基于包络学习和分级结构一致性机制的不平衡集成算法

李 帆[1],张小恒[1,2],李勇明[1*],王 品[1]
(1. 重庆大学微电子与通信工程学院,重庆 400030;2. 重庆广播电视大学,重庆400052)

摘 要: 集成方法是不平衡学习方法的重要分支,然而,现有不平衡集成方法均作用于原样本而没考虑样本的结构信息,因此其效能仍然有限. 样本的结构信息包括局部和全局结构信息. 为了解决上述问题,本文提出了一种基于深度样本包络网络(Deep Instance Envelope Network, DIEN)和分级结构一致性机制(Hierarchical Structure Consistency Mechanism, HSCM)的不平衡集成学习算法. 该算法在考虑局部流形和全局结构信息的情况下,通过多层样本聚类,生成高质量的多层包络样本,从而实现类平衡化. 首先,算法基于样本近邻拼接和模糊C均值聚类算法,设计DIEN来挖掘样本的结构信息,得到深度包络样本. 然后,设计局部流形结构度量和全局结构分布度量来构建HSCM用于增强层间样本的分布一致性. 接着,将DIEN和HSCM结合起来,构建出优化后的深度样本包络网络——DH (DIEN with HSCM). 之后,将基分类器应用于包络样本. 最后,设计bagging集成学习机制来融合基分类器的预测结果. 文末组织了多组实验,采用了十多个公共数据集和有代表性的相关算法进行验证比较. 实验结果表明,本文算法在AUC(Area Under Curve), F-measure等四个性能指标上显著最优.

关键词: 不平衡学习;包络学习;分级结构一致性机制;局部流形结构度量;全局结构分布度量
基金项目: 国家自然科学基金(No.61771080, No.U21A20448);中央高校基本科研业务费(No.2022CDJJJ-003)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2024)03-0751-11
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220712

Imbalanced Ensemble Algorithm Based on Envelope Learning and Hierarchical Structure Consistency Mechanism

LI Fan[1], ZHANG Xiao-heng[1,2], LI Yong-ming[1*], WANG Pin[1]
(1. *College of Communication Engineering, Chongqing University, Chongqing* 400030, *China*;
2. *Chongqing Radio & TV University, Chongqing* 400052, *China*)

Abstract: Ensemble methods have become an important branch of imbalanced learning. However, the existing imbalanced ensemble methods all rely on the original instances without considering the structure information of the instances, so their effectiveness is still limited. The research shows that the structure information of instances includes local and global structure information. In order to solve the above problem, this paper proposes an imbalanced ensemble algorithm based on deep instance envelope network (DIEN) and hierarchical structure consistency mechanism (HSCM). Considering the local manifold and global structure information, the algorithm generates high-quality deep envelope instances to achieve class balance. Firstly, based on the instance neighborhood concatenation and fuzzy c-means clustering algorithm, the DIEN is designed to mine the structure information of instances, obtaining the deep envelope instances. Then, the local manifold structure measure and global structure distribution measure are designed to construct the HSCM to enhance the distribution consistency of interlayer instances. Next, DIEN and HSCM are combined to construct the optimized deep instance envelope network—DH (DIEN with HSCM). Then, the base classifier is applied to the deep envelope instances. Finally, the bagging ensemble learning mechanism is designed to fuse the prediction results of the base classifier to obtain the final results. At the end of this paper, several groups of experiments are organized. More than 10 public datasets and representative related algorithms are used for verification. Experimental results show that the proposed algorithm is significantly better in four perfor-

收稿日期:2022-06-20;修回日期:2022-11-03;责任编辑:梅志强
*通讯作者:李勇明

mance metrics, such as AUC (Area Under Curve) and F-measure.

Key words: imbalanced learning; envelope learning; hierarchical structure consistency mechanism; local manifold structure measure; global structure distribution measure

Foundation Item(s): National Natural Science Foundation of China (No.61771080, No.U21A20448); Central University Basic Scientific Research Operation Cost Special Fund (No.2022CDJJJ-003)

1 引言

类不平衡问题广泛存在于数据分析和挖掘等领域中,一直是研究热点和最具挑战性的问题之一. 类不平衡主要指的是各类别样本的数目不一样. 在分类阶段,当遇到类不平衡时,分类器通常倾向于多数类,因此难以正确分类少数类,导致分类性能变差[1]. 在许多实际应用中,如医疗诊断[2]、信用卡检测[3]和目标识别[4]等,随着数据集的规模变大,类不平衡问题的负面影响也越来越严重.

现有解决类不平衡问题的方法主要分为数据级方法、算法级方法、集成方法和特征选择方法等[5]. 集成方法是指将集成学习与数据级或算法级方法相结合[6]. 大多数集成方法会修改基于bagging和boosting方法中的采样步骤(即数据级). 此外,一些算法会修改样本错误分类的代价(即算法级),但这些算法不容易确定样本误分类的代价[7,8]. 相反,由于数据采样和集成训练步骤之间的独立性,将数据级方法与集成学习结合就相对简单[7]. 研究表明,数据级集成方法为解决类不平衡问题提供了一种有效方案. 例如,SMOTE(Synthetic Minority Oversampling Technique)可以与boosting和bagging结合得到SMOTEBoost (SMOTE with adaBoost)[9], SMOTEBagging (SMOTE with Bagging)[10]. 此外,生成性对抗网络(Generative Adversarial Networks, GANs)也可用于生成少数类样本,以提高合成数据的质量[11]. 欠采样集成方法也取得了明显的改进效果,如UnderBagging (Undersamping with Bagging)[12], RUSBoost (Random Undersampling with AdaBoost)[13], BalancedBagging[14], EasyEnsemble 和 BalanceCascade[15]. 混合数据级方法也可以嵌入到bagging和boosting框架中. Tsai等人[16]将抽样方法与集成分类器相结合,该抽样方法是将聚类与样本选择相结合. Liu等人[17]提出了一种欠采样自动协调数据硬度的不平衡分类框架. Yang等人[18]提出了一种基于密度的欠采样方法的混合优化集成分类器框架. Chen等人[19]开发了一种混合数据级集成框架.

目前大多数数据级不平衡集成方法只考虑重采样来构造类平衡样本,然而,重采样针对样本个体本身,没有挖掘样本间的结构信息,不能生成新样本,也无法很好地代表原样本,难以更好的提升原样本质量. 因此,这些集成方法可能会受到原采样样本的影响,从而导致多样性不佳,不平衡学习性能有限.

2 相关工作

由于目前大多数数据级不平衡集成方法只考虑重采样来构造类平衡样本,没有挖掘样本间的结构信息,因此我们需要研究一种新的数据级集成方法:该方法可以挖掘样本的结构信息,生成新的代表性样本,从而克服原样本的不足,提高不平衡学习性能. 近年来有学者提出了多层聚类算法[20]. 但是,它的目的仍然是聚类,而不是生成新样本;并且其也没有考虑层间样本的分布等.

现有的聚类算法中,由Bezdek提出的模糊均值聚类(Fuzzy C-Means, FCM)是最常见的软聚类算法之一[21],能使样本更加多样化,信息更加丰富. 虽然多层聚类可以用于样本变换以获得更多信息,但层间样本之间的分布往往不一致,从而会影响样本生成(变换)的质量. 因此,本文考虑探索层间样本的结构信息,从而保持层间样本的一致性. 样本的结构信息包括局部流形结构和全局结构信息[22,23]. 基于此,我们将同时挖掘样本的局部流形和全局结构信息,用于生成高质量新样本. 最大平均差异(Maximum Mean Discrepancy, MMD)在域适应[24]中被广泛用于测量全局分布的差异. 流形学习被用来衡量局部结构分布的差异[23].

基于上述分析,本文提出一种基于深度样本包络网络(Deep Instance Envelope Network, DIEN)和分级结构一致性机制(Hierarchical Structure Consistency Mechanism, HSCM)的不平衡集成算法IEDH (Imbalanced Ensemble algorithm based on DIEN and HSCM). 首先,该算法设计了深度样本包络网络来挖掘样本的结构信息,得到深度包络样本,包括样本邻域拼接(Instance Neighborhood Concatenation, INC)和基于多层模糊均值聚类(Multilayer Fuzzy C-Means, MlFCM)的深度包络样本生成. 然后,提出了HSCM来保持层间样本的分布一致性. 其次,将DIEN和HSCM结合起来,构建最终优化的深度样本包络网络——DH网络. 之后,将基分类器应用于包络样本. 最后,设计bagging集成学习机制来融合各基分类器的预测结果,以获得最终结果.

