控制与决策
Control and Decision

基于变分贝叶斯推断最优高斯混合模型的自适应不均衡数据综合采样法

刘金平, 杨本芳, 周嘉铭, 徐鹏飞

引用本文:
刘金平,杨本芳,周嘉铭,徐鹏飞. 基于变分贝叶斯推断最优高斯混合模型的自适应不均衡数据综合采样法[J]. 控制与决策, 2023, 38(6): 1653−1660.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.1350

基于变分贝叶斯推断最优高斯混合模型的自适应不均衡数据综合采样法

刘金平, 杨本芳, 周嘉铭, 徐鹏飞†

(湖南师范大学 智能计算与语言信息处理湖南省重点实验室， 长沙 410081)

摘 要: 实际的分类数据往往是分布不均衡的. 传统的分类器大都会倾向多数类而忽略少数类, 导致分类性能恶化. 针对该问题提出一种基于变分贝叶斯推断最优高斯混合模型 (varition Bayesian-optimized optimal Gaussian mixture model, VBoGMM) 的自适应不均衡数据综合采样法. VBoGMM 可自动衰减到真实的高斯成分数, 实现任意数据的最优分布估计; 进而基于所获得的分布特性对少数类样本进行自适应综合过采样, 并采用 Tomek-link 对准则对采样数据进行清洗以获得相对均衡的数据集用于后续的分类模型学习. 在多个公共不均衡数据集上进行大量的验证和对比实验, 结果表明: 所提方法能在实现样本均衡化的同时, 维持多数类与少数类样本空间分布特性, 因而能有效提升传统分类模型在不均衡数据集上的分类性能.

关键词: 不均衡数据；高斯混合模型；变分推断；自适应综合采样法

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.1350 **开放科学 (资源服务) 标识码 (OSID):**

引用格式: 刘金平, 杨本芳, 周嘉铭, 等. 基于变分贝叶斯推断最优高斯混合模型的自适应不均衡数据综合采样法 [J]. 控制与决策, 2023, 38(6): 1653-1660.

Adaptive synthetic sampling of imbalanced data based on variation Bayesian-optimized Gaussian mixture model

LIU Jin-ping, YANG Ben-fang, ZHOU Jia-ming, XU Peng-fei†

(Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing，Hunan Normal University，Changsha 410081，China)

Abstract: In actual pattern classification tasks, the processing data is generally imbalanced. Traditional pattern classification models tend to learn towards the majority class and ignore the minority class samples, leading to classifier performance deterioration. This paper proposes an adaptive synthetic sampling method for the imbalanced data processing using a variation Bayesian optimized GMM (VBoGMM) estimation method. The VBoGMM can automatically attenuate to the real number of Gaussian components to achieve the optimal estimation of any distribution. Based on the spatial distribution characteristics of unbalanced data sets, the adaptive synthetic sampling is performed, and the Tomek-link approach is further adopted to clean over-sampling samples to obtain a relatively balanced data set for the subsequent classifier learning. A large number of comparative experiments have been carried out on multiple public imbalanced data sets. Experimental results show that the proposed method can achieve relatively balanced samples while maintaining their spatial distribution characteristics of majority and minority samples, thus effectively improving the performance of traditional classifiers on various uneven data sets.

Keywords: imbalanced data；Gaussian mixture model；variational inference；adaptive synthetic sampling

0 引 言

分类是机器学习中的基本任务[1]. 大多数经典的分类模型往往假定待处理数据具有均匀分布特性, 或者具有某种特定的分布特性[2]. 然而, 在实际的工程应用中, 待处理数据的真实分布特性往往是未知的, 并且一般会出现某一类或几类样本比其他类样本多的情况[3]. 传统的分类器模型往往倾向于多数类而忽略少数类以获得更高的检测率, 导致分类器整体性能

收稿日期: 2021-08-02；**录用日期:** 2022-02-25.

基金项目: 国家自然科学基金项目 (61971188).

†通讯作者. E-mail: xupf@hunnu.edu.cn.

*本文附带电子附录文件, 可登录本刊官网该文“资源附件”区自行下载阅览.

恶化. 因而,如何对不均衡数据进行处理以有效提高少数类的分类准确率和分类器的整体性能已成为机器学习领域的热点问题[4]. 目前,不均衡数据集的处理方法可归结为3类: 特征层[5-7]、分类算法层[8-9]以及数据预处理层[10]处理方法.

数据层处理方法通过调整数据样本的分布减少不均衡数据对分类性能的影响,因其具有较低的时间复杂度以及较高的分类性能优化结果在不均衡数据处理中应用广泛. 最基本的数据层处理方法为采样法,包括欠采样和过采样法. 本文从少数类样本的过采样方法着手对不均衡数据集进行处理.

过采样方法分为随机过采样和非随机过程采样. 随机过采样是指重复已知采样,包括简单重复[11]、线性插值[12]、非线性插值[13]等. 早年经典的基于少数类的过采样(synthetic minority oversampling technique, SMOTE)方法[10]在对少数类进行过采样时采用了线性插值,以增加合成样本的信息量和合理性,然而SMOTE获得的插值数据缺乏结构稳定性. 因而,He等[14]提出了一种自适应综合采样法(adaptive synthetic sampling, ADASYN),通过计算样本间的不均衡比实现自适应采样. 近年来,还相继出现了一些新的数据生成方法. 比如: Liu等[15]将少数类数据的采样问题转化为数据估计问题; Razavi-Far等[16]将缺失值填充思想应用于数据过采样中,通过期望值最大法(expectation-maximization, EM)对缺失值进行估计,实现了故障数据集(少数类样本)的有效生成. 然而,上述过采样方法仅考虑样本数量对分类器性能的影响,忽略了少数类样本的分布特性. 在对少数类样本进行过采样时,如果破坏了样本的原始空间分布特性,则会导致训练集和测试集存在空间分布上的结构偏差,从而影响分类器的泛化性能.

