基于谱聚类和成对数据表示的多层感知机分类算法

刘树栋 魏嘉敏
(中南财经政法大学信息与安全工程学院 武汉 430073)

摘 要 面向类别不均衡数据集的分类学习一直是数据挖掘和机器学习领域的研究热点。数据级、算法级和集成方法是目前解决类别不均衡学习的 3 种主流方法,其中欠抽样是类别不均衡学习一种常用的数据级解决方法,其缺点在于容易丢失多数类中部分有用信息。文中将谱聚类引入到成对数据表示的多数类欠抽样过程中,首先利用谱聚类方法,对多数类样本进行聚类,根据聚类簇大小和簇内样本点与少数类样本点的平均距离,在每个聚类簇内抽取不同个数有代表性的样本,并将簇内样本点之间及所有少数类样本点两两成对表示,从而有效降低了所有样本成对数据表示中两两组合而导致的数据暴涨问题,同时避免了随机抽样而可能导致的有效信息丢失问题。最后在 9 组 UCI 数据集上验证了所提算法的有效性。

关键词 多层感知机,分类,欠抽样,谱聚类,不均衡学习
中图法分类号 TP311 **文献标识码** A

Multilayer Perceptron Classification Algorithm Based on Spectral Clustering and Simultaneous Two Sample Representation

LIU Shu-dong WEI Jia-min
(School of Information and Security Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China)

Abstract Classification learning from imbalanced datasets is always one of hot topics in data mining and machine learning domains. Data-level, algorithm-level and ensemble solutions are three main methods so far for addressing imbalanced learning. Undersmapling, which is one of data-level solutions, is widely utilized in many imbalanced learning scenarios. However, its drawback is discarding potentially useful majority data instances. In this paper, spectral clustering was introduced to take sample of the majority class instances so as to build simultaneous two sample representation. Firstly, all majority class instances are divided into many different clusters by spectral clustering analysis, different numbers of representative samples are extracted from different clusters according to the size of each cluster and the average distance between the minority class instances are generated simultaneous and each cluster, then two sample representation with the extracted instances are generated simultaneous from clusters and the minority class instances. The proposed method not only alleviates the issue of data explosion in simultaneous two sample representation, but also avoids the loss of useful information in random sampling. Finally, several experiments certificate its validity on nine groups of datasets from UCI.

Keywords Multilayer perceptron, Classification, Under-sampling, Spectral clustering, Imbalanced learning

1 引言

分类算法是数据挖掘和机器学习技术的重要研究内容之一,其主要任务是在训练样本上学习分类模型,并对未知样本的类标签进行预测,目前已有大量分类模型成功应用于不同领域的分类任务中。例如朴素贝叶斯(Naive Bayes, NB)在垃圾邮件识别上成功应用,逻辑斯蒂回归(Logistic Regression, LR)在银行贷款信用评分中的应用,支持向量机(Support Vector Machine, SVM)在网络谣言识别中的应用,等。

传统的分类模型是基于如下假设:数据集中各类的数目是均衡的[1]。然而在不同应用领域的具体分类问题中,这种假设是不成立,即数据集中存在类别不平衡问题(class imbalance)。类别不均衡是指[2]:在一个数据集中一类样本数目特别多,称为多数类(也称为负类, negative class),另一类样本数目特别少,称为少数类(也称为正类, positive class),且二者不平衡比率(Imbalanced Ratio, IR)较大。例如网络入侵检测数据中入侵访问记录要比正常访问少得多。肿瘤检测数据集中恶性肿瘤只占总体的小部分,搜索引擎点击预测中点击的网页往往只占很小的比例。传统的分类模型在处理这些分类任务时,往往对多数类样本的分类效果较好,对少数类样本的分类效果却很差。解决上述问题的算法一般统称为类别不均衡学习算法[3-6]。

类别不均衡学习一直是数据挖掘和机器学习领域研究的热点和难点之一,近年来备受关注。数据挖掘和机器学习领

本文受国家自然科学基金项目(61602518, 71872180),中南财经政法大学中央高校基本科研业务费专项资金(2722019JCG074, 2722019JCT035)资助。