3 方法

3.1 深度样本包络网络

3.1.1 样本近邻拼接

定义数据 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$, $\boldsymbol{x}_n\in\mathbf{R}^s$,其中 n 表示

样本的个数,s 表示维度. 对于样本 $\boldsymbol{x},\boldsymbol{x}\in\boldsymbol{X}$,我们采用欧几里德距离度量获得它的 K 个最近邻样本:

$$d(\boldsymbol{x},\boldsymbol{x}_i)=\sqrt{(\boldsymbol{x}-\boldsymbol{x}_i)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{x}_i)}\ ,1\leqslant i\leqslant n \tag{1}$$

令 $NN_K(\boldsymbol{x},\boldsymbol{X},K)=\{nn_{\boldsymbol{x}}^i\,|\,nn_{\boldsymbol{x}}^i\in\boldsymbol{X}\}_{i=1}^K$ 表示样本 $\boldsymbol{x}$ 的 K 近邻样本,我们用下式定义 K 最近邻搜索的样本:

$$NN_K(\boldsymbol{x},\boldsymbol{X},K)=\boldsymbol{A} \tag{2}$$

其中 $\boldsymbol{A}$ 是满足以下条件的矩阵:

$$\boldsymbol{A}\subseteq\boldsymbol{X},\forall\boldsymbol{x}_p\in\boldsymbol{A},\boldsymbol{x}_q\in\boldsymbol{X}-\boldsymbol{A},\ d(\boldsymbol{x},\boldsymbol{x}_p)\leqslant d(\boldsymbol{x},\boldsymbol{x}_q) \tag{3}$$

获得样本的 K 个最近邻样本 $\boldsymbol{A}$ 后,与原样本进行拼接,形成近邻包络样本 $\boldsymbol{x}_{\mathrm{e}}$,如式(4):

$$\boldsymbol{x}_{\mathrm{e}}=\boldsymbol{x}\oplus\boldsymbol{A} \tag{4}$$

其中 $\oplus$ 表示拼接算子. 因此,样本经过拼接后,原数据 $\boldsymbol{X}$ 变换成新的包络样本数据 $\boldsymbol{X}_{\mathrm{e}}=\{\boldsymbol{x}_{1\mathrm{e}},\boldsymbol{x}_{2\mathrm{e}},\cdots,\boldsymbol{x}_{n\mathrm{e}}\}$,$\boldsymbol{x}_{n\mathrm{e}}\in\mathbf{R}^{(K+1)s}$.

3.1.2　深度包络样本生成

FCM (Fuzzy C-Means) 聚类生成 c 个簇,从而构建包络样本 $\boldsymbol{X}_{\mathrm{e}}=\{\boldsymbol{x}_{1\mathrm{e}},\boldsymbol{x}_{2\mathrm{e}},\cdots,\boldsymbol{x}_{n\mathrm{e}}\},\boldsymbol{x}_{n\mathrm{e}}\in\mathbf{R}^{(K+1)s}$ 相应的簇原型 $\boldsymbol{V}_{\mathrm{e}}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_c\},\boldsymbol{v}_c\in\mathbf{R}^{(K+1)s}$,即聚类包络样本,FCM 的目标函数如下:

$$\min J\left(\boldsymbol{U},\boldsymbol{V}_{\mathrm{e}}\right)=\sum_{i=1}^{c}\sum_{p=1}^{n}u_{ip}^m d_{ip}^2,\ \ \mathrm{s.t.}\sum_{i=1}^{c}u_{ip}=1 \tag{5}$$

其中 u_{ip} 表示样本 $\boldsymbol{x}_{p\mathrm{e}}$ 属于第 i 个簇的隶属程度,划分矩阵 $\boldsymbol{U}=\left(u_{ip}\right)_{c\times n}$,$d_{ip}=\left\|\boldsymbol{x}_{p\mathrm{e}}-\boldsymbol{v}_i\right\|$ 表示样本 $\boldsymbol{x}_{p\mathrm{e}}$ 到聚类中心 $\boldsymbol{v}_i$ 的欧氏距离. m 为模糊系数,通常取值为 $2(m>1)$. 通过优化式(5),可得到原型以及划分矩阵的求解:

$$u_{ip}=\frac{1}{\sum\limits_{j=1}^{c}\left[\dfrac{d_{ip}}{d_{jp}}\right]^{\frac{2}{m-1}}},\quad \boldsymbol{v}_i=\frac{\sum\limits_{p=1}^{n}\left(u_{ip}\right)^m\boldsymbol{x}_{p\mathrm{e}}}{\sum\limits_{p=1}^{n}\left(u_{ip}\right)^m} \tag{6}$$

MlFCM 是基于单层 FCM 聚类(Single layer Fuzzy C-Means, SlFCM)实现的. 主要步骤如下:原数据 $\boldsymbol{X}_{\mathrm{e}}$ 通过式(6)得到新的聚类中心 $\boldsymbol{V}_{\mathrm{e}}{}^1=\left\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{c_1}\right\},\boldsymbol{v}_{c_1}\in\mathbf{R}^{(K+1)s}$,聚类中心作为新数据再通过聚类可再次得到新聚类中心 $\boldsymbol{V}_{\mathrm{e}}{}^2=\left\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{c_2}\right\},\boldsymbol{v}_{c_2}\in\mathbf{R}^{(K+1)s}$. 以此类推,当实现 L 层 FCM 聚类,可获得多层包络样本 $\boldsymbol{V}_{\mathrm{e}}^L=\left\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{c_L}\right\}$.

$$\boldsymbol{v}_{i^L}=\frac{\sum\limits_{i^{L-1}=1}^{c_{L-1}}\left(u_{i^Li^{L-1}}\right)^m\boldsymbol{v}_{i^{L-1}}}{\sum\limits_{i^{L-1}=1}^{c_{L-1}}\left(u_{i^Li^{L-1}}\right)^m},\ i^L=1,2,\cdots,c_L \tag{7}$$

式(7)建立了层间样本之间的变换关系,本节将 INC 和 MlFCM 结合起来构建深度样本包络网络 DIEN. 整体方案如图 1 所示. DIEN 的伪码描述如算法 1 所示.

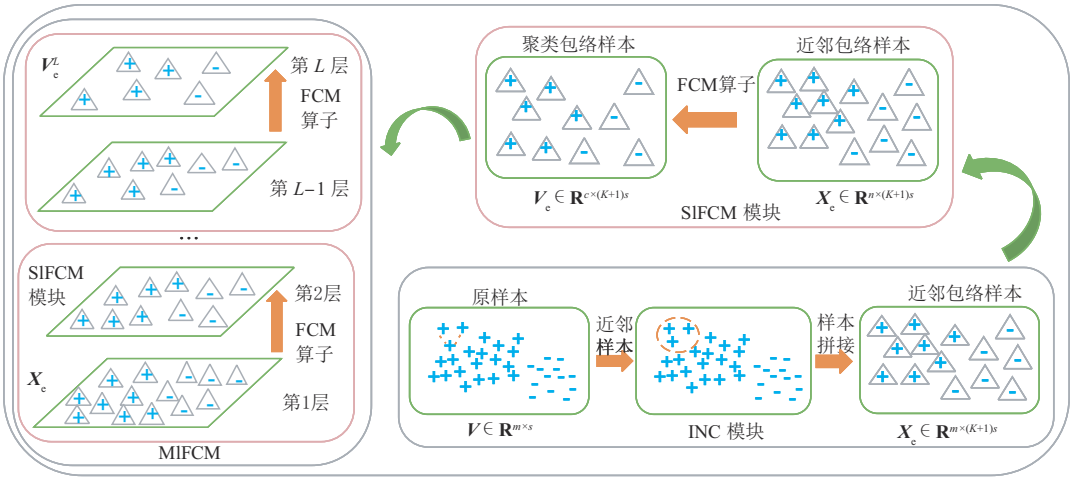


图 1　深度样本包络网络(DIEN)

3.2　分级结构一致性机制

为了提高层间深度包络样本的分布一致性,本文提出了分级结构一致性机制(HSCM). 主要步骤如下:将聚类前后的样本转置表示层间的包络样本 $\boldsymbol{X}_{\mathrm{e}}\in\mathbf{R}^{(K+1)s\times n},\boldsymbol{V}_{\mathrm{e}}\in\mathbf{R}^{(K+1)s\times c}$,$n,c$ 表示样本的个数,通过投影矩阵 $\boldsymbol{P}\in\mathbf{R}^{(K+1)s\times d}$ 将数据映射到一个潜在的公共子空间,即将数据空间 $\mathbf{R}^{(K+1)s}$ 映射到潜在的子空间 $\mathbf{R}^d$ 中 ($(K+1)s\geqslant d$). 在公共子空间中,基于 $\boldsymbol{V}_{\mathrm{e}}$ 和引入的矩阵 $\boldsymbol{G}\in\mathbf{R}^{c\times n}$ 来生成过渡数据 $\boldsymbol{X}_{\mathrm{M}}\in\mathbf{R}^{(K+1)s\times n}$,通过最小化 $\boldsymbol{X}_{\mathrm{M}}$ 和 $\boldsymbol{X}_{\mathrm{e}}$ 之间的局部和全局分布,来保持 DIEN 中的层间样本分布的一致性.

