控制与决策
Control and Decision

基于数据分布特性的代价敏感宽度学习系统

徐鹏飞， 王敏， 刘金平， 唐朝晖， 马天雨

引用本文:
徐鹏飞, 王敏, 刘金平, 等. 基于数据分布特性的代价敏感宽度学习系统[J]. 控制与决策, 2021, 36(7): 1686−1692.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.1484

基于数据分布特性的代价敏感宽度学习系统

徐鹏飞[1], 王 敏[1], 刘金平[1†], 唐朝晖[2], 马天雨[1]

(1. 湖南师范大学 信息科学与工程学院，长沙 410081；2. 中南大学 自动化学院，长沙 410083)

摘 要: 宽度学习系统(broad learning system, BLS)作为深度神经网络的替代框架,具有快速自适应模型结构选择和在线增量学习能力,被认为是知识发现和数据工程领域中一种极具前途的技术. 传统的BLS主要应用于数据分布均衡且误分类代价相同的模式分类任务,但大多数实际应用的数据是非均衡分布的,如网络入侵监测、医疗诊断、信用卡欺诈检测等. 基于此,提出一种基于数据分布特性的代价敏感BLS (data distribution-based cost-sensitive-BLS, DDbCs-BLS),解决数据分布不均、误分代价不同的模式分类任务. DDbCs-BLS在充分考虑数据统计分布特性的基础上寻找代价敏感型BLS分类器的最佳分类边界,保证少数类样本信息不被丢失,从而提高BLS在各类数据集上的模式分类性能. 在多种公共数据集(包括均衡和不均衡数据集)上进行大量的验证性和对比性实验,结果表明DDbCs-BLS能有效确定分类边界线的最佳位置,无论是在均衡数据集还是在不均衡数据集上均能获得更好的分类性能.

关键词: 非均衡数据；代价敏感；宽度学习系统；自适应模型结构选择；增量学习

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.1484 **开放科学(资源服务)标识码(OSID):**

引用格式: 徐鹏飞,王敏,刘金平,等. 基于数据分布特性的代价敏感宽度学习系统[J]. 控制与决策, 2021, 36(7): 1686-1692.

Data distribution-based cost-sensitive broad learning system

XU Peng-fei[1], *WANG Min*[1], *LIU Jin-ping*[1†], *TANG Zhao-hui*[2], *MA Tian-yu*[1]

(1. College of Information Science and Engineering，Hunan Normal University，Changsha 410081，China；2. College of Automation，Central South University，Changsha 410083，China)

Abstract: Broad learning system (BLS) provides a flexible modeling framework, which is a potential substitute of deep neural network models. Due to its fast adaptive ability of automatic model structure selection and online incremental learning strategies, BLS is referred to as a promising technology in the field of knowledge discovery and data engineering. However, traditional BLS model are mainly aimed at pattern classification tasks with approximately even-distributed data and equal misclassification cost. In real applications, most of pattern recognition tasks are unevenly-distributed, such as credit card fraud detection, network intrusion detection, medical diagnosis, etc. In this paper, a data distribution-based cost-sensitive-BLS (DDbCs-BLS) is proposed for solving the problem of pattern classification tasks with imbalance data and varying misclassification costs on different classes. The DDbCs-BLS can achieve the best classification boundary by adopting the cost sensitive BLS learners, and ensure the lossless of the information of sparse classes, so as to ensure the classification performance of the BLS classifier in various data sets. The DDbCs-BLS is validated on multiple public data sets (including balanced and imbalanced data sets). Extensive validation and comparative results show that the DDbCs-BLS can effectively determine the best location of the classification boundary line, consequently, it can achieve better classification performance on both balanced and imbalanced data sets.

Keywords: imbalanced dataset；cost sensitive；broad learning system；adaptive model structure selection；incremental learning

0 引 言

随着多层神经网络结构及其学习方法的不断发展,深度神经网络在诸如计算机视觉[1-2]、工业故障检测[3-4]、音视频识别[5-7]等多个领域得到广泛应用. 例

收稿日期: 2019-10-23；**修回日期:** 2020-01-16.

基金项目: 国家自然科学基金项目(61971188)；湖南省自然科学基金项目(2018JJ3349)；湖南省教育厅优秀青年项目(19B364)；湖南省知识产权战略推进专项项目(2019F012K)；湖南省研究生科研创新项目(CX20190415).

责任编委: 阳春华.

†通讯作者. E-mail: ljp@hunnu.edu.cn.

如,生成对抗网络(GAN)可被用于图像生成的超分辨率任务[8],利用深度置信网络(DBN)进行深度特征提取[9],基于递归神经网络(RNN)对视频图像目标进行监控[10]等. 尽管深层结构如此强大,但仍面临着一些问题: 复杂结构和大量的超参数,使得多数网络受到训练过程耗时的影响. 为解决这一问题, Chen等[11]提出了宽度学习系统(broad learning system, BLS),通过扩展神经网络的宽度而不是深度来解决高维数据问题,不仅克服了深度学习训练过程耗时的缺点,还可以快速增量构建网络模型.

