文章编号: 1001-0920 (2014) 05-0815-06 **DOI:** 10.13195/j.kzyjc.2013.0209

基于概率分布估计的混合采样算法

曹 鹏, 李 博, 栗 伟, 赵大哲

(东北大学 a. 信息科学与工程学院，b. 医学影像计算教育部重点实验室，沈阳 110004)

摘 要: 在类别不均衡的数据中, 类间和类内不均衡性问题都是导致分类性能下降的重要因素. 为了提高不均衡数据集下分类算法的性能, 提出一种基于概率分布估计的混合采样算法. 该算法依据数据概率分别对每个子类进行采样以保证类内的均衡性; 并扩大少数类的潜在决策域和减少多数类的冗余信息, 从而同时从全局和局部两个角度改善数据的平衡性. 实验结果表明, 该算法提高了传统分类算法在不均衡数据下的分类性能.

关键词: 不均衡数据学习；类内不均衡；混合采样；概率分布估计

中图分类号: TP273 **文献标志码:** A

Hybrid sampling algorithm based on probability distribution estimation

CAO Peng, *LI Bo*, *LI Wei*, *ZHAO Da-zhe*

(a. College of Information Science and Engineering，b. Key Laboratory of Medical Image Computing of Ministry of Education，Northeastern University，Shenyang 110004，China．Correspondent：CAO Peng，E-mail：cao.p@neusoft.com)

Abstract: In the class imbalanced data distribution, both the between-class and within-class imbalance issues are critical factors to decrease the performance. To improve the performance of classifier algorithm on the imbalanced data, a hybrid sampling algorithm based on probability distribution estimation is proposed. The approach re-samples the data of subclass to balance the distribution in each class based on probability distribution estimation. Moreover, it expands the decision region of minority class and removes the redundant information of majority class, so as to solve the imbalance issues from both global and local perspectives simultaneously. Experimental results show that the proposed method improves the classification performance for imbalanced data.

Key words: imbalanced data learning；within-class imbalance；hybrid sampling；probability distribution estimation

0 引 言

不均衡数据分类问题是机器学习领域的研究主题之一, 受到越来越多的重视[1]. 数据分布的不均衡性广泛存在于现实应用中, 如文本分类、医疗诊断等领域. 在不均衡数据分类时, 各个类别的样本数目存在较大的差异, 导致不同类别的样本对于训练算法提供的信息不对称; 因为传统分类器都是基于准确率最大化来进行训练, 所以常常忽略了数量小的类别信息, 从而影响了传统分类器的分类结果.

关于不均衡数据的研究可分为基于数据层面[2-3]和基于分类算法层面[4-5]两大类. 数据层面的方法主要是通过改变训练数据的分布来降低数据的不均衡性; 算法层面主要是改进原有算法的工作原理或设计对不均衡分布不敏感的新算法, 提升少数类的识别率. 上述两类方法之间的关系及效果的比较目前仍不是十分清楚, 但基于数据层面的方法更加灵活, 是解决不均衡问题最简单有效的技术手段.

目前该领域已提出了很多数据层面的重采样算法, 但是仍有两点不足: 1) 通过升采样或降采样方法可以令数据分布尽量均衡化. 然而, 理论和实验研究表明, 分类器性能的下降不能只归咎于多数类与少数类之间的不均衡性, 数据中存在的小区块问题也是导致分类器性能不佳的重要原因[6]. 小区块即为那些覆盖样本数量偏少的子概念. 传统的分类算法本身的归纳偏置对这些含有少量样本的子概念无法进行有效的学习, 如决策树和关联规则分类算法等. 而这些小区块正是由于类内的若干子概念之间的样本分布不均衡造成的, 即类内不均衡[7]. 类间和类内不平衡是

收稿日期: 2013-02-27；**修回日期:** 2013-05-10.

基金项目: 国家自然科学基金项目(61001047)；中央高校基本科研业务费专项资金项目(N110618001).

作者简介: 曹鹏(1982−), 男, 博士生, 从事数据挖掘的研究；赵大哲(1960−), 女, 教授, 博士生导师, 从事数据挖掘、医学影像处理等研究.