3.2.1　局部流形结构度量

本文的局部结构度量(Local Manifold Structure Metric, LMSM)可以定义为

算法1 深度样本包络网络(DIEN)

输入：原数据$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，近邻样本的个数$K$，每一层聚类的数目$c_1,\cdots,c_L$，聚类的层数$L$，模糊系数$m$，迭代次数$w$以及阈值$\varepsilon$

输出：生成的深度包络样本$\boldsymbol{V}_{\mathrm{e}}{}^{L}$.

步骤：

1. 通过式(2)得到原始样本的最近邻样本;
2. 通过式(4)拼接原始样本得到近邻包络样本数据;
3. For $l=1:L$
4. 随机初始化划分矩阵$\boldsymbol{U}$;
5. $w\leftarrow 1$;
6. 重复
7. 通过式(6)更新新样本$\boldsymbol{V}_{\mathrm{e}}{}^{l(w)}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{c_l}\}$;
8. $w\leftarrow w+1$;
9. 直到$\left|J(\boldsymbol{U},\boldsymbol{V}_{\mathrm{e}})^{(w+1)}-J(\boldsymbol{U},\boldsymbol{V}_{\mathrm{e}})^{(w)}\right|<\varepsilon$;
10. 返回$\boldsymbol{V}_{\mathrm{e}}{}^{l}$，并且$\boldsymbol{V}_{\mathrm{e}}{}^{l}$作为下一层的输入;
11. End
12. 返回深度包络样本$\boldsymbol{V}_{\mathrm{e}}{}^{L}$.

$$\begin{aligned}L_{\mathrm{LMSM}(\mathrm{F}_{X_{\mathrm{M}}},\mathrm{F}_{X_{\mathrm{e}}})}&=\sum_{h,i}^{n}\boldsymbol{S}_{hi}\left\|\varphi(\boldsymbol{x}_h)-\varphi(\boldsymbol{x}_{i\mathrm{e}})\right\|_2^2\\&=\mathrm{Tr}\left(\varphi(\boldsymbol{X}_{\mathrm{M}})\boldsymbol{D}\left(\varphi(\boldsymbol{X}_{\mathrm{M}})^{\mathrm{T}}\right)\right)\\&\quad+\mathrm{Tr}\left(\varphi(\boldsymbol{X}_{\mathrm{e}})\boldsymbol{D}\left(\varphi(\boldsymbol{X}_{\mathrm{e}})^{\mathrm{T}}\right)\right)\\&\quad-2\mathrm{Tr}\left(\varphi(\boldsymbol{X}_{\mathrm{M}})\boldsymbol{S}\left(\varphi(\boldsymbol{X}_{\mathrm{e}})^{\mathrm{T}}\right)\right)\end{aligned}\tag{8}$$

其中$\mathrm{F}_{X_{\mathrm{M}}}$和$\mathrm{F}_{X_{\mathrm{e}}}$表示$\boldsymbol{X}_{\mathrm{M}}$和$\boldsymbol{X}_{\mathrm{e}}$的分布，$\varphi(\boldsymbol{x}_h)\in\varphi(\boldsymbol{X}_{\mathrm{M}})=\varphi(\boldsymbol{V}_{\mathrm{e}})\boldsymbol{G}$，$\varphi(\boldsymbol{x}_{i\mathrm{e}})\in\varphi(\boldsymbol{X}_{\mathrm{e}})$，$\varphi$表示通用的隐式变换，$\mathrm{Tr}(\bullet)$表示矩阵的迹，$\boldsymbol{D}_{hh}=\sum_{i=1}^{n}\boldsymbol{S}_{hi}$转换为矩阵形式后$\boldsymbol{D}$是对角矩阵，$\boldsymbol{S}$是亲和矩阵，可通过下式进行计算：

$$\boldsymbol{S}_{hi}=\begin{cases}1,\ \text{若}\ \boldsymbol{x}_h\in NN_K(\boldsymbol{x}_{i\mathrm{e}})\ \big\|\ \boldsymbol{x}_{i\mathrm{e}}\in NN_K(\boldsymbol{x}_h)\\0,\ \text{其他}\end{cases}\tag{9}$$

令投影矩阵$\boldsymbol{P}^{\mathrm{T}}=\boldsymbol{\Theta}^{\mathrm{T}}\varphi(\boldsymbol{X}_{\mathrm{r}})^{\mathrm{T}}$，投影矩阵$\boldsymbol{P}$表示为$\boldsymbol{X}_{\mathrm{r}}$的某种线性组合，其中$\boldsymbol{\Theta}\in\mathbf{R}^{(c+n)\times d}$表示线性组合系数矩阵，$\boldsymbol{X}_{\mathrm{r}}=\left[\boldsymbol{V}_{\mathrm{e}},\boldsymbol{X}_{\mathrm{e}}\right]\in\mathbf{R}^{(K+1)s\times(c+n)}$，则投影后的$\boldsymbol{X}_{\mathrm{e}}$，$\boldsymbol{V}_{\mathrm{e}}$可以表示为$\boldsymbol{\Theta}^{\mathrm{T}}\varphi(\boldsymbol{X}_{\mathrm{r}})^{\mathrm{T}}\varphi(\boldsymbol{X}_{\mathrm{e}})$，$\boldsymbol{\Theta}^{\mathrm{T}}\varphi(\boldsymbol{X}_{\mathrm{r}})^{\mathrm{T}}\varphi(\boldsymbol{V}_{\mathrm{e}})$. 因此，经过投影映射后，式(8)可以表示为：

$$\begin{aligned}&\min_{\boldsymbol{\Theta},\boldsymbol{G}}\frac{1}{n^2}\mathrm{Tr}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}\boldsymbol{D}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}\right)^{\mathrm{T}}\right)\\&\quad+\frac{1}{n^2}\mathrm{Tr}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{e}}\boldsymbol{D}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{e}}\right)^{\mathrm{T}}\right)\\&\quad-\frac{2}{n^2}\mathrm{Tr}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}\boldsymbol{S}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{e}}\right)^{\mathrm{T}}\right)\end{aligned}\tag{10}$$

其中$\boldsymbol{\Psi}_{\mathrm{v}}=\varphi(\boldsymbol{X}_{\mathrm{r}})^{\mathrm{T}}\varphi(\boldsymbol{V}_{\mathrm{e}})$，$\boldsymbol{\Psi}_{\mathrm{e}}=\varphi(\boldsymbol{X}_{\mathrm{r}})^{\mathrm{T}}\varphi(\boldsymbol{X}_{\mathrm{e}})$是核矩阵.

3.2.2 全局结构分布度量

全局结构分布度量(Global Structure Distribution Metric，GSDM)通过约束生成过渡数据$\boldsymbol{X}_{\mathrm{M}}$，间接地减少$\boldsymbol{X}_{\mathrm{e}}$和$\boldsymbol{V}_{\mathrm{e}}$之间的分布差异，如下所示：

$$L_{\mathrm{GSDM}(\mathrm{F}_{X_{\mathrm{M}}},\mathrm{F}_{X_{\mathrm{e}}})}=\frac{1}{n}\sum_{h=i=1}^{n}\left\|\varphi(\boldsymbol{x}_h)-\varphi(\boldsymbol{x}_{i\mathrm{e}})\right\|_2^2\tag{11}$$

同样地，经过投影映射，式(11)可以表示为如下：

$$\min_{\boldsymbol{\Theta},\boldsymbol{G}}\frac{1}{n}\left\|\boldsymbol{\Theta}^{\mathrm{T}}(\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}-\boldsymbol{\Psi}_{\mathrm{e}})\mathbf{1}\right\|_2^2\tag{12}$$

其中$\mathbf{1}$表示元素全为1的列向量.

