DOI:10.3969/j.issn.1672-5565.20161019001

高斯混合模型的上采样分析

沈乐阳*,孙廷凯
(南京理工大学 计算机科学与工程学院, 南京 210094)

摘 要:在机器学习问题中,类别不平衡问题严重影响一些标准分类器的性能。因此,解决类别不平衡问题尤为重要。上采样是解决类不平衡问题的常用方法,其通过合成新的少数类样本来平衡类的分布。在文中,使用一种基于高斯混合模型的上采样方法来解决不平衡学习问题。通过高斯混合模型来模拟少数类的分布,在此基础上使用高斯模型来生成新的少数类样本。在UCI类别不平衡数据集上的实验结果表明,所提出的方法能够缓解类不平衡所带来的负面影响并帮助提升分类性能。

关键词:不平衡学习;支持向量机;高斯混合模型;上采样
中图分类号:TP181 **文献标志码**:A **文章编号**:1672-5565(2017)02-084-06

A new over-sampling algorithm by gaussian mixture model

SHEN Leyang*, SUN Tingkai
(*School of Computer Science and Engineering*, *Nanjing University of Science and Technology*, *Nanjing* 210094, *China*)

Abstract: It's significant to solve the class-imbalance problems which have a serious impact on the performance of standard classifiers in machine learning problems. Over-sampling is a popular method in dealing with class-imbalance problems, which attempts to balance the sizes of different classes by generating additional samples for minority class. We propose a new over-sampling algorithm that synthesizes new additional samples for minority classes by the Gaussian mixture model. Comparing with several state-of-art related methods on UCI datasets, the experimental results demonstrate that the proposed over-sampling algorithm can reduce the side effect of the class imbalance and help improve the classification performance.

Keywords: Imbalance learning; Support vector machine; Gaussian mixture model; Over-sample

在很多分类问题中,数据集往往是很不平衡的,一些类的样本数目远远高于其他类,这就形成了不平衡学习问题。许多传统的机器学习和数据挖掘算法假设目标类具有相似的先验概率,但是许多现实应用都严重违背了这种假设,如网络入侵检测、欺诈检测、文本分类、风险管理和医学诊断等。从现实应用的角度看,不平衡问题主要体现在两个方面:数据本身的不平衡和外部因素导致的不平衡[1]。

类别不平衡严重影响分类器的预测性能。传统的机器学习方法没有考虑到类别不平衡的问题,易于产生偏向并忽略少数类样本,导致少数类样本被错分[1]。例如,在一个多数类和少数类比例为99的问题中,学习算法为了最小化错误率可能将所有的样本都预测为多数类,这种情况下少数类都被错分。此外,类别不平衡也与代价敏感学习紧密相关。对于一些问题,错分少数类的代价往往高于错分多数类。

目前,不平衡学习的研究已经取得了相当大的进展[2]。目前的解决方法大致可分为3类:基于采样的方法[3]、基于学习的方法(如敏感学习[4])和结合采样和学习的方法[5]。在上述3种方法中,基于采样的方法是最基本的策略,如下采样[2]和上采样[6]。采样方法通过改变样本的数量和分布来平衡不同类的样本,这种方法对于不平衡学习问题往往有较好的效果。上采样对于不平衡学习问题是一种较有效的方法。这种方法利用原始的少数类样本来合成新的少数类样本,从而增加少数类样本数目,

收稿日期:2016-10-19;**修回日期**:2016-12-00.
基金项目:国家自然科学基金(61373062,61371040).
*通信作者:沈乐阳,男,硕士研究生,研究方向:模式识别与生物信息学;E-mail: shenleyang@gmail.com.

平衡样本分布。但是上采样方法的主要问题有两个:一方面,上采样扩大了训练集,导致训练和预测的耗时增加;另一方面,上采样仅仅是复制原始少数类样本,这导致了某些样本的重复,可能会出现过拟合的问题。

本文使用上采样方法解决类别不平衡的二分类问题。目前,很多上采样方法已经被提出,如随机上采样(random over-sample,ROS)、合成少数类上采样技术(synthetic minority over-sampling,SMOTE)[6]、自适应合成采样(adaptive synthetic sampling,ADASYN)[7]和严格合成少数类上采样技术(critical SMOTE,CSMOTE)[5]。本文提出一种基于高斯混合模型的上采样方法来生成新的少数类样本。高斯模型被广泛用于分类或表示数据。因此,本文使用高斯模型来模拟少数类样本的分布,在此基础上合成新的少数类样本。

