误差和的情况下，模型会被多数类中包含的信息影响，这样会造成生成模型更擅长生成多数类样本而忽略对少数类样本的建模，最终影响生成的少数类样本的质量。因此，本文中提出增加模型对少数类样本的生成效果的关注度。在本文模型中，由于 classifier 容易受到不平衡数据集的影响，因此本文中提出对 classifier 的 loss 函数提出修改，即使用 F1 值的近似函数作为该分类器的 loss 函数。因为 F1 值的计算过程中，TN 并未在公式中出现，因而 F1 值更注重少数类的分类效果，但是 F1 的计算过程中涉及到 sgn 函数，从而使得真实的 F1 值并不能直接作为 loss 函数，因其是不可导的，所以本文中采用的是近似 F1 值。

设 h(x)是样本通过分类器后的原始结果，则最终的预测类标时根据 h(x)和对应的阈值的大小关系的符号函数的结果。

$$h(x) = sigmoid(wx + b) \tag{2-24}$$

$$\mathrm{y_{pred}} = sgn\big(h(x)\big) \tag{2-25}$$

假设$\mathrm{y_{true}}$是样本的真实类标，$\mathrm{y_{pred}}$为 classifier 的分类结果，则精确率和召回率可以表达为公式(2-26)(2-27)：

$$precision = \frac{y_{true}y_{pred}}{y_{pred}^2} \tag{2-26}$$

$$recall = \frac{y_{true}*y_{pred}}{y_{true}^2} \tag{2-27}$$

$$\mathrm{F1} = \frac{1}{\frac{1}{precision}+\frac{1}{recall}} = \frac{2y_{true}*y_{pred}}{y_{true}^2+y_{pred}^2} \tag{2-28}$$

所有的网络结构均采用 MLP，每个网络结构的 loss 函数如下：

$$\mathrm{loss_c} = -\log(\mathrm{P(c_r|x_r)}) \tag{2-29}$$

$$\mathrm{loss}_{KL} = KL(q(z|x_r, c_r)||P_z) \tag{2-30}$$

$$\mathrm{loss_D} = -(\log\big(D(x_r)\big) + \log\Big(1 - D\big(x_f\big)\Big) + \log(1 - D(x_p))) \tag{2-31}$$

$$L_{GD} = \frac{1}{2}||\frac{1}{m}\sum_i^m f_D(x_r) - \frac{1}{m}\sum_i^m f_D(x_p)||_2^2 \tag{2-32}$$

$$L_{GC} = \frac{1}{2}\sum_{c_i} || f_C^{c_i}(x_r) - f_C^{c_i}(x_p) ||_2^2 \tag{2-33}$$

$$\mathrm{L_F} = 1 - \frac{2*y_{true}*y_{pred}}{y_{true}^2+y_{pred}^2} \tag{2-34}$$

$$L_G = \frac{1}{2}(||x_r - x_f||_2^2 + ||f_D(x_r) - f_D(x_f)||_2^2 + ||f_C(x_r) - f_C(x_f)||_2^2) \tag{2-35}$$

最终的 loss 函数定义如下：

$$\mathrm{loss_c} = L_C + \lambda_5 * L_F \tag{2-36}$$

$$\mathrm{loss_D} = L_D \tag{2-37}$$

$$\mathrm{loss_G} = \lambda_2 * L_G + \lambda_3 * L_{GD} + \lambda_4 * L_{GC} \tag{2-38}$$

$$\mathrm{loss_E} = \lambda_1 * L_{KL} + \lambda_2 * L_G \tag{2-39}$$

2.2.3 实验验证

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。数据集中存在着值缺失的情况，为了保证数据集的完整性，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，采用了归一化，将其进行缩放，公式如下所示：

$$x_{inew} = \frac{x_i - \bar{x}}{s} \tag{2-40}$$

其中$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,$s = \sqrt{\left(\frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x})\right)^2}$

