AdaBoostRS:高维不平衡数据学习的集成整合

杨平安 林亚平 祝团飞
(湖南大学信息科学与工程学院 长沙 410000)

摘 要 机器学习中类不平衡分布问题包含了不同类之间数据样本的偏差分布,导致学习过程更偏向于多数类。而高维数据的稀疏性使得分类的偏差更加明显,因此对于高维不平衡数据,维度灾难与类不平衡分布这两个挑战性问题相互叠加在一起,使得解决高维不平衡问题变得更为困难。针对这一问题,文中提出结合随机子空间和 SMOTE 过采样技术的 AdaBoost 集成方法(AdaBoost ensemble of Random subspace and SMOTE,AdaBoostRS)来处理高维不平衡数据的分类。具体地,AdaBoostRS 通过随机子空间选取部分特征来训练每个分类器,以增加分类样本的多样性和降低高维数据的维度,然后通过 SMOTE 方法对降维数据的少数类进行线性插值,以解决类不平衡问题。基于 8 个高维不平衡的标准时间序列数据集进行实验,结果表明,以 F-measure、G-mean 与 AUC 3 个性能指标来进行评判,AdaBoostRS 优于传统的集成学习方法。

关键词 高维不平衡,随机子空间,SMOTE,AdaBoost
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/jsjkx.180901813

AdaBoostRS:Integration of High-dimensional Unbalanced Data Learning

YANG Ping-an LIN Ya-ping ZHU Tuan-fei
(College of Information Science and Engineering,Hunan University,Changsha 410000,China)

Abstract The class imbalance problem in machine learning contains a skewed distribution of data samples among different classes,resulting in a learning bias toward the majority class. In high-dimensional data,the sparseness of the data makes the classification bias more obvious. For high-dimensional unbalanced data,the two challenging problems of dimensional disaster and class imbalance distribution are superimposed,making it more difficult to solve high-dimensional imbalance problems. This paper proposed an AdaBoost integration method combining random subspace and SMOTE oversampling technology,named AdaBoostRS (AdaBoost ensemble of Random subspace and SMOTE),to deal with the classification of high-dimensional unbalanced data. AdaBoostRS trains each classifier by selecting partial features in a random subspace to increase the diversity of the classification samples and reduce the dimensions of the high-dimensional data. Then a few classes of dimensionality reduction data are linearly interpolated through the SMOTE method to solve the class imbalance problem. The experiment is based on 8 high-dimensional unbalanced standard time series dataset. The results show that AdaBoostRS is superior to the traditional integrated learning method in terms of three performance indicators of F-measure,G-mean and AUC.

Keywords High-dimensional imbalance,Random subspace,SMOTE,AdaBoost

1 引言

不平衡学习问题是指在不同类之间样本分布不均衡,即多数类样本明显多于少数类样本。在实际分类问题中,经常会遇到不平衡数据集,例如疾病诊断[1]、网络入侵检测[2]、软件缺陷预测[3]、人脸识别[4]、文本分类[5]等,它们都被列为不平衡性问题[6-8]。如果数据样本仅来自两个类,则包含大多数样本的类被称为多数类,其他的被称为少数类。在上述应用领域中,少数类才是至关重要的。

对于传统的分类算法,如决策树[9]、贝叶斯网络[10]、支持向量机[11]、神经网络[12]等,类不平衡问题是[13-14]一个难点。由于学习分类器对多数类产生严重偏置[15-16],因此导致少数类的分类性能很差。然而,在不平衡学习领域中识别稀有的少数样本通常是至关重要的,在现实生活中,少数类的错误分类往往会导致严重的后果。例如,在医学诊断中,将癌症患者(少数类样本)错误分类为非癌症患者[17],将会导致不可估量的后果,类似的还有直升机故障监测[18]、财务欺诈预测[19]等。

目前解决不平衡问题的方法主要有:代价敏感、数据重采样与集成学习。

1)代价敏感学习算法是给少数类样本分配大的误分类代

到稿日期:2018-09-27 返修日期:2018-12-16

杨平安(1995-),女,硕士生,主要研究方向为机器学习、数据挖掘等,E-mail:ypingan@hnu.edu.cn;**林亚平**(1956-),男,博士,教授,博士生导师,主要研究方向为计算机网络、云安全和机器学习等,E-mail:yplin@hun.edu.cn(通信作者);**祝团飞**(1987-),男,博士,CCF 会员,主要研究方向为云安全和机器学习等。