Chen等[12]已经证明BLS具有通用的逼近能力,提出一种基于正则化稳健的BLS来学习不确定的数据模型[13]. Xu等[14]基于BLS提出一种时间序列预测模型(recurrent BLS)用于多变量预测时间序列. BLS具有通用的逼近能力、良好的泛化,且拥有时间记忆和灵活的重塑过程,被认为是一种极具前途的深度学习替代方法. 然而,传统的BLS在分类问题上默认数据样本是类别分布均衡或者错分类的代价相同,但实际应用中的大部分模式分类问题面临的都是一些非均衡数据,而且少数类往往占据更重要的地位,如网络入侵检测、医疗诊断、网络欺诈监测等[15-17]. 为此,本文提出一种基于数据分布特性的代价敏感宽度学习系统,旨在提供一种简单、高效的模式分类方法,在快速收敛、成本节约的基础上针对非均衡数据获得较高的准确率,主要贡献如下:

1) 针对数据非均衡分布问题,将通用的代价敏感学习框架引入BLS,提出代价敏感宽度学习系统(cost-sensitive-BLS, Cs-BLS),将传统的BLS扩展到不均衡数据集的模式分类任务.

2) 在Cs-BLS中考虑样本数据的分布特性,提出基于数据分布特性的代价敏感宽度学习系统(data distribution-based cost-sensitive BLS, DDbCs-BLS),根据不同类别数据的分布特性确定最佳分类边界,使少数类样本不被边缘化,有效应对各种分布特性数据集的模式分类任务.

1　BLS基本结构

宽度学习系统是以随机向量函数链接神经网络(random vector functional link neural network, RVFL)为载体,通过神经节点的增量实现网络横向扩展的一种单层神经网络学习系统,模型结构如图1所示.

BLS使用n组特征映射窗口和m组增强节点取代RVFL中输入与输出之间的直接连接,每组特征映射窗口包含N个神经节点,每组增强节点包含M个神经节点.

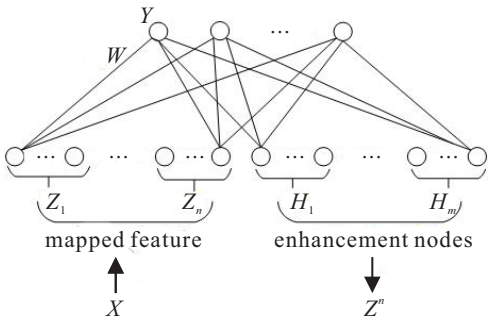


图1　BLS基本结构

首先,样本矩阵$\boldsymbol{X}=[x_1,x_2,\ldots,x_L]^{\mathrm{T}}_{L\times D}$经过如下公式的线性变换形成特征映射窗口$\boldsymbol{Z}_i$:

$$\boldsymbol{Z}_i=\phi(\boldsymbol{X}\boldsymbol{W}_{zi}+\boldsymbol{\beta}_{zi})_{L\times N},\ i=1,2,\ldots,n. \tag{1}$$

其中: L为样本总数, D为样本$x_l\in\boldsymbol{R}^{1\times D}$的维度;权重矩阵$\boldsymbol{W}_{zi}\in\boldsymbol{R}^{D\times N}$和偏置矩阵$\boldsymbol{\beta}_{zi}\in\boldsymbol{R}^{L\times N}$随机产生,各特征映射窗口的线性函数可以不同;所有特征映射窗口$\boldsymbol{Z}_i$拼接组成特征映射层$\boldsymbol{Z}^n$为

$$\boldsymbol{Z}^n=[\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_n]_{L\times N_n}. \tag{2}$$

然后,特征映射层$\boldsymbol{Z}^n$经过非线性变换形成增强节点,第j组增强节点的输出如下所示:

$$\boldsymbol{H}_j=\delta(\boldsymbol{Z}^n\boldsymbol{W}_{hj}+\boldsymbol{\beta}_{hj})_{L\times M},\ j=1,2,\ldots,m. \tag{3}$$

其中:权重矩阵$\boldsymbol{W}_{hj}\in\boldsymbol{R}^{N_n\times M}$和偏置矩阵$\boldsymbol{\beta}_{hj}\in\boldsymbol{R}^{L\times M}$随机产生,各增强节点的非线性函数可以不同;所有增强节点$\boldsymbol{H}_j$拼接组成增强层$\boldsymbol{H}^m$为

$$\boldsymbol{H}^m=[\boldsymbol{H}_1,\ldots,\boldsymbol{H}_m]_{L\times M_m}. \tag{4}$$

最后,将特征映射层$\boldsymbol{Z}^n$和增强层$\boldsymbol{H}^m$直接连接到BLS的输出端,不妨设$\boldsymbol{A}=[\boldsymbol{Z}^n|\boldsymbol{H}^m]$, 则BLS的最终估计输出为

$$\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{W}. \tag{5}$$

其中: $\boldsymbol{A}\in\boldsymbol{R}^{L\times(N_n+M_m)}$为样本矩阵$\boldsymbol{X}$变换后的输入矩阵, $\boldsymbol{Y}\in\boldsymbol{R}^{L\times c}$为样本矩阵$\boldsymbol{X}$对应的预测值, c为样本类别数, $\boldsymbol{W}\in\boldsymbol{R}^{(N_n+M_m)\times c}$为输出权重矩阵.