不均衡数据学习的两个不同侧面, 它们可能会同时出现, 都将影响分类器的性能[8]. 2) 现有的采样方法或者是简单地通过随机方式进行复制或删除样本, 或者是基于某些启发式的规则插入样本或过滤噪音样本, 都没有遵循数据本身的分布规律, 当所设计的样本生成规则与潜在真实类分布不完全一致时, 将不可避免地向训练样本集内引入噪声, 并扭曲数据的空间分布.

鉴于上述问题, 本文提出一种基于概率分布估计的混合采样算法 (HPS), 从全局和局部两个角度对不均衡问题进行深入的剖析和处理. 该算法使用高斯混合模型[9]对两个类分别进行概率分布参数估计和数据分解, 并根据估计的概率函数对类中的每个子类数据进行有效的采样: 在保证多数类的空间结构不变的情况下去除冗余的信息; 根据样本的分布规律对少数类合成更加准确真实的新样本, 挖掘和扩展潜在的决策区域. 通过实验验证, 该算法有效改善了不均衡数据分布问题, 提升了传统分类器的分类性能.

1 相关工作

数据重采样算法是通过改变原有训练集 D, 使分类器在新数据集 D^* 上能够提高性能. 采样的目的是找到采样过程 $S: D \to D^*$. 采样方法可进一步分为降采样和升采样. 降采样是通过减少多数类样本来提高少数类的分类性能, 但容易丢失多数类的一些重要信息, 使分类器无法进行充分的学习. 虽然很多改进方法通过一定的规则, 有选择地去掉对分类作用不大的多数类样本, 如 NCL[10]算法, 但在复杂的数据分布中很难准确地确定噪音及冗余样本, 导致很多重要信息被删除.

升采样技术则是通过增加少数类的样本来提高少数类的分类性能, 最简单的升采样方法是随机复制少数类样本, 缺点是没有给少数类增加任何新的信息, 会使分类器学到的决策域变小, 导致过学习. 较高级的升采样算法则采用一些启发式的技巧, 有选择地插入少数类样本, 如 SMOTE 算法[3], 对每个少数类样本随机选出几个邻近样本, 并且在该样本与这些邻近样本的连线上随机取点, 生成无重复的新的少数类样本. 但 SMOTE 在插入新样本时未考虑多数类的分布, 插入了很多噪音, 造成过泛化. 为了防止合成的新样本侵入到多数类空间中, 在 SMOTE 基础之上 Barua 等[11]提出了一种基于聚类的升采样算法 —— CBSO 算法, 先对少数类样本进行聚类, 再划分成若干簇区域, 在对少数类进行采样时可保证新样本在同一个簇区域内.

以上的升采样算法虽在一定程度上提高了少数类的识别率, 但其采样策略只是针对样本之间的连线进行插入, 仍有大量重要的空间未能开采, 而且没有考虑类内的不均衡性问题. 文献 [8] 发现了类内不均衡问题的严重性, 提出了一种基于 K-means 的升采样算法 KOS 来解决类内和类间的不均衡性问题, 利用 K-means 对两类样本分别进行聚类, 在每个簇内使用随机升采样, 使同类中的所有簇数据以及两类整体数据都达到均衡化. 该算法虽然达到了整体和局部的均衡性, 但却引入了另外的问题: 1) 对两类数据都只采用了随机升采样技术, 造成冗余数据增多, 增加了数据处理的时间消耗; 2) 聚类的数目需要人工设置, 而且很难找到最佳的簇个数.

2 基于概率分布的混合采样算法

为了使采样算法生成的新样本空间 D^* 更加符合数据的真实分布, 并改善类间与类内的不均衡分布, 该算法基于高斯混合模型 (GMM) 概率分布, 分别对多数类和少数类进行采样. 即首先通过 GMM 概率分布对两类样本分布进行模拟和数据分解, 再根据学习到的概率分布函数进行样本的重采样. GMM 可视为由 L 个高斯分布以一定比例混合而成, 每个高斯成分用均值 μ 和协方差矩阵 σ 来决定其几何特征, 如下式所示:

$$p(x) = \sum_{l=1}^{L} p(l)p(x|l) = \sum_{l=1}^{L} \pi_l N(\mu_l, \sigma_l), \qquad (1)$$

其中 $\sum_{l=1}^{L} \pi_l = 1$.