1　方法

1.1　高斯混合模型

高斯混合模型(gaussian mixture model, GMM)是单高斯模型的延伸,能够较好地描述数据的密度分布。GMM 假设所有的数据点都是由有限个高斯分布生成的,通常作为概率分布的参数模型使用。对于基于 GMM 的分类系统,模型训练的主要目的就是估计参数使得高斯混合分布能够较好地匹配训练集中特征向量的分布。GMM 的参数主要是在先验模型的基础上使用最大期望算法(expectation maximization,EM)进行估计[8]。GMM 可以认为是 M 个单一高斯概率密度函数的加权平均,其概率密度分布函数为

$$p(x)=\sum_{i=1}^{M}\pi_i N(x\mid\mu_i,\Sigma_i).$$

式中:x 为服从高斯混合分布的随机变量;M 为高斯分量的数目;π_i 为第 i 个高斯分量的权重;$N(x\mid\mu_i,\Sigma_i)$ 为第 i 个高斯分量的概率密度函数;μ_i、Σ_i 分别为第 i 个高斯分量的均值和方差。

1.2　主成分分析和核主成分分析

主成分分析(principal component analysis,PCA)是常用的线性降维方法,能够有效地从高维数据中提取重要信息[9]。PCA 通过线性投影将高维数据映射到低维空间中表示,并期望在所投影的维度上的数据方差最大。PCA 追求在降维后最大化保持数据的内在信息,通过在投影方向上的数据方差衡量该方向的重要性。PCA 最初被用来分析多元数据,但现在已经被广泛应用到其他方面,如去噪信号、盲源分离和数据压缩等。

核主成分分析(kernel principal component analysis,KPCA)[10]是主成分分析的非线性扩展。PCA 从高维空间到低维空间的映射是线性的,对于非线性映射往往无能为力。KPCA 通过使用核技巧来实现非线性的降维,被广泛使用于多种领域,如去噪、压缩和结构预测等。

1.3　支持向量机

支持向量机(support vector machine, SVM)已被广泛用于多种领域[11]。本文选择使用支持向量机作为基本的学习模型来评估所提出方法的有效性。下面简单介绍支持向量机的基本思想。

给定样本集 $\{(x_i,y_i)\}_{i=1}^{N}$,其中 $x_i\in R^d$,$y\in\{+1,-1\}$,它们分别是第 i 个样本的特征向量和相对应的标签,而+1 和−1 分别代表正类和负类的标签。

SVM 寻找满足分类要求并拥有最大间隔的划分超平面,即寻找最小化 $\frac{1}{2}\|w\|^2$ 并满足如下约束的参数 w 和 b:

$$y_i\cdot(\boldsymbol{w}^{\mathrm{T}}\cdot x_i+b)\geqslant 1,i=1,2,\ldots,N.$$

式中:$\boldsymbol{w}$ 为超平面的法向量,$\|w\|^2$ 是 $\boldsymbol{w}$ 的欧几里得范数。

本文使用 LIBSVM[12]工具构建模型,并选择被广泛使用的高斯核函数 $K(x_i,y_i)=\mathrm{e}^{-\|x_j-x_i\|^2/2\sigma^2}$ 作为核函数。本文采用 LIBSVM 软件中基于交叉验证的网格搜索策略优化正则参数和核参数。LIBSVM 的最新下载地址为 https://www.csie.ntu.edu.tw/~cjlin/libsvm/。

1.4　Tomek links

如果两个属于不同类别的样本点都是彼此的最近邻,那么他们就构成了一个 Tomek link[13]。记点对 (x_i,x_j) 的距离为 $d(x_i,x_j)$,其中 $x_i\in S_{\min}$,$x_j\in S_{\max}$。如果不存在点 x_k 使得 $d(x_i,x_k)<d(x_i,x_j)$ 或 $d(x_j,x_k)<d(x_i,x_j)$,那么点对 (x_i,x_j) 就是一个 Tomek link。如果点对属于 Tomek link,则该点对属于噪声或边界点。因此,本文利用 Tomek links 方法来清洗采样后类间的数据交叠,建立类簇分布良好的训练集,以此提高分类性能。