表 2-3 数据集描述

Index	Dataset	Samples	Attributes	Minority	Imbalance ratio
1	breast-w	699	9	241	1.90
2	vehicle	846	18	199	3.25
3	segment-challenge	1500	19	205	6.32
4	Diabetes	768	8	268	1.87
5	Ionosphere	351	34	126	1.79

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，因此传统的精度指标不足以分类器在不平衡分类中的表现，在二分类中，常常使用混淆矩阵来评估分类器的性能，其定义如表 2-4：

表 2-4 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正类的样本数，目前出现了一些新的不平衡数据的分类评价指标，例如 F-value 和 G-mean、AUC 值等方法。在不平衡分类的极端情况，AUC 值是不可靠的，在相同的分布下，不平衡率越高，则 AUC 值会越高。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：

$$F\text{-}value=\frac{(1+\beta^2)\times recall\times precision}{\beta^2\times recall+precision} \tag{2-41}$$

其中，准确率$\text{precision}=\frac{TP}{TP+FP}$，召回率$\text{recall}=\frac{TP}{TP+FN}$,$\beta$取值为[0,+∞]。本实验中取$\beta$=1，此时的 F-value 表示召回率和准确率权重一致。

gmean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：

$$gmean=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}} \tag{2-43}$$

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

本文中对比了基于分布的采样算法 NDO 和经典的插值算法 SMOTE，VAE 中采用 5 层本文中对比了不同的过采样算法，基于分布的 NDO 算法，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，其中 CVGA 是一次 10 折交叉验证的结果，其余每个算法都计算了 10 次 10 折交叉验证的平均分类结果。

本文中针对传统的过采样算法中难以保持样本分布的问题，提出了基于 VAE 的过采样方法，该方法改变了传统的过采样思路，将连续特征和离散特征分开生成，提高了合成样本的实际意义；从表 2-5 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少数类分类效果。同时，表 2-6 和表 2-7 中的实验数据显示，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，能够提高整体的分类效果，且在个别数据集上的提升空间较大，这些实验数据显示，该方法能够保证样本分布，生成更加合理的样本。

在 CVGA 方法中，考虑到加入不同类别样本需要更细致的评判标准，要求生成器对不同类别的样本生成均达标，因而加入了 classifier 网络；但加入该网络反馈对实验结果的影响还需要验证。从表 2-5 中可以看到，第 3 个数据集的实验结果有显著提高，这表明该方法仍具有一定的合理性；但在试验中增加采样率反而会降低少数类的分类准确度，

这表明 KL 散度收敛有限，还有可能是 GAN 的收敛问题导致的，因此在接下来的工作中，需要对该算法进行更合理的改进，增加生成样本的信息量。

表 2-5 不同算法的F – min对比

	100%				200%				300%			
	CVGA	VAE	NDO	SMO	CVGA	VAE	NDO	SMO	CVGA	VAE	NDO	SMO
1	**94.75**	94.22	**94.38**	**94.38**	94.54	**94.99**	94.38	94.38	94.54	**94.99**	94.38	94.38
2	55.32	**58.07**	55.66	56.26	54.56	**58.75**	56.63	56.45	50.43	**59.36**	56.27	56.45
3	**80.91**	**66.49**	65.47	62.44	**70.74**	**69.56**	66.55	61.15	57.31	**70.86**	61.90	61.15
4	60.79	**66.61**	65.93	66.27	59.76	**67.99**	66.74	66.33	58.87	**66.58**	65.59	66.33
5	68.27	**87.02**	82.34	80.54	67.15	**87.62**	82.63	82.71	65.62	**86.20**	81.44	82.71