刘树栋(1984—),男,博士,讲师,CCF 会员,主要研究方向为机器学习与推荐系统,E-mail:bupt.mymeng@gmail.com(通信作者)。

域的主流会议和期刊都曾以此题目举办过研讨会和专刊，并在 ICDM'05 会议上把类别不均衡问题列为了数据挖掘领域亟待解决的十大难题之一[7]。时至今日，学术界和产业界对类别不均衡问题的研究热情仍然没有消退，随着大数据和深度学习技术的发展而呈现逐渐升温的趋势[6,8]。面向大数据处理平台的不平衡算法、虚拟样本生成的新算法、类间不平衡比率的不同权衡策略等问题成为目前大数据不均衡学习研究的难点[8]。目前解决类别不均衡学习的已有方法主要可以分为如下 3 种[9-10]：

(1)数据级解决方法(data-level solution)。此方法采用不同形式的采样技术平衡数据集中的类别分布，使其适合标准分类学习算法的基本要求，本质上属于数据预处理范畴，涉及的采样方法主要包括随机抽样[11]和选择性抽样，其中选择性抽样可以分为对少数类样本人工合成的过采样(over-sampling)[12-13]、对多数类样本选择性抽取的欠采样(under-sampling)[14]和过抽样与欠抽样相结合的混合抽样方法(hybrid sampling)[8]。

(2)算法级解决方法(algorithm-level solution)。此方法不会对训练数据集进行过多的预处理，直接针对数据集内类别不均衡问题设计新的分类器，代表性方法是代价敏感学习算法[10,15]，此算法不再以样本选择的整体误差最小为训练目标，而是以整体样本分类误差最小为优化目标，实现对分类准则和分类算法的修正。此外还有一类学习(one-class learning)[16-19]、主动学习(active learning)[20-23]和基于极端学习机的学习算法[24-26]。

(3)集成解决方法(ensemble solution)。此方法主要通过多基分类器和标准集成策略(例如 bagging，boosting，stacking 等)的融合，提高少数类样本的分类准确度。集成学习作为解决类别不均衡问题的重要技术手段之一，一直以来都是类别不均衡分类学习的研究热点，近年来人们提出了多种面向类别不均衡数据集的集成学习方法。文献[3]将其简单地划分为基于装袋集成学习(bagging-based ensemble)、基于提升的集成学习(boosting-based ensemble)和基于混合集成学习(hybrid-based ensemble)三大类。由于具有融合多种分类器的灵活性等优点，基于堆叠的集成学习(stacking-based ensemble)在近年来得到了快速发展，取得了一些研究成果[27-28]。此外，从集成策略在类别不均衡学习过程的切入点上讲，集成学习方法还可以分为数据级集成、特征级集成和算法级集成三大类。

绝大多数分类算法是以所有样本 x 的特征空间 $X\in R^n$ 为输入，其中 $x_i=(x^{i\,(1)},x_i^{(2)},\cdots,x_i^{(n)})\in X$ 是第 i 个样本的特征表示，所有样本 x 的类别标签 Y 为输出，训练一个从特征空间 $X\in R^n$ 到输出空间 Y 的模型函数，满足 $f(X)=Y$。Dumpala 等[27]提出一种适用于不均衡学习的成对数据表示方法，分类模型的输入不再是单一样本 x_i，而是一对样本$\{x_i,x_j\}$，输出这对样本的类标记$\{y_i,y_j\}$。这种成对数据表示方式将分类模型原来的 n 维输入向量扩展到了 $2n$ 维，扩大了训练集的样本数量，为分类模型学习类内和类间变量特征提供了一种更好的途径。然而，在成对数据的组合过程中，随机性的样本选择会可能会导致训练集过于庞大，特别是在大数据环境下，此问题会更加凸显，从而增加分类模型的训练时长，降低模型分类效率。针对上述问题，本文从样本选择的角度出发，将训练数据集进行谱聚类，从聚类簇中选择有代表性的样本进行成对组合数据。