样本的空间分布特性,特别是少数类样本的空间分布对于分类器性能具有重要作用. 一些学者在进行不均衡数据处理时已在一定程度上考虑了样本的空间分布信息[3, 17]. 比如: Chen[18]提出了一种基于正态分布的过采样方法; Li等[19]提出了一种基于Weibull分布的过采样方法; Zhang等[20]提出了一种基于高斯混合模型(Gaussian mixture model, GMM)的SMOTE方法. 上述方法虽然在一定程度上考虑了样本的空间分布信息,但并未在数据的分布模型拟合时进行最优化处理. 模型参数选择(比如GMM中的高斯分量数)以及参数的初始值极大影响了最终分布模型的获取,无法保证所获得的分布模型能充分表现数据的空间分布特性.

变分推断[21]是机器学习领域中一类基于贝叶斯估计和近似计算复杂积分的技术,广泛应用于各种复杂模型的推断. 使用变分贝叶斯推断对数据的GMM进行优化求解,相较于传统的EM算法,具有计算开销小并可实现GMM的最优化估计,利于在大规模数据集中应用. 此外,传统的过采样法还可能产生模糊分类边界的新样本点. 依据样本分布信息进行过采样可围绕样本类别进行自适应插值,在一定程度上减少了过采样对于分类精度的影响. 为了维护分类边界,本文采用Tomek-link[22]选择并删除两类样本的最邻对,进一步优化过采样处理结果.

针对传统过采样法因不能充分利用不均衡数据的分布信息而导致后续的分类模型精度下降的问题,本文提出一种基于变分贝叶斯推断最优化高斯混合模型(variation Bayesian-optimized GMM, VBoGMM)的自适应不均衡数据综合采样法(VBoGMM-based adaptive synthetic sampling, VBoGMM-sampling),即首先采用VBoGMM拟合少数类样本的分布特性,再根据少数类样本的分布信息自适应生成样本,最后依据Tomek-link对规则对所有样本均衡化清洗,获得相对均衡的数据集. 在多个不均衡数据集上进行的大量验证性和对比性实验,验证了本文算法的优越性.

1 相关工作

1.1 GMM

GMM实际上是高斯分量的线性叠加,即

$$p(\boldsymbol{x}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})=\sum_{k=1}^{K}\pi_k\mathcal{N}(x|\mu_k,\Sigma_k). \tag{1}$$

其中: K为相应的高斯分量数, $\boldsymbol{\pi}=\{\pi_1,\pi_2,\dots,\pi_K\}$为混合权重系数向量$\Big($满足$0\leqslant\pi_k\leqslant 1$且$\sum_{k=1}^{K}\pi_k=1\Big)$, $\boldsymbol{\mu}=\{\mu_1,\mu_2,\dots,\mu_K\}$, $\boldsymbol{\Sigma}=\{\Sigma_1,\Sigma_2,\dots,\Sigma_K\}$, $\mathcal{N}(\boldsymbol{x}|\mu_k,\Sigma_k)$代表均值向量为$\mu_k$、协方差矩阵为$\Sigma_k$的高斯分布.

概率模型参数估计最常用的方法为极大似然估计(maximum likelihood estimation, MLE). 给定N个训练样本$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_N\}$, GMM对应的负对数似然函数$L(\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})$为

$$L(\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})=-\ln\prod_{n=1}^{N}p(\boldsymbol{x}_n|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})=$$
$$-\sum_{n=1}^{N}\ln\Big\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}|\mu_k,\Sigma_k)\Big\}. \tag{2}$$

由于式(2)的对数中含有求和表达式,直接对式(2)中各参量求偏导以进行参数求解无法获得各参数的闭式解. 因而, GMM参数求解常采用一种交叉迭代求解方法——EM方法来最小化$L(\boldsymbol{X};\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})$.

EM有效降低了多参数MLE的求解复杂度. 然而,式(1)所示的GMM有相当多的参数组合(共有$O(D!)$种可能,其中D代表模型中参数的个数[23])可以获得等同的概率密度估计结果,即式(2)所示的负对数似然函数有多个局部极值,并且式(2)所示的对数似然函数存在奇异点. 比如,如果某个高斯分量的协方差矩阵行列式接近0,则获得无穷的概率密度值,严重影响参数的优化计算. 因而,EM和MLE一样,极易陷入模型的局部极值,并且EM需要人为确定其高斯分量数K和选定各高斯分量的参数的初始值. 不恰当的参数选取极易出现数据过拟合或者模型收敛慢的情形. 同时,EM和MLE 都无法利用已有的先验信息,且对模型中各高斯分量初始值敏感,无法保证获得最优的分布模型[24].

1.2 ADASYN算法及Tomek-Link

SMOTE[10]通过线性插值进行过采样,插值空间位于原数据空间. 然而,SMOTE产生新的少数类样本时,只是简单地在同类近邻之间插值,并没有考虑到少数类样本周围多数类样本的分布情况,并且对每个少数类样本合成数量相同的样本,这使得基于SMOTE的过采样法存在一定的盲目性.

He等[14]提出的ADASYN通过统计少数类样本周围多数类样本的情况,对每个少数类样本进行自适应的过采样. 与SMOTE相比,ADASYN自适应决定每一个少数类样本周围应合成的样本数目,能在一定程度上保证过采样后样本的边界不被破坏.

Tomek-link是一类样本配对规则,通过特定规则区分类别边界样本对,然后适量删除样本对使样本分类边界更为清晰,以提升分类器分类准确性[22].

定义1 (Tomek-link对)　假设样本点$\boldsymbol{x}_i$与$\boldsymbol{x}_j$的距离为$d(\boldsymbol{x}_i,\boldsymbol{x}_j)$;如果不存在第3个样本$\boldsymbol{x}_l$使得$d(\boldsymbol{x}_i,\boldsymbol{x}_l)<d(\boldsymbol{x}_i,\boldsymbol{x}_j)$或者$d(\boldsymbol{x}_j,\boldsymbol{x}_l)<d(\boldsymbol{x}_i,\boldsymbol{x}_j)$成立,则认为$d(\mathrm{x}_i,\mathrm{x}_j)$为一个Tomek-link对.