由于 GMM 是一种对数据的真实分布进行模拟和逼近的半参数表达模型, 可以近似于任意的数据分布, 假定两类样本数据遵从具有某种参数的高斯混合分布, 从而使用 GMM 分别对两类的分布进行参数估计. 对 GMM 进行参数估计的常用方法是 EM, 但由于其需要人工指定聚类的个数, 并且其对初始值点较为敏感, 本文中使用 Figueiredo-Jain (FJ) 解法[12]来对参数进行估计. 该算法可以自动确定高斯模型的最佳个数, 从而更加准确地计算出模型中的参数. HPS 算法流程如下.

Step 1 计算两类样本的采样数量.

若数据集 D 中两类样本数量分别为 M_{maj} 和 M_{min}, 则混合采样数量 $N = (M_{\mathrm{maj}} - M_{\mathrm{min}}) \times \gamma$, $\gamma = 1$. 升采样的数量为 $N_{\mathrm{min}} = N \times R_{\mathrm{hs}}$, 降采样的数量为 $N_{\mathrm{maj}} = N \times (1 - R_{\mathrm{hs}})$, R_{hs} 为混合采样的比例系数.

Step 2 过滤数据集中的噪音样本.

为了防止噪音样本对概率分布估计造成影响, 首先对样本进行预处理. 根据最近邻域思想, 当样本数据的 M 个最近邻域样本中超过 4/5 为相反的类别 ($M = 5$) 时, 该样本是噪音的可能性较大, 应进行过滤. 过滤后生成新数据集 D', $D' = \mathrm{filter}(D)$.

Step 3　对两类样本分别进行高斯混合建模.

利用 FJ 解法对 D' 中的两类样本分别进行概率密度估计, 每个样本被分到概率最大的簇中, 完成数据聚类分割的工作, 并获得概率密度函数参数. 确定了两类样本的采样数量和数据子类后, 对两类样本分别采样以解决类间和类内的不均衡性问题.

Step 4　对少数类进行升采样.

Step 4.1　确定少数类各个子类的采样数量.

为解决少数类内部各个子类(簇)之间的均衡性, 每个子类的升采样数量应与子类中的样本数量成反比. 利用下式计算每个子类的升采样数量:

$$N_{\min}^{i}=\left(\frac{1}{\text{size}_{\min}^{i}}\Big/\sum_{j=1}^{S_{\min}}\frac{1}{\text{size}_{\min}^{j}}\right)\times N_{\min}. \tag{2}$$

其中: $\text{size}_{\min}^{i}$ 为少数类中 i-th 子类的样本数量, $S_{\min}$ 为少数类中子类的个数.

Step 4.2　对少数类的子类进行升采样.

在每个子类中使用该子类的高斯密度函数进行升采样. 由于边界数据对少数类的识别作用较大, 为了扩展少数类的决策域, 需要对少数类样本有选择地采样, 即重点对位于边界区域的少数类样本进行采样. 根据文献 [13] 的思想, 计算子类中每个样本 x_k 的采样权重 r_k, 即

$$r_k=\frac{1}{1+\exp(-\alpha\times\delta_k)}, \tag{3}$$

其中 δ_k 是 x_k 的 K_1 个邻域中多数类的个数. δ_k 越大, 说明该样本是边界样本的概率越大, 所以采样的权重越大 (由文献 [21] 可设为 0.25). 之后根据下式对 r_k 进行归一化:

$$\hat{r_k}=r_k\Big/\sum_{j=1}^{\text{size}_{\min}^{i}}r_j. \tag{4}$$

再由归一化的采样权重, 计算每个样本 x_k 的采样数量 g_k, 有

$$g_k=N_{\min}^{i}\times\hat{r_k}. \tag{5}$$

因而边界区域的少数类样本采样的数量较其他区域更多. 基于每个样本 x_k 进行升采样时, 为了在扩展潜在空间的同时避免引入噪音, 限定在其 K_2 最近邻域的区域之内合成 g_k 个新样本.