2　基于谱聚类和成对数据表示的多层感知机分类算法

2.1　成对数据表示的多层感知机

成对数据表示的多层感知机[27]由如下 3 步组成：成对数据表示、分类器的训练、分类器测试。

(1)类别不均衡中的成对数据表示

设不均衡学习二分类问题中多数类样本集合和少数类的样本集合分别为 Ma 和 Mi，其样本个数分别为 M 和 N，多层感知机的输入是任意两个样本的成对组合$\{x_i,x_j\}\in\{X,X\}$，$X=X_M\cup X_N$，其中 X_M 为多数类样本集，X_N 为少数类样本集。输出是两个样本的类标签$\{y_i,y_j\}\in\{Y,Y\}$，$Y=\{y_1,y_2\}$。由于存在类别不平衡问题，原始训练集中多数类样本数目远远大于少数类数目，即 $M\gg N$。成对数据表示使得组合数据集有如下 4 种类型：

$$\{x_i,x_j\,|\,x_i,x_j\in X_M\}$$
$$\{x_i,x_j\,|\,x_i,x_j\in X_N\}$$
$$\{x_i,x_j\,|\,x_i\in X_M,x_j\in X_N\}$$
$$\{x_i,x_j\,|\,x_i\in X_N,x_j\in X_M\}$$

且将分类算法的训练数据规模从 $M+N$ 扩展到了$(M+N)^2$。从包含少数类信息的训练样本上讲，把原始数据集中只有 N 个训练样本扩展到了组合数据集的 N^2+2NM 个组合样本，有效地扩充了包含少数类样本信息的组合样本，但是只包含多数类样本信息的组合样本也从原始数据集中的 M 个扩展到了 M^2 个，因此类别不均衡问题仍然存在。为了降低这种组合数量，可以从多数类样本集中随机选择出 $M_N=N$ 个样本与其他多数类样本进行组合，从而使这种组合样本数目从 M^2 个降低到 $M_N\cdot N$。

(2)成对数据表示的多层感知机模型学习

多层感知机模型以成对组合$\{x_i,x_j\}$为输入，以类标签组合$\{y_i,y_j\}$为输出，用 $sigmod(\cdot)$函数作为输出层的激活函数，以交叉熵作为损失函数($\{y_i',y_j'\}$为$\{x_i,x_j\}$的真实类标签)。

$$loss=\frac{1}{2}\sum_{k=i,j}(y_k\ln y_k'+(1-y_k)ln(1-y_k'))$$

选择合适的学习率(实验过程中学习率设置为 0.001)，采用梯度下降法对多层感知机进行参数学习。

(3)成对数据表示的多层感知机预测

成对数据表示的多层感知机预测过程借鉴集成学习的思想，利用投票机制进行优化选择。首先预选择一组类标签已知的参考样本$\{x_{r1},x_{r2},\cdots,x_{rR}\}$，将待预测样本 x_p 与这组参考样本$\{x_{r1},x_{r2},\cdots,x_{rR}\}$进行成对组合，依次将这些成对组合输入到感知机模型中，得到一组关于 x_p 的预测结果$\{y_{p1},y_{p2},\cdots,y_{pR}\}$，最终通过投票的方式把这组结果中的大多数预测类标签作为对样本 x_p 的标签预测。

2.2　基于谱聚类和成对数据表示的多层感知机分类算法

(1)谱聚类

谱聚类(Spectral clustering)是一种基于图论的聚类方法，主要用于将带权无向图划分为多个最优子图，使得子图内部尽量相似而子图之间距离尽量远。本文将所有聚类样本点

作为构造加权无向图 $G=\{V,E\}$ 的顶点集 V，任意两个顶点 $v_i,v_j\in V$ 之间的连接权重 $\omega_{ij}=\omega_{ji}$，采用它们之间的相似性进行计算，若 $\omega_{ij}=0$，则表示顶点 v_i,v_j 之间不存在连边。图 G 的加权连接矩阵定义为 $W=(\omega_{ij})$，$i,j=1,2,\cdots,n$，图 G 中任意一个顶点 $v_i\in V$ 的度及图的度矩阵分别为：

$$d_i=\sum_{j=1}^{n}\omega_{ij}$$

$$D=diag(d_1,d_2,\cdots,d_n)$$

图 G 对应的拉普拉斯矩阵 L 定义为：

$$L=I-D^{-1/2}WD^{-1/2}$$

设聚类样本集簇划分个数为 $m\leqslant n$，求拉普拉斯矩阵 L 前 m 个特征值对应的特征向量 $e_1,e_2,\cdots,e_m$，以这些特征向量作为列生成矩阵 $U\in R^{n\times m}$，并对矩阵 U 中的每一行进行标准化，得到 U'，即 $u'_{ij}=u_{ij}/(\sum_j u_{ij}^2)^{\frac{1}{2}}$，然后对矩阵 U' 中每一行向量进行 K-means 聚类，最终得到聚类簇 $C_1,C_2,\cdots,C_m$。