求解输出权重矩阵$\boldsymbol{W}$的目标优化函数如下所示:

$$\arg\min_{\boldsymbol{W}}:\|\boldsymbol{A}\boldsymbol{W}-\boldsymbol{Y}\|_2^2+\lambda\|\boldsymbol{W}\|_2^2, \tag{6}$$

其中λ为正则化参数.

2　所提出的方法

本节先简要推导Cs-BLS,并分析数据的分布特性;然后详细介绍所提出的DDbCs-BLS,并对方法进行简单的理论分析.

2.1　Cs-BLS

代价敏感学习利用代价矩阵的不同错误分类成本来解决不均衡数据集的模式分类问题[18].

典型的代价敏感矩阵如下所示:

$$C=\begin{bmatrix}0 & C_{12} & \cdots & C_{1L}\\ C_{21} & 0 & \cdots & C_{2L}\\ \vdots & \vdots & \ddots & \vdots\\ C_{L1} & C_{L2} & \cdots & 0\end{bmatrix}_{L\times L},$$

其中C_{ij}表示将样本x_i误判为样本x_j的代价. 通常,少数类误判为多数类的代价高于多数类误分为少数类的代价.

在式(6)的BLS优化目标函数中,Cs-BLS为误差函数引入一个代价敏感矩阵,如下所示:

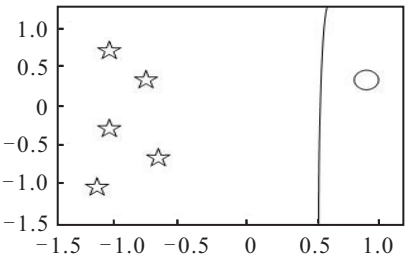
$$\arg\min_{\boldsymbol{W}}:\lambda\|\boldsymbol{W}\|_2^2+\boldsymbol{C}\|\boldsymbol{AW}-\boldsymbol{Y}\|_2^2, \tag{7}$$

使少数类获得一个较大权值,提高分类器对少数类的敏感度,解决非均衡样本的有效分类问题,其中代价敏感矩阵$\boldsymbol{C}$随机产生.

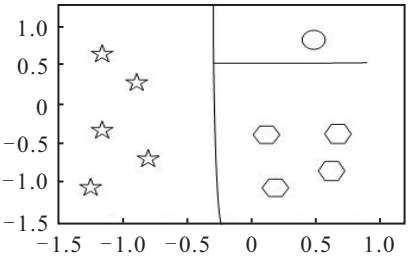
2.2 DDbCs-BLS

在使用Cs-BLS进行模式分类时,若样本数据分布不均衡,则多数类会凭借自身数量优势积累训练误差,忽略少数类积累训练误差,导致多数类边界距离最大化,即分类边界线偏向少数类.

例如,在图2所示的Cs-BLS分类示意图中,发现分类边界线偏向于少数类,这种少数类边界的模糊性会加大分类难度,降低少数类的泛化性能.



(a) 二分类



(b) 三分类

maj class min class

图2 Cs-BLS分类示意

显然,适当将分类边界向中间移动是一种较为理想的结果. 为此,DDbCs-BLS根据文献[19]中加权ELM的思想,为少数类样本分配一个较大的权值,以提高少数类积累训练误差的影响. 为多数类分配一个较小的权值,以降低多数类积累训练误差的影响,即再次为每个样本x_i赋予一个权值T_{ii},具体计算过程如下所示:

$$T_{ii}=\begin{cases}\dfrac{\varepsilon}{\mathrm{Class}(x_i)}, & \mathrm{Class}(x_i)>\mathrm{AVG(class)};\\ \dfrac{1}{\mathrm{Class}(x_i)}, & \text{otherwise.}\end{cases} \tag{8}$$

其中:$0<\varepsilon\leqslant 1$,$\mathrm{Class}(x_i)$表示与样本x_i同类别的样本总数,$\mathrm{AVG(class)}$表示样本类别的平均数.

从本质上,权值T_{ii}决定了用户寻求的重新均衡程度,以及边界向多数阶级推进的程度,但其值并不是越大越好. 例如,在图3(a)和(d)中,$\varepsilon=0.2$,出现与式(7)相似的情况,关注点放在多数类,少数类的泛化性能下降;在图3(c)和(f)中,$\varepsilon=1$,更多的关注误差最小化,边界是弯向少数类,如果数据非均衡度提高,边界线会更加偏向少数类,少数类性能提高的并不明显;在图3(b)和(e)中,$\varepsilon=0.6$,分类边界线偏向多数类,边界线偏于平滑,没有出现明显弯曲现象,具有更好的分类特性.