获得Tomek-link对后可进行: 1)欠采样,将Tomek-link对中多数类的样本剔除; 2)数据清洗,将Tomek-link对中的两个样本剔除. 本文对经过采样后的新数据样本点,依据Tomek-link对进行均衡化清洗,获得相对均衡的分类结构保持的数据集用于分类器学习.

1.3 变分推断

基于给定的观测点$\boldsymbol{x}$对模型参量(向量)$\boldsymbol{z}$进行估计,在贝叶斯观点下要获得后验概率$p(\boldsymbol{z}|\boldsymbol{x})$的最优逼近值$q(\boldsymbol{z})$,其可以通过最小化$p(\boldsymbol{z}|\boldsymbol{x})$和$q(\boldsymbol{z})$的KL散度$\mathrm{KL}(p||q)$得到. 然而,由于$p(\boldsymbol{z}|\boldsymbol{x})$是未知的,难以直接计算. 变分推断通过引入相应变量的先验分布并设定合适的$q(\boldsymbol{z})$的形式来最优化逼近$p(\boldsymbol{z}|\boldsymbol{x})$.

根据贝叶斯公式,有

$$\ln p(\boldsymbol{X})=\ln\frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}-\ln\frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})}. \tag{3}$$

对式(3)两边求关于$\boldsymbol{Z}$的期望可得

$$\begin{aligned}\ln p(\boldsymbol{X})=&\int_{\boldsymbol{Z}}\ln p(\boldsymbol{X})q(\boldsymbol{Z})\mathrm{d}\boldsymbol{Z}=\\&\underbrace{\int_{\boldsymbol{Z}}q(\boldsymbol{Z})\ln\frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}\mathrm{d}\boldsymbol{Z}}_{L(q)}+\\&\underbrace{\Big\{-\int_{\boldsymbol{Z}}q(\boldsymbol{Z})\ln\frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})}\mathrm{d}\boldsymbol{Z}\Big\}}_{KL(p||q)}.\end{aligned} \tag{4}$$

在式(4)中,由于左边$\ln p(\boldsymbol{X})$固定,则最小化右边$\mathrm{KL}(p||q)$等价于最大化右边第1项$L(q)$. 因而,$L(q)$也常称为对数边缘似然($\ln p(\boldsymbol{X})$)的下界(evidence lower bound, ELOB)[23],即变分下界.

2 VBoGMM-sampling

本节详细介绍VBoGMM的基本原理和本文提出的VBoGMM-sampling的具体步骤.

2.1 GMM的变分推断

设有N个样本点集,$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}(\boldsymbol{x}_i\in R^{D\times1})$,独立采样于某个GMM. 为实现对应GMM的估计,为每个样本点$\boldsymbol{x}_n$引入一个K维的0-1向量隐变量$\boldsymbol{z}_n=(z_{nk})_{k=1}^{K}\Big(\sum\limits_k z_{nk}=1, z_{nk}=1$的位置用来记录$\boldsymbol{x}_n$属于对应的高斯分量$\Big)$. 为描述方便,记$\boldsymbol{\theta}=\{\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}\}$,其中$\boldsymbol{\Lambda}=\{\Sigma_1^{-1},\Sigma_2^{-1},\ldots,\Sigma_K^{-1}\}$为$K$个高斯分量的精度矩阵. 在变分推断中,假设$q(\boldsymbol{\theta})$是可分解的,即

$$q(\boldsymbol{\theta})=q(\boldsymbol{\theta}_1)q(\boldsymbol{\theta}_2)\ldots q(\boldsymbol{\theta}_F). \tag{5}$$

那么$\boldsymbol{\theta}_i$的最优估计结果为[23]

$$\hat{q}(\boldsymbol{\theta}_i)=\frac{\exp\{\boldsymbol{E}_{\theta_j|j\neq i}[\ln p(\boldsymbol{X},\boldsymbol{\theta})]\}}{\int\exp\{\boldsymbol{E}_{\theta_j|j\neq i}[\ln p(\boldsymbol{X},\boldsymbol{\theta})]\}\mathrm{d}\boldsymbol{\theta}_i}, \tag{6}$$

其中$\boldsymbol{E}[\cdot]$代表数学期望.

2.2 VBoGMM

如何选取高斯混合模型分量值K,对模型估计结果有重大影响. 传统的估计方法对高斯分量数K值往往根据经验给定,偶然性较大.

采用变分推断方法求解时,通过最大化变分下界($L(q)$)可确定具有K个高斯分量数的混合模型的后验分布. 然而,该方法需要在较大范围内,基于各个不同的K值建立起最优的基于变分估计GMM模型,即计算各种值对应的$L(q)$,然后选取最大$L(q)$对应的K值作为相应的最优变分GMM模型. 很明显,该方法虽然可以获得一个最好的K值,但是计算复杂度过高,难以实际应用.

受Corduneanu等[25]的启发,在进行变分贝叶斯优化GMM求解时,将高斯分量的混合参数$\boldsymbol{\pi}$当作超参数,首先选择较大的高斯成分数K,然后计算$\boldsymbol{\pi}$条件下的似然函数$p(\boldsymbol{X}|\boldsymbol{\pi})$,采用类似于Type II型极大边缘似然函数方法[23]获取$p(\boldsymbol{X}|\boldsymbol{\pi})$的极大值,获得相应的混合因子,最终当某个高斯分量实际上不存在时,其对应的混合参量π_k将自动衰减为零.