(2)基于谱聚类和成对数据表示的多层感知机分类算法

利用上述谱聚类方法对多数类样本进行聚类分析，选取有代表性的样本与全部少数类样本进行成对组合，形成新的成对训练数据集，输入到多层感知机模型中进行训练学习，具体步骤如下：

1)在原始多数类样本集上建立相似矩阵 $S\in R^{M\cdot M}$，其中 M 是原始数据集中多数类样本的个数。

2)预设聚类划分的簇数。例如，如果原始数据集中多数类样本的数量远远大于少数类样本的数量，可将聚类簇数设置为原始数据集中少数类样本的个数，即令 $m=n$。在实验过程中采用 R 中的 speccalt 包对每个数据集进行自动聚类分析，并确定聚类簇个数。

3)采用谱聚类对原始数据集中的多数类样本进行聚类簇划分，形成聚类簇 $C_1,C_2,\cdots,C_N$。

4)选择每一个聚类簇中有代表性的样本点。根据每个聚类簇的大小和簇中所有样本点与原始数据集中少数类样本点的平均距离选取样本。设每个聚类簇大小为 $|C_1|,|C_2|,\cdots,|C_N|$，聚类簇中所有样本点与原始数据集中少数类样本点的平均距离为 $Dis_{C_1,M_i},Dis_{C_2,M_i},\cdots,Dis_{N,M_i}$，定义每个聚类簇选取样本的比重为 $Ratio_{C_j}$：

$$Ratio_{C_j}=\frac{|C_j|}{M}\cdot\frac{Dis_{C_j,M_i}}{\sum_j Dis_{C_j,M_i}}$$

在每个聚类簇抽取样本的个数为：

$$SSize_{C_j}=M\cdot\frac{Ratio_{C_j}}{\sum_j Ratio_{C_j}}$$

5)在每一个聚类簇 $C_j(j=1,2,\cdots,N)$ 中，选择前 $SSize_{C_j}$ 个离聚类中心距离最小的样本点形成该簇的抽样样本集 $Sample_{C_j}$。

6)把每个聚类簇的抽样集中样本点两两成对组合，并将其与所有少数类样本点进行成对组合，形成成对数据表示的多层感知机训练样本集。

(3)算法分析

与文献[27]中成对数据表示对多数类样本进行随机选择的方法相比，本文利用谱聚类方法抽取每个聚类簇中有代表性的样本点，有效避免了随机欠抽样可能存在的有效信息丢失问题，在成对数据表示过程中，只让同簇内多数类样本进行两两组合，不同簇之间的多数类样本不进行两两组合，有效降低了两两组合而导致训练数据量成倍增长的问题，从而有效降低了算法复杂度。

算法的计算复杂度主要集中在谱聚类算法和多层感知机模型上，其中拉普拉斯矩阵特征分解的时间复杂度一般为 $O(n^3)$，可以采用近似加权核的方法将拉普拉斯矩阵特征分解的时间复杂度降到小于 $O(n^2)$[28]，而多层感知机模型的时间复杂度为 $O(n^2)$，因此算法的总时间复杂度为 $O(n^2)$。

3 实验设置与结果分析

3.1 评价指标

在类别不均衡分类学习中，常常采用 F-measure 和 G-mean 作为评价指标。需要首先定义分类结果的混淆矩阵，如表 1 所列。

表 1 分类结果的混淆矩阵

	预测的正例	预测的负例
实际的正例	真正率(TP)	假负率(FN)
实际的负例	假正率(FP)	正负例(TN)