价，从而提高分类器的学习性能。Victoria 等[20]提出使用模糊规则和代价敏感学习技术来处理大规模的不平衡数据。Bartosz 等[21]构建了一种基于代价敏感决策树集成的融合算法，通过 ROC 分析来选择被估计的成本矩阵。Maciej 等[22]引入了两种经验代价敏感算法，其中一种是结合采样、代价敏感和支持向量机的算法，另一种则是将代价比作超参数，且在训练最终模型前需要对其进行优化。

2）数据重采样的方法是重新调整数据集以达到类平衡，其中两种最主要的重采样方法是欠采样和过采样。过采样的方法是通过创建新的少数类样本来解决类不平衡问题，而欠采样方法通过减少多数类样本的数量来达到类平衡。在以往的研究中，最著名的过采样算法是由 Chawla 等[23]提出的 SMOTE 方法。近期，也有不少研究者讨论了基于 SMOTE 方法的其他欠采样方法，如 bSMOTE[24]和 V-synth[25]方法。随机欠采样（RUS）[26]的方法也被许多研究者应用，例如 Vorraboot 等[27]提出了数据过滤技术，即将大部分样本分为安全区域、边缘区域和噪声区域，仅将来自安全区域的样本作为训练样本。还有研究者提出了一种新的基于蚁群优化算法的欠采样方法[28]等。

3）集成学习在解决不平衡问题时应用非常广泛，这主要是因为它能够显著地提高单个分类器的性能[29]。其中使用最广泛的是 Yoav[30]提出的 Boosting 算法，它已经被应用于许多著名的集成算法中，如 SMOTEBoost[31]，RUSBoost[32]，EasyEnsemble[33]，EUSboost[34]。后来 Sun 等[35]提出了一种新的不平衡数据分类的集成方法，该方法是将不平衡数据集转换为多个平衡子集，对每个子集进行训练得到基分类器，文献[36]对该方法进行了应用。Krawczyk 等[37]创建了一个名为 PUSBE 的集成算法，它包含采样、修剪和 Boosting 技术，该算法首先将数据分为非重叠区域、边界区域和重叠区域，然后用不同的分类器对不同区域进行训练，并探讨了重叠区域和非重叠区域的不平衡情况。Zieba 等[38]提出了一种支持向量机的集成，使得每个用于构建基本分类器的训练集更加均衡。

由于现实世界中不仅仅存在低维不平衡数据，许多重要的应用领域中数据特征呈现高维与类不平衡分布，例如放射性威胁分类[39]、基因功能注释[40]以及文本分类[41]等。但是应用现有的研究直接对高维不平衡数据进行操作将面临一系列困难。因为处于高维空间中的少数类样本的分布会更加稀疏，使得分类算法难以对多数类和少数类进行区分；且高维数据还会带来维度灾难的问题，维度灾难会使大多数学习算法的计算开销随着维度的升高而呈指数增长[42-43]。这些问题给原本不平衡的数据带来了更大的挑战。

本文针对高维不平衡数据的问题，提出了 AdaBoostRS 方法，它是在 SMOTEBoost 算法的基础上增加一个随机子空间，以更加高效地解决高维不平衡数据的问题。该方法的关键体现在通过随机子空间随机选取部分特征来训练每个分类器，增加分类样本的多样性，与此同时随机子空间还能用来降低维度，以此来降低维度灾难的风险。然后通过 SMOTE 方法对降维数据的少数类进行过采样，以此来处理类不平衡分布的问题。本文对 8 个高维不平衡的时间序列公共数据集进行了实验，结果表明 AdaBoostRS 在 F-measure，G-mean 与 AUC 3 个评价指标上优于传统的集成学习方法。

本文第 2 节对 SMOTE 算法与随机子空间的算法进行了介绍；第 3 节对本文提出的方法 AdaBoostRS 进行了介绍；第 4 节给出实验与结果分析；最后总结全文。