Step 5　对多数类进行降采样.

Step 5.1　确定多数类各个子类的采样数量.

对于多数类, 每个子类的降采样数量应与子类中的样本数量成正比, 即样本数量多的子类其降采样的数量也多, 从而保证了多数类内部各个子类的均衡. 每个子类的降采样数量计算如下:

$$N_{\text{maj}}^{i}=\left(\text{size}_{\text{maj}}^{i}\Big/\sum_{j=1}^{S_{\text{maj}}}\text{size}_{\text{maj}}^{j}\right)\times N_{\text{maj}}. \tag{6}$$

其中: $\text{size}_{\text{maj}}^{i}$ 为多数类 i-th 子类的样本数量, S_{maj} 为多数类中子类的个数.

Step 5.2　对多数类的子类进行降采样.

对于多数类的每个子类中, 需要在保持子类的空间结构信息不被破坏的前提下减少冗余样本. 因每个子类的高斯分布中, 处于中心位置的区域样本分布较其他区域更为稠密, 故应具有更高的降采样几率. 计算每个样本 x_k 的高斯概率值并归一化, 每个样本被移除的几率与每个样本的归一化高斯概率值成正比, 从而在压缩多数类的同时保留了具有代表性的样本.

通过图 1 可以直观地了解 HPS 采样的原理以及与 SMOTE 采样的区别. 原数据分布 D 如图 1 (a) 所示 (三角为少数类, 圆形为多数类). 图 1 (b) 为 SMOTE 采样后的数据分布 D_{SM}^{*} (矩形为新合成的少数类样本), 可以看到, 基于样本连线插值的 SMOTE 算法没有考虑多数类的数据分布, 生成很多噪音数据. 图 1 (c) 为对两类分别进行 GMM 建模及分解的结果: 多数类有 3 个簇, 样本数量为 20, 15, 10; 少数类有 2 个簇, 样本数量为 10, 5. 图 1 (d) 为 HPS 采样分布 D_{HPS}^{*}. 两类样本的采样总数量 $N=M_{\text{maj}}-M_{\min}=30$, 若 $R_{\text{hs}}=0.5$, 则两类的采样数量为 $N_{\text{maj}}=N_{\min}=15$.

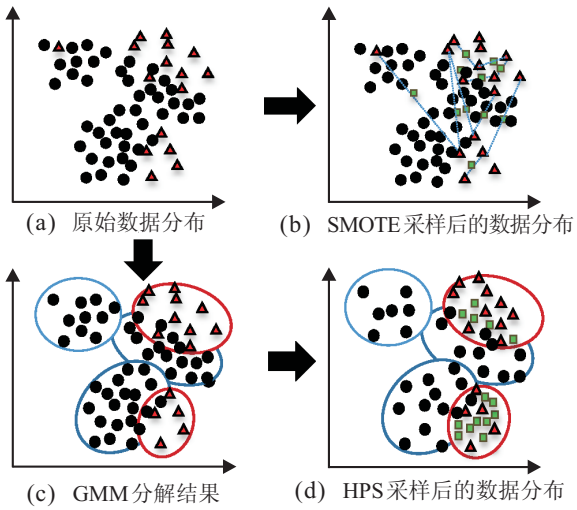

图 1　基于概率分布的混合采样算法图示

由式 (2) ~ (6) 计算每个子类的采样数量并进行重采样, 采样后多数类各个子类的数量分别为 13, 10, 7; 少数类的数量均为 15, 同时缓解了类间和类内的不均衡. 另外可以看到: 对于多数类, 在保证每个子类的样本空间结构不发生变化的前提下, 减少了冗余数据; 而对于少数类中的子类区域, 增加有效的少数类信息, 扩展了数据空间并避免了插入的样本侵入多数类空间, 这与基于全局采样的 SMOTE 算法的采样原理有着本质上的区别. 由于本文算法是基于局部的数据空间, 即以簇区域为单元进行采样的, 可以获得更

加合理的数据分布, 文献 [14] 也证明了从局部数据入手进行处理可以达到更好的采样效果.