正例预测值(查准率)定义为：

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

真正例率(召回率或敏感度)定义为：

$$Recall=TPR=\frac{TP}{TP+FN}$$

真负例率(特异性)定义为：

$$TNR=\frac{TN}{TN+FP}$$

F-measure 定义为：

$$F\text{-}measure=\frac{2\times Precision\times Recall}{Precision+Recall}$$

G-mean 定义为：

$$G\text{-}mean=\sqrt{TPR\times TNR}$$

3.2 实验数据集

本文选用来自国际机器学习标准数据库 UCI 的 9 组不同领域数据集进行实验分析，实验数据描述如表 2 所列，其中 IR(imbalanced ratio)表示不均衡比例，即数据集中多数类样本数目与少数类样本数目的比值。

表 2 实验数据描述

数据集	样本数	特征数	IR
glass0	214	9	1.8
pima	768	8	1.9
vehicle0	846	18	3.2
yeast3	1484	8	8.1
ecoli4	336	7	15.8
pageblock	472	10	15.8
glass5	314	9	22.8
yeast6	1484	8	41.4
abalone9	4174	8	129.4

3.3 实验结果分析

为了比较本文提出的算法在类别不均衡数据集上的分类效果，选择如下方法：1)基于过抽样方法，有合成少数类过抽样技术(Synthetic Minority Oversampling Technique，SMOTE)[29]的 SVM、C4.5 和 MLP 分类方法(SSVM，SC45，SMLP)；2)基于欠抽样的方法，有压缩最近邻规则(Condensed Nearest Neighbor，CNN)[30] KNN 和 C4.5 分类方法(CKNN 和 CC45)；3)代价敏感的分类方法，有 GSVM 算

法[31]；4)集成学习方法，有 EUSBoost 算法[32]、RUSBoot 算法[33]和随机抽样的成对数据表示的多层感知机模型(S2SMLP)[27]。本文提出的方法记为 SS2SMLP。

(1)成对数据表示中多数类样本的个数选择

本文首先将多数类样本进行谱聚类，根据聚类簇大小在每个聚类簇中选择有代表性的多数类样本，实现对多数类样本欠采样目标，相比随机欠采样，此方法目的性更强，被选择的样本更有代表性。聚类个数的确定是解决此问题的首要工作，采用 R 语言中的 speccalt 包对每个数据集进行自动聚类分析，为每个数据集确定聚类个数。例如 ecoli4 数据集中的多数类样本将被聚类成 8 个簇，而 abalone9 数据集中的多数类样本将被聚类成 19 个簇。

依据本文提出的算法，从每个聚类簇中选择有代表性的 M_N 个多数类样本与少数类样本进行两两组合，构造成对数据表示数据集作为多层感知机的输入，多层感知机的隐含层数为 1，学习率为 0.001，迭代次数为 5。

(2)F-measure 和 G-mean 结果的对比分析

实验过程中，采用五折交叉验证的方法对分类效果进行验证，计算每种分类算法的 *F-measure* 值和 *G-mean* 值如表 3 和表 4 所列。从中可以看出，在低不平衡比例($IR<10$)数据集上，本文方法相比其他对比方法优势不是特别显著，因为实验中使用的这几个低不平衡数据集的数据量较少，基于谱聚类的样本选择优势不明显，但本文方法在 F-measure 和 G-mean 上的表现仍然与对比方法中的最优值持平。在高不平衡比例($IR>10$)数据集上，本文方法在 *F-measure* 和 *G-mean* 值上要优于其他对比方法。

表 3　不同数据集的 *F-measure* 值的实验结果对比

	glass0	pima	vehicle0	yeast3	ecoli4	pageblock	glass5	yeast6	abalone9
CKNN	0.60	0.58	0.85	0.63	0.72	0.64	0.67	0.14	0.01
CC45	0.60	0.65	0.87	0.75	0.49	0.79	0.76	0.17	0
SSVM	0.63	0.63	0.63	0.64	0.72	0.55	0.48	0.26	0.02
SMLP	0.60	0.57	0.59	0.61	0.55	0.75	0.35	0.12	0.04
SC45	0.69	0.64	0.87	0.75	0.70	0.77	0.67	0.36	0.05
EUSBoost	0.72	0.66	0.89	0.71	0.57	0.77	0.60	0.22	0.03
RUSBoost	0.72	0.66	0.91	0.69	0.57	0.80	0.45	0.23	0.03
GSVM	0.57	0.67	0.92	0.66	0.71	0.69	0.62	0.26	0.05
S2SMLP	0.68	0.69	0.92	0.72	0.72	0.79	0.75	0.23	0.05
SS2SMLP	0.72	0.69	0.94	0.78	0.74	0.82	0.77	0.38	0.06