2　相关算法简介

2.1　SMOTE 算法

SMOTE 算法的主要思想是通过在特性空间中创建新的合成实例以拓宽少数类的分布区域。算法的核心思想如下。

1）对于少数类中的每一样本，找到它的 k 个最近的少数类邻居。

2）随机从 k 个近邻居中选择一个样本 x^j，利用下式生成一个合成样本：

$$x^{syn}=x^i+(x^j-x^i)*\gamma$$

其中，x^i 是所考虑的少数类样本，x^j 是从 x^i 的最接近的少数邻居中随机选择的一个样本；γ 是一个随机向量，其中每个元素来自区间[0，1]。

SMOTE 在处理低维不平衡分类的问题中已经得到广泛的应用，但如果直接处理高维不平衡数据的分类，则容易造成过泛化问题[44]，这是由于少数类样本在高维空间中的稀疏分布更明显，而 SMOTE 是为每个样本找到 k 个最近邻样本，因此在两个样本之间进行插值时更容易产生噪声样本，从而不能产生理想的合成样本[45]，又由于在高维上存在维度灾难的问题，使得区分多数类和少数类变得更加困难，给不平衡分类增加了难度，这些问题对分类器的性能产生负面影响。

2.2　随机子空间算法

随机子空间方法是通过随机选取部分特征而不是所有的特征来训练每个分类器，从而降低了每个分类器之间的相关性，增加了分类样本的多样性。随机子空间方法是一种成功的分类器集成方法[46]，相比单个分类器，它对噪声和冗余信息更加健壮。算法的具体步骤如算法 1 所示。

算法 1　随机子空间算法

1. 输入数据集 $X=\{x_1,x_2,\cdots,x_n\mid x_i\in R^d,i=1,2,\cdots,n\}$。
2. 数据初始化。初始化随机子空间的数目 q，随机子空间的维数 p（$p<d$）。
3. 利用随机函数生成 q 个随机二元向量 $r^t\in R^d(t=1,\cdots,q)$，其中对第 j 个元素有 $r_j^t=\{0,1\}(j=1,\cdots,d)$，且 r^t 满足约束条件 $\sum_{j=1}^{d}r_j^t=p$。
4. 利用交叉函数生成 q 个随机子空间。
5. 输出生成的 q 个随机子空间 $S^1,S^2,\cdots,S^q$。

3　AdaBoostRS：随机子空间＋SMOTE 的 AdaBoost 集成

本节将随机子空间、SMOTE 与 AdaBoost 进行结合，提出 AdaBoostRS 算法，其伪代码如算法 2 所示。具体地，第 1 步完成对数据集的输入；第 2－4 步是对样本的权重及所需参数进行初始化；第 5－22 步是整个算法的核心，其中第 7 步根据算法 1 对每一次加权抽样数据 D_k 进行随机子空间的划分，第 8 步对划分至相对低维的数据集 S^j 进行 SMOTE 插值，第 9—19 步对每一次合成的新样本 S^{symj} 进行 AdaBoost 的集成学习，第 20 步是依次进行循环直到迭代终止；第 23 步是对所有的低维平衡 S^{symj} 上的基分类器进行 Bagging 集成；

第 24 步输出最终的分类结果。

算法 2 AdaboostRS 算法

1. 输入数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$。
2. 对新合成的样本进行权重的初始化 $w=\{w_j=1/N|j=1,2,\cdots,N\}$。
3. 令 T 表示不同的数据集。
4. 令 k 表示不同的数据集。
5. for i=1 to T do
6. for j=1 to q do
7. 对数据集 D_k 使用随机子空间法分别生成 $S^1,S^2,\cdots,S^q$；
8. 对数据集 D_k 上的 S^j 依次进行 SMOTE 插值，得到新的合成样本 S^{symj}；
9. 根据 w，通过对 S^{symj} 进行抽样(有放回)产生训练集 $S^{symj'}$；
10. 在 $S^{symj'}$ 上对训练集 S^{symj} 中的所有样本分类；
11. 用 C_i 对训练集 S^{symj} 中的所有样本分类；
12. 计算加权误差 $\varepsilon_i=\frac{1}{N}[\sum w_j\delta(C_i(x_j)\neq y_j)]$；
13. if $\varepsilon_i>0.5$ then
14. $w=\{w_j=1/N|j=1,2,\cdots,N\}$(重新设定 N 个样本的权值)
15. $\alpha=0$
16. 返回步骤 6
17. end if
18. $\alpha=\frac{1}{2}\ln\frac{1-\varepsilon_i}{\varepsilon_i}$
19. 更新每个样本的权值
20. 返回步骤 5
21. end for
22. end for
23. $c^*(x)=\text{argmax}\sum_{j=1}^{T}\alpha_j\delta(C_j(x)=y)$
24. 输出最终的分类结果。