为了方便描述,重新定义$\boldsymbol{\theta}=\{\boldsymbol{\mu},\boldsymbol{\Lambda},\boldsymbol{Z}\}$,此时真实的对数边缘似然的积分下限变为

$$L(q)=\int q(\boldsymbol{\theta})\ln\frac{p(\boldsymbol{X},\boldsymbol{\theta}|\boldsymbol{\pi})}{q(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}. \tag{7}$$

式(7)中有

$$\ln p(\boldsymbol{X},\boldsymbol{\theta}|\boldsymbol{\pi})=$$
$$\ln p(\boldsymbol{Z}|\boldsymbol{\pi})+\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})+\ln p(\boldsymbol{\mu},\boldsymbol{\Lambda}), \tag{8}$$

且

$$p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})=\prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}(\boldsymbol{x}_n|\mu_k,\boldsymbol{\Lambda}_k^{-1})^{z_{nk}}. \tag{9}$$

在混合参数$\boldsymbol{\pi}$下,隐变量$\boldsymbol{Z}$的分布为一个K元Bernoulli分布,其概率密度表达式如下:

$$p(\boldsymbol{Z}|\boldsymbol{\pi})=\prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}. \tag{10}$$

高斯分布的均值向量和精度矩阵的共轭先验分布并不能写成独立的边缘分布形式,但根据Bayes公式,其总可以写成$p(\boldsymbol{\mu},\boldsymbol{\Lambda})=p(\boldsymbol{\Lambda})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})$. 研究表明,Guassian-Wishart分布为共轭先验分布[23],即

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda})=p(\boldsymbol{\Lambda})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})=$$
$$\prod_{k=1}^{K}\mathcal{N}(\mu_k|\boldsymbol{m}_0,(\beta_0\Lambda_k)^{-1})\mathcal{W}(\Lambda_k|\boldsymbol{W}_0,v_0), \tag{11}$$

其中$\mathcal{W}(\Lambda_k|\mathrm{W}_0,v_0)$代表Wishart分布.

将式(9)~(11)代入(8),再根据(6)可知$\hat{q}_i(\boldsymbol{\theta}_i|\boldsymbol{\pi})$具有与其先验分布$p(\boldsymbol{\theta}_i)$相同的概率密度形式,即

$$\hat{q}_1(\boldsymbol{Z}|\boldsymbol{\pi})=\prod_{n=1}^{N}\prod_{k=1}^{K}\rho_{nk}^{z_{nk}}, \tag{12}$$

$$\hat{q}_2(\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{\pi})=\hat{q}_{21}(\boldsymbol{\mu}|\boldsymbol{\Lambda})\hat{q}_{22}(\boldsymbol{\Lambda})=$$
$$\prod_{k=1}^{K}\mathcal{N}(\mu_k|\boldsymbol{m}_k,(\beta_k\boldsymbol{\Lambda}_k)^{-1})W(\boldsymbol{\lambda}_k|\boldsymbol{W}_k,v_k). \tag{13}$$

在实际求解中,这些参数因子的最优估计结果是关联耦合的,因而在进行GMM模型学习时,需要通过迭代求解,直到所获得的结果收敛. 因此,可以参考EM的求解方式来交替优化求解,以实现混合参数最优化.

1) E-Step: 采用式(12)和(13)获得$\ln(p(\boldsymbol{X}|\boldsymbol{\pi}))$的最优变分下界$L(q)$.

E-Step是一个迭代求解过程,直到参数收敛或者达到预设迭代步骤(或者$L(q)$小于某个阈值)说明已经获得$\ln(p(\boldsymbol{X}|\boldsymbol{\pi})$的最大变分下界.

2) M-Step: 对$L(q)$求关于$\boldsymbol{\pi}$的偏导数,令其等于0,通过重新评估$\boldsymbol{\pi}$来最大化$L(q)$. $\boldsymbol{\pi}$的更新规则为

$$\hat{\pi}_k=\frac{\sum_{n=1}^{N}\boldsymbol{E}[z_{nk}]}{N}=\frac{\sum_{n=1}^{N}\rho_{nk}}{N}=\frac{N_k}{N}. \tag{14}$$

如果考虑$\boldsymbol{\pi}$的先验概率$p(\boldsymbol{\pi})=\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0)=C(\boldsymbol{\alpha}_0)\prod_{k=1}^{K}\pi_k^{\alpha_0-1}$,则

$$\hat{\pi}_k=\frac{N_k+\alpha_k}{N+K\alpha_0}, \tag{15}$$

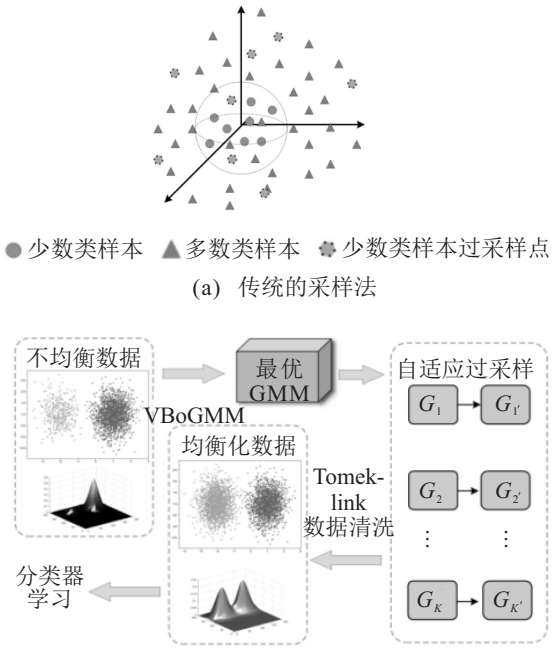
其中$\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0)$代表Dirchlet分布.

通过迭代E-Step和M-Step可实现GMM的最优估计,即先设置较大的高斯成分数K,经过迭代计算后,多余的高斯成分的混合权重参量π_k将趋于0. 本文通过设定相应的阈值,当π_k小于对应阈值时,其对应的高斯成分将被舍弃,从而能基于数据的分布特性获得最佳的GMM估计结果.