3.2.3 联合优化

本文提出的HSCM旨在调整层间深度包络样本的分布，并保留局部流形结构. 因此，通过投影矩阵，在公共子空间中，通过结合LMSM(10)、GSDM(12)和低秩约束(Low Rank Constraint，LRC)正则[25]，最小化$\boldsymbol{X}_{\mathrm{e}}$和$\boldsymbol{V}_{\mathrm{e}}$之间的局部和全局分布差异，如下所示：

$$\begin{aligned}&\min_{\boldsymbol{\Theta},\boldsymbol{G}}\frac{1}{n^2}\mathrm{Tr}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}\boldsymbol{D}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}\right)^{\mathrm{T}}\right)\\&\quad+\frac{1}{n^2}\mathrm{Tr}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{e}}\boldsymbol{D}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{e}}\right)^{\mathrm{T}}\right)\\&\quad-\frac{2}{n^2}\mathrm{Tr}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}\boldsymbol{S}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{e}}\right)^{\mathrm{T}}\right)\\&\quad+\frac{\lambda}{n}\left\|\boldsymbol{\Theta}^{\mathrm{T}}(\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}-\boldsymbol{\Psi}_{\mathrm{e}})\boldsymbol{I}\right\|_2^2+\lambda_1\|\boldsymbol{G}\|_*\\&\text{s.t}\quad\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}\boldsymbol{\Theta}=\boldsymbol{I}\end{aligned}\tag{13}$$

其中$\boldsymbol{\Psi}=\varphi(\boldsymbol{X}_{\mathrm{r}})^{\mathrm{T}}\varphi(\boldsymbol{X}_{\mathrm{r}})$是核矩阵，$\|\boldsymbol{G}\|_*$为矩阵$\boldsymbol{G}$的低秩约束，$\boldsymbol{I}$为单位矩阵. λ,λ_1是平衡参数. 通过引入辅助矩阵$\boldsymbol{J}$，基于增广拉格朗日函数，问题式(13)可重新表达为：

$$\begin{aligned}&\min_{\boldsymbol{\Theta},\boldsymbol{G},\boldsymbol{J}}\frac{1}{n^2}\Big(\mathrm{Tr}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}\boldsymbol{D}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}\right)^{\mathrm{T}}\right)\\&\quad+\mathrm{Tr}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{e}}\boldsymbol{D}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{e}}\right)^{\mathrm{T}}\right)\\&\quad-2\mathrm{Tr}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}\boldsymbol{S}\left(\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{e}}\right)^{\mathrm{T}}\right)\Big)\\&\quad+\frac{\lambda}{n^2}\boldsymbol{\Theta}^{\mathrm{T}}(\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}\boldsymbol{I}(\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G})^{\mathrm{T}}-\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G}\boldsymbol{I}\left(\boldsymbol{\Psi}_{\mathrm{e}}\right)^{\mathrm{T}}\\&\quad-\boldsymbol{\Psi}_{\mathrm{e}}\boldsymbol{I}(\boldsymbol{\Psi}_{\mathrm{v}}\boldsymbol{G})^{\mathrm{T}}+\boldsymbol{\Psi}_{\mathrm{e}}\boldsymbol{I}(\boldsymbol{\Psi}_{\mathrm{e}})^{\mathrm{T}})\boldsymbol{\Theta}+\lambda_1\|\boldsymbol{J}\|_*\\&\quad+\mathrm{Tr}\left(\zeta_1{}^{\mathrm{T}}(\boldsymbol{G}-\boldsymbol{J})\right)+\frac{\delta}{2}\left(\|\boldsymbol{G}-\boldsymbol{J}\|_{\mathrm{F}}^2\right)\end{aligned}\tag{14}$$

其中ζ_1表示拉格朗日乘子，δ为惩罚因子. $\boldsymbol{\Theta},\boldsymbol{J},\boldsymbol{G}$优化过程可采用变量交替策略，即优化其中一个变量时，固定剩下两个变量. 具体的优化过程可以参考文献[23]. 当$\boldsymbol{\Theta}$获得最优解时，通过式子$\boldsymbol{V}_{\mathrm{e}}'=\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{v}}$获得新的深度包络样本集. HSCM的总体方案如图2所示，HSCM的伪代码描述如算法2所示.

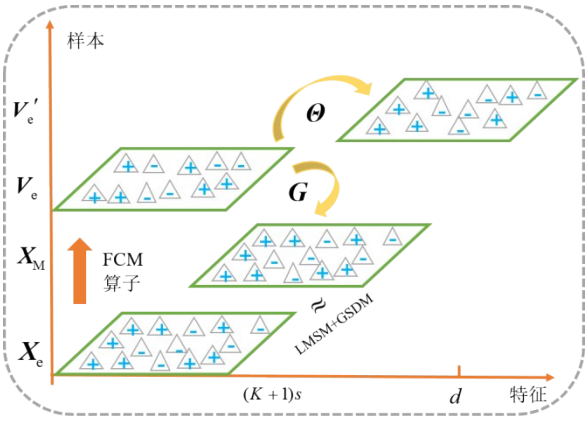


图2　分级结构一致性机制 (HSCM)

算法2　分级结构一致性机制(HSCM)

输入: 层间深度包络样本 $\boldsymbol{X}_e, \boldsymbol{V}_e$

输出: 生成的深度包络样本 $\boldsymbol{V}'_e$

步骤:

1. 初始化: $\boldsymbol{J}=\boldsymbol{G}=\boldsymbol{0}$;

2. 计算 $\boldsymbol{\Psi}=\varphi(\boldsymbol{X}_r)^{\mathrm{T}}\varphi(\boldsymbol{X}_r), \boldsymbol{\Psi}_v=\varphi(\boldsymbol{X}_r)^{\mathrm{T}}\varphi(\boldsymbol{V}_e), \boldsymbol{\Psi}_e=\varphi(\boldsymbol{X}_r)^{\mathrm{T}}\varphi(\boldsymbol{X}_e)$, $\boldsymbol{X}_r=[\boldsymbol{V}_e,\boldsymbol{X}_e]$;

3. 优化 $\boldsymbol{\Theta},\boldsymbol{J},\boldsymbol{G}$ 并计算 $\boldsymbol{V}_e'=\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Psi}_v$ 当 $\boldsymbol{\Theta}$ 最优时;

4. 返回 $\boldsymbol{V}'_e$

3.3　整体算法

本文基于特征权重进行子集划分[26]. 假设训练集中多数类样本 $\boldsymbol{X}_{maj}(maj=1,2,\cdots,n_1)\in\mathbf{R}^{n_1\times s}$ 的数量为 n_1,少数类样本 $\boldsymbol{X}_{min}\in\mathbf{R}^{n_2\times s}$ 的数量为 n_2,则多数类中每个样本的特征加权和可以按下式计算:

$$y=\sum_{f=1}^{s}w_{majf}\boldsymbol{x}_{majf},\quad w_{majf}=\boldsymbol{x}_{majf}\Big/\sum_{f=1}^{s}\boldsymbol{x}_{majf} \tag{15}$$

其中 $\boldsymbol{x}_{majf}$ 为样本 $\boldsymbol{x}_{maj}$ 的第 f 个特征的值,w_{majf} 为第 f 个特征的权重. 每个样本由等式(15)给出一个索引值,并且多数类样本按照相应的 y 值进行升序排列,然后将排序后的多数类样本划分为 Q 个子集,每个子集中的多数类样本的数目为 n_2. 具体地说,将第1个到第 n_2 个的有序数据作为第1个子集,第 n_2+1 到第 $2n_2$ 个的有序数据作为第2个子集,以此类推得到 Q 个子集. 将每个子集中的多数类样本与原始少数类样本进行融合,得到平衡的训练集. 我们可以找到子集的数量和不平衡度(Imbalance Ratio, IR)之间的关系,如下式:

$$Q=\lfloor n_1/n_2\rfloor\leqslant \mathrm{IR} \tag{16}$$

其中 $\lfloor\ \rfloor$ 表示向下取整. 因此不同不平衡数据集划分子集的个数与不平衡度IR相关.

