基于主曲线的不均衡在线贯序极限学习机研究

王金婉[1] 毛文涛[1,2] 王礼云[1] 何 玲[1]
(河南师范大学计算机与信息工程学院 新乡 453007)[1]
(河南省高校"计算智能与数据挖掘"工程技术研究中心 新乡 453007)[2]

摘 要 针对现有机器学习算法难以有效提高不均衡在线贯序数据中少类样本分类精度的问题,提出了一种基于主曲线的不均衡在线贯序极限学习机。该方法的核心思路是根据在线贯序数据的分布特性,均衡各类别样本,以减少少类样本合成过程中的盲目性,主要包括离线和在线两个阶段。离线阶段采用主曲线分别建立各类别样本的分布模型,利用少类样本合成过采样算法对少类样本过采样,并根据各样本点到对应主曲线的投影距离分别为其设定相应大小的隶属度,最后根据隶属区间削减多类和少类虚拟样本,进而建立初始模型。在线阶段对贯序到达的少类样本过采样,并根据隶属区间均衡贯序样本,进而动态更新网络权值。通过理论分析证明了所提算法在理论上存在损失信息上界。采用UCI标准数据集和实际澳门气象数据进行仿真实验,结果表明,与现有典型算法相比,该算法对少类样本的预测精度更高,数值稳定性更好。

关键词 在线贯序极限学习机,不均衡数据,主曲线,少类样本合成过采样
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.3.012

Imbalanced Online Sequential Extreme Learning Machine Based on Principal Curve

WANG Jin-wan[1] MAO Wen-tao[1,2] WANG Li-yun[1] HE Ling[1]
(College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China)[1]
(Engineering Technology Research Center for Computing Intelligence & Data Mining in Henan Province, Xinxiang 453007, China)[2]

Abstract Many traditional machine learning methods tend to get biased classifier which leads to lower classification precision for minor class in sequential imbalanced data. To improve the classification accuracy of minor class, a new imbalanced online sequential extreme learning machine based on principal curve was proposed. The core idea of the method is to get balanced samples based on the distribution features of online sequential data, reducing the blindness in the process of synthetic minority, which contains two stages. In offline stage, the principal curve is introduced to establish the distribution model of two kinds of samples. Over-sampling is done by using SMOTE for minor class. Then the membership degree of each sample is set according to the projection distance respectively, and the majority and virtual minor samples are deleted according to the under interval. Then the initial model is established. In online stage, over-sampling is done by using SMOTE for online sequential minor samples, getting the balanced samples according to the under interval. Then network weight is updated dynamically. The proposed algorithm has upper bound of the loss of information through the theoretical proof. The experiment was taken on three UCI datasets and the real-world air pollutant forecasting dataset, which shows that the proposed method outperforms the traditional methods in terms of prediction accuracy and numerical stability.

Keywords Online sequential extreme learning machine, Imbalanced data, Principal curve, Synthetic minority oversampling

1 引言

实际工程应用中,往往存在着大量类别严重不均衡的在线贯序数据。例如,按顺序到达(每小时或每天)的气象监测数据中,空气质量良好的天数会远远多于空气严重污染的天数。传统的学习方法在解决不均衡数据的分类问题时,对少类样本的识别率远远低于多类样本,易造成"虚假"学习现象。例如,对于一个二分类问题:多类样本为90个,少类样本为10个,此时,即使多类样本全部预测正确而少类样本全部预测错误,总体的分类精度仍可达到90%,很显然这一结果对

到稿日期:2015-03-20 返修日期:2015-06-20 本文受国家自然科学基金(U1204609),河南省基础与前沿技术研究计划项目(132300410430)资助。

王金婉(1991-),女,硕士,主要研究方向为机器学习、模式识别,E-mail:wjwhtu@qq.com;毛文涛(1980-),男,博士,副教授,硕士生导师,主要研究方向为机器学习、弱信号检测,E-mail:maowt.mail@gmail.com;王礼云(1986-),女,硕士,主要研究方向为智能信息处理;何 玲(1990-),女,硕士,主要研究方向为泛化性分析。

少类样本是"虚假"的。因此，提高在线贯序不均衡数据中少类样本的分类精度具有重要的理论和工程意义。

目前针对不均衡数据分类的处理方法主要有两类[1]：(1)数据预处理方法，即通过数据过采样和欠采样来降低类别的不均衡程度，从而提高分类性能；(2)基于算法的处理方法，即通过修改现有机器学习算法，使分类器更有利于少类样本，例如，调节各类样本之间的代价敏感学习[2]、加权支持向量机[3]、集成学习[4]等。数据预处理方法中，对多类样本的欠采样易造成对样本集的有损压缩，导致分类信息的丢失；而过采样技术通过复制或内插的方法来增加少类样本的数量，保留了原始样本集全部分类信息。目前较为常用的过采样技术是少类样本合成过采样技术(Synthetic Minority Over-sampling Technique，SMOTE)[5]。然而，SMOTE 存在一定的盲目性，即新生成的合成样本不符合原始数据分布，特别是在样本集失衡严重的情况下，容易造成严重的样本混合现象，影响分类效果。为解决该问题，杨智明[1]提出了一种改进的过采样方法——自适应 SMOTE，该方法根据样本集内部特征，自适应调整 SMOTE 方法近邻选择策略，控制合成样本质量。针对支持向量机(Support Vector Machine，SVM)在解决非平衡数据分类问题上的局限性，曾志强[6]提出了一种基于核 SMOTE 的分类算法，其所合成的样本质量明显高于 SMOTE 算法，从而有效提高了 SVM 在非平衡数据集上的分类效果。Jeatrakul[7]提出了一种将 SMOTE 算法和互补神经网络相结合来提高少类样本分类精度的方法。Zhai[8]提出了改进的 SMOTE 方法，即采用分层过滤机制来处理噪声数据，将少类选择策略和动态分布密度相结合，从而改进数据分布的不均衡程度。