1.5　提出的方法

上采样方法的关键问题是如何生成有效的新的样本。高斯混合模型基于多变量正态分布,假设数据集是多个混合在一起的多元高斯分布,从而用极大似然估计的思想来将数据聚类,达到描述数据分布的目的。尽管高斯混合模型常用于聚类,但其有效地描述了数据的密度分布。此外,在实际应用中高斯混合模型的使用十分广泛,高斯分布很常见,

很少出现不符合其假设的应用场景。因此,本文利用该思想,单独对少数类的样本进行聚类,从而得到少数类样本的数据分布。在此基础上,利用得到的模型随机生成新的少数类样本,达到采样的目的。

在整个采样算法中有两个需要注意的问题。首先,在利用高斯混合模型对少数类样本的分布进行模拟之前,需要对样本数据有一定的了解。虽然高斯混合模型并不假设数据集到底是由多少个多元高斯分布叠加而成的,但是如果能够知道这个信息,算法能够更快速准确地学习到数据的结构。总之,能够利用的信息越多,算法的效果就会越好。因此,在使用高斯混合模型对少数类样本进行模拟前应充分了解数据的特性。其次是采样比例的问题。本文采样并不需要达到绝对的平衡。对于不平衡比例较大的数据集,如果采样比例过大,同样会造成分类器性能的降低。因此,对于不同类别比例的数据集,需要采样的数目也不一样。在本文中,使用增长比例来衡量采样的数目。记原始数据集多数类样本与少数类样本的比例为 u,上采样后多数类样本与少数类样本的比例为 v,则增长比例为 $\frac{u-v}{v}$。此外,在最后的筛选环节仍然会剔除一些样本,因此需要稍微增加增长比例。在本文的实验中,一般将增长比例设为 2。

记训练集为 $S = S_{\min} \cup S_{\max}$,其中,$S_{\min}$、$S_{\max}$ 分别为少数类的样本集和多数类的样本集。本文提出的方法主要是通过上采样合成新的少数类样本来获得一个相对平衡的训练集,记为 S_{new}。记 α 为采样系数,控制生成样本的数目。β 为置信度,决定生成样本是否可信。

本文提出方法的主要流程如下。

步骤 1 利用高斯混合模型对少数类样本 $S_{\min}$ 进行建模,得到模型 G_{model}。生成的模型用于接下来合成样本。

$$G_{\text{model}} \leftarrow \text{Model}(S_{\min}).$$

步骤 2 记少数类样本的数目为 $N_{\min}$,利用高斯混合模型随机生成新的少数类样本,记采样得到的样本集为

$$S_{\text{sample}} \leftarrow \text{Sample}(G_{\text{model}}, \alpha \cdot N_{\min}).$$

步骤 3 经过上述的采样本文得到了新的少数类样本,但由于上采样经常会引入数据交叠的问题,因此必须对训练集进行清洗。本文选择 Tomek links 技术进行清洗,去除由采样引入的数据交叠为

$$S_{\text{new}} \leftarrow \text{Tomeklinks}(S, S_{\text{sample}}).$$

1.6 评价指标

在处理类别不平衡问题时,衡量性能的指标也有所不同。本文中,使用查准率(precision)、查全率(recall)、*F*-Measure 和 *G*-mean 来综合衡量分类器的性能,定义如下:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$F - \text{Measure} = \frac{(1+\beta^2) \cdot \text{Recall} \cdot \text{Precision}}{\beta^2 \cdot \text{Recall} + \text{Precision}},$$

$$G - \text{mean} = \sqrt{\frac{\text{TP}}{\text{TP} + \text{FN}} \times \frac{\text{TN}}{\text{TN} + \text{FP}}}.$$

式中:β 为衡量查准率对查全率相对重要性的系数,通常为 1;TP(true positive)、FP(false positive)、TN(true negative)和 FN(false negative)分别为正确预测为正类的数目、错误预测为正类的数目、正确预测为负类的数目和错误预测为负类的个数,其性能可用混淆矩阵(confusion matrix)来表示,如图 1 所示。

		预测结果:正类	预测结果:负类
真实情况	正类	TP (true positive)	FN (false negative)
	负类	FP (false positive)	TN (true negative)

图 **1** 混淆矩阵

Fig.1 Confusiton matrix

但是上述指标都是基于阈值的,选取不同的阈值可以得到不同的结果。因此本文还使用了另外一种评价指标 AUC(area under roc curve),即 ROC(receiver operating characteristic)曲线下的面积。与上述 4 种指标不同,AUC 是与阈值无关的且与分类器的性能成正比,因此本文选择 AUC 来衡量分类器的总体预测性能。