接着,通过上述子集划分与融合(Division & Fusion, D&F)方法得到 Q 个平衡训练集后,将平衡后的训练集 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_Q$ 分别输入到各自的 L 层DH网络进行训练,得到深度包络训练集 $\boldsymbol{V'}_{1e}{}^{L},\boldsymbol{V'}_{2e}{}^{L},\cdots,\boldsymbol{V'}_{Qe}{}^{L}$. 然后,将测试集通过训练后的DH网络得到深度包络测试集 $\boldsymbol{V'}_{t1}{}^{L},\boldsymbol{V'}_{t2}{}^{L},\cdots,\boldsymbol{V'}_{tQ}{}^{L}$. 最后,深度包络训练集和测试集分别进行分类模型训练和预测,并通过投票机制确定测试样本的最终预测结果. 具体流程图如图3所示.

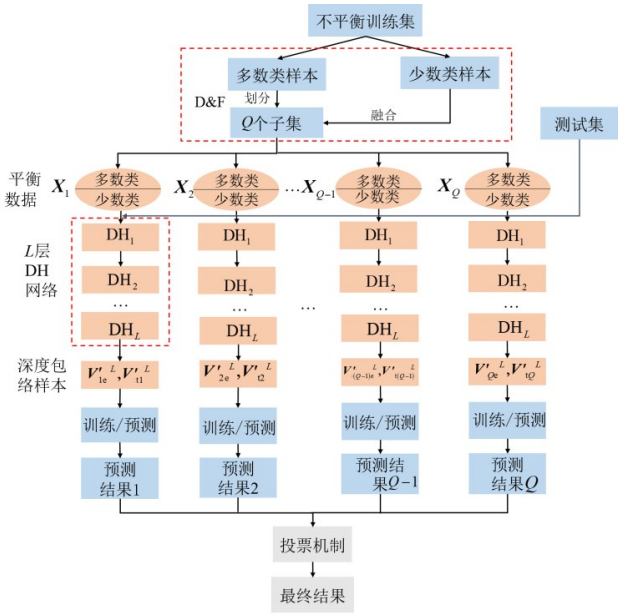


图3　本文IEDH算法

4　实验结果与分析

4.1　实验条件

实验中使用了18个公共的不平衡数据集. 为了验证本文方法的有效性,本文算法与多组相关算法进行了性能对比. 这包括:(1)将本文算法与MlFCM和普通bagging算法进行了比较. (2)由于本文算法是基于Bagging集成,并且Bagging和Boosting属于同源集成算法,即基分类器相同的,所以将本文算法与7个经典不平衡集成算法进行了比较,包括:RUSBoost(RBO)[14]、SMOTEBoost(SBO)[9]、Underbagging(UBAG)[12]、SMOTEBagging(SBAG)[10]、BalancedBagging(BBAG)[14]、EasyEnsemble[15]、BalanceCascade[15]. 这些方法都是独特的代表性不平衡集成方法,包含了bagging、boosting和两者混合并且都是数据级集成方法. (3)将本文算法与四种最先进的不平衡集成算法进行了比较,包括:CBIS(Cluster-based Instance Selection)[16]、SPE(Self-paced Ensemble)[17]、HOEC(Hybrid Optimal Ensemble Classifier)[18]、HD-ensemble(Hybrid Data-level Ensemble)[19]. 这些不平衡集成算法都是数据级集成算法,与本文算法一致,并且都是最新的发表在相关权威期刊或顶级会议上. 另外本文还分析了本文算法中基分类器的多样性以及性能(本

文完整实验结果与分析参见链接：https://pan.baidu.com/s/15lZ9GztB95ySrNwEmtrCfA，提取码：1111）. 实验中，决策树为集成方法中较流行的基分类器[16~19]，因此，本文使用决策树 C4.5 作为基分类器，采用 5 倍交叉验证（5 fold Cross Validation，5-CV）进行比较. 为了消除随机性的影响，实验对每个数据集重复 5-CV 程序 10 次.

4.1.1 数据

本文使用的数据集来源于 KEEL（Knowledge Extraction based on Evolutionary Learning）公共数据库，表 1 总结了这些数据集的相关信息，包括特征数、样本数、类别之间的样本数量以及不平衡度.

表 1 实验数据

数据集	特征数	样本数	少数类样本数	多数类样本数	不平衡度	数据集	特征数	样本数	少数类样本数	多数类样本数	不平衡度
Haberman	3	306	81	225	2.78	Glass2	9	214	17	197	11.59
Vehicle1	18	846	217	629	2.9	Shuttle-c0-vs-c4	9	1 829	123	1 706	13.87
Ecoli1	7	336	77	259	3.36	Yeast-1-vs-7	8	459	30	429	14.3
Ecoli2	7	336	52	284	5.46	Ecoli4	7	336	20	316	15.8
Ecoli3	7	336	35	306	8.6	Yeast-1-4-5-8-vs-7	8	693	30	663	22.1
Page-blocks0	10	5 472	559	4 913	8.79	Yeast4	8	1 484	51	1 433	28.10
Yeast-2-vs-4	8	514	51	463	9.08	Yeast-1-2-8-9	8	947	30	917	30.57
Vowel0	10	988	90	898	9.98	Yeast5	8	1 484	44	1440	32.73
Glass016vs2	9	192	17	175	10.29	Yeast6	8	1 484	35	1 449	41.4

4.1.2 评价指标和非参数统计检验

为了评估方法的性能，本文使用 AUC、F-M（F-measure）、G-M（Geometric Mean）和 Matthews 相关系数 Mcc（Matthews correlation coefficient）作为实验结果的评价指标[16~19]，并计算每种方法在数据集上相应的平均结果. 另外，本文使用了 Kappa（κ）值来评估集成方法中基分类器的多样性[27]，从而表明样本包络网路构建的各样本空间具有较好的多样性，有利于提升集成学习的性能.

为了检验算法间是否存在显着差异，本文采用了非参数统计检验[28,29]. 首先，假设竞争算法之间表现相似，没有显着差异（零假设）. 该检验方法主要目的是确定是否会拒绝无效假设. 拒绝这一假设意味着算法之间存在显着差异.

4.1.3 参数设置

本文算法主要涉及三个参数：（1）K，用于确定 INC 的最近邻数，（2）L，用于确定 DH 网络的层数，（3）Q，如式（16）定义，用于确定基分类器的个数. 在本研究中，令 $K=3, L=3, Q=\lfloor \mathrm{IR} \rfloor$，每层聚类数目=聚类前样本数目 −1，并且聚类涉及的迭代次数 $w=50$，阈值 $\varepsilon=1\times10^{-5}$ 以及模糊系数 $m=2$. $\lambda=\lambda_1=1, \delta=0.5$，核函数使用高斯核函数. 实验中的所有结果都是在该设置下获得的. 对于上述 7 个经典不平衡集成算法，基分类器的个数为 $\lfloor \mathrm{IR} \rfloor$，其他参数为默认值. 其中对于 SMOTEBoost 和 SMOTEBagging，近邻样本个数为 3.

4.2 消融实验-DH 网络的有效性

深度样本包络生成机制是本文算法的重要创新点. 本节采用消融法来验证 DH 网络的有效性，即将本文算法与 MlFCM 和不进行任何预处理的 bagging（Bgging+None）集成算法进行比较. MlFCM 是指原数据集进行多层 FCM 聚类. 以 Ecoli2、Ecoli3、Yeast-1-4-5-8-vs-7、Yeast5 为例，它们分别代表低、高不平衡度数据集. 对比结果见表 2.