然而，上述研究多集中于离线数据，很少涉及在线贯序到达的不均衡数据。提高在线贯序不均衡数据的分类精度，关键是：(1)为在线贯序数据选择一个合适的基本算法；(2)如何描述在线贯序到达数据的分布特性。作为神经网络的一个分支，Huang[9]提出的极限学习机(Extreme Learning Machine，ELM)是一种快速的单隐层前向神经网络学习算法，在解决回归估计和模式识别问题时该算法具有"极端"快速的特点。为了将 ELM 推广到在线学习和增量学习，Liang 等人[10]提出了在线贯序极限学习机(Online Sequential Extreme Learning Machine，OS-ELM)，它是一种更快、更准确的学习算法。而主曲线具有数据信息保持良好的优点，且能根据已知数据的走向趋势描绘出未知数据的走向趋势。因此，本文提出了一种基于主曲线的不均衡在线贯序极限学习机(Principal Curve Imbalance Online Sequential Extreme Learning Machine，PCI-OSELM)算法，同时从数据策略和算法策略入手，通过引入主曲线，分别提取各类别样本的分布特性，并利用 SMOTE 算法对少类样本过采样。根据各样本点到对应主曲线的投影距离分别为其赋以相应大小的隶属度，进而根据隶属区间削减多类和少类虚拟样本，从而动态更新网络权值。同时受文献[11]的启发，在理论上给出了在线过程中舍弃样本的损失信息上界，最后采用 UCI 标准数据集和实际澳门气象数据进行仿真实验，结果表明所提算法可有效提高不均衡在线贯序数据中少类样本的分类精度。

2 背景介绍

2.1 在线贯序极限学习机

Huang 等[7]从理论上严格证明了对 N 个严格分离的样本和根据任意连续概率分布容易产生 $R^n \mapsto R$ 中的 (ω_i, b_i) 值，如果激活函数 $g: R \mapsto R$ 无限可微，则隐层输出矩阵 H 及 $\| H\beta - T \| = 0$ 以概率 1 存在。与传统的神经网络方法不同的是，该算法的输入权值和隐层阈值均随机选取，而输出权值可以直接计算得到。整个过程一次完成，无需迭代。但在实际应用中，所有数据可能不是一次性添加到网络中。当新数据添加到网络时，ELM 算法会把新数据和旧数据放到一起重新训练网络，因此会花费很长时间。为解决这一问题，文献[10]把序列学习思想应用于 ELM 算法并提出了 OS-ELM 算法。在该算法中，数据可以一个一个或一块一块地添加到网络中，并且原先的数据学习完成后就会被抛弃不再使用。下面简单介绍 OS-ELM 算法[10]。

Step 1　初始化阶段

从 M 中选取部分数据集 $M_0 = \{(\boldsymbol{x}_i, t_i), i = 1, 2, \cdots, N_0\}$，其中 $N_0 \geqslant L$。

(1)随机选取输入权值 $\boldsymbol{w}_i$ 和 b_i，$i = 1, 2, \cdots, L$，计算隐层输出矩阵 $\boldsymbol{H}_0$。

(2)计算初始输出权值 $\boldsymbol{\beta}^0 = \boldsymbol{D}_0 \boldsymbol{H}_0^{\mathrm{T}} \boldsymbol{T}_0$，其中 $\boldsymbol{D}_0 = (\boldsymbol{H}_0^{\mathrm{T}} \boldsymbol{H}_0)^{-1}$，$\boldsymbol{T}_0 = [t_1, t_2, \cdots, t_{N_0}]^{\mathrm{T}}$。

(3)置 $k = 0$。

Step 2　序列学习阶段

(1)学习第 $k+1$ 个数据：$d_{k+1} = (\boldsymbol{x}_{N_0+k+1}, \boldsymbol{t}_{N_0+k+1})$。

(2)计算新学习数据的隐层输出矩 $\boldsymbol{H}_{k+1}$。

$$\boldsymbol{H}_{k+1} = [g(\boldsymbol{w}_1 \cdot \boldsymbol{x}_{N_0+k+1} + b_1) \cdots g(\boldsymbol{w}_L \cdot \boldsymbol{x}_{N_0+k+1} + b_L)]_{1 \times L} \tag{1}$$

令 $\boldsymbol{T}_{k+1} = [t_{N_0+k+1}]^{\mathrm{T}}$。

(3)计算输出权值 $\boldsymbol{\beta}^{k+1}$。

$$\begin{aligned} \boldsymbol{D}_{k+1} &= \boldsymbol{D}_k - \boldsymbol{D}_k \boldsymbol{H}_{k+1}^{\mathrm{T}} (\boldsymbol{I} + \boldsymbol{H}_{k+1} \boldsymbol{D}_k \boldsymbol{H}_{k+1}^{\mathrm{T}})^{-1} \boldsymbol{H}_{k+1} \boldsymbol{D}_k \\ \boldsymbol{\beta}^{k+1} &= \boldsymbol{\beta}^k + \boldsymbol{D}_{k+1} \boldsymbol{H}_{k+1}^{\mathrm{T}} (\boldsymbol{T}_{k+1} - \boldsymbol{H}_{k+1} \boldsymbol{\beta}^k) \end{aligned} \tag{2}$$