如果样本集$\boldsymbol{X}$中的点带有少量的高斯噪声,则相当于在$p(\boldsymbol{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})$中引入了一个新的高斯分量,但该分量的混合权重参数较小. 由于VBoGMM能通过变分优化对混合参数进行自动优化,让多余的或者轻微的高斯成分对应的混合权重自动趋于0,从而对噪声有较好的抑制作用.

2.3 VBoGMM-sampling流程

传统过采样法的缺陷及VoGMM-sampling方法流程如图1所示.



(a) 传统的采样法

(b) VoGMM-sampling流程

图1 传统过采样法的缺陷及VoGMM-sampling方法流程

在处理不均衡数据时,传统的数据层处理方法往往会忽略样本间原有的分布信息(如图1(a) 所示),从而可能会破坏样本间的分布信息,影响多数类和少数类样本的边界,导致分类结果恶化. 如果以GMM拟合产生的子类为基础对样本进行有约束地过采样,则能够充分利用样本间的空间分布特性,维持少数类样本的原始分布特性.

本文提出的VBoGMM-sampling首先采用VBoGMM获得不均衡数据中少数类样本的最优分布特性,再根据GMM的优化求解结果,针对每一个少数类的高斯分量分别进行自适应过采样;最后,依据Tomek-link规则对样本进行清洗,得到均衡化处理后的数据集. VBoGMM-sampling流程示意如图1(b)所示.

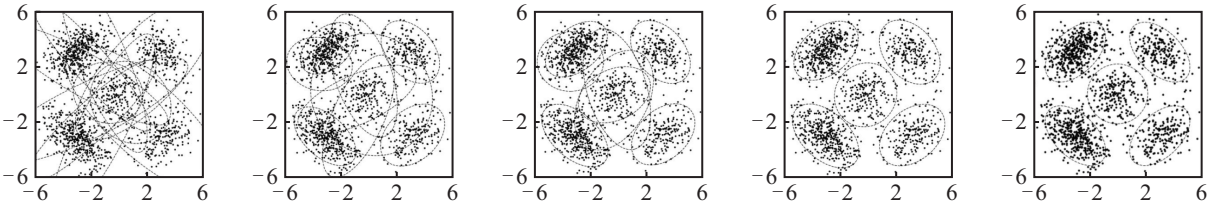
3　仿真实例

本节采用3个简单的数值实例验证VBoGMM的有效性.

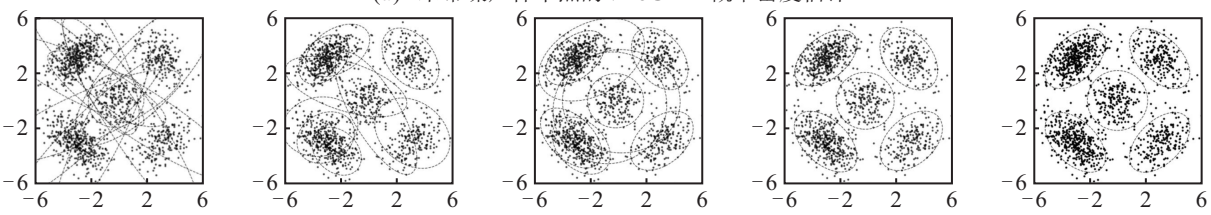
1) VBoGMM概率密度函数拟合. 设某GMM由5个高斯分量组成. 5个高斯分量的均值向量和协方差矩阵分别为$(0,0),(3,-3),(3,3),(-3,3),(-3,-3)$和$\begin{bmatrix}1&0\\0&1\end{bmatrix}$, $\begin{bmatrix}1&0.5\\0.5&1\end{bmatrix}$, $\begin{bmatrix}1&-0.5\\-0.5&1\end{bmatrix}$, $\begin{bmatrix}1&0.5\\0.5&1\end{bmatrix}$, $\begin{bmatrix}1&-0.5\\-0.5&1\end{bmatrix}$,混合权重为$[0.149,0.149,0.149,0.285\,7,0.285\,7]$,共生成了1 500个样本.

采用VBoGMM进行概率密度拟合时,将K值初始化为10. 图2(a)显示了基于VBoGMM的求解结果. 当某个π_k值小于1×10^{-5}时表明该高斯分量是可忽略的,图2中将不对该高斯分量进行显示. 从图2的迭代结果可以看出,当迭代70次时,所获得的GMM密度与真实的概率模型高度一致. 结果表明,VBoGMM能有效、正确地对少数类样本进行快速分布拟合,可用于对少数类样本的分析.

为进一步验证所提方法对噪声的鲁棒性,图2(b)显示了带有噪声的样本点的混合高斯模型的概率拟合结果. 样本点采用图2(a)中同样的混合高斯模型生成相应的数据点,但是各数据点受一定的噪声污染. 从图2(b)可以看出,所提方法也能在有限迭代中迅速恢复样本的真实分布模型.