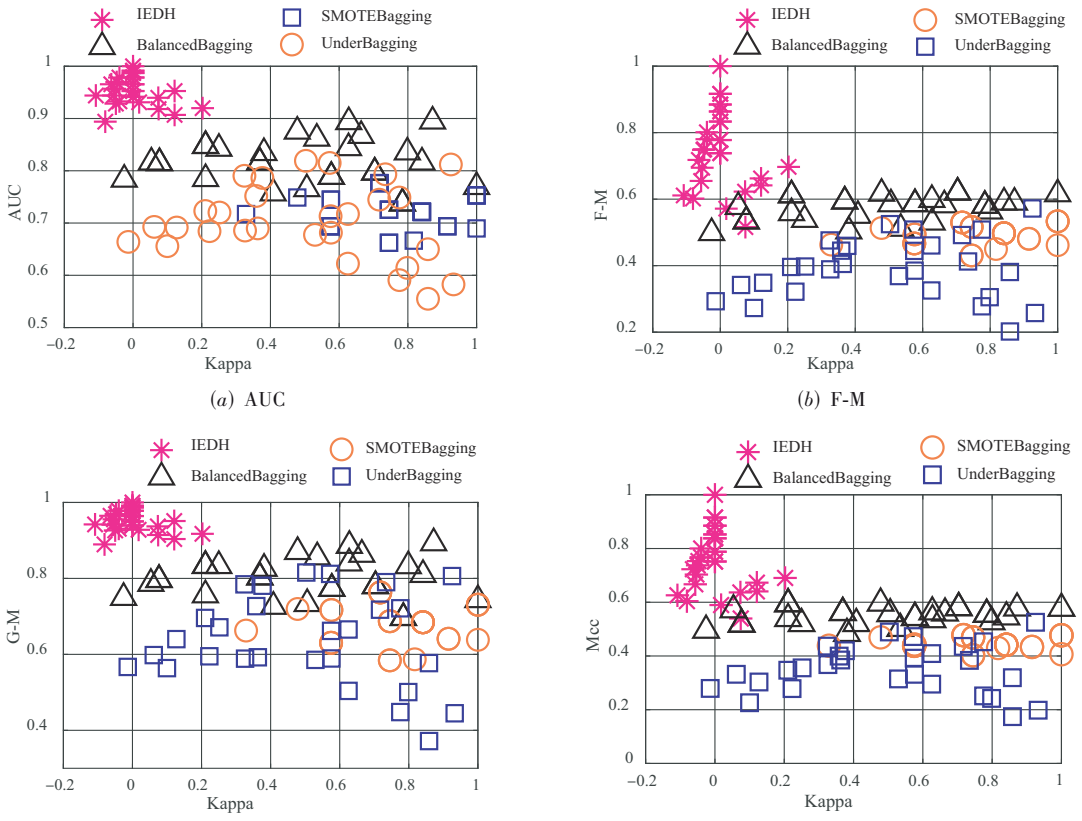
表 2 消融法

数据集	评价指标	Bagging+None	MlFCM	IEDH
Ecoli2	AUC	0.713 1	0.743 3	**0.936 2**
	F-M	0.398 4	0.424 2	**0.827 9**
	G-M	0.661 6	0.708 6	**0.927 6**
	Mcc	0.319 4	0.355 3	**0.820 1**
Ecoli3	AUC	0.862 8	0.781 5	**0.957 0**
	F-M	0.502 2	0.356 9	**0.733 7**
	G-M	0.857 9	0.756 9	**0.955 0**
	Mcc	0.493 7	0.346 8	**0.743 0**
Yeast-1-4-5-8-vs-7	AUC	0.583 8	0.518 5	**0.795 9**
	F-M	0.102 0	0.085 8	**0.180 4**
	G-M	0.551 4	0.438 8	**0.720 5**
	Mcc	0.070 2	0.017 0	**0.241 5**
Yeast5	AUC	0.949 0	0.863 9	**0.980 2**
	F-M	0.381 7	0.184 9	**0.943 7**
	G-M	0.947 5	0.852 9	**0.975 1**
	Mcc	0.460 2	0.272 2	**0.930 0**

从表 2 可以看出，与 MlFCM 和 Bgging+None 方法相比，本文算法在四个评价指标上的性能均最优，且相比其它两种算法具有显着的改进. 这表明 DH 网络能够

生成高质量有效样本. 大多数情况下,MIFCM 算法的性能最差. 可能的原因是,现有 MIFCM 算法主要是用于无监督聚类,聚类数需要实现设定. 当聚类数设定不当的话,则聚类性能将差强人意.

受 Kappa-AUC 图[27]的启发,本文通过 Kappa-AUC、F-M、G-M 和 Mcc 图分析不平衡集成方法中基分类器的多样性和性能,用于不平衡数据分类集成方法中的基分类器应具有较高的 AUC、F-M、G-M 和 Mcc 值和足够高的多样性. 以数据集 Ecoli3 为例,图 4 对比了本文算法 IEDH、BalancedBagging, SMOTEBagging 和 UnderBagging 下基分类器的多样性和相应的分类性能.



(a) AUC　(b) F-M　(c) G-M　(d) Mcc

图 4　基于 Kappa-AUC、F-M、G-M 和 Mcc 图的多样性和性能分析(Ecoli3 数据集)

较小的 Kappa 值表示集成系统中基分类器的高度多样性,较高的 AUC、F-M、G-M 和 Mcc 值表示基分类器出色的分类性能. 因此,位于图中左上角的点表示集成系统中基分类器更高的多样性和更好的性能. 图 4 可以看出,本文算法所获得的结果都位于图中的左上角,表明本文算法中基分类器具有高多样性和高分类性能.

4.3　算法对比

4.3.1　与经典的不平衡集成算法对比

为了验证本文算法的有效性,本节将提出算法与 7 种经典的不平衡集成算法对比. 表 3 列出了对比结果. 其中,每个数据集下的最佳结果以粗体显示.

表 3 中的实验结果表明,在四个评价指标上,IEDH 算法比其它不平衡集成方法都有了极大的提升. 对于 AUC 和 G-M 指标,本文方法在 17 个数据集上均取得了最好的性能;对于 F-M 和 Mcc 指标,本文方法分别在 14 和 16 个数据集上取得了最佳性能. 所以,本文算法在大多数不平衡数据集中都实现了最佳性能. 以数据集 Yeast5 为例,它的 IR 很高,样本数不到 1 000 个,本文算法的 AUC 值、F-M、G-M 和 Mcc 均最高. 另外,以 Ecoli3 为例,本文算法的 AUC 值、F-M、G-M 和 Mcc 值分别为 0.957 0,0.733 7,0.955 0,0.743 0,相较第 2 名的性能提高了 8.66%,8.83%,8.81%,12.79%. 为了更直观地展示提出算法的性能,图 5 记录了不同算法下的 PR(Precision-Recall)曲线,以数据集 Ecoli3 和 Yeast-1-4-5-8-vs-7 为例,Ecoli3, Yeast-1-4-5-8-vs-7 分别代表低、高不平衡度数据集.

从图 5(a)可以看出,7 种经典不平衡集成算法的

表3 不平衡集成方法的对比结果

数据集	评价指标	RBO	SBO	UBAG	SBAG	BBAG	EasyEnsemble	BalanceCascade	IEDH
Haberman	AUC	0.532 9	0.574 1	0.594 7	0.520 0	0.588 9	0.560 6	0.519 5	**0.669 4**
	F-M	0.305 0	0.406 2	0.430 1	0.304 0	0.421 6	0.401 0	0.340 4	**0.509 1**
	G-M	0.475 5	0.568 1	0.588 2	0.467 5	0.578 8	0.555 2	0.506 5	**0.662 6**
	Mcc	0.062 4	0.136 7	0.173 1	0.049 7	0.169 9	0.110 4	0.036 2	**0.314 1**
Vehicle1	AUC	0.665 1	0.702 9	0.780 3	0.726 2	0.751 0	**0.791 2**	0.761 2	0.733 0
	F-M	0.495 5	0.555 6	0.646 7	0.590 1	0.615 3	**0.658 7**	0.625 5	0.573 2
	G-M	0.632 2	0.690 6	0.779 1	0.715 3	0.747 2	**0.790 3**	0.758 8	0.721 0
	Mcc	0.340 3	0.396 0	0.512 0	0.445 4	0.470 0	**0.529 3**	0.482 9	0.406 8
Ecoli1	AUC	0.843 1	0.861 5	0.877 0	0.880 0	0.871 7	0.883 9	0.881 7	**0.924 7**
	F-M	0.758 0	0.773 3	0.771 4	0.803 8	0.774 4	0.780 7	0.779 6	**0.804 2**
	G-M	0.836 8	0.857 1	0.875 2	0.877 0	0.869 3	0.881 8	0.880 3	**0.920 9**
	Mcc	0.690 4	0.708 4	0.704 5	0.747 8	0.708 2	0.717 9	0.714 8	**0.804 8**
Ecoli2	AUC	0.901 4	0.840 2	0.892 9	0.877 5	0.890 1	0.870 2	0.871 6	**0.936 2**
	F-M	**0.834 3**	0.702 1	0.771 9	0.825 5	0.772 8	0.743 2	0.772 2	0.827 9
	G-M	0.895 5	0.832 5	0.888 5	0.865 7	0.886 2	0.865 0	0.860 7	**0.927 6**
	Mcc	0.813 8	0.647 9	0.733 8	0.813 7	0.733 3	0.700 0	0.742 6	**0.820 1**
Ecoli3	AUC	0.787 0	0.770 0	0.869 3	0.767 2	0.859 4	0.870 4	0.856 7	**0.957 0**
	F-M	0.591 4	0.555 4	0.620 6	0.587 4	0.622 2	0.624 6	0.645 4	**0.733 7**
	G-M	0.759 1	0.740 5	0.864 8	0.733 4	0.853 5	0.866 9	0.849 7	**0.955 0**
	Mcc	0.558 8	0.512 1	0.594 1	0.550 2	0.592 8	0.598 0	0.615 1	**0.743 0**
Page-blocks0	AUC	0.874 8	0.937 9	0.956 7	0.938 9	0.951 5	0.957 0	0.953 2	**0.992 1**
	F-M	0.781 0	0.841 3	0.812 0	0.862 9	0.814 9	0.812 5	0.857 3	**0.934 2**
	G-M	0.867 8	0.937 1	0.956 6	0.937 9	0.951 4	0.956 9	0.952 9	**0.992 1**
	Mcc	0.758 4	0.825 1	0.799 3	0.847 7	0.800 7	0.799 8	0.843 4	**0.929 2**