(4)置 $k = k+1$，返回 Step 2。

2.2 主曲线

Hastie 和 Stuetzle 于 1984 年提出了主曲线[12]的概念。主曲线是通过数据集"中间"光滑无参数的曲线或曲面，是线性主成分的非线性推广，也是嵌入高维数据空间的非欧空间的低维流形描述。不同于传统的非线性回归方法，主曲线具有两个明显的优点：(1)数据信息的保持性好；(2)可有效勾勒出原始信息的轮廓，即数据集是"云"，主曲线是该数据集的"骨架"。目前应用最广泛的是 K 主曲线。主曲线的算法步骤可概括为[11]：

步骤 1　令初始曲线 $f^0(\lambda)$ 为 X 的第一主成分，设 $j = 0$；

步骤 2　(投影步)对所有 $\boldsymbol{x} \in \boldsymbol{R}^d$，求投影指标：

$$\lambda_{f^{(j)}}(\boldsymbol{x}) = \max\{t: \| \boldsymbol{x} - f(\lambda) \| = \min_{\tau} \| \boldsymbol{x} - f(\tau) \| \} \tag{3}$$

步骤 3　(期望步)定义 $\boldsymbol{x}$ 在 f 上的投影点为：$f^{(j+1)}(\lambda) = E[X | \lambda_{f^{(j)}}(X) = \lambda)]$；

步骤 4　如果 $1 + \dfrac{\Delta(f^{(j+1)})}{\Delta(f^{(j)})}$ 小于某个阈值，则停止(其中 $\Delta(f^{(j)})$ 表示点 $\boldsymbol{x}$ 到曲线 f 的欧氏平方距离)；否则，令 $j = j+1$，转步骤 2。

2.3 SMOTE 算法

Chawla. N[13]提出的 SMOTE 算法是一种新型的过采样技术。不同于传统的样本复制过采样方法，其通过产生新的少类样本，并控制新生成样本数量和分布来达到平衡样本集

的目的，它可以有效解决传统过采样方法由于决策区间过小而引起的分类器过拟合问题，是目前常用的过采用方法之一。对于少类样本点 x_1，从少类样本集合中随机选择一个样本，设其为 x_2，x_1 与 x_2 之间对应属性 j 上的差值 $diff_j=x_{2j}-x_{1j}$。SMOTE 方法使用式(4)将差值 $diff_j$ 与一个[0,1]内的随机数相乘之后，同原始的属性向量中对应的属性 x_{1j} 相加，生成一个新的属性值 f_{1j}。将得到的 m 个属性值[f_{11}，…，f_{1m}]组成一个向量，从而产生一个新的少类样本 $\boldsymbol{f}_1$。

$$\begin{aligned} f_{1j}&=x_{1j}+diff_j\times rand[0,1]\\ &=x_{1j}+(x_{2j}-x_{1j})\times rand[0,1] \end{aligned} \tag{4}$$

按照预先设定的过采样率反复执行以上过程，并将新产生的少类合成样本加入到原始数据集中，共同训练得到最终的分类器。

3 PCI-OSELM 算法

为减少少类样本合成过程中的盲目性，提高其分类精度，本文同时从数据策略和算法策略两个角度出发，提出一种基于主曲线的不均衡在线贯序极限学习机，主要分为离线和在线两个阶段。

3.1 离线阶段

为叙述方便，首先给出几个定义。设某类别样本集合 $S=\{\boldsymbol{x}_i,i=1,2,\cdots,n\}$，其中，$\boldsymbol{x}_i$ 表示 m 维向量，维数大小代表样本特征个数。曲线 f 为基于 S 的主曲线，d_i 表示样本 $\boldsymbol{x}_i$ 到主曲线 f 的投影距离。

定义 1(隶属度) 指样本 $\boldsymbol{x}_i$ 符合样本集 S 分布特性的程度。计算方法如式(5)所示：

$$\begin{cases} D=\max(d_1,d_2,\cdots,d_n)\\ \mu_i=\dfrac{d_i}{D} \end{cases} \tag{5}$$

定义 2(隶属区间) 基于隶属度，给出隶属区间 δ 的计算方法，如式(6)所示：

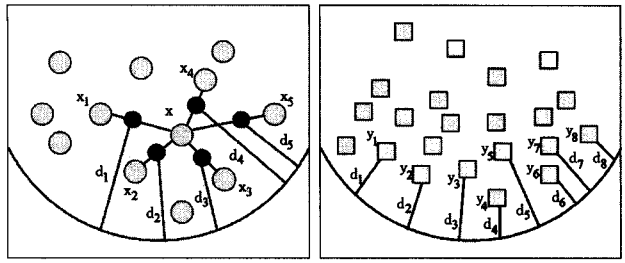
$$\delta=[\min(\mu_i),\mathrm{mean}(\mu_i)] \tag{6}$$

在初始离线阶段，首先采用主曲线分别提取多类和少类样本的分布特性；其次利用 SMOTE 方法对少类样本过采样；然后分别计算多类和少类样本的隶属度，并根据隶属区间分别削减多类样本和少类虚拟样本，得到均衡的离线样本集，进而建立初始模型。具体步骤如下。

(1)对初始离线样本集 $D=\{(\boldsymbol{x}_i,t_i)\mid i=1,2,\cdots,N\}$，分别构建少类样本集 $S=\{(\boldsymbol{x}_i,t_i)\mid i=1,2,\cdots,n\}$ 和多类样本集 $M=\{(\boldsymbol{x}_i,t_i)\mid i=1,2,\cdots,m\}$ 的主曲线 f_1 和 f_2。

(2)对于少类样本点 $\boldsymbol{x}_i,i=1,2,\cdots,n$，在其附近随机选择 k 个少类样本点 $\boldsymbol{x}_j,j=1,2,\cdots,k$。利用 SMOTE 算法对少类样本过采样，如式(4)所示，得到过采样后的少类样本集 $S_1=\{(\boldsymbol{x}_i,t_i)\mid i=1,2,\cdots,n,\cdots,n+n_1\}$，其中，$n_1$ 表示过采样的虚拟少类样本个数。