4 实验与分析

本节将评估 AdaBoostRS 方法的有效性。4.1 节给出了实验数据集、评价指标及比较的方法，4.2 节给出了实验结果与分析。

4.1 实验设置

4.1.1 实验数据集

为评价 AdaBoostRS 算法对高维不平衡数据集分类问题的有效性，本文选择 8 个高维不平衡的标准时间序列数据集作为实验数据。我们从 UCR 时间序列知识库[47-48]中选择了 8 个高维不平衡数据集，表 1 列出了这些数据集的特征。从表 1 中可以看出，这些数据集具有广泛不同数量的样本和特征。

表 1 不平衡时间序列集

Table 1 Unbalanced time series set

数据集名称	训练样本	测试样本	特征数	类别定义	训练分布
DistalPhalanx OutlineCorrect(DPOC)	276	600	80	\0\1	\115\161
ECG200(ECG)	100	100	96	\-1\1	\31\69
Earthquakes(EQ)	139	322	512	\0\1	\104\35
Lighting2(LT)	60	61	637	\-1\1	\20\40
Phalanges OutlinesCorrect(POC)	1800	858	80	\0\1	\628\1172
ProximalPhalanx OutlineCorrect(PPOC)	600	291	80	\0\1	\194\406
Strawberry(SB)	370	613	235	\1\2	\132\238
Wafer(WF)	1000	6164	152	\-1\1	\97\903

4.1.2 评价指标

对于二分类问题，可将样例根据其真实类别与学习器预测类别的组合划分为：真阳性样本(TP，正确预测成正类)；真阴性样本(TN，正确预测为负类)；假阳性样本(FP，错误预测成正类)；假阴性样本(FN，错误预测为负类)。根据这 4 项指标可以计算得到 $Precision$，TRP 和 TNR 的值，具体公式如下：

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

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

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

上述公式还可以派生得到以下几个常用的评价指标，如 F-measure[49]与 G-mean[50]，计算公式如下：

$$F\text{-}measure=\frac{(1+\beta)^2*Precision*Recall}{(\beta^2*Precision*Recall)}$$

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

其中，β 用来调节 $Precesion$ 和 $Recall$ 的相对重要性，通常情况下 β 设置为 1。

上述指标中，Precision 用于评价预测为正类的实例的可信度，Recall 用于评价有多大比例的正类实例被正确预测，F-measure 用于权衡 Precision 和 Recall。G-mean 反映了分类器识别少数类和多数类的平衡程度。F-measure 和 G-mean 是两种常用的综合评价指标，另一种常用的综合指标是接受者操作特征曲线下的面积(AUC)，与 F-measure 和 G-mean 不同，AUC 不依赖于分类器的具体决策阈值。本文分别使用 F-measure，G-mean 与 AUC 3 个指标来评估算法的分类性能。

4.1.3 实验对比方法

我们使用 AdaBoost，SmoteBoost，RAdaBoost(随机子空间＋AdaBoost)方法与本文提出的 AdaBoostRS 方法进行比较。其中，AdaBoostRS 所使用的参数都是默认设置，SMOTE 算法的邻域值 K 设置为 5[51]，随机子空间的个数 p 设置为 10[46]，维数 q 设置为原样本维数的一半[46]，AdaBoost 的迭代次数为 25。