将所有样本的权值T_{ii}加入式(7)后,可得DDbCs-BLS目标优化函数

$$\arg\min_{\boldsymbol{W}}:\lambda\|\boldsymbol{W}\|_2^2+\boldsymbol{Q}\|\boldsymbol{AW}-\boldsymbol{Y}\|_2^2. \tag{9}$$

其中:$\boldsymbol{Q}=\boldsymbol{TC}$,$\boldsymbol{T}=\mathrm{diag}(T_{ii}),i=1,2,\ldots,L$.

求解式(9)的目标优化函数,可得DDbCs-BLS的最终输出权重矩阵$\boldsymbol{W}$为

$$\boldsymbol{W}=(\lambda\boldsymbol{I}+\boldsymbol{A}^{\mathrm{T}}\boldsymbol{QA})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{QY}. \tag{10}$$

2.3 DDbCs-BLS增量学习

假设现有样本为$\boldsymbol{X}_{L_0}$,依据式(10),现有样本的输出权重矩阵为

$$\boldsymbol{W}_0=(\lambda\boldsymbol{I}+\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{Q}_0\boldsymbol{A}_0)^{-1}\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{Q}_0\boldsymbol{Y}_0. \tag{11}$$

设工作方阵$\boldsymbol{P}_0=(\lambda\boldsymbol{I}+\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{Q}_0\boldsymbol{A}_0)^{-1}$,则有

$$\boldsymbol{P}_0^{-1}=\lambda\boldsymbol{I}+\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{Q}_0\boldsymbol{A}_0, \tag{12}$$

$$\boldsymbol{W}_0=\boldsymbol{P}_0\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{Q}_0\boldsymbol{Y}_0. \tag{13}$$

当一批新样本$\boldsymbol{X}_{L_1}$加入在线学习时,设$\boldsymbol{A}_1$为新样本的变换输入矩阵,依据式(10),现有样本合并新样本后的输出权重矩阵为

$$\begin{aligned}&\boldsymbol{W}_1=\\ &\left(\lambda\boldsymbol{I}+\begin{bmatrix}\boldsymbol{A}_0\\ \boldsymbol{A}_1\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{Q}_0 & 0\\ 0 & \boldsymbol{Q}_1\end{bmatrix}\begin{bmatrix}\boldsymbol{A}_0\\ \boldsymbol{A}_1\end{bmatrix}\right)^{-1}\times\\ &\begin{bmatrix}\boldsymbol{A}_0\\ \boldsymbol{A}_1\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{Q}_0 & 0\\ 0 & \boldsymbol{Q}_1\end{bmatrix}\begin{bmatrix}\boldsymbol{Y}_0\\ \boldsymbol{Y}_1\end{bmatrix}=\\ &(\lambda\boldsymbol{I}+\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{Q}_0\boldsymbol{A}_0+\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{A}_1)^{-1}(\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{Q}_0\boldsymbol{Y}_0+\\ &\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{Y}_1)=\\ &(\boldsymbol{P}_0^{-1}+\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{A}_1)^{-1}(\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{Q}_0\boldsymbol{Y}_0+\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{Y}_1).\end{aligned} \tag{14}$$

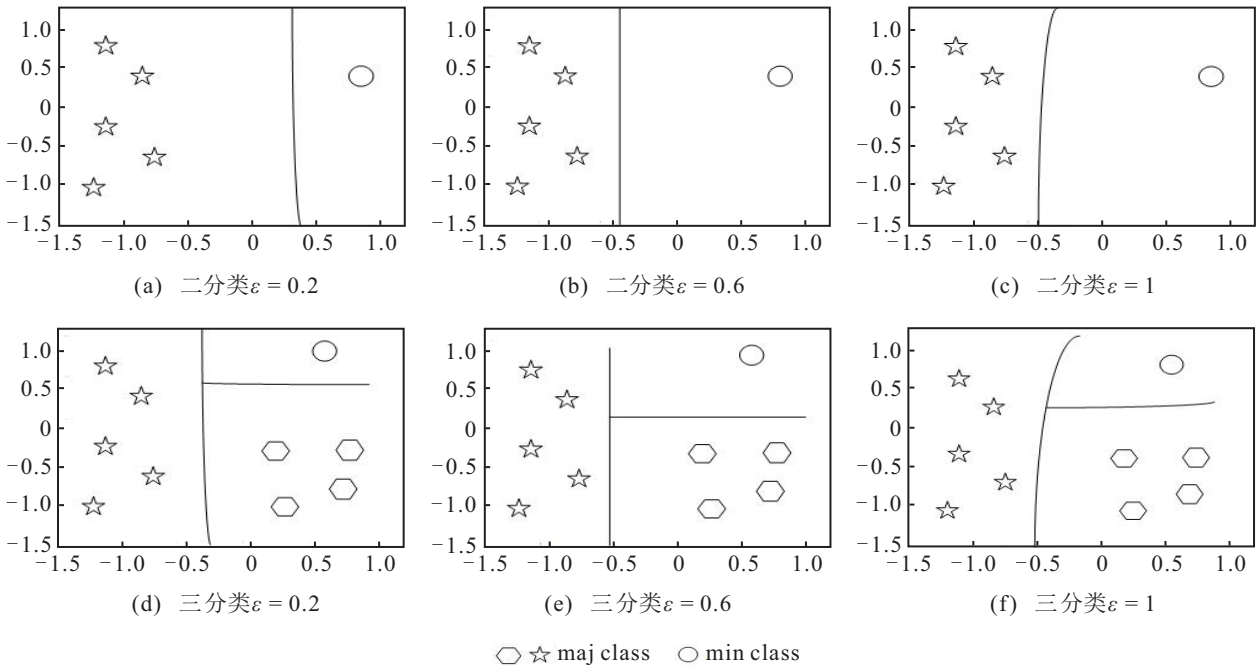


图 3　DDbCs-BLS分类示意

设工作方阵$\boldsymbol{P}_1 = (\boldsymbol{P}_0^{-1} + \boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{A}_1)^{-1}$,则有