(3)根据式(3)，分别计算过采样后的少类样本集 S_1 中的每个样本 $\boldsymbol{x}_i(i=1,2,\cdots,n,\cdots,n+n_1)$ 到主曲线 f_1 的投影距离 d_i，以及多类样本集 M 中的每个样本 $\boldsymbol{x}_i(i=1,2,\cdots,m)$ 到主曲线 f_2 的投影距离 d_i。如图 1 所示，其中，空心圆圈表示原始少类样本点，实心圆圈表示过采样的虚拟少类样本点，方框表示多类样本点。



(a)少类样本　　(b)多类样本

图 1　样本点到主曲线的投影距离示意图

(4)根据式(5)和式(6)分别计算 S_1 和 M 中各样本对应的隶属度 μ_i，并分别得到集合 S_1 和 M 对应的隶属区间 δ_1 和 δ_2。

(5)分别判断过采样的 n_1 个虚拟少类样本和 m 个多类样本是否在对应的隶属区间 δ_1 和 δ_2 内。若不在，则从集合 S_1 或 M 中剔除相应样本。最终得到均衡合理的少类样本集 $S_2=\{(\boldsymbol{x}_i,t_i)\mid i=1,2,\cdots,n,\cdots,n+p\}$ 和多类样本集 $M_1=\{(\boldsymbol{x}_i,t_i)\mid i=1,2,\cdots,m-q\}$。其中，$p$ 为合理的少类虚拟样本个数，q 为根据隶属区间剔除的多类样本个数。

(6)合并样本集 S_2 和 M，得到均衡的初始训练样本集 $D_0=\{(\boldsymbol{x}_i,t_i)\mid i=1,2,\cdots,N_0\}$，最后建立初始训练模型。

给定隐层激活函数 $g(x)$和隐层神经元个数 L，随机选取输入权值 $\boldsymbol{w}_i$ 和偏置 $b_i,i=1,2,\cdots,L$，计算隐层输出矩阵：

$$\boldsymbol{H}_0=\begin{bmatrix} g(\boldsymbol{w}_1\cdot\boldsymbol{x}_1+b_1) & \cdots & g(\boldsymbol{w}_L\cdot\boldsymbol{x}_1+b_L)\\ g(\boldsymbol{w}_1\cdot\boldsymbol{x}_2+b_1) & \cdots & g(\boldsymbol{w}_L\cdot\boldsymbol{x}_2+b_L)\\ \vdots & & \vdots\\ g(\boldsymbol{w}_1\cdot\boldsymbol{x}_{N_0}+b_1) & \cdots & g(\boldsymbol{w}_L\cdot\boldsymbol{x}_{N_0}+b_L) \end{bmatrix}_{N_0\times L} \tag{7}$$

输出向量为 $\boldsymbol{T}_0=[t_1\quad t_2\quad\cdots\quad t_{N_0}]^{\mathrm{T}}$，输出权值为

$$\boldsymbol{\beta}_0=\boldsymbol{H}_0^{+}\boldsymbol{T}_0 \tag{8}$$

式中：

$$\boldsymbol{H}_0^{+}=(\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{H}_0)^{-1}\boldsymbol{H}_0^{\mathrm{T}} \tag{9}$$

令 $\boldsymbol{M}_0=(\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{H}_0)^{-1}$，则式(9)即为 $\boldsymbol{H}_0^{+}=\boldsymbol{M}_0\boldsymbol{H}_0^{\mathrm{T}}$。

置 $k=0$。

3.2 在线贯序阶段

对贯序到达的样本数据，首先判断其类别。若是多类样本，则判断其对应隶属度是否在多类样本对应的隶属区间 δ_2 内，若在，则更新网络权值，否则剔除该样本，不更新；若新到样本为少类样本，则首先采用 SMOTE 算法过采样，得到 k 个虚拟少类样本点，判断这 k 个虚拟样本点对应的隶属度是否在对应的隶属区间 δ_1 内，剔除不在 δ_1 区间内的虚拟少类样本，利用贯序到达的少类样本和在 δ_1 区间内的虚拟样本一起更新网络权值。

设第 $k+1$ 步贯序到达的新样本块为 $\Omega_{k+1}=\{(\boldsymbol{x}_i,t_i)\}$，$i=k+N_0+1,\cdots,k+N_0+Block$，其中 $Block$ 表示第 $k+1$ 步添加的数据个数。易知，Ω_{k+1} 为不均衡贯序样本块，则首先根据样本的分布特性，均衡各类别样本，具体步骤如下。

(1)对新样本块 Ω_{k+1}，根据类别将其分为多类样本块 I_y 和少类样本块 I_x。

(2)对多类样本块 I_y，计算 I_y 内各样本到对应主曲线 f_2 的投影距离，进而计算隶属度，剔除隶属度不在多类样本对应的隶属区间 δ_2 内的样本点，得到多类样本块样本 I_y'。

(3)对少类样本块 I_x，利用 SMOTE 算法对 I_x 中每个样本过采样，得到虚拟少类样本，计算各个虚拟少类样本点到对应主曲线 f_1 的投影距离，计算隶属度，把隶属度在少类样本对应的隶属区间 δ_1 内的虚拟样本点并入 I_x 中，最终得到少类样本块 I_x'。