(a)　未带噪声样本点的VBoGMM概率密度估计



(b)　带噪声样本点的VBoGMM概率密度估计

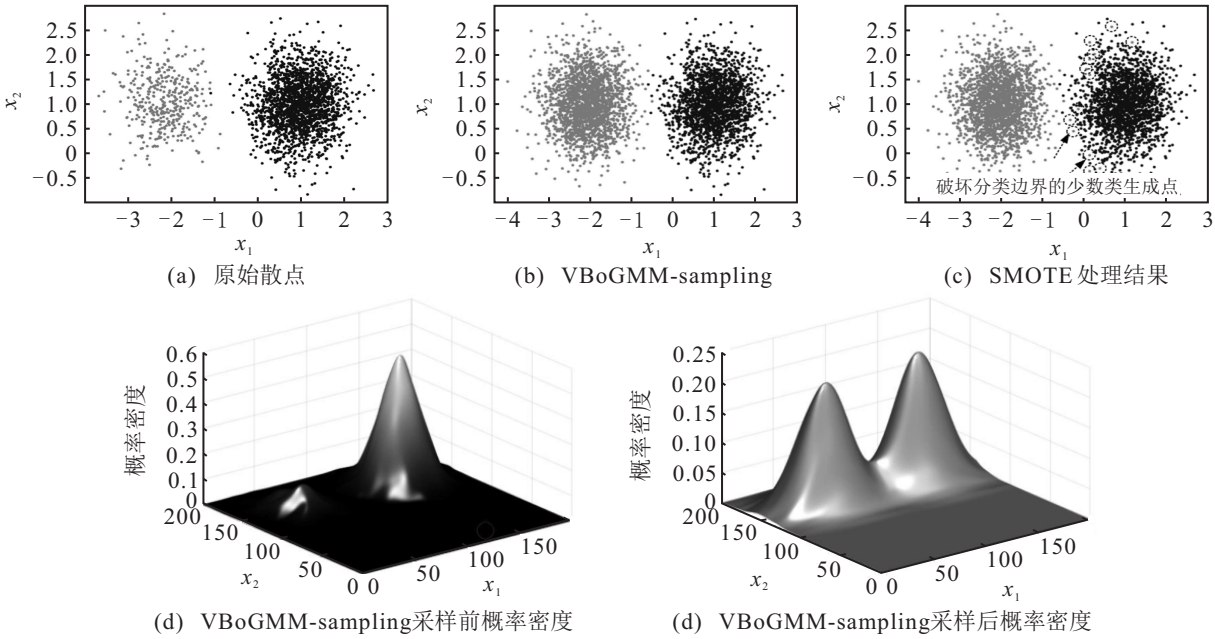
图2　VBoGMM概率密度函数拟合

2) 2个相互独立且服从高斯分布的样本集. 设两个独立高斯分布的样本点集的中心分别为$(-1,1)$、$(1,1)$,初始方差分别设置为0.38、0.38,这两个样本集的点集数目分别为300和2 000,其初始散点图、VBoGMM-sampling和SMOTE方法的采样结果,以及VBoGMM-sampling方法采样前后样本集的概率密度估计结果如图3所示.

从图3可以明显看出,SMOTE方法生成的少数类点会对分类边界产生较大影响,极有可能破坏分类边界,导致分类器偏置. 而使用VBoGMM-sampling进行少数类样本过采样后,新增数据点集中在少数类样本簇,并未影响分类边界;同时,采样前后数据依然服从高斯分布,说明VBGMM-sampling有效保持了采样前后数据分布的稳定性. 采样后样本方差为0.37、0.38,也在一定程度上反映了样本点集的数据特性保持稳定.

3) 两组分布不均衡的随机分类数据集.

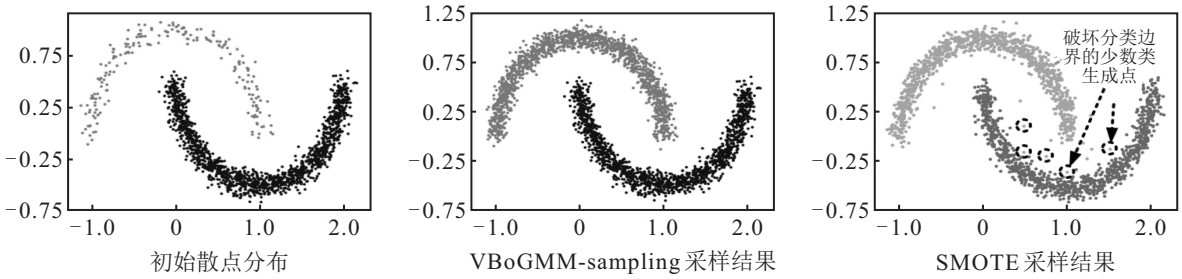
① 数据集A: 共包含1 300个样本点,4维特征,其中2维为有效特征,2维为冗余特征. 采用VBoGMM-sampling、SMOTE的采样处理结果如图4(a)所示.



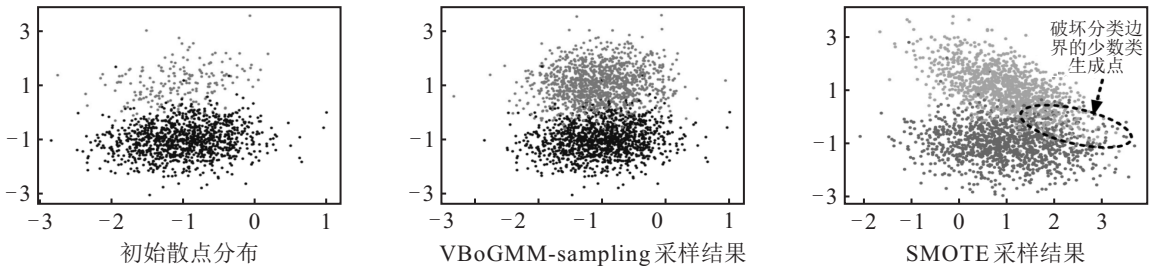
(a) 原始散点 (b) VBoGMM-sampling (c) SMOTE处理结果

(d) VBoGMM-sampling采样前概率密度 (d) VBoGMM-sampling采样后概率密度

图3 两个独立高斯样本集采样前后散点图及概率密度估计结果



(a) 随机分类数据集A采样前后散点分布对比



(b) 随机分类数据集B采样前后散点分布对比

图4 传统过采样法的缺陷及VoGMM-sampling方法流程

仿真数据集A中两类样本在交汇处存在分类边界不明确、模糊化等问题. 采用VBoGMM-sampling采样后,样本点仍集中于少数类样本区域,这是由于VBGMM-sampling先对少数类样本进行分析,插值过程围绕着少数类高斯分布群,使新增样本点集中于少数类样本附近.