2 结果与分析

2.1 采样前后特征值比较

为了进一步了解本文使用的采样方法,采用 KPCA 方法对采样前后的数据进行主成分分析。为了更好地展现实验结果,本文使用二维数据进行实验。在该实验中,使用了 banana 数据集、同心圆数据集和 3 个高斯分布构成的数据集进行实验。本文分别对采样前后的数据进行 KPCA 主成分分解,并

根据不同的主成分画出等高线图。

在 banana 数据集上的结果如图 2 所示,可以发现在图 1 中采样前第 1 个主成分很好地将数据分成了两部分,第 2 个主成分将数据分成 3 部分,而第 3 个和第 4 个主成分分别将数据分成:4 部分和 5 部分,与采样后的情况基本一致。对于特征值而言,采样前后基本不变,尤其是第 1 个主成分,仅仅相差 0.02%。

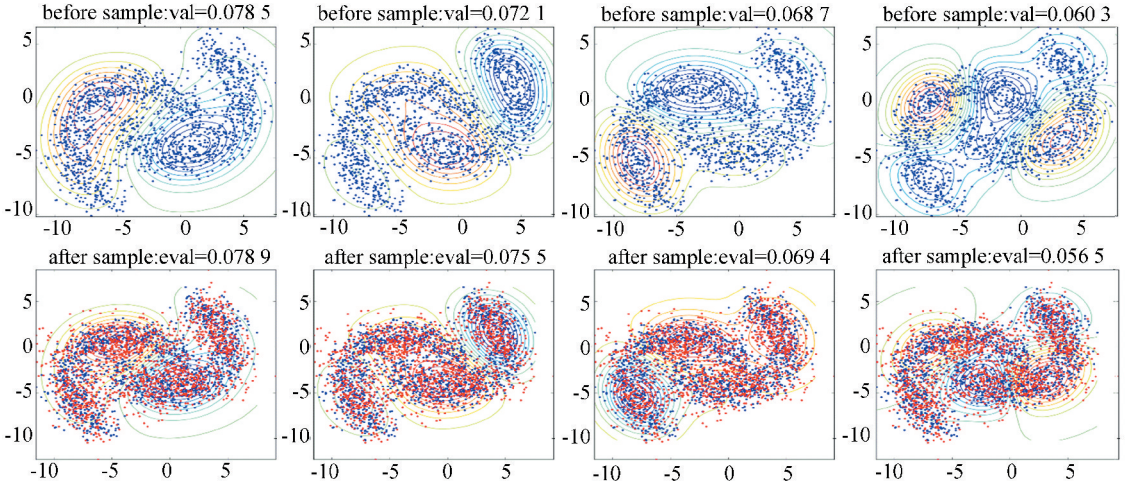

图 **2**　**banana** 数据集采样前后主成分对比

Fig.2　Comparisons of principal components on before-sampling and after-samping banana dataset

在同心圆数据集上的实验结果如图 3 所示。本文发现在同心圆上的结果不同于 banana 数据集,在采样前后主成分对数据的划分不是完全相同。第 1 个主成分在采样前后都将数据划分成了两部分,并且特征值也比较接近,两部分的中心分布也比较相似。而在第 2 个和第 3 个在采样前将数据分别划分为 2 部分和 3 部分,但是在采样后分别将数据划分为 3 部分和 2 部分。由此可以发现采样改变了数据的分布,但是特征值还是比较接近的。第 4 个主成分在采样前后都将数据划分成了 4 部分,各部分中心的分布也比较相似,特征值变化也不是很大。总体而言,采样前、后有的分布还是比较类似的,没有发生比较大的变化。