(4)合并 I_x'和 I_y'，得到新的均衡的贯序样本块 $\Phi_{k+1}=\{(\boldsymbol{x}_i,t_i)\}$，$i=k+N_0+1,\cdots,k+N_0+\varphi$。其中，$\varphi$ 为 I_x'和 I_y'的样本数目之和。则均衡的贯序样本块 Φ_{k+1} 对应的神经元矩阵为 $\boldsymbol{H}_{\Phi_{k+1}}=[\boldsymbol{h}_{k+N_0+1}\ \boldsymbol{h}_{k+N_0+2}\ \cdots\ \boldsymbol{h}_{k+N_0+\varphi}]$。此时，隐层输出矩阵为 $\boldsymbol{H}_{k+1}=[\boldsymbol{H}_k^{\mathrm{T}}\ \boldsymbol{H}_{\Phi_{k+1}}^{\mathrm{T}}]^{\mathrm{T}}$，输出向量为 $\boldsymbol{T}_{k+1}=[\boldsymbol{T}_k^{\mathrm{T}}\ \boldsymbol{T}_{\Phi_{k+1}}^{\mathrm{T}}]^{\mathrm{T}}$。根据式(10)更新网络权值。

$$\boldsymbol{\beta}_{k+1}=\boldsymbol{H}_{k+1}^{+}\boldsymbol{T}_{k+1} \tag{10}$$

其中，$\boldsymbol{H}_{k+1}^{+}=(\boldsymbol{H}_{k+1}^{\mathrm{T}}\boldsymbol{H}_{k+1})^{-1}\boldsymbol{H}_{k+1}^{\mathrm{T}}$。令 $\boldsymbol{P}_{k+1}=(\boldsymbol{H}_{k+1}^{\mathrm{T}}\boldsymbol{H}_{k+1})^{-1}$，则有

$$\boldsymbol{H}_{k+1}^{+}=\boldsymbol{P}_{k+1}\boldsymbol{H}_{k+1}^{\mathrm{T}} \tag{11}$$

因为

$$\begin{aligned}\boldsymbol{H}_{k+1}^{\mathrm{T}}\boldsymbol{H}_{k+1}&=[\boldsymbol{H}_k^{\mathrm{T}}\ \boldsymbol{H}_{\Phi}^{\mathrm{T}}][\boldsymbol{H}_k^{\mathrm{T}}\ \boldsymbol{H}_{\Phi}^{\mathrm{T}}]^{\mathrm{T}}\\&=\boldsymbol{H}_k^{\mathrm{T}}\boldsymbol{H}_k+\boldsymbol{H}_{\Phi}^{\mathrm{T}}\boldsymbol{H}_{\Phi}\end{aligned} \tag{12}$$

即

$$\boldsymbol{P}_{k+1}^{-1}=\boldsymbol{P}_k^{-1}+\boldsymbol{H}_{\Phi}^{\mathrm{T}}\boldsymbol{H}_{\Phi} \tag{13}$$

对式(13)两端求逆，根据 Sherman-Morrison 矩阵求逆引理可得 $\boldsymbol{P}_{k+1}$的递推表达式[10]：

$$\boldsymbol{P}_{k+1}=(\boldsymbol{P}_k^{-1}+\boldsymbol{H}_{\Phi}^{\mathrm{T}}\boldsymbol{H}_{\Phi})^{-1}=\boldsymbol{P}_k-\frac{\boldsymbol{P}_k\boldsymbol{H}_{\Phi}^{\mathrm{T}}\boldsymbol{H}_{\Phi}\boldsymbol{P}_k}{\boldsymbol{I}+\boldsymbol{H}_{\Phi}\boldsymbol{P}_k\boldsymbol{H}_{\Phi}^{\mathrm{T}}} \tag{14}$$

因此，$\boldsymbol{P}_{k+1}$可以在 $\boldsymbol{P}_k$ 的基础上计算得到，从而大大缩减了计算量。将式(14)代入式(10)即得到 $\boldsymbol{H}_{k+1}^{+}$，并根据式(10)更新网络权值得到 $\boldsymbol{\beta}_{k+1}$。

PCI-OSELM 算法流程如图 2 所示。

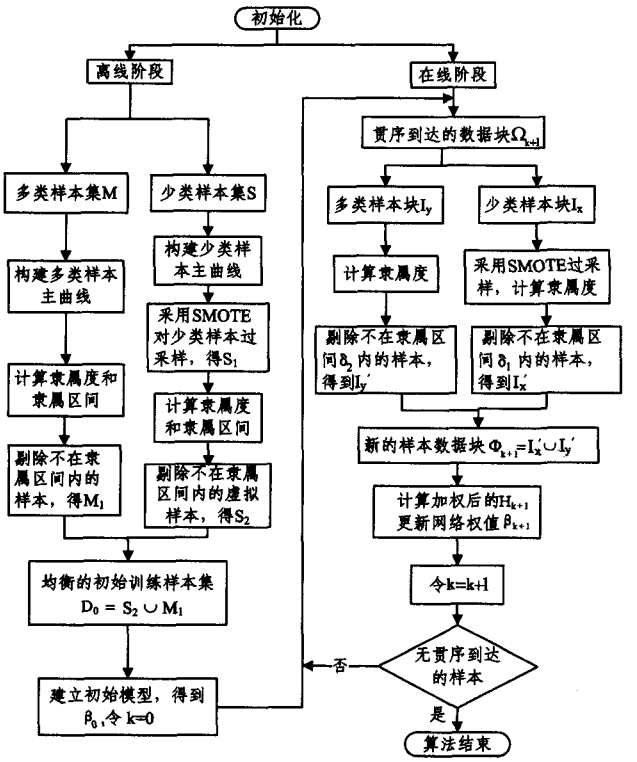


图 2　PCI-OSELM 算法流程

4　理论分析

由于熵可以表示一个数据集所包含的信息量，因此，本文从信息熵的角度给出在线过程中根据隶属度剔除样本的整体损失信息上界。限于篇幅，仅以多类样本为例，证明其损失信息上界，少类样本的信息损失上界可依此推导得出。对贯序到达的多类样本块 I_y，分别计算 I_y 内各样本对应的隶属度 u_i，剔除不在 δ_2 内的样本，得到 I_y'。由式(5)、式(6)可知，u_i 越大，对应样本到主曲线的投影距离越大，剔除的可能性越大，即该样本在 I_y 中所占权重越小。由此给出如下定义。