3 实验评估

3.1 数据集描述

为了评估算法的性能, 选择 9 组具有不同不均衡比例的 UCI 数据集进行测试, 数据集详细信息如表 1 所示.

表 1 实验数据集描述

数据集 (+)	样本数	特征数	不均衡比例/%
Glass (tableware)	214	9	4
Letter (Z)	20 000	16	4
Page (2, 3, 4, 5)	5 473	10	10
Segment (bricface)	2 310	19	14
Transfusion (yes)	748	4	24
Vehicle (opel)	846	18	25
German (1)	1 000	20	30
Pima (positive)	768	8	35
Spambase (spam)	4 601	57	40

3.2 HPS 算法的性能验证

为了验证 HPS 算法对不均衡数据学习的有效性, 使用几种常用的不均衡数据处理算法进行性能比较, 如 SMOTE (SM), SMOTE+NCL (SML), KOS, CBSO. 其中: SMOTE 的采样率 $R_{\rm os}$ 设为 200%; KOS 算法中聚类个数 K 设为 2; CBSO 算法中 β 设为 1, 即代表升采样数量等于两类数量的差额. 在 HPS 算法中, 为了避免减少过多的信息, $R_{\rm hs}$ 设置为 0.7. 参数 K_1 和 K_2 是分别用来控制样本采样权重和新样本合成区域大小的. 对于高维数据, 根据欧氏距离确定的邻域并不能准确描述样本之间的相对位置, 所以 K_1 选取过大会获得不准确的权重值; 而且由于处于决策区域的样本周围会有较多的多数类样本, 为避免引入噪音 K_2 也不宜设置过大. 通过多组实验发现, 邻域参数 K_1 和 K_2 均设为 5 时可获得较好且较稳定的采样性能. 本实验使用不均衡数据学习最常用的决策树 C4.5 算法作为基分类器.

分类器评测指标直接影响着分类器的性能, 传统的准确率对于不均衡数据的评价不再有效, 这里将采用不均衡分类的评测指标 GM (Geometric mean) 和 AUC (Area under the ROC curve). 其中: GM 是综合衡量两类准确率的指标 ($\mathrm{GM} = (\mathrm{ACC}_{\rm maj} \times \mathrm{ACC}_{\rm min})^{1/2}$), 只有两类的准确率都较高时才能得到较高的 GM 值; AUC 是另一个有效的不均衡数据分类性能评价手段, 由于 ROC 曲线作为分类器评估的可视化技术得到了广泛应用, AUC 能以定量的方式表示 ROC 曲线对应的分类器性能. 所有算法都使用 10 折交叉验证. 由于升采样算法本身具有一定的随机性, 对于所有升采样算法, 在每次交叉验证时, 对训练集分别执行 5 次采样操作, 并在采样后的训练数据集分别构建分类器并测试, 最后将 5×10 次结果计算均值作为该算法的无偏结果, 实验结果如表 2 所示.

表 2 不同算法的分类性能比较结果

数据集	指标	C4.5	SM	SML	KOS	CBSO	HPS
Glass	GM	0.859	0.923	0.931	0.897	**0.933**	0.876
	AUC	0.968	0.995	**0.997**	0.977	**0.997**	0.994
Letter	GM	0.928	0.935	0.935	0.919	0.934	**0.955**
	AUC	0.986	0.992	0.996	0.980	0.995	**0.998**
Page	GM	0.728	0.814	**0.843**	0.764	0.827	0.833
	AUC	0.816	0.839	0.863	0.818	0.845	**0.869**
Segment	GM	0.979	0.988	0.994	0.975	0.991	**0.996**
	AUC	0.998	**0.999**	**0.999**	0.995	**0.999**	**0.999**
Transfusion	GM	0.526	0.577	0.583	0.579	0.581	**0.596**
	AUC	0.751	0.755	0.766	0.752	0.771	**0.792**
Vehicle	GM	0.642	0.688	0.712	0.704	0.694	**0.751**
	AUC	0.935	0.989	0.991	0.989	0.991	**0.994**
German	GM	0.618	0.641	0.653	0.605	**0.662**	0.656
	AUC	0.715	0.728	0.739	0.722	**0.764**	0.757
Pima	GM	0.579	0.649	0.660	0.644	0.637	**0.668**
	AUC	0.744	0.766	0.767	0.788	0.760	**0.806**
Spambase	GM	0.766	0.845	0.849	0.821	0.853	**0.892**
	AUC	0.951	0.972	0.969	0.967	0.977	**0.984**