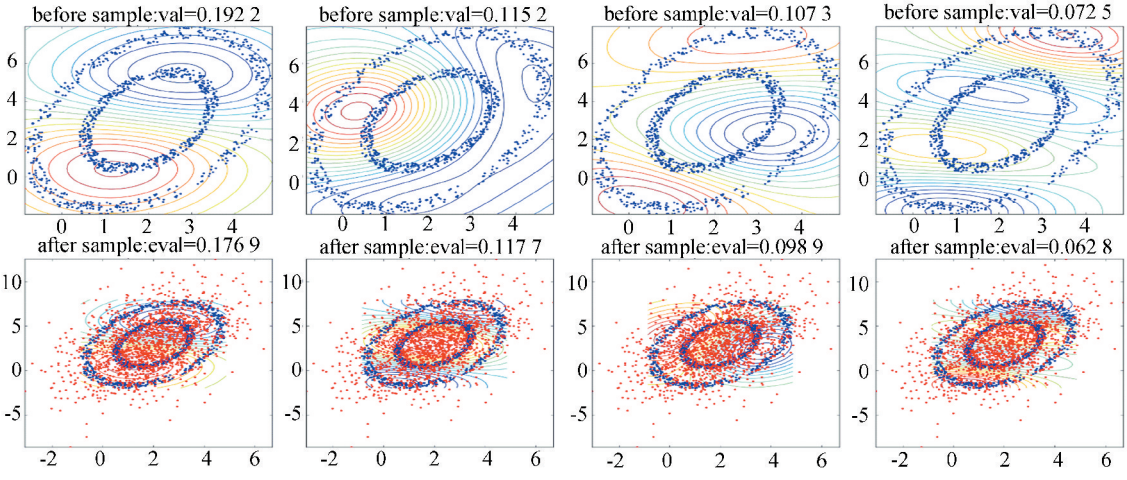

图 **3**　同心圆数据集采样前后主成分对比

Fig.3　Comparisons of principal components on before-sampling and after-samping circle datasets

在 3 个高斯分布合成的数据集上的实验结果如图 4 所示。本文可以很明显地发现在 4 个主成分上采样前后的分布基本相同,第 1 个和第 2 个主成分都将数据很好地划分成了 3 部分,并且特征值也很相近。而第 3 个和第 4 个主成分都将数据划分成了 4 部分,但是中心位置的分布略有不同。但是就采样前后的比较而言,两个主成分的结果基本相同。

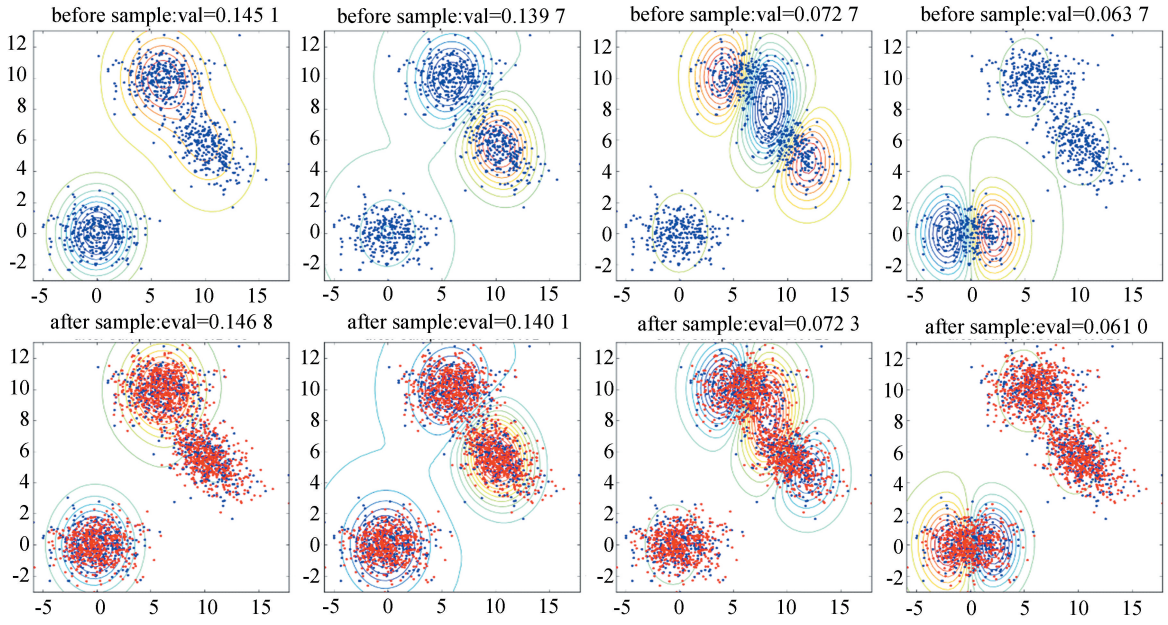

图 4 3 个高斯分布合成的数据集采样前后主成分对比

Fig.4 Comparisons of principal components on before-sampling and after-samping gaussian datasets