定义 3(样本权重)　给定集合 $D=\{(\boldsymbol{x}_i,t_i)\mid i=1,2,\cdots,n\}$，已知样本$(\boldsymbol{x}_i,t_i)$对应的隶属度为 u_i，则其所占的样本权重定义为

$$w_i=1-u_i \tag{15}$$

易知，每一次贯序学习阶段所剔除的样本集为 $\Psi=I_y-I_y'=\{(\boldsymbol{x}_i,t_i),i=1,2,\cdots,m\}$，由式(15)可得 Ψ 对应的总体样本权重为$\sum_{i=1}^{m}w_i=\sum_{i=1}^{m}(1-u_i)=\sum_{i=1}^{m}(1-\frac{d_i}{D})=m-\frac{1}{D}\sum_{i=1}^{m}d_i$。

定理 1　令 $H(\Psi)$表示在线欠采样过程中多类样本对应的损失信息量，则 $H(\Psi)\leqslant(m-\frac{1}{D}\sum_{i=1}^{m}d_i)\log(m/(m-\frac{1}{D}\sum_{i=1}^{m}d_i))$，且损失信息 $H(\Psi)$的上界仅与 Ψ 中所有样本到主曲线的投影距离之和$\sum_{i=1}^{m}d_i$ 有关。

证明：根据熵的定义，有 $H(\Psi)=-\sum_{i=1}^{m}w_i\log w_i$。根据最大熵原理，当 w_i 都取相同的值$(m-\frac{1}{D}\sum_{i=1}^{m}d_i)/m$ 时，$H(\Psi)$达到最大值，则有

$$\begin{aligned}H(\Psi)&\leqslant-\sum_{i=1}^{M}(m-\frac{1}{D}\sum_{i=1}^{m}d_i)/m\log((m-\frac{1}{D}\sum_{i=1}^{m}d_i)/m)\\&=(m-\frac{1}{D}\sum_{i=1}^{m}d_i)\log(m/(m-\frac{1}{D}\sum_{i=1}^{m}d_i))\end{aligned} \tag{16}$$

由式(16)可知，$H(\Psi)$的上界仅与$\sum_{i=1}^{m}d_i$ 即 Ψ 中各样本对应的投影距离之和有关，且$\sum_{i=1}^{m}d_i$ 越大，上界越小。

定理 1 从理论上证明了根据样本点到主曲线的投影距离设定隶属度，进而根据隶属度均衡样本的合理性。考虑极端情况，若 Ψ 中样本到主曲线的投影距离之和(即$\sum_{i=1}^{m}d_i$)趋近于无穷大，则对应的信息损失上界趋近于无穷小，这意味着根据隶属度挑选样本对整体样本信息的影响可忽略不计。

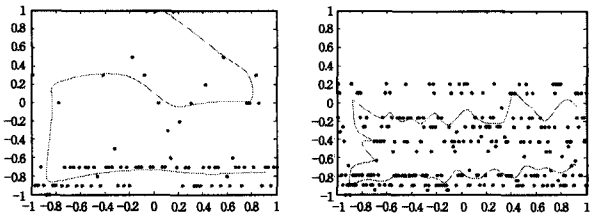
5　仿真实验

为验证所提算法的有效性，本文采用 UCI 数据集和实际气象数据[14]进行仿真实验，分别采用 ELM、OS-ELM 与 SMOSELM 与本文所提算法进行对比。其中，SMOSELM 为少类样本合成在线贯序极限学习机(Synthetic Minority Online Sequential Extreme Learning Machine，SMOSELM)，其基本思想是仅用 SMOTE 算法对少类样本过采样来解决样本的不均衡问题。所有实验结果均为重复 100 次所得结果的平均值。在训练前，所有样本被线性归一化到[−1,1]之间。

5.1　UCI 标准数据集

选择 Pima、Banknote 和 Blood 3 个 UCI 数据集[15]进行仿真实验。

对于离线样本，分别构建 3 个标准数据集描绘多类和少类样本大致轮廓的主曲线，如图 3 所示(限于篇幅，仅以 Blood 数据集为例)。



(a)少类样本主曲线　　(b)多类样本主曲线

图 3　Blood 数据集多类和少类样本的主曲线

采用 SMOTE 对少类样本过采样，根据隶属区间分别削减多类样本和过采样后的少类样本，进而得到均衡的样本集，如表 1 所列。

表 1　离线阶段的样本集数目变化情况

数据集	特征数	少类样本数		多类样本数	
		均衡前	均衡后	处理前	处理后
Pima	8	47	169	253	168
Banknote	4	66	224	344	213
Blood	4	76	178	272	167

利用均衡后的离线训练样本集，建立初始模型。给定隐层激活函数为 RBF 核函数，隐层节点数分别为 30、20、25，运行 100 次，4 种模型的平均性能如表 2－表 4 所列。

表 2　Pima 数据集

	PCI-OSEM	OS-ELM	ELM	SMOSELM
训练时间	0.3647	0.3528	0.0620	0.2583
测试时间	0.0053	0.0095	0.0051	0.0019
少类训练精度	**89.07%**	44.13%	43.94%	87.26%
多类训练精度	90.31%	97.46%	97.03%	86.23%
少类测试精度	**63.84%**	29.88%	29.39%	54.29%
多类测试精度	88.29%	96.26%	95.81%	88.24%
总体训练精度	89.39%	88.56%	88.51%	86.74%
总体测试精度	84.16%	84.93%	84.52%	82.47%