通过实验对比发现, HPS 在多数数据集上均优于其他采样算法. 由于 HPS 算法是基于概率密度函数进行采样, 在少数类样本数量充足的条件下, 可准确估计出真实的概率密度函数, 从而发现潜在的数据空间并合成准确的新样本. 不均衡数据中少数类一般分为绝对稀缺和相对稀缺. HPS 在少数类样本数量不充足的情况下, 即绝对稀缺时, 将无法准确获得少数类样本的概率分布参数, 导致可能插入不准确的样本, 无法达到最佳的采样效果. 如 Glass 数据集中少数类只有 9 个样本, 所以 HPS 算法在该数据集中较其他算法性能有所下降. 但不均衡比例同为 4% 的 Letter 数据集含有充足的样本, 属于相对稀缺, HPS 算法可以达到理想的结果. 对于绝对稀缺的数据学习问题一直是数据挖掘中重点研究的课题之一, 目前仍没有有效的解决方法[15].

从实验结果还可以看到 SMOTE 和 CBSO 在大多数数据集下提升了分类性能, 但基于样本连线的插入机制受到了很多限制, 而且只从全局不均衡性的角度处理数据, 未能很好地保证不均衡数据采样质量. SML 结合了两种采样的优势, 要普遍好于 SMOTE, 但其中的 NCL 降采样操作在复杂空间下不可避免地删除了多数类的重要信息, 导致保留的样本无法反映原始数据的分布. KOS 算法虽同时考虑了两个不均衡性问题, 但由于采用随机升采样造成数据过

拟合, 而且 K 值的固定设置对于某些数据集并不准确, 影响了采样的效果.

3.3 采样率 $R_{\rm hs}$ 对分类性能的影响及优化

采样算法中的采样率决定着采样的性能, 从而影响着分类准确性, 但是最佳的采样率很难通过经验获得. 本文选取不均衡度为 30% 的 German 数据集进行演示, 通过选取 HPS 算法中不同的采样率比例 $R_{\rm hs}$ 来观察其对分类性能的影响, 如图 2 所示.

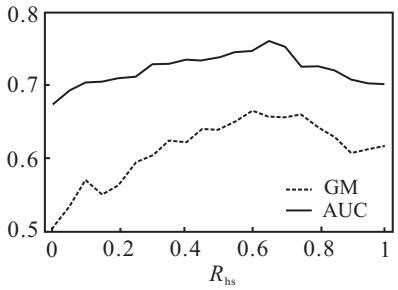

图 2　改变采样率 $R_{\rm hs}$ 对 German 数据集的分类影响

通过图 2 曲线可以看到 $R_{\rm hs}$ 直接影响着分类性能: 当 $R_{\rm hs}=0$ 时, 只有降采样对多数类进行处理以达到类别均衡的目的, 但可能漏掉潜在重要的数据; $R_{\rm hs}=1$ 时只有升采样起作用, 合成大量的新样本, 造成分类模型的过拟合. HPS 算法在 $R_{\rm hs}=0.65$ 和 $R_{\rm hs}=0.6$ 时分别达到了最高的 AUC 和 GM 值, 这也证明了混合采样可以避免单一采样的缺陷.