2.2 UCI 数据集上的实验结果

为了验证本文所提出的方法,使用 UCI(university of california, irvine)机器学习库中的一些不平衡数据集来进行实验[5],数据集及其特性见表 1。

表 1 数据集描述

Table 1 Summary description of datasets

数据集	样本数	属性数	少类数目	多类数目	多类与少类比例
Ecoli1	336	7	77	259	3.36
Ecoli3	336	7	35	301	8.60
Iris0	150	4	50	100	2.00
Yeast1vs7	459	8	30	429	14.30
Yeast1458vs7	693	8	30	663	22.10

本文将比较几种针对不平衡数据集的方法在上述数据集上的性能,包括 SVM 方法、CSMOTE 方法、随机下采样集成方法(RUSBoost)[14]、HardEnsemble(HE_S)[5]方法及所提出的 GOS 方法。RUSBoost 结合采样和 boosting 集成方法,是 SMOTEBoost[15] 的变种方法。SMOTEBoost 利用 SMOTE 合成少数类样本,而 RUSBoost 则使用了随机下采样来实现样本的平衡。HardEnsemble 方法结合了上采样和下采样的方法来减少不同采样带来的负面影响,从而提高分类的性能。

首先,本文考察 AUC 指标,其结果见表 2。在 Yeast1458vs7 数据集上,GOS 的 AUC 值高于其他 4 种方法,但在 Yeast1vs7 数据集上,GOS 的 AUC 值远远低于其他方法。在其他 3 种数据集中,GOS 的 AUC 值略优由于 CSMOTE,但稍稍低于 SVM 和 HE_S 方法。而从多数类和少数类的比例角度来看,GOS 方法在比例较低和较高的情况下的 AUC 值往往较高,而在比例值为 14.30 的 Yeast1vs7 数据集上的表现较差。

表 2 SVM、CSMOTE、RUSBoost 和 HE_S 方法在 AUC 值上的比较

Table 2 AUC comparisons between SVM, CSMOTE, RUSBoost and HE_S for serveral datasets

AUC	SVM *	CSMOTE *	RUSBoost *	HE_S *	GOS
Ecoli1	0.959 7	0.958 7	0.959 4	0.956 7	0.959 0
Ecoli3	0.951 3	0.942 5	0.946 8	0.949 9	0.949 2
Iris0	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0
Yeast1vs7	0.795 7	0.854 2	0.815 5	0.852 0	0.727 5
Yeast1458vs7	0.681 6	0.697 4	0.709 5	0.701 4	0.717 5

注:* 数据来自文献[5].

而不同方法在几种数据集上的 *F*-Measure 值和 *G*-mean 值的比较结果见表 3、4。本文发现 GOS 方法的 *F*-Measure 值优于其他 4 种方法,尤其在 Yeast1vs7 数据集上,其结果远远高于其他 4 种方法。在几种数据集上,GOS 方法的 *F*-Measure 值最少比其他方法高 1%。而对于 *G*-mean 值而言,GOS 方法在 Ecoli3 和 Yeast1458vs7 数据集上高于其他方法,在 Yeast1vs7 数据集上仅仅低于 CSMOTE 方法。而在 Ecoli 数据集中,GOS 方法的 *G*-mean 值低于 RUSBoost 和 HE_S 方法。总体而言,GOS 在不同平衡比例的数据集上的 *F*-Measure 值有很好的表现,高于其他 4 种方法,并且在 AUC 值和 *G*-mean 值上也有不俗的表现。而从多数类和少数类的比例角度来看,GOS 方法在不同比例数据上都有较好的表

现,尤其在比例较高的情况下,该方法的 *F*-Measure 值和 *G*-mean 值远远高于其他方法。

表 3　SVM、CSMOTE、RUSBoost 和 HE_S 方法在 *F*-Measure值上的比较

Table 3　*F*-Measure comparisons between SVM, CSMOTE, RUSBoost and HE_S for serveral datasets

F-Measure	SVM *	CSMOTE *	RUSBoost *	HE_S *	GOS
Ecoli1	0.746 0	0.738 8	0.745 8	0.726 6	0.795 0
Ecoli3	0.540 9	0.497 6	0.523 9	0.535 8	0.552 2
Iris0	1.000 0	1.000 0	0.995 8	1.000 0	1.000 0
Yeast1vs7	0.149 3	0.378 5	0.224 8	0.367 2	0.526 3
Yeast1458vs7	0.146 9	0.178 6	0.199 7	0.186 8	0.213 7

注:*数据来自文献[5].