表 3　Banknote 数据集

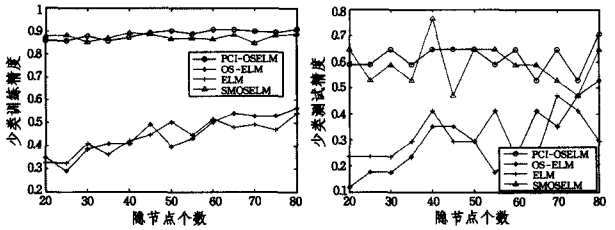
	PCI-OSEM	OS-ELM	ELM	SMOSELM
训练时间	0.5343	0.4368	0.0539	0.5020
测试时间	0.0059	0.0032	0.0057	0.0058
少类训练精度	**99.91%**	99.77%	99.61%	99.84%
多类训练精度	98.91%	99.72%	99.85%	98.13%
少类测试精度	**99.96%**	99.71%	98.57%	99.93%
多类测试精度	99.73%	99.79%	99.69%	99.19%
总体训练精度	99.42%	99.72%	99.81%	98.97%
总体测试精度	99.89%	99.78%	99.51%	99.30%

表 4　Blood 数据集

	PCI-OSEM	OS-ELM	ELM	SMOSELM
训练时间	0.2803	0.2109	0.0380	0.2761
测试时间	0.0140	0.0069	0.0057	0.0068
少类训练精度	**78.49%**	35.91%	33.90%	75.41%
多类训练精度	58.27%	95.21%	94.94%	68.58%
少类测试精度	**82.93%**	39.57%	34.69%	71.02%
多类测试精度	43.46%	91.35%	91.28%	59.95%
总体训练精度	68.26%	79.74%	79.68%	72.52%
总体测试精度	51.27%	80.73%	80.68%	62.21%

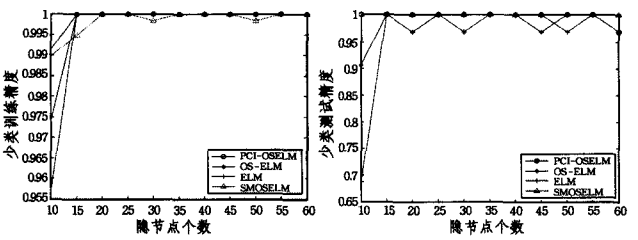
可以看出，尽管 PCI-OSELM 的整体精度不是最高的，但其少类训练精度和少类测试精度均明显高于其它 3 种算法，这表明在不均衡在线贯序数据的分类问题上，PCI-OSELM 对少类样本的识别率更高；且除 ELM 外，其余 3 种算法的训练时间均为一个数量级，进一步表明，PCI-OSELM 在不影响算法复杂度的情况下可有效提高少类样本的分类精度。

为进一步验证所提算法对提高少类样本分类精度的有效性，图 4－图 6 给出了 PCI-OSELM、OS-ELM、ELM 和 SMOSELM 4 种算法分别在 3 个标准数据集上少类训练精度和少类测试精度随隐节点变化的变化情况。



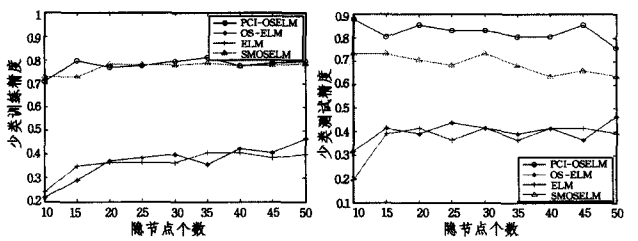
(a)少类训练精度　　(b)少类测试精度

图 4　随着隐节点的变化，4 种算法在 Pima 上的精度变化



(a)少类训练精度　　(b)少类测试精度

图 5　随着隐节点的变化，4 种算法在 Banknote 上的精度变化



(a)少类训练精度　　(b)少类测试精度

图 6　随着隐节点的变化，4 种算法在 Blood 上的精度变化

从图 4－图 6 可以看出，与其它 3 种算法相比，PCI-OSELM 的少类训练精度和少类测试精度更高，这进一步表明所提算法可有效提高少类样本的分类精度；且随着隐层节点的变化，PCI-OSELM 的曲线变化较为平稳，表明与其它 3 种算法相比，所提算法的数值稳定性更好。

5.2　空气污染预测

在空气质量监测等实际问题中，数据往往具有在线序列到达的特点，且空气质量良好的天数远远多于空气严重污染的天数，因此之是一种典型的不均衡在线贯序问题。由于采集数据的局限性，本文采用澳门气象局网站上公布的空气质量数据[14]进行仿真实验。

给定训练数据集 $D=(\boldsymbol{x},t)$，$\boldsymbol{x}$ 表示输入向量，即当天的 PM10、SO_2、NO_2、O_3 的浓度值，即 $\boldsymbol{x}=(d(\mathrm{PM10}),d(SO_2),d(NO_2),d(O_3))$，$t$ 是输出变量即第二天的 PM10 值，即 $t=d+1(\mathrm{PM10})$。

为验证 PCI-OSELM 在解决实际不均衡在线贯序分类问题上的有效性，利用 2010 年到 2012 年氹仔格兰德气象站收集的序列数据进行仿真实验。其中，2010 年的数据作为初始离线训练样本，2011 年的数据作为在线训练样本，2012 年的数据作为测试样本。