4.2 实验结果与分析

采用不同方法对表 1 的不平衡数据集进行实验，结果如表 2－表 4 所列，分别是 4 种比较方法在不同数据集下的 $F\text{-}measure$ 值、$G\text{-}mean$ 值以及 AUC 值。其中，性能最好的结果加粗显示，且每次所得结果都是通过 10 次交叉验证得到的。

表 2 4 种方法在 8 个数据集上的 *F-measure* 值

Table 2 *F-measure* values of 4 methods on 8 datasets

data	AdaBoost	SmoteBoost	RAdaBoost	AdaBoostRS
DPOC	0.9337	0.9330	0.9374	**0.9378**
ECG	0.9532	0.9580	0.9536	**0.9620**
EQ	0.7325	**0.7506**	0.7312	0.7445
LT	0.9153	0.9235	0.9245	**0.9341**
POC	0.8329	0.8241	0.8295	**0.8343**
PPOC	0.9853	0.9853	0.9853	**0.9853**
SB	0.9694	0.9658	0.9564	**0.9703**
WF	0.8688	**0.8688**	0.8688	0.8535

表 3　4 种方法在 8 个数据集上的 *G-mean* 值

Table 3　*G-mean* values of 4 methods on 8 datasets

data	AdaBoost	SmoteBoost	RAdaBoost	AdaBoostRS
DPOC	0.3500	**0.4492**	0.1443	0.3792
ECG	0.9305	0.9297	0.9256	**0.9388**
EQ	0.5455	0.5806	0.5497	**0.5807**
LT	0.1882	**0.2989**	0.1892	0.2859
POC	0.4358	0.4700	0.3910	**0.4730**
PPOC	0.9733	0.9733	0.9733	**0.9733**
SB	0.8430	0.8237	0.7268	**0.8543**
WF	0.1465	0.1507	0.1543	**0.2053**

表 4　4 种方法在 8 个数据集上的 *AUC* 值

Table 4　*AUC* values of 4 methods on 8 datasets

data	AdaBoost	SmoteBoost	RAdaBoost	AdaBoostRS
DPOC	0.6388	**0.6750**	0.5898	0.6481
ECG	0.9324	0.9326	0.9270	**0.9419**
EQ	0.6485	0.6716	0.6487	**0.6722**
LT	0.5461	0.5752	0.5338	**0.5814**
POC	0.5196	0.5349	0.5037	**0.5357**
PPOC	0.9347	**0.9599**	0.9479	**0.9599**
SB	0.8127	0.7899	0.7150	**0.8244**
WF	0.7984	**0.7984**	0.7984	0.7720

从 3 种评价指标的结果来看，本文提出的方法在大部分数据集上取得了很好的结果，说明本文方法能更加有效地处理高维不平衡分类问题。

为了更方便地比较不同方法之间的性能，本文计算了 4 种方法的平均排名，结果如图 1 所示。在大多数情况下，AdaBoost 与 RAdaBoost 是两种处理高维不平衡问题最差的方法，这是因为它们没有考虑到少数类的局部分布。而本文提出的 AdaBoostRS 方法比 SmoteBoost 方法的排名更好，验证了上文提到的使用 SMOTE 方法直接处理高维不平衡问题会出现过泛化等问题，导致分类性能变差。

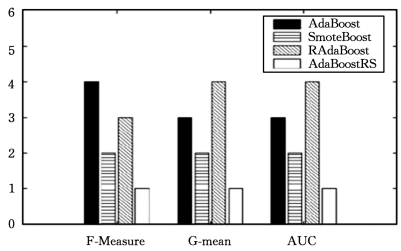

图 1　各方法在不同指标上的平均排名

Fig. 1　Average ranking of each method under different indicators

由图 1 可知，本文提出 AdaBoostRS 方法平均排名第一，虽然其他 3 种方法也能在一定程度上解决高维不平衡的分类问题，但本文所提方法是最有效的。

结束语　针对高维不平衡分类的问题，本文提出了随机子空间与 SMOTE 的 AdaBoost 集成方法（即 AdaBoostRS）。本文首先理论分析了该方法的可行性，然后对 8 个高维不平衡的标准时间序列数据集进行了有效性验证。但随机子空间的方法会大大增加空间复杂度，耗费时间较长，因此对该算法进行进一步优化是我们下一步的工作。

参 考 文 献

[1] PARVIN H, BEHROUZ M B, HOSEIN A. Detection of cancer patients using an innovative method for learning at imbalanced datasets[C] // International Conference on Rough Sets and Knowledge Technology. Springer, Berlin, Heidelberg, 2011.