最佳采样率依赖于具体的数据分布, 为了使 HPS 算法达到最佳性能, 本文对每个数据集在训练数据中进行交叉验证来获取最优采样率 $R_{\rm hs}$. 采样率取值范围设置为 [0.1, 1], 步长为 0.1, 对于每一个 $R_{\rm hs}$ 值, 采用 3.2 节中介绍的方法进行验证, 最终选取交叉验证最好的 $R_{\rm hs}$ 值作为该数据集的最佳采样率. 同样也对 SMOTE 算法进行升采样率的寻优, 采样率选取范围为 [50%, $2\times R_{\max}$], 其中 $R_{\max}$ 为使两类数据相等时的采样率, 步长设置为 50%, 优化过程同上. 对于 HPS 算法, 分别使用 GM 和 AUC 作为评价指标来指导采样率的优化, GM 分类结果和最佳采样率如表 3 所示.

表 3　优化 SM 和 HPS 算法采样率的 GM 性能比较结果

数据集	SM		$\rm HPS_{GM}$		$\rm HPS_{AUC}$	
	GM	$R_{\rm os}$/%	GM	$R_{\rm hs}$	GM	$R_{\rm hs}$
Glass	**0.944**	350	0.889	0.85	0.881	0.8
Letter	0.948	450	**0.966**	0.85	0.963	0.75
Page	0.859	300	**0.867**	0.75	**0.867**	0.75
Segment	0.995	300	**0.998**	0.55	0.997	0.45
Transfusion	0.577	200	**0.598**	0.75	**0.598**	0.75
Vehicle	0.706	100	**0.751**	0.7	**0.751**	0.7
German	**0.672**	300	0.667	0.6	0.661	0.65
Pima	0.659	150	0.668	0.7	**0.673**	0.8
Spambase	0.845	200	0.899	0.65	**0.906**	0.6

由表 3 可以发现: 在多个数据集下, HPS 算法的优化结果好于 SMOTE 算法; 而对于数据集 Pima 和 Spambase, 在AUC指标的指导优化下得到了更好的 GM 结果, 说明 AUC 在某些数据集上, 对不均衡数据具有更好的评估和优化作用, 文献 [4, 16] 也同样支持这一结论.

3.4 HPS 对噪音的鲁棒性测试

数据集中不可避免地具有很多噪音数据, 噪音数据是指样本中含有错误的值, 包括特征属性的错误和类别标签的错误[17]. 由于 C4.5 决策树分类器本身的训练机制 (根据信息增益率进行特征选择和分裂) 具有一定的属性噪音的抑制能力, 类别标签噪音具有更强的分类器阻碍性, 这里重点考虑含有错误类标签的样本噪音. 为了系统地验证 HPS 算法对噪音数据的鲁棒性, 实验中人工加入噪音数据, 并调整噪音的级别程度. 该实验使用其他文献中相同的实验方法来对原始数据集注入不同程度的噪音数据[18]. 在给定的噪音级别 $l\%$下, 每个样本以概率为 $l\%$ 的可能性出现相反的类别, 即噪音级别 $l\%$ 越大, 样本类别被反转的可能性越大, 噪音出现的概率也越大.

随机选取 German 数据集进行测试, 表 4 列出了在具有不同噪音级别的不均衡数据下的实验比较结果. 可以发现 HPS 相对于其他算法具有更强的抗噪性, 特别是在噪音级别较高的情况下, 这完全归功于 HPS 算法在对数据进行分布估计和采样之前进行的过滤操作, 同时采样过程中的样本权重设置操作, 降低了噪音数据对采样和分类学习的影响.