表 4　不同数据集的 *G-mean* 值的实验结果对比

	glass0	pima	vehicle0	yeast3	ecoli	pageblock	glass5	yeast6	abalone
CKNN	0.60	0.63	0.91	0.84	0.91	0.86	0.93	0.70	0.40
CC45	0.60	0.74	0.92	0.90	0.84	0.93	0.93	0.33	0
SSVM	0.65	0.64	0.72	0.70	0.90	0.68	0.85	0.87	0.54
SMLP	0.66	0.65	0.76	0.88	0.88	0.87	0.91	0.78	0.73
SC45	0.76	0.72	0.94	0.90	0.95	0.94	0.97	0.82	0.39
EUSBoost	0.78	0.73	0.96	0.92	0.89	0.95	0.97	0.82	0.60
RUSBoost	0.77	0.72	0.96	0.92	0.89	0.95	0.94	0.84	0.70
GSVM	0.65	0.74	0.96	0.90	0.95	0.93	0.97	0.88	0.76
S2SMLP	0.75	0.74	0.94	0.90	0.93	0.93	0.95	0.82	0.74
SS2SMLP	0.78	0.80	0.94	0.92	0.95	0.96	0.97	0.90	0.77

结束语　为了解决不平衡学习多层感知机分类算法中成对数据表示中成对组合而导致的数据量暴涨问题，本文提出一种基于谱聚类的欠抽样方法，从每个聚类簇中选择不同个数的多数类样本，一方面降低了随机欠抽样造成的多数类中部分信息丢失的可能性，另一方选择有代表性的样本作为成对数据表示的数据源，有效降低了整个算法的复杂度。在标准的 UCI 数据集上，与其他方法相比，本文提出的方法在 F-measure 和 G-mean 两个评价指标上均取得较好的结果。如何在大数据和高维数据上提高基于成对数据表示模型的效率是我们下一步需要关注的重点内容。

参 考 文 献

[1] PROBOST F. Machine learning from imbalanced data set 101 [C]// Proceedings of Workshop on Learning from Imbalanced Data Set (AAAI'00). Palo Alto, CA: AAAI, 2000: 1-3.

[2] CHAWLA N V, JAPKOWICZ N, KOLCZ A. Editorial: special issue on learning from imbalanced data sets[J]. SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets, 2004, 6(1): 1-6.

[3] GALAR M, FERNANDEZ A, BARRENCHEA E, et al. A review on ensembles for the class imbalance problem: Bagging-, Boosting-, and hybrid-based approaches[J]. IEEE Transaction on Systems, Man and Cybernetics, 2012, 42(4): 463-484.

[4] KRAWCZYK B. Learning from imbalanced data: open challenge and future directions[J]. Progress in Artificial Intelligence, 2016, 5(4): 1-12.

[5] ROY A, CRUZ R M O, CAVALCANI G D C. A study on combining dynamic selection and data preprocessing for imbalanced learning[J]. Neurocom-puting, 2018, 286: 179-192.

[6] GUO H, LI Y, JENNIFER S, et al. Learning from class-imbalanced data: review of methods and applications[J]. Expert Systems with Applications, 2017, 73: 220-239.

[7] YANG Q, WU X. 10 challenging problems in data mining research[J]. International Journal of Information Technology and Decision Making, 2006, 5(4): 597-604.

[8] FERNANDEZ A, RIO S, CHAWLA N V, et al. An insight into imbalanced big data classification: outcomes and challenges[J]. Complex Intelligent Systems, 2017, 3(2): 105-120.

[9] GUERMAZI R, CHAABANE I, HAMMAMI M. AECID: asymmetric entropy for classifying imbalanced data[J]. Information Sciences, 2018, 467: 373-397.