PR曲线被提出算法IEDH的曲线“包住”,说明提出算法IEDH的性能优于7种经典不平衡集成算法. 图5(*b*)中,虽然提出算法IEDH的PR曲线与其他算法的曲线发生了交叉,但是提出算法的“平衡点”要高于其他算法,也说明提出算法IEDH性能优于其他算法. 另外,本节还计算并分析了每种方法在这些数据集上的AUC、F-M、G-M和Mcc的平均排名. 表现最好的方法排名第1(记为1),而表现最差的方法排名第8(记为8). 排名分数越小,表明算法性能更好. 图6给出了每种方法在18个数据集上的平均排名,本文方法在AUC、F-M、G-M和Mcc指标上的平均排名均最低,分别为1.333、2.222、1.444和1.944,表明其具有最佳性能.

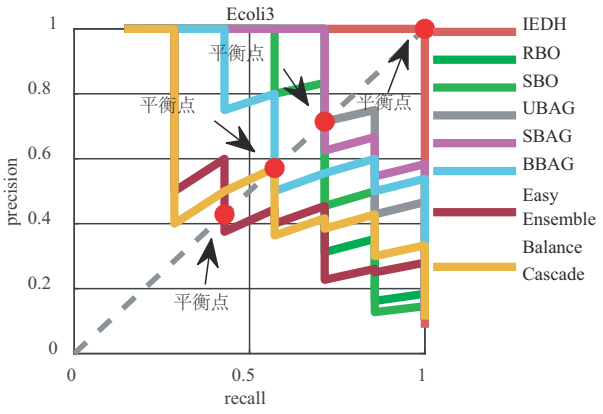
将本文方法视为对照方法,进行Holm检验,以比较本文方法与其他不平衡集成方法在平均排名上的统计显着性. 结果见表4. 从表4可以看出,所有的等价性假设都被拒绝了,这表明本文方法比其他方法的优越性显着有效. 总的来说,从表3、表4的结果来看,本文算法比其他不平衡集成学习方法显着更优.

表4 Holm's检验结果

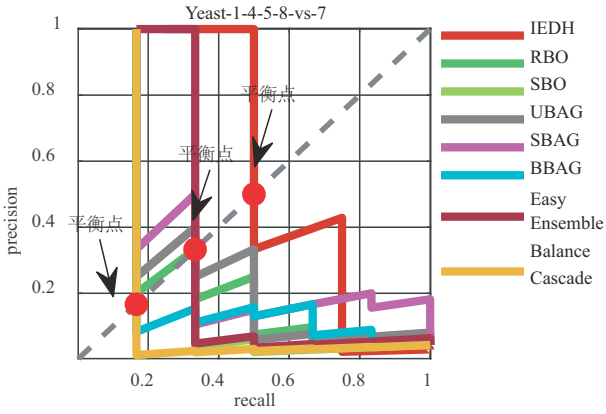
$\alpha_{0.05}$	AUC		F-M		G-M		Mcc	
	方法	P值	方法	P值	方法	P值	方法	P值
0.007 3	RUSBoost	1.35×10^{-17}	RUSBoost	2.73×10^{-5}	SMOTEBagging	2.58×10^{-18}	RUSBoost	4.67×10^{-7}
0.008 5	SMOTEBagging	2.48×10^{-17}	EasyEnsemble	1.23×10^{-4}	RUSBoost	1.66×10^{-17}	BalancedBagging	1.03×10^{-5}
0.010 2	SMOTEBoost	5.55×10^{-15}	UnderBagging	1.64×10^{-4}	SMOTEBoost	1.38×10^{-14}	EasyEnsemble	1.96×10^{-5}
0.012 7	BalanceCascade	1.17×10^{-8}	BalanceCascade	2.18×10^{-4}	BalanceCascade	5.42×10^{-8}	UnderBagging	6.82×10^{-5}
0.016 9	BalancedBagging	3.77×10^{-7}	BalancedBagging	2.88×10^{-4}	BalancedBagging	2.42×10^{-7}	BalanceCascade	1.24×10^{-4}
0.025 3	EasyEnsemble	7.50×10^{-5}	SMOTEBoost	1.80×10^{-3}	UnderBagging	2.78×10^{-4}	SMOTEBoost	2.97×10^{-4}
0.05	UnderBagging	2.42×10^{-4}	SMOTEBagging	1.76×10^{-2}	EasyEnsemble	1.21×10^{-3}	SMOTEBagging	8.91×10^{-3}

表5　CBIS, HD-ensemble, HOEC, SPE 和 IEDH 之间的比较结果

数据集	Haberman				Vehicle1				Ecoli1			
评价指标	AUC	F-M	G-M	Mcc	AUC	F-M	G-M	Mcc	AUC	F-M	G-M	Mcc
CBIS	0.648 0	—	—	—	0.825 0	—	—	—	0.957 0	—	—	—
HD-Ensemble	—	—	—	—	—	—	—	—	—	—	—	—
HOEC	0.624 2	—	—	—	0.759 6	—	—	—	0.881 6	—	—	—
SPE	0.600 2	0.438 2	0.593 1	0.179 2	0.774 4	0.639 2	0.773 1	0.501 1	0.863 3	0.784 6	0.858 8	0.724 0
IEDH	0.669 4	0.509 1	0.662 6	0.314 1	0.733 0	0.573 2	0.721 0	0.406 8	0.924 7	0.804 2	0.920 9	0.804 8
数据集	Ecoli2				Ecoli3				Page-blocks0			
评价指标	AUC	F-M	G-M	Mcc	AUC	F-M	G-M	Mcc	AUC	F-M	G-M	Mcc
CBIS	0.934 0	—	—	—	0.933 0	—	—	—	0.987 0	—	—	—
HD-Ensemble	—	—	—	—	—	—	—	—	—	—	—	—
HOEC	0.912 8	—	—	—	0.873 4	—	—	—	0.929 4	—	—	—
SPE	0.899 2	0.806 7	0.893 8	0.778 7	0.836 8	0.613 7	0.825 9	0.579 1	0.932 4	0.862 4	0.930 9	0.847 2
IEDH	0.936 2	0.827 9	0.927 6	0.820 1	0.957 0	0.733 7	0.955 0	0.743 0	0.992 1	0.934 2	0.992 1	0.929 2



(*a*)　基于 Ecoli3 的 PR 曲线



(*b*)　基于 Yeast-1-4-5-8-vs-7 的 PR 曲线

图5　不同算法下的 PR 曲线

4.3.2　与最新的不平衡集成算法对比

表5对比了前述的四种相关新算法. 结果表明,本文算法在 AUC、F-M、G-M 和 Mcc 方面提供了更好的性能,表明本文算法优于四种方法. 以 Glass2 为例,本文提出的算法的 AUC 值为 0.876 9,相较其他四种算法的性能

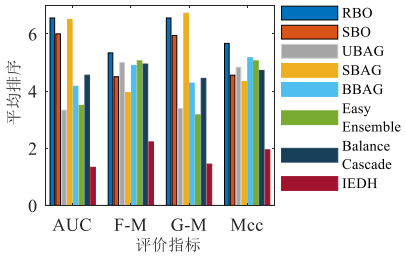


图6　算法性能的平均排名

提高了 11.09%,1.04%,9.73% 和 15.17%. 本节采用 Wilcoxon 配对符号秩检验,结果如表6所示. 表6中的统计检验结果表明算法间存在显着性差异. 进一步说明本文

表6　Wilcoxon pairwise 检测结果

对比组	评价指标	*R*+	*R*−	P值	假设(0.05)
IEDH vs CBIS	AUC	123	30	0.027 7	拒绝
	F-M	—	—	—	—
	G-M	—	—	—	—
	Mcc	—	—	—	—
IEDH vs HOEC	AUC	53	2	0.005 9	拒绝
	F-M	—	—	—	—
	G-M	—	—	—	—
	Mcc	—	—	—	—
IEDH vs HD-Ensemble	AUC	51	4	0.013 7	拒绝
	F-M	—	—	—	—
	G-M	55	0	0.002 0	拒绝
	Mcc	—	—	—	—
IEDH vs SPE	AUC	162	9	0.000 9	拒绝
	F-M	131	40	0.047 5	拒绝
	G-M	160	11	0.001 2	拒绝
	Mcc	144	27	0.010 8	拒绝

算法明显优于四种最先进的不平衡集成方法.