表 4　调整类别标签噪音级别的实验结果

噪音级别/%	C4.5	SM	SML	KOS	CBSO	HPS
0	0.715	0.728	0.739	0.722	0.764	0.757
10	0.709	0.721	0.719	0.722	0.751	0.749
20	0.685	0.722	0.715	0.707	0.743	0.745
30	0.669	0.707	0.701	0.691	0.739	0.739
40	0.655	0.685	0.691	0.684	0.711	0.722
50	0.638	0.681	0.678	0.672	0.685	0.701

4 结　论

为了提升不均衡数据的采样性能, 本文提出了一种基于概率分布的混合采样算法. 该算法根据估计的数据分布规律对每个子类进行采样, 在保证类间均衡化的同时, 也分别对两类数据内部的不均衡性进行改进, 从而更好地改进了不均衡数据的采样效果, 提升了分类性能. 实验结果表明该算法在处理不均衡数据集时具有更高的分类精度. 下一步的工作是: 1) 研究如何结合集成算法来提升不均衡学习的泛化能力; 2) 本实验中对参数 γ 默认为 1, 未来的实验中将对参数 γ 进行调整优化以获得最佳的参数值.

参考文献(References)

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

[2] 陶新民, 张冬雪, 付丹丹, 等. 基于谱聚类欠取样的不均衡数据SVM分类算法[J]. 控制与决策, 2012, 27(12): 1761-1768.
(Tao X M, Zhang D X, Fu D D, et al. The SVM classifier for unbalanced data based on spectrum cluster-based under-sampling approaches[J]. Control and Decision, 2012, 27(12): 1761-1768.)

[3] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. J of Artificial Intelligence Research, 2002, 6(1): 321-357.

[4] Cao P, Zhao D, Zaiane O. An optimized cost-sensitive SVM for imbalanced data learning[C]. Proc of the 17th Pacific-Asia Conf on Knowledge Discovery and Data Mining. Gold Coast, 2013: 280-292.

[5] 陈刚, 冯丹. 一种新的模糊规则权重方法的非平衡数据分类问题的研究[J]. 控制与决策, 2012, 27(1): 104-108.
(Chen G, Feng D. Research on a new method for fuzzy rule weights in imbalanced data classification problem[J]. Control and Decision, 2012, 27(1): 104-108.)

[6] Weiss G. The impact of small disjuncts on classifier learning[J]. Annals of Information Systems, 2010, 8(1): 193-226.

[7] Jo T, Japkowicz N. Class imbalances versus small disjuncts[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 40-49.

[8] Japkowicz N. Concept-learning in the presence of between-class and within-class imbalances[C]. Proc of Advances in Artificial Intelligence. Adelaide, 2001: 67-77.

[9] Titterington D M, Smith A F M, Makov U E. Statistical analysis of finite mixture distributions[M]. New York: John Wiley Sons, 2001.

[10] Laurikkala J. Improving identification of difficult small classes by balancing class distribution[C]. Proc of AI in Medicine in Europe: Artificial Intelligence Medicine. Cascais, 2001: 63-66.

[11] Barua S, Md I, Kazuyuki M. A novel synthetic minority oversampling technique for imbalanced data set learning[C]. Proc of the 18th Int Conf on Neural Information Processing. Shanghai, 2011: 735-744.

[12] Figueiredo M A T, Jain A K. Unsupervised learning of finite mixture models[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(3): 381-396.

[13] Chen S, He H, Garcia E A. RAMO Boost: Ranked minority oversampling in Boosting[J]. IEEE Trans on Neural Networks, 2010, 21(10): 1624-1642.

[14] Cieslak D A, Chawla N V. Start globally, optimize locally, predict globally: Improving performance on imbalanced data[C]. Proc of IEEE Int Conf on Data Mining. Pisa, 2008: 43-152.

[15] Weiss G M. Mining rare cases[C]. Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Germany: Springer-Verlag, 2005: 765-776.

[16] Chawla N V, Cieslak D A, Hall L O, et al. Automatically countering imbalance and its empirical relationship to cost[J]. Data Mining and Knowledge Discovery, 2008, 17(2): 225-252.

[17] Zhu X q, Wu X D. Class noise vs attribute noise: A quantitative study[J]. Artificial Intelligence Review, 2004, 22(3): 177-210.

[18] Anyfantis D, Karagiannopoulos M, Kotsiantis S, et al. Robustness of learning techniques in handling class noise in imbalanced datasets[C]. Proc of the 4th IFIP Int Conf on Artificial Intelligence Applications and Innovations(AIAI'07). Athens, 2007: 21-28.

(责任编辑: 李君玲)