[10] WU F, JING X, SHIN S, et al. Multiset feature learning for highly imbalanced data classification[C]// Proceedings of the thirty-first AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 1583-1589.

[11] LOYOLA-GONZALEZ O, MARTINEZ-TRINIDAD J F, CARRASCO-OCHOA J A. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases[J]. Neurocomputing, 2016, 175: 935-947.

[12] LIN C, HSIEH T, LIN Y, et al. Minority Oversampling in Kernel Adaptive Subspaces for Class Imbalanced Datasets[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(5): 950-962.

[13] SHAHEE S A, ANANTHAKUMAR U. An adaptive oversampling technique for imbalanced datasets[C]// Proceedings of IEEE International Conference on Data Mining (ICDM'18). NJ: IEEE, 2018: 1-16.

[14] LIN W, TSAI C, HU Y, et al. Clustering-based undersampling in class-imbalanced data[J]. Information Sciences, 2017, 409/410: 17-26.

[15] LI F, ZHANG X, ZHANG X, et al. Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets [J]. Information Sciences, 2018, 422: 242-256.

[16] DECHERCHI S, ROCCHIA W. Import vector domain description: a kernel logistic one-class learning algorithm[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(7): 1722-1729.

[17] FERNANDEZ-FRANCOS D, FONTENLA-ROMERO O, ALONSO-BETANZOS A. One-class convex hull-based algorithm for classification in distributed environments [J]. IEEE Transactions on Systems, Man and Cybernetics, 2017, 99: 1-11.

[18] SUN J, SHAO J, HE C. Abnormal event detection for video surveillance using deep one-class learning[J]. Multimedia Tools and Applications, 2017, 3: 1-15.

[19] ERFANI S M, REJASEGARAR S, KARUNA-SEKERA S, et al. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning[J]. Pattern Recognition, 2016, 58(C): 121-134.

[20] FERDOWSI Z, GHANI R, SETTIMI R. Online active learning with imbalanced Classes[C]// Proceedings of IEEE 13th International Conference on Data Mining (ICDM'13), NJ: IEEE, 2013: 1043-1048.

[21] ZHANG X, YANG T, SRINIVASAN P. Online asymmetric active learning with imbalanced data[C]// Proceedings of 22th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'16). New York: ACM. 2016: 2055-2064.

[22] RAMIREZ-LOAIZA M, SHARMA M, KUMAR G, et al. Active learning: An empirical study of common baselines[J]. Data Mining and Knowledge Discovery, 2017, 31: 287-313.

[23] ZHANG Y, ZHAO P, CAO J, et al. Online adaptive asymmetric active learning for budgeted imbalanced data[C]// Proceedings of 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'18). New York: ACM. 2018: 2768-2777.

[24] LI K, KONG X, LU Z. Boosting weighted ELM for imbalanced learning[J]. Neurocomputing, 2014, 128: 15-21.

[25] YU H, SUN C, YANG X, et al. ODC-ELM: optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data[J]. Knowledge-Based Systems, 2016, 92: 55-70.

[26] DING S, MIRZA B, LIN Z, et al. Kernel based online learning for imbalance multi- class classification[J]. Neurocomputing, 2018, 277: 139-148.

[27] DUMPALA S H, CHAKRABORTY R, KOPPARAPU SK. A novel data representation for effective learning in class imbalanced scenarios[C]// Proceedings of the Twenty-seventh International Joint Conference on Artificial Intelligence. 2018: 2100-2106.

[28] 贾洪杰, 丁世飞, 史忠植. 求解大规模谱聚类的近似加权核 k-means 算法[J]. 软件学报, 2015, 26(11): 2836-2846.

[29] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority oversampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[30] HART P. The condensed nearest neighbor rule [J]. IEEE Transactions on Information Theory, 1968, 14: 515-516.

[31] TANG Y, ZHANG Y, CHAWLA N V, et al. SVMs modeling for highly imbalanced classification [J]. IEEE Transactions on Systems, Man, and Cybernetics, 2009, 39(1): 281-288.

[32] GALAR M, FERNANDEZ A, BARRENECHEA E, et al. Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling [J]. Pattern Recognition, 2013, (12): 3460-3471.

[33] 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, 2010, 40(1): 185-197.