$$\boldsymbol{P}_0^{-1} = \boldsymbol{P}_1^{-1} - \boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{A}_1. \tag{15}$$

从而输出权重矩阵$\boldsymbol{W}_1$又可以改写为

$$\begin{aligned}&\boldsymbol{W}_1 = \\ &\boldsymbol{P}_1\boldsymbol{P}_0^{-1}\boldsymbol{P}_0\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{Q}_0\boldsymbol{Y}_0 + \boldsymbol{P}_1\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{Y}_1 = \\ &\boldsymbol{P}_1(\boldsymbol{P}_1^{-1} - \boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{A}_1)\boldsymbol{W}_0 + \boldsymbol{P}_1\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{Y}_1 = \\ &\boldsymbol{W}_0 + \boldsymbol{P}_1\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{Q}_1(\boldsymbol{Y}_1 - \boldsymbol{A}_1\boldsymbol{W}_0).\end{aligned} \tag{16}$$

因此,当第k批样本$\boldsymbol{X}_{L_k}$加入后,可以递推得到

$$\boldsymbol{W}_k = \boldsymbol{W}_{k-1} + \boldsymbol{P}_k A_k^{\mathrm{T}}\boldsymbol{Q}_k(\boldsymbol{Y}_k - \boldsymbol{A}_k\boldsymbol{W}_{k-1}), \tag{17}$$

其中工作方阵$\boldsymbol{P}_k = (\boldsymbol{P}_{k-1}^{-1} + \boldsymbol{A}_k^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{A}_k)^{-1}$.

2.4　算法流程

综合上述,DDbCS-BLS算法流程描述如下.

初始化阶段:

step 1:　随机初始化权重矩阵$\boldsymbol{W}_{zi}$和偏置矩阵$\boldsymbol{\beta}_{zi}$,根据式(1)计算特征映射窗口$\boldsymbol{Z}_i$,将所有$\boldsymbol{Z}_i$组成特征映射层$\boldsymbol{Z}^n$;

step 2:　随机初始化权重矩阵$\boldsymbol{W}_{hj}$和偏置矩阵$\boldsymbol{\beta}_{hj}$,根据式(3)计算增强节点$\boldsymbol{H}_j$,将所有$\boldsymbol{H}_j$组成增强层$\boldsymbol{H}^m$;

step 3: 将特征映射层$\boldsymbol{Z}^n$和增强层$\boldsymbol{H}^m$拼接组成初始样本的变换输入矩阵$\boldsymbol{A}_0$;

step 4:　随机初始化代价敏感矩阵$\boldsymbol{C}$,根据式(8)计算对角矩阵$\boldsymbol{T}$,计算初始样本的权值矩阵$\boldsymbol{Q}_0 = \boldsymbol{TC}$;

step 5: 计算初始样本的工作方阵$\boldsymbol{P}_0$,根据式(13)计算初始样本的输出权重矩阵$\boldsymbol{W}_0$.

增量学习阶段:

step 6:　当新一批样本$\boldsymbol{X}_{L_k}$加入时,重复step 1～step 4,计算新样本的变换输入矩阵$\boldsymbol{A}_k$和权值矩阵$\boldsymbol{Q}_k$;

step 7: 依据工作方阵$\boldsymbol{P}_{k-1}$,递推更新加入新样本后的工作方阵$\boldsymbol{P}_k$;

step 8: 根据式(17),递推更新加入新样本后的输出权重矩阵$\boldsymbol{W}_k$;

step 9: return $\boldsymbol{P}_k$和$\boldsymbol{W}_k$.

2.5　理论分析

BLS样本矩阵$\boldsymbol{X}$为连续值,所以式(6)在该定义域是可导的,故式(6)有唯一最小解$\boldsymbol{W}$,根据式(9)推出的$\boldsymbol{W}$也是唯一解. 由于加入新样本$\boldsymbol{X}_{L_k}$后的样本矩阵仍然为连续值,增量学习模型在该定义域仍是可导的,故增量学习模型的输出权重矩阵$\boldsymbol{W}_k$有唯一最小解.