表 4　SVM、CSMOTE、RUSBoost 和 HE_S 方法在 *G*-mean 值上的比较

Table 4　*G*-mean comparisons between SVM, CSMOTE, RUSBoost and HE_S for serveral datasets

G-mean	SVM *	CSMOTE *	RUSBoost *	HE_S *	GOS
Ecoli1	0.810 2	0.851 9	0.920 3	0.929 4	0.906 9
Ecoli3	0.825 9	0.783 5	0.830 0	0.808 2	0.912 9
Iris0	1.000 0	1.000 0	0.925 6	1.000 0	1.000 0
Yeast1vs7	0.297 3	0.761 0	0.398 8	0.728 2	0.746 1
Yeast1458vs7	0.411 3	0.587 5	0.604 6	0.588 7	0.693 5

注:*数据来自文献[5].

3　结　语

本文中,对一个新的上采样方法进行了研究,该方法基于高斯混合模型合成新的少数类样本,在此基础上使用 Tomek links 技术对新生成的样本进行筛选,最终得到相对平衡的训练集样本。在 UCI 不平衡数据集上对 GOS 方法进行实验,并和其他预测方法相比较。实验结果表明,该方法有助于缓解类不平衡,并提升分类的准确性。

参考文献 (References)

[1] CHAWLA N V, JAPKOWICZ N, KOTCZ A. Editorial: special issue on learning from imbalanced data sets[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1): 1-6. DOI:10.1145/1007730.1007733.

[2] HE Haibo, GARCIA E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284. DOI: 10.1109/TKDE.2008.239.

[3] ESTABROOKS A, JO T, JAPKOWICZ N. A multiple resampling method for learning from Imbalanced data sets[J]. Computational Intelligence, 2004, 20(1): 18-36. DOI: 10.1111/j.0824-7935.2004.t01-1-00228.x.

[4] ZHOU Zhihua, LIU Xuying. On multi-class cost-sensitive learning[J]. Computational Intelligence, 2010, 26(3): 232-257. DOI: 10.1111/j.1467-8640.2010.00358.x.

[5] NANNI L, FANTOZZI C, LAZZARINI N. Coupling different methods for overcoming the class imbalance problem [J]. Neurocomputing, 2015, 158: 48-61. DOI: 10.1016/j.neucom.2015.01.068.

[6] 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. DOI:10.1613/jair.953.

[7] HE Haibo, BAI Yang, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]//Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). HongKang: IEEE, 2008: 1322-1328. DOI: 10.1109/IJCNN.2008.4633969.

[8] ULUKAYA S, ERDEM C E. Gaussian mixture model based estimation of the neutral face shape for emotion recognition [J]. Digital Signal Processing, 2014, 32: 11-23. DOI: 10.1016/j.dsp.2014.05.013.

[9] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016. 229-232.
ZHOU Zhihua. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 229-232.

[10] LI Junbao, GAO Huijun. Sparse data-dependent kernel principal component analysis based on least squares support vector machine for feature extraction and recognition [J]. Neural Computing and Applications, 2012, 21(8): 1971-1980. DOI: 10.1007/s00521-011-0600-z.

[11] DIOSAN L, ROGOZAN A, PECUCHET J P. Improving classification performance of support vector machine by genetically optimising kernel shape and hyper-parameters [J]. Applied Intelligence, 2012, 36(2): 280-294. DOI: 10.1007/s10489-010-0260-1.

[12] CHANG C C, LIN C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27. DOI: 10.1145/1961189.1961199.

[13] ZHANG Jing, CAO Peng, GROSS D P, et al. On the application of multi-class classification in physical therapy recommendation [J]. Health Information Science and Systems, 2013, 1(1): 15. DOI: 10.1186/2047-2501-1-15.

[14] SEIFFERT C, KHOSHGOFTAAR T M, HULSE J V, et al. RUSBoost: A hybrid approach to alleviating class imbalance[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2010, 40(1): 185-197. DOI: 10.1109/TSMCA.2009.2029559.

[15] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting[C]//European Conference on Principles of Data Mining and Knowledge Discovery. Berlin: Springer, 2003: 107-119. DOI: 10.1007/978-3-540-39804-2_12.