对 2010 年初始训练样本，根据 3.1 节所述方法均衡各类

别样本，最终得到均衡的训练样本集，如表 5 所列。

表 5　均衡样本前后 2010 年样本数目

类别	少类样本数	多类样本数	少类百分比	多类百分比
处理前	31	334	8.49%	91.51%
处理后	131	154	45.97%	54.03%

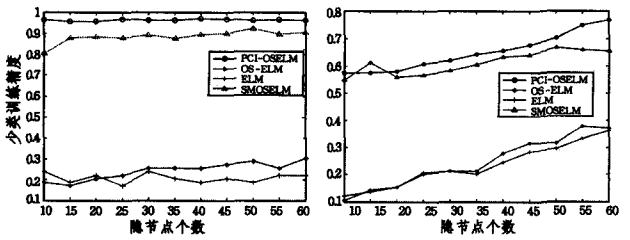
给定隐层激活函数为 RBF 核函数，隐节点个数为 30，表 6 所列为运行 100 次所得结果的平均值。

表 6　澳门数据实验结果

	PCIOSEM	OSELM	ELM	SMOSELM
训练时间	0.5661	0.4126	0.0486	0.6970
测试时间	0.0055	0.0057	0.0123	0.0025
少类训练精度	**96.64%**	23.89%	20.66%	80.23%
多类训练精度	89.26%	99.14%	99.18%	80.11%
少类测试精度	**68.09%**	50.45%	56.31%	62.04%
多类测试精度	85.66%	97.98%	98.06%	86.02%
总体训练精度	92.95%	93.08%	92.71%	84.73%
总体测试精度	83.20%	91.22%	91.23%	82.68%

从表 6 可以看出，在解决实际不均衡在线贯序分类问题时，PCI-OSELM 的少类训练精度和少类测试精度仍最高，分别比其它 3 种算法提高了 72.75%、75.98%、16.41% 和 17.64%、11.78%、6.05%，进一步证明了所提算法对提高少类样本分类精度的有效性。尽管两种经典的 OSELM 和 ELM 算法的总体测试精度均达到 90%以上，且高于 PCI-OSELM，但这一结果在实际不均衡在线贯序数据的分类问题上并无实际意义。

图 7 所示为随隐层节点个数的变化，4 种算法的少类训练精度和少类测试精度的变化情况。不难发现，PCI-OSELM 对少类样本的分类精度均明显高于其它 3 种算法；且随着隐层节点的变化，PCI-OSELM 的精度变化曲线较为平缓，这进一步表明 PCI-OSELM 不仅对少类样本的识别能力更高，而且数值稳定性良好。



(a)少类训练精度变化　　(b)少类测试精度变化

图 7　随着隐层节点的变化，少类训练精度和少类测试的精度变化

结束语　为解决不均衡在线贯序数据的分类问题，提高少类样本的分类精度，本文同时从数据和算法两个角度出发，提出一种基于主曲线的不均衡在线贯序极限学习机。通过引入主曲线，提取在线贯序数据的分布特性，同时根据隶属区间削减多类样本和采用 SMOTE 算法过采样的少类虚拟样本，减少了少类样本合成过程中的盲目性。实验结果表明，本文所提算法在不影响计算复杂度的前提下，对少类样本的预测精度更高，数值稳定性更好。

参 考 文 献

[1] Yang Zhi-ming, Qiao Li-yan, Peng Xi-yuan. Research on Data-ming Method for Imbalanced Dataset Based on Improved SMOTE[J]. Acta Electronica Sinica, 2007, 35(12A): 22-26(in Chinese)
杨智明，乔立言，彭喜元. 基于改进 SMOTE 的不平衡数据挖掘方法研究[J]. 电子学报，2007，35(12A)：22-26

[2] Fu Zhong-liang. Cost-sensitive Ensemble Learning Algorithm for Multi-label Classification Problems[J]. Acta Automatica Sinica, 2014(6): 1075-1085(in Chinese)
付忠良. 多标签代价敏感分类集成学习算法[J]. 自动化学报，2014(6)：1075-1085

[3] Zeng Hui. Research on Improved Weighted Support Vector Machine and Application In Fault Diagnosis Method[D]. Guangzhou: South China University of Technology, 2010(in Chinese)
曾辉. 改进加权支持向量机的研究及在故障诊断中的应用[D]. 广州：华南理工大学，2010

[4] Zhang Chun-xia, Zhang Jiang-she. A Survey of Selective Ensemble learning Algorithm[J]. Chinese Journal of Computers, 2011, 34(8): 1399-1410(in Chinese)
张春霞，张讲社. 选择性集成学习算法综述[J]. 计算机学报，2011，34(8)：1399-1410

[5] Rok B, Lara L. SMOTE for high-dimensional class-imbalanced data[J]. BMC Bioinformatics, 2013, 14(1): 1-16

[6] Zeng Zhi-qiang, Wu Qun, Liao Bei-shui, et al. A Classification Method For Imbalance Data Set Based on Kernel SMOTE[J]. Acta Electronica Sinica, 2009, 37(11): 2489-2495(in Chinese)
曾志强，吴群，廖备水，等. 一种基于核 SMOTE 的非平衡数据集分类方法[J]. 电子学报，2009，37(11)：2489-2495

[7] Jeatrakul P, Wong KW, Fung C C. Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE Algorithm[M]// Neural Information Processing. Models and Applications, 2010: 152-159

[8] Zhai Y, Ma N, Ruan D. An effective over-sampling method for imbalanced data sets classification[J]. Chinese Journal of Electronics, 2011, 20(3): 489-494

[9] Huang G-B, Zhou H, Ding X, et al. Extreme Learning Machine for Regression and Multiclass[