在表6中,$R+$是第1种算法优于第2种算法的数据集的排名之和,$R-$是第2种算法优于第1种算法的排名之和. 可以发现$R+$总是大于$R-$,而且所有的P值都小于0.05.P值<0.05意味着四组对比的等值假设均被拒绝. 因此,IEDH算法明显优于四种最新的不平衡集成方法.

5 结论

不平衡学习很重要,但现有的方法大都是基于原始样本,忽略了样本之间的结构信息. 为了解决这个问题,本文提出了一种基于包络学习和分级结构一致性机制的不平衡集成学习算法,通过构造分级样本空间进行集成学习,来提高不平衡学习效能. 实验部分采用了十多个不平衡数据集和相关算法进行了验证对比. 实验结果表明,本文提出的不平衡集成方法显着优于其他经典的和最新的不平衡学习方法. 虽然本文算法是有效的,但仍有一些工作要做. 未来可以考虑更多的聚类算法、更多的数据集和更多的域适应方法进行本文算法的进一步验证和应用推广.

参考文献

[1] 李艳霞, 柴毅, 胡友强, 等. 不平衡数据分类方法综述[J]. 控制与决策, 2019, 34(4): 673-688.
LI Y X, CHAI Y, HU Y Q, et al. Review of imbalanced data classification methods[J]. Control and Decision, 2019, 34(4): 673-688. (in Chinese)

[2] 翟云, 王树鹏, 马楠, 等. 基于单边选择链和样本分布密度融合机制的非平衡数据挖掘方法[J]. 电子学报, 2014, 42(7): 1311-1319.
ZHAI Y, WANG S P, MA N, et al. A data mining method for imbalanced datasets based on one-sided link and distribution density of instances[J]. Acta Electronica Sinica, 2014, 42(7): 1311-1319. (in Chinese)

[3] 欧阳震诤, 罗建书, 胡东敏, 等. 一种不平衡数据流集成分类模型[J]. 电子学报, 2010, 38(1): 184-189.
OUYANG Z Z, LUO J S, HU D M, et al. An ensemble classifier framework for mining imbalanced data streams [J]. Acta Electronica Sinica, 2010, 38(1): 184-189. (in Chinese)

[4] QI C R, SU H, NIEBNER M, et al. Volumetric and multi-view CNNs for object classification on 3D data[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 5648-5656.

[5] KOTSIANTIS S B, KANELLOPOULOS D, PINTELAS P E. Handling imbalanced datasets: A review[J]. GESTS International Transactions on Computer Science & Engineering, 2005, 30(1): 25-36.

[6] GALAR M, FERNANDEZ A, BARRENECHEA E, et al. A review on ensembles for the class imbalance problem: Bagging- , boosting- , and hybrid-based approaches[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2012, 42(4): 463-484.

[7] 于重重, 田蕊, 谭励, 等. 非平衡样本分类的集成迁移学习算法[J]. 电子学报, 2012, 40(7): 1358-1363.
YU C C, TIAN R, TAN L, et al. Integrated transfer learning algorithmic for unbalanced samples classification[J]. Acta Electronica Sinica, 2012, 40(7): 1358-1363. (in Chinese)

[8] SUN Y, KAMEL M S, WONG A K C, et al. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(12): 3358-3378.

[9] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting[C]//Knowledge Discovery in Databases: PKDD 2003. Berlin: Springer, 2003: 107-119.

[10] WANG S, YAO X. Diversity analysis on imbalanced data sets by using ensemble models[C]//2009 IEEE Symposium on Computational Intelligence and Data Mining. Piscataway: IEEE, 2009: 324-331.

[11] Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks[J]. Expert Systems with Applications, 2018, 91: 464-471.

[12] RAGHUWANSHI B S, SHUKLA S. UnderBagging based reduced kernelized weighted extreme learning machine for class imbalance learning[J]. Engineering Applications of Artificial Intelligence, 2018, 74: 252-270.

[13] 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.

[14] HIDO S, KASHIMA H, TAKAHASHI Y. Roughly balanced bagging for imbalanced data[J]. Statistical Analysis and Data Mining: the ASA Data Science Journal, 2009, 2 (5/6): 412-426.

[15] 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.

[16] TSAI C F, LIN W C, HU Y H, et al. Under-sampling

class imbalanced datasets by combining clustering analysis and instance selection[J]. Information Sciences, 2019, 477: 47-54.

[17] LIU Z N, CAO W, GAO Z F, et al. Self-paced ensemble for highly imbalanced massive data classification[C]// 2020 IEEE 36th International Conference on Data Engineering (ICDE). Piscataway: IEEE, 2020: 841-852.

[18] YANG K X, YU Z W, WEN X, et al. Hybrid classifier ensemble for imbalanced data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(4): 1387-1400.

[19] CHEN Z, DUAN J, KANG L, et al. A hybrid data-level ensemble to enable learning from highly imbalanced dataset[J]. Information Sciences, 2021, 554: 157-176.

[20] PEDRYCZ W, AL-HMOUZ R, BALAMASH A S, et al. Hierarchical granular clustering: An emergence of information granules of higher type and higher order[J]. IEEE Transactions on Fuzzy Systems, 2015, 23(6): 2270-2283.

[21] BEZDEK J C, EHRLICH R, FULL W. FCM: The fuzzy c-means clustering algorithm[J]. Computers & Geosciences, 1984, 10(2/3): 191-203.

[22] KANG Z, PENG C, CHENG Q. Clustering with adaptive manifold structure learning[C]//2017 IEEE 33rd International Conference on Data Engineering (ICDE). Piscataway: IEEE, 2017: 79-82.

[23] ZHANG L, WANG S S, HUANG G B, et al. Manifold criterion guided transfer learning via intermediate domain generation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(12): 3759-3773.

[24] LONG M S, CAO Y, WANG J M, et al. Learning transferable features with deep adaptation networks[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. New York: ACM, 2015: 97-105.

[25] KANAMORI T, HIDO S, SUGIYAMA M. Efficient direct density ratio estimation for Non-stationarity adaptation and outlier detection[C]//Proceedings of the 21st International Conference on Neural Information Processing Systems. New York: ACM, 2008: 809-816.

[26] SHEN Y H, PEDRYCZ W, CHEN Y, et al. Hyperplane division in fuzzy C-means: Clustering big data[J]. IEEE Transactions on Fuzzy Systems, 2020, 28(11): 3032-3046.

[27] XU Y H, YU Z W, CHEN C L P, et al. Adaptive subspace optimization ensemble method for high-dimensional imbalanced data classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(5): 2284-2297.

[28] GARCÍA S, FERNÁNDEZ A, LUENGO J, et al. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power[J]. Information Sciences, 2010, 180(10): 2044-2064.

[29] HOLM S. A simple sequentially rejective multiple test procedure[J]. Scandinavian Journal of Statistics, 1979, 6 (2): 65-70.

作者简介

李　帆　男，1993年出生于湖北省．博士研究生．主要研究领域为非平衡数据处理、机器学习．

E-mail: 979940181@qq.com

张小恒　男，1980年生，四川达州人．博士研究生，副教授．主要研究领域为医学信号处理、机器学习．

E-mail: 7818320@qq.com

李勇明　男，1976年生于四川．博士，教授，博士生导师．主要研究领域为医学信号处理、机器学习．中国电子学会会员编号：E190020470M.

E-mail: yongmingli@cqu.edu.cn

王　品　女，1979年出生于江苏省．博士，副教授，硕士生导师．主要研究领域为图像处理与识别．

E-mail: wangpin@cqu.edu.cn