② 数据集B: 共包含1 300个样本点,有2维特征. 采用VBoGMM-sampling、SMOTE进行采样处理,结果如图4(b)所示.

由图4可知,使用VBoGMM-sampling对随机生成的分类数据进行采样后,样本的分布信息依然保持稳定,类别之间的边界也并未被破坏,表明VBoGMM-sampling在处理各类分布的不均衡数据时,预期性能保持稳定.

4 实验验证

将所提出的VBoGMM-sampling方法与分类器组合进行不均衡集分类实验,以验证本文方法的优劣. 本文选用朴素贝叶斯分类器(naive Bayes classifier, NBC)和随机森林(random forest, RF)[23]这两类有代表性的分类方法. 分类性能评价标准包括G-mean, F-measure, Precision以及Recall[1].

4.1 实验1

选用5组来自UCI的不均衡数据集进行实验. 表1列出了数据集的基本信息. Vehicle为车辆数据集,

根据车辆的轮廓进行分类; Glass为玻璃分类数据集, 通过氧化物的含量对6种玻璃进行分类; Inosphere为电离层数据集, 通过样本属性对雷达反馈进行分类; Abalone数据集为鲍鱼数据集, 通过环数预测鲍鱼年龄; Seg数据集通过样本属性对场景进行辨别. 表1中IR (imbalance rate)为多数类与少数类样本数目的比值.

表 1　数据集基本信息描述

名称	样本数	属性	少数类样本数	多数类样本数	IR
Vehicle	846	18	212	634	2.99
Glass	214	11	41	115	2.80
Inosphere	351	34	225	126	1.79
Abalone	731	7	42	689	16.4
Seg	2 310	19	330	1 980	6

实验结果如表2所示. 经本文方法处理后, 无论采用NBC还是采用RF分类器, F-measure和G-mean相较于SMOTE方法都有不同程度地提高, 说明经VBoGMM-sampling处理后, 分类器整体分类性能以及对少数类样本的分类精度显著提高, 本文提出的方法能有效地处理数据类不均衡的问题, 从而提高分类器的性能.

表 2　分类实验结果

数据集	过采样方法	NBC分类性能		RF分类性能	
		G-mean	F-Measure	G-mean	F-Measure
Vehicle	SMOTE	82.1	79.6	82.2	80.4
	VBoGMM-sampling	82.4	81.4	85.4	83.6
Glass	SMOTE	74.4	75.2	76.4	74.5
	VBoGMM-sampling	75.1	77.3	79.4	80.3
Inosphere	SMOTE	72.4	73.4	73.8	74.4
	VBoGMM-sampling	75.4	75.2	77.5	78.6
Abalone	SMOTE	78.4	78.9	78.2	75.3
	VBoGMM-sampling	80.1	80.9	82.4	78.8
Seg	SMOTE	80.3	79.9	81.2	82.4
	VBoGMM-sampling	81.4	80.5	84.5	85.6

4.2　实验2

在金融风险领域, 分辨、预防信用卡欺诈一直受到广受关注. 本节进一步采用信用卡欺诈检测数据集(CreditCard)进行实验.

CreditCard数据集(https://www.kaggle.com/mlg-ulb/creditcardfraud)为欧洲持卡人的一些信用卡交易数据(2013年9月), 包含高达28万次交易(原始数据文件约150.83 M), 但其中仅492次为欺诈行为. 对于此类严重不均衡的数据集进行分类处理时, 若不对样本进行一定的均衡化处理, 则导致分类器性能严重下降, 特别是难以有效检测出数据集中的少数类样本.

CreditCard数据集包含了31维特征, 除了“Time”和“Amount (记录了交易金额)”两个特征外, 其余的维特征是使用主成分分析后获得的主成分变量信息. 将数据集按比例8:2划分为训练集和测试集, 分别使用简单随机过采样、SMOTE、基于传统GMM模型的ADASYN方法(简写为GMM-sampling)以及VBoGMM-sampling对训练集进行过采样处理, 得到相应的均衡化处理后的结果用于分类器学习.

采用不同方法对数据集进行均衡化处理后, 再分别训练NBC、RF分类器进行信用卡欺诈检测实验. 测试集中共含有54 114条交易数, 其中欺诈性交易条数为100条. 表3详细列出了实验recall和误判数(FP)结果.

表 3　信用卡欺诈检测结果

过采样方法	NBC检测器		RF检测器	
	recall	误判数	recall	误判数
随机过采样	0.88	1 285	0.90	1 189
SMOTE	0.89	923	0.90	922
GMM-sampling	0.90	906	0.91	886
VBoGMM-sampling	0.9	496	0.92	472

从表3可以看出, 无论是采用NBC还是RF作为分类检测器, 采用4种过采样方法进行数据集均衡化处理后, 信用卡欺诈检测的recall值都比较接近, 基本可达到0.9, 但很明显, 简单随机过采样和SMOTE方法对于普通交易的误判数明显高于GMM-sampling和本文提出的VBoGMM-sampling. 特别是采用VBGMM-sampling处理后, 无论采用NBC还是RF分类器, 其误判数都小于490条, 低于简单随机过采样方法的误判数目的一半.

随机过采样从样本少的类别中随机抽样, 该方法虽然简单, 但只通过随机复制, 不仅忽略了少数类样本周围多数类样本的分布情况, 还导致样本分布信息破坏, 加大了分类器过拟合的可能性. 变分推断从贝叶斯公式出发, 将后验推断问题巧妙地转化为优化问题进行求解, 具有快速收敛性和高度可扩展性. 使用变分推断求解高斯混合模型, 相较于传统的EM算法求解, 具有计算开销小、收敛快的优点. 在处理CreditCard等大规模数据集时, 性能良好.

综上可知, 采用VBoGMM-sampling能更有效地学习少数类样本间的分布信息, 减少过采样对于分类精度的影响, 进而大大降低相应数据类型的误判可能性, 有效提高相应的分类器的检测性能.