[2] CIESLAK D A, CHAWLA N V, STRIEGEL A. Combating imbalance in network intrusion datasets[C] // GrC. 2006: 732-737.

[3] JING X Y, WU F, DONG X W, et al. An improved SDA based defect prediction framework for both within-project and cross-project class-imbalance problems[J]. IEEE Transactions on Software Engineering, 2017, 43(4): 321-339.

[4] ZHANG Y, ZHOU Z H. Cost-sensitive face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(10): 1758-1769.

[5] LIU C L, HSAIO W H, LEE C H, et al. Semi-supervised text classification with universumlearning[J]. IEEE Transactions on Cybernetics, 2016, 46(2): 462-473.

[6] LIU X Y, WU J X, ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009, 39(2): 539-550.

[7] SÁEZ J A S, LUENGO J, STEFANOWSKI J, et al. SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering [J]. Information Sciences, 2009, 21(9): 184-203.

[8] HE H B, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge & Data Engineering, 2009, 21(9): 1263-1284.

[9] ALBERTO C, ZAFRA A, VENTURA S. Weighted data gravitation classification for standard and imbalanced data[J]. IEEE Transactions on Cybernetics, 2013, 43(6): 1672-1687.

[10] DANIELE C R, PORTINALE L. Dynamic Bayesian networks for fault detection, identification, and recovery in autonomous spacecraft[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2015, 45(1): 13-24.

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

[12] KANG Q, HUANG B Y, ZHOU M C. Dynamic behavior of artificial Hodgkin-Huxley neuron model subject to additive noise [J]. IEEE Transactions on Cybernetics, 2016, 46(9): 2083-2093.

[13] ZHANG X W, HU B G. A new strategy of cost-free learning in the class imbalance problem[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(12): 2872-2885.

[14] LIU X Y, ZHOU Z H. The influence of class imbalance on cost-sensitive learning[C] // Sixth International Conference on Data Mining (ICDM'06). IEEE, 2006: 970-974.

[15] WEISS, GARY M. Mining with rarity: a unifying framework [J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1): 7-19.

[16] PRATI, RONALDO C, BATISTA G E, et al. Class imbalances versus class overlapping: an analysis of a learning system behavior[C] // Mexican International Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2004.

[17] RAO, BHARAT R, KRISHNAN S, et al. Data mining for improved cardiac care[J]. ACM SIGKDD Explorations Newsletter 2006, 8(1): 3-10.

[18] JAPKOWICZ, NATHALIE, MYERS C, et al. A novelty detection approach to classification[M]. Morgan Kaufmann Publishers Inc, 1995.

[19] DI MARTINO M, DECIA F, MOLINELLI J, et al. Improving Electric Fraud Detection using Class Imbalance Strategies [C]// ICPRAM. 2012:135-141.

[20] VICTORIA L, SARA D R, MANUEL B J, et al. Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data [J]. Fuzzy Sets and Systems, 2015(258):5-38.

[21] BARTOSZ K, WOĆNIAK M, SCHAEFER G. Cost-sensitive decision tree ensembles for effective imbalanced classification[J]. Applied Soft Computing, 2014(14):554-562.

[22] MACIEJ Z, TOMCZAK J M. Boosted SVM with active learning strategy for imbalanced data[J]. Soft Computing, 2015, 19(12): 3357-3368.

[23] 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.

[24] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]// International Conference on Intelligent Computing. Springer, Berlin, Heidelberg, 2005.

[25] YOUNGW A, NYKL S L, WECKMAN G R, et al. Using Voronoi diagrams to improve classification performances when modeling imbalanced datasets[J]. Neural Computing and Applications, 2015, 26(5):1041-1054.