表 2-6 不同算法的F – maj对比

	100%				200%				300%			
	CVGA	VAE	NDO	SMO	CVGA	VAE	NDO	SMO	CVGA	VAE	NDO	SMO
1	**96.96**	96.77	**96.89**	**96.89**	96.85	**97.22**	96.89	96.89	96.85	**97.21**	96.89	96.89
2	**76.09**	**74.40**	72.06	72.35	**82.71**	**76.43**	71.98	71.87	**84.96**	**78.22**	71.73	71.90
3	**96.29**	91.54	**91.79**	89.69	**95.84**	**92.81**	92.22	89.41	**95.17**	**93.30**	92.47	89.03
4	78.96	**80.30**	79.73	80.25	78.15	**78.12**	77.78	76.71	**77.41**	**76.60**	74.95	74.48
5	88.85	**93.53**	89.62	87.79	88.49	**93.98**	89.84	88.56	88.14	**93.49**	90.07	89.45

表 2-7 不同算法的gmean对比

	100%				200%				300%			
	CVGA	VAE	NDO	SMO	CVGA	VAE	NDO	SMO	CVGA	VAE	NDO	SMO
1	**96.64**	96.04	**96.35**	**96.35**	96.43	**96.47**	96.35	96.35	96.43	**96.45**	96.35	96.35
2	72.63	**75.00**	72.71	73.27	70.51	**75.79**	73.50	73.18	64.67	**76.31**	73.30	73.34
3	**93.88**	**90.60**	89.55	88.91	78.90	**91.47**	89.23	89.36	65.38	**91.78**	89.41	89.01
4	69.20	**74.04**	73.57	73.84	68.43	**74.89**	74.07	73.01	67.63	**73.66**	73.24	73.03
5	72.47	**88.83**	86.47	85.32	71.51	**89.08**	86.66	85.99	70.28	**87.81**	86.86	86.98

3 后期拟完成的研究工作及进度安排

后期需要完成：

（1）增加以 IGIR 为评价指标的样本筛选；

（2）调整 CVAE-GAN 的模型结构和参数，提高分类样本质量；

（3）在近似 F1 值计算的过程中，提高近似程度，减缓不平衡对模型的影响。

后期进度安排如下：

2018 年 5 月——2018 年 7 月：参考其他模型的 loss 函数设计，对模型进行改进。分析不同网络在 CVAE-GAN 中所起到的作用，并做实验对比。

2018 年 7 月——2018 年 9 月：进行系统总体测试，总结。

2018 年 9 月——2018 年 12 月：总结研究结论，撰写毕业论文，准备答辩。

4 存在问题及解决方案

4.1 存在问题与困难

（1）IGIR 的计算需要遍历整个训练集，新样本评价时，理论上需要对每个新加入的样本和当前候选样本的距离再进行计算，计算复杂度较高；IGIR 的定义不具有对称性，则需要对新训练集不断进行迭代。

（2）CVAE-GAN 模型的训练还需要优化：目前的实验结果中显示，该模型下产生的样本质量还十分不够，表现为增加采样样本数反而会降低少数类分类效果。

4.2 解决方案

针对目前存在的问题和困难，提出以下解决方案：

（1）将过程简化，考虑到过采样主要是为了增加原始训练集中的信息，忽略候选样本对新加入的样本的影响，考虑两个集合间的样本距离，仔细观察样本分布同分类结果间的关系。

（2）查看相关文献，仔细研究 CVAE-GAN 算法的原理，研究优化不平衡分类问题中的过采样算法的主要思想，查阅相关资料，寻求文献原作者或指导老师帮助。认真设计改进过采样算法，采用不同的方法进行试验。与指导老师和同学多交流，寻找新思路，认真学习神经网络相关知识。考虑样本的潜在编码的真实分布同假设分布间的关系，和 GAN 中的模型坍塌问题。

5 如期完成全部论文工作的可能性

本课题目前已经完成了利用 VAE 和 CVAE-GAN 进行过采样的算法实现，并给出实验证明其正确性。接下来除了给出更多的实验结果证明，还要对基于 VAE 的过采样算法进行优化，争取提高合成样本的质量和最终的分类效果。

在样本的选择中，利用 IGIR 对样本进行衡量，其定义已经完成，之后的工作中将两种方案整合即可。

综上所述，论文能够按期完成，并取得一定的研究成果。

导师确认学生是否按意见修改：

是　　否

导师签字：
签字日期：