5　结　论

本文提出的基于变分贝叶斯的最优化GMM建模方法VBoGMM能将多余的高斯成分的混合权重自动衰减为零, 实现任意未知分布的最优分布建模. 进而, 基于样本的空间分布特性, 对数据中的少数类样本进行自适应过采样, 以改善数据的不均衡度, 方便后续分类模型学习. 在多种公共数据集上进行

了大量的验证性和对比性实验,结果表明,本文提出的不均衡数据处理方法能有效获得数据内部真实的分布特性,有效提高了传统分类模型的分类性能. 后续将进一步深入研究基于生成模型的少数类样本的模式分类方法,比如将生成对抗的思想引入到不均衡数据集处理中,以获得更好的分类性能.

参考文献(References)

[1] Tharwat A. Classification assessment methods[J]. Applied Computing and Informatics, 2021, 17(1): 168-192.

[2] He H B, Garcia E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[3] Liu J P, Zhou J M, He J B, et al. Spectral clustering-fused adaptive synthetic oversampling approach for imbalanced data processing[J]. CAAI Transactions on Intelligent Systems, 2020, 15(4): 732-739.

[4] Dong Q, Gong S G, Zhu X T. Imbalanced deep learning by minority class incremental rectification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(6): 1367-1381.

[5] Hou X D, Zhang T, Ji L, et al. Combating highly imbalanced steganalysis with small training samples using feature selection[J]. Journal of Visual Communication and Image Representation, 2017, 49(11): 243-256.

[6] Viegas F, Rocha L, Gonçalves M, et al. A genetic programming approach for feature selection in highly dimensional skewed data[J]. Neurocomputing, 2018, 273: 554-569.

[7] Moayedikia A, Ong K L, Boo Y L, et al. Feature selection for high dimensional imbalanced class data using harmony search[J]. Engineering Applications of Artificial Intelligence, 2017, 57: 38-49.

[8] Li Q J, Mao Y B, Wang Z Q. Research on boosting-based imbalanced data classification[J]. Computer Science, 2011, 38(12): 224-228.

[9] Liu M, Xu C, Luo Y, et al. Cost-sensitive feature selection by optimizing F-measures[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2018, 27(3): 1323-1335.

[10] 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: 321-357.

[11] Olken F, Rotem D. Random sampling from databases: A survey[J]. Statistics and Computing, 1995, 5(1): 25-42.

[12] Blagus R, Lusa L. Improved shrunken centroid classifiers for high-dimensional class-imbalanced data[J]. BMC Bioinformatics, 2013, 14(1): 64.

[13] Zhang C K, Zhou Y, Guo J W, et al. Research on classification method of high-dimensional class-imbalanced datasets based on SVM[J]. International Journal of Machine Learning and Cybernetics, 2019, 10(7): 1765-1778.

[14] He H B, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]. 2008 IEEE International Joint Conference on Neural Networks. Hong Kong, 2008: 1322-1328.

[15] Liu S G, Zhang J, Xiang Y, et al. Fuzzy-based information decomposition for incomplete and imbalanced data learning[J]. IEEE Transactions on Fuzzy Systems, 2017, 25(6): 1476-1490.

[16] Razavi-Far R, Farajzadeh-Zanjani M, Saif M. An integrated class-imbalanced learning scheme for diagnosing bearing defects in induction motors[J]. IEEE Transactions on Industrial Informatics, 2017, 13(6): 2758-2769.

[17] 刘金平, 周嘉铭, 刘先锋, 等. 基于聚类簇结构特性的自适应综合采样法在入侵检测中的应用[J]. 控制与决策, 2021, 36(8): 1920-1928.
(Liu J P, Zhou J M, Liu X F, et al. Toward intrusion detection via cluster structure-based adaptive synthetic sampling approach[J]. Control and Decision, 2021, 36(8): 1920-1928.)

[18] Chen S H. A generalized Gaussian distribution based uncertainty sampling approach and its application in actual evapotranspiration assimilation[J]. Journal of Hydrology, 2017, 552: 745-764.

[19] Li D C, Hu S C, Lin L S, et al. Detecting representative data and generating synthetic samples to improve learning accuracy with imbalanced data sets[J]. PLoS One, 2017, 12(8): e0181853.

[20] Zhang T L, Yang X. G-SMOTE: A GMM-based synthetic minority oversampling technique for imbalanced learning[EB/OL]. 2018, arXiv: 1810.10363.

[21] Blei D M, Kucukelbir A, McAuliffe J D. Variational inference: A review for statisticians[J]. Journal of the American Statistical Association, 2017, 112(518): 859-877.

[22] Devi D, Biswas S K, Purkayastha B. Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance[J]. Pattern Recognition Letters, 2017, 93: 3-12.

[23] Bishop C M. Pattern Recognition and Machine Learning[M]. New York: Springer, 2006: 179-485.

[24] Andrieu C, Freitas N D, Doucet A, et al. An introduction to MCMC for machine learning[J]. Machine Learning, 2003, 50(1): 5-43.

[25] Corduneanu A, Bishop C M. Variational bayesian model selection for mixture distributions[C]. Proceedings Eighth International Conference on Artifificial Intelligence and Statistics. Morgan Kaufmann, 2001: 27-34.

作者简介

刘金平(1983−), 男, 教授, 博士, 从事复杂工业过程智能监测、人工智能与机器学习的研究, E-mail: ljp202518@163.com;

杨本芳(1999−), 男, 硕士生, 从事智能信息处理的研究, E-mail: 739156765@qq.com;

周嘉铭(1996−), 男, 硕士生, 从事计算机视觉、模式识别的研究, E-mail: zhouj1am1ng@qq.com;

徐鹏飞(1976−), 男, 副教授, 博士, 从事计算机视觉、模式识别等研究, E-mail: xupf@hunnu.edu.cn.