[26] LIU X Y, WU J, ZHOU Z H. Exploratory Under-sampling for class-imbalance learning, bioinformatics[J]. Proceedings of the IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(2):539-550.

[27] VORRABOOT P, RASMEQUAN S, CHINNASARN K. Improving classification rate constrained to imbalanced data between overlapped and non-overlapped regions by hybrid algorithms [J]. Neurocomputing, 2015(152):429-443.

[28] YU H L, NI J, ZHAO J. ACOSampling: an ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data[J]. Neurocomputing, 2013(101):309-318.

[29] YIN Q Y, ZHANG J S, ZHANG C X, et al. A novel selective ensemble algorithm for imbalanced data classification based on exploratory undersampling[J]. Mathematical Problems in Engineering, 2014, 71(3):741-764.

[30] YOAV F. Boosting a weak learning algorithm by majority[J]. Information and Computation, 1995, 121(2):256-285.

[31] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. [J]. Lecture Notes in Computer Science, 2003, 2838:107-119.

[32] SEIFFERT C, KHOSHGOFTAAR T M, VAN HULSE J, 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.

[33] LIU X Y, WU J, ZHOU Z H. Exploratory Under-sampling for class-imbalance learning, bioinformatics [J]. Proceedings of the IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(2):539-550.

[34] NANNI L, FANTOZZI C, LAZZARINI N. Coupling different methods for overcoming the class imbalance problem[J]. Neurocomputing, 2015, 158:48-61.

[35] SUN Z, SONG Q, ZHU X. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition, 2015, 48:1623-1637.

[36] DÍEZ-PASTOR J F, RODRÍGUEZ J J, GARCÍA-OSORIO C, et al. Random balance: ensembles of variable prors classifiers for imbalanced data[J]. Knowledge-Based Systems, 2015, 85:96-111.

[37] KRAWCZYK B, SCHAEFER G. An improved ensemble approach for imbalanced classification problems[C]// IEEE, International Symposium on Applied Computational Intelligence and Informatics. IEEE, 2013:423-426.

[38] ZIEBA M, TOMCZAK J M. Boosted SVM with active learning strategy for imbalanced data[J]. Soft Computing, 2015, 19(12): 3357-3368.

[39] BELLINGER C, JAPKOWICZ N, DRUMMOND C. Christopher Drummond. Synthetic Oversampling for Advanced Radioactive Threat Detection[C]// 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). IEEE, 2015: 948-953.

[40] MATHIEU B, SEKI K, UEHARA K. Tackling class imbalance and data scarcity in literature-based gene function annotation[C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011.

[41] NGUWI Y Y, CHO S Y. Support vector self-organizing learning for imbalanced medical data[C]// International Joint Conference on Neural Networks(IJCNN 2009). IEEE, 2009:2250-2255.

[42] NASRABADI, NASSER M. Pattern recognition and machine learning[J]. Journal of electronic imaging, 2007, 16(4):049901.

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

[44] BELLINGER C, DRUMMOND C, JAPKOWICZ N. Manifold-based synthetic oversampling with manifold conformance estimation[J]. Machine Learning, 2018, 107(3):605-637.

[45] CUI Y, MA H, SAHA T. Improvement of power transformer insulation diagnosis using oil characteristics data preprocessed by SMOTEBoosttechnique[J]. IEEE Transactions on Dielectrics and Electrical Insulation, 2014, 21(5):2363-2373.

[46] GU J, JIAO L, LIU F, et al. Random subspace based ensemble sparse representation[J]. Pattern Recognition, 2018(74):544-555.

[47] KEOGH E, XI X, WEI L C A. Ratanamahatana. UCRTime Series Classification/ClusteringPage[OL]. http://www.cs.ucr.edu/~eamonn/time_series_data.

[48] WEI L, KEOGH E J. Semi-Supervised Time Series Classification [C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006: 748-753.

[49] GAO J W, LIANG J Y. Research and advancement of classification method of imbalanced data sets[J]. Computer Sciense, 2008, 35:10-13.

[50] LI K W, YANG L, LIU W Y, et al. Unbalanced Data Classification Method Based on RSBoost Algorithm[J]. Computer Science, 2015, 42(9):249-252.

[51] 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.