假设特征映射层和增强层分别有N_n和M_m个神经节点,现有样本的总数为L,样本的维数为D,则计算一个特征向量的时间是$O(N_nM_mD)$,所有样本的时间复杂度约为$O(LN_nM_mD)$. 设新加入样本$\boldsymbol{X}_{L_k}$的数量为L_k,递推更新W_k的主要计算是式(17)的第2项,时间复杂度约为$O(L_kN_nM_mD)$;随着新样本的不断加入,将有$L_k << L$,递推更新时间复杂度远小于非递推所需时间复杂度$O(LN_nM_mD + L_kN_nM_mD)$.

3 实验验证

为验证所提方法的有效性,在多种公共数据集进行验证性和对比性实验. 实验环境如下:处理器型号为Intel (R) Core (TM) i5-4210 U CPU @1.70 GHz,运行内存为4 GB;操作系统为64位Windows;实现语言为Python 3.7.

BLS和DDbCs-BLS设置相同实验参数:特征映射层和增强层的神经节点个数分别为150和200,正则化系数$\lambda = 2^{-30}, \varepsilon = 0.618$,非线性函数$\delta = \tanh(x)$表示为

$$\tanh(x) = \frac{2}{1+\exp(-2x)} - 1. \tag{18}$$

3.1 验证性实验

3.1.1 均衡数据集

MNIST数据集由黑白手写0～9数字图像组成,每个数字尺寸为28×28的图像,其中训练集有60 000张图像,测试集有10 000张图像.

NORB数据集是以不同照明及摆放方式摄制50种不同的3D玩具模型图像,其中训练集有24 300张图像,测试集有24 300张图像.

表1 MNIST和NORB数据集accuracy %

	BLS	DDbCs-BLS
MNIST	98.74	99.07
NORB	89.24	89.31

从表1的结果可以看出,DDbCs-BLS分别以99.07 %和89.31 %的准确度展示了更好的性能,比传统BLS分别高0.33 %和0.04 %.

3.1.2 非均衡数据集

本组实验选用UCI的8组非均衡数据集,如表2所示,其中IR为不均衡率. Adult与CMC为普通的分类数据集,其余6组为医学应用数据集.

表2 UCI的8组非均衡数据集基本信息

名称	样本数	特征数	类别	训练集	测试集	IR
CMC	1 473	10	3	1 031	442	3.42
Adult	15 003	14	2	10 503	4 500	3.08
Blood	671	5	2	470	201	3.21
Pima	768	8	2	538	230	1.87
Breast1	683	9	2	479	204	1.86
Breast2	569	30	2	399	170	1.61
Haberman	306	3	2	214	92	2.78
ILPD	583	10	2	408	175	2.49

针对非均衡数据集,分类准确度accuracy已不合适,为此本组实验使用G-mean值来衡量整体的分类性能,其定义如下:

$$G\text{-mean} = \sqrt{\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}}, \tag{19}$$

其中各参数如表3所示. 通常,只有在多数类和少数类的分类精度同时都高时,G-mean值才会更大.

表3 混淆矩阵

	预测正类	预测负类
真正类	TP	FN
真负类	FP	TN

两种方法的G-mean值如表4所示,DDbCs-BLS明显优于传统BLS,仅在Breast2数据集中的训练准确率比BLS低0.1 %,在Adult数据集上训练准确率要比BLS高5.7 %,在Pima数据集训练和测试的准确率分别比BLS高6.5 %和7.9 %.

表4 两种方法在不同数据集上的G-mean

dataset	方法	G-mean / %	
		train	test
CMC	BLS	80.1	79.2
	DDbCs-BLS	83.4	82.3
Adult	BLS	73.4	73.2
	DDbCs-BLS	79.1	77.6
Blood	BLS	70.4	72.3
	DDbCs-BLS	73.4	74.2
Pima	BLS	71.2	68.3
	DDbCs-BLS	77.7	76.2
Breast1	BLS	78.4	78.9
	DDbCs-BLS	80.1	78.9
Breast2	BLS	84.3	83.8
	DDbCs-BLS	84.2	84.0
Haberman	BLS	76.6	75.3
	DDbCs-BLS	79.9	81.0
ILPD	BLS	75.9	74.4
	DDbCs-BLS	77.3	76.9

3.1.3 实验结果与分析

通过上述两组对比实验可以发现:在均衡数据集上,DDbCs-BLS不会降低传统BLS的分类精度;而在非均衡数据集上,DDbCs-BLS较传统BLS获得了更为理想的分类性能. 其主要原因是:DDbCs-BLS利用代价敏感学习使数据得到均衡,并充分考虑数据集的分布特性,进而得到分类模型的最佳边界线,使每个类别都能最大限度地得到正确分类.

3.2 对比性实验

3.2.1 数据集与评价标准

本组实验选用KEEL非均衡数据库的8组数据集,如表5所示,其中训练集与测试集按照7:3的比例随机划分. 计算F-value、G-mean、AUC三个评价指标值,并与CWsRF[20]、WOS-ELM[21]、CS-CNN[22]等非

均衡方法进行比较.

表5　KEEL的8组非均衡数据集基本信息

数据集	特征	样本	少数类	多数类	IR
Segment	19	2 308	329	1 979	6.02
Vehicle0	18	846	199	647	3.25
Vehicle1	18	846	217	629	2.90
Vehicle2	18	846	218	628	2.88
Vehicle3	18	846	212	634	2.99
Yeast1	8	1 484	429	1 055	2.46
Yeast3	8	1 484	161	1 321	8.10
Page-block0	10	5 472	559	4 913	8.79

F-value是查全率和查准率的调和均值,可以衡量少数类样本的精确度,其定义如下:

$$F\text{-value}=\frac{(1+\eta^2)\text{Precision}\times\text{Recall}}{\eta^2\times(\text{Recall}+\text{Precision})} \tag{20}$$

其中: $\text{Precision}=\dfrac{\text{TP}}{\text{TP}+\text{FP}}$,表示真正类占所有预测正类的比例; $\text{Recall}=\dfrac{\text{TP}}{\text{TP}+\text{FN}}$,表示真正类占实际正类的比例;参数$\eta$用来调节Precision和Recall的相对重要度,一般取值为1. 通常,F-value数值越大,少数类样本的分类精度越高.

AUC是由ROC曲线演变而来,常用曲线下面积AUC来代替ROC曲线作为评价方法,提供了评价模型平均性能的另一种方法,它表示预测的正例排在负例前面的概率,AUC值越大,模型越好.

3.2.2　实验结果与分析

1)　F-value值,如表6所示. DDbCs-BLS的F-value明显高于CWsRF和CS-CNN,仅仅在Vehicle3数据集上比WOS-ELM低0.19 %. 在Page-block0和Vehicle1数据集上分别比WOS-ELM高4.68 %和2.91 %. 即本文方法能更好地细化出数据的分类边界线,使少数类样本的分类精度更高.

表6　4种方法在不同数据集上的F-value　%

Datasets	CWsRF	WOS-ELM	CS-CNN	DDbCs-BLS
Segment	98.43	98.70	98.62	99.71
Vehicle0	98.02	96.97	97.68	99.32
Vehicle1	85.21	83.54	86.32	86.45
Vehicle2	99.11	98.73	97.56	99.32
Vehicle3	83.95	85.01	82.77	84.82
Yeast1	76.96	76.95	75.45	78.29
Yeast3	93.22	92.91	92.46	93.56
Page-block0	96.78	93.40	96.06	98.08

2)　G-mean值,如表7所示. 在8组数据集上DDbCs-BLS所获得的G-mean值都要比CWsRF和WOS-ELM高. 虽然在Vehicle1数据集上略比CS-CNN低1.11 %,但在其他数据集上的G-mean值有不同程度的提高,其中在Page-block0与Segment数据集上的G-mean值要分别比CS-CNN方法高2.2 %和0.96 %.

表7　4种方法在不同数据集上的G-mean　%

Datasets	CWsRF	WOS-ELM	CS-CNN	DDbCs-BLS
Segment	96.33	96.59	96.66	97.62
Vehicle0	96.59	96.97	97.68	98.02
Vehicle1	84.92	82.42	86.42	85.31
Vehicle2	98.21	95.67	97.31	98.55
Vehicle3	82.75	83.41	82.77	83.82
Yeast1	75.46	75.59	74.54	76.91
Yeast3	92.12	91.90	91.61	93.65
Page-block0	94.97	92.82	95.66	97.68

3)　AUC值,如表8所示. 在8组数据集上,本文方法的AUC值基本高于其他3种方法,说明本文方法的平均性能得到显著提升.

表8　4种方法在不同数据集上的AUC　%

Datasets	CWsRF	WOS-ELM	CS-CNN	DDbCs-BLS
Segment	96.31	96.98	96.33	97.61
Vehicle0	96.76	97.07	97.64	97.92
Vehicle1	85.02	83.02	85.23	85.31
Vehicle2	97.93	95.76	97.64	98.65
Vehicle3	82.89	83.56	82.77	83.82
Yeast1	75.35	76.61	75.94	77.00
Yeast3	92.21	92.00	91.96	93.03
Page-block0	95.79	92.97	96.46	97.01

4　结　论

本文提出了一种基于数据分布特性的代价宽度学习框架DDbCs-BLS,以BLS为理论基础,首先采用代价敏感矩阵对非均衡数据进行加权调整比例,然后通过分析数据的分布特性确定最佳分类边界线,提高了方法的稳定性. 在多种公共数据集(包括均衡数据集和不均衡数据集)上进行了大量的验证性和对比性实验,表明DDbCs-BLS一定程度上提高了BLS对非均衡数据分类的性能,能够在不降低BLS对均衡数据分类精度的同时,保证分类器对少数类的正确分类. 下一步的研究工作是将DDbCs-BLS在复杂工业过程中进行工业验证与应用,以进一步改进提高所提方法的性能.

参考文献(References)

[1]　刘金平, 何捷舟, 唐朝晖. 基于WCGAN的矿物浮选泡沫图像光照不变颜色提取[J]. 自动化学报, 2019, DOI: https://doi.org/10.16383/j.aas.c190330.
(Liu J P, He J Z, Tang Z H. WCGAN-based illumination-invariant color measuring of mineral flotation frothImages[J]. Acta Automatica Sinica, 2019, DOI: https://doi.org/10.16383/j.aas.c190330.)

[2]　He X D, Deng L. Deep learning for image-to-text

generation: A technical overview[J]. IEEE Signal Processing Magazine, 2017, 34(6): 109-116.

[3] 陈桥, 丁宝苍, 王雅楠, 等. 基于目标跟踪的双层结构工业预测控制[J]. 控制与决策, 2017, 32(5): 797-803.
(Chen Q, Ding B C, Wang Y N, et al. Double-layered industrial predictive control based on target tracking[J]. Control and Decision, 2017, 32(5): 797-803.)

[4] 丁进良, 杨翠娥, 陈远东, 等. 复杂工业过程智能优化决策系统的现状与展望[J]. 自动化学报, 2018, 44(11): 1931-1943.
(Ding J L, Yang C E, Chen Y D, et al. Research progress and prospects of intelligent optimization decision making in complex industrial process[J]. Acta Automatica Sinica, 2018, 44(11): 1931-1943.)

[5] Tassi A, Khirallah C, Vukobratovic D. Resource allocation strategies for network-coded video broadcasting services over LTE-advanced[J]. IEEE Transactions on Vehicular Technology, 2015, 64(5): 2186-2192.

[6] Rao Y M, Lu J W, Zhou J. Learning discriminative aggregation network for video-based face recognition and person re-identification[J]. International Journal of Computer Vision, 2019, 127(6/7): 701-718.

[7] 黄雅婷, 石晶, 许家铭, 等. 鸡尾酒会问题与相关听觉模型的研究现状与展望[J]. 自动化学报, 2019, 45(2): 234-251.
(Huang Y T, Shi J, Xu J M, et al. Research advances and perspective on the cocktail party problem and related auditory models[J]. Acta Automatica Sinica, 2019, 45(2): 234-251.)

[8] Song D Y, Chandolu A, Stojanovic N. Effect of impurity incorporation on emission wavelength in cathodoluminescence spectrum image study of GaN pyramids grown by selective area epitaxy[J]. Journal of Applied Physics, 2008, 104(6): 064309.

[9] Li J J, Xi B B, Li Y S, et al. Hyperspectral classification based on texture feature enhancement and deep belief networks[J]. Remote Sensing, 2018, 10(3): 396.

[10] Liu F, Chen Z G, Wang J. Video image target monitoring based on RNN-LSTM[J]. Multimedia Tools Applications, 2019, 78(4): 4527-4544.

[11] Chen C L P, Liu Z L. Broad learning system: An effective and efficient incremental learning system without the need for deep architecture[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(1): 10-24.

[12] Chen C L P, Liu Z L, Feng S. Universal approximation capability of broad learning system and its structural variations[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(4): 1191-1204.

[13] Jin J W, Philip Chen C L. Regularized robust broad learning system for uncertain data modeling[J]. Neurocomputing, 2018, 322: 58-69.

[14] Xu M, Han M, Chen C L P. Recurrent broad learning systems for time series prediction[J]. IEEE Transactions on Cybernetics, 2018, 9(10): 1-13.

[15] Ogunleye A, Wang Q G, Marwala T. Integrated learning via randomized forests and localized regression with application to medical diagnosis[J]. IEEE Access, 2019, 7: 18727-18733.

[16] Papamartzivanos D, Gomez Marmol F, Kambourakis G. Introducing deep learning self-adaptive misuse network intrusion detection systems[J]. IEEE Access, 2019, 7: 13546-13560.

[17] Vinayakumar R, Alazab M, Soman K P, et al. Deep learning approach for intelligent intrusion detection system[J]. IEEE Access, 2019, 7(4): 41525-41550.

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

[19] Zong W, Huang G B, Chen Y Q. Weighted extreme learning machine for imbalance learning[J]. Neurocomputing, 2013, 101: 229-242.

[20] Zhu M, Xia J, Jin X Q, et al. Class weights random forest algorithm for processing class imbalanced medical data[J]. IEEE Access, 2018, 6: 4641-4652.

[21] Mirza B, Lin Z P, Tou K A. Weighted online sequential extreme learning machine for class imbalance learning[J]. Neural Processing Letters, 2013, 38(3): 465-486.

[22] Khan S H, Hayat M, Bennamoun M. Cost-sensitive learning of deep feature representations from imbalanced data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 29(8): 3573-3587.

作者简介

徐鹏飞(1977−), 男, 副教授, 博士, 从事智能信息处理等研究, E-mail: xupf@hunnu.edu.cn;

王敏(1995−), 女, 硕士生, 从事智能算法的研究, E-mail: 201870291208@smail.hunnu.edu.cn;

刘金平(1983−), 男, 副教授, 博士, 从事复杂工业过程自动化监控等研究, E-mail: ljp@hunnu.edu.cn;

唐朝晖(1965−), 男, 教授, 博士, 从事复杂工业过程建模与故障诊断等研究, E-mail: zhtang@csu.edu.cn;

马天雨(1978−), 男, 讲师, 博士, 从事复杂工业过程建模及优化控制等研究, E-mail: mty@hunnu.edu.cn.

(责任编辑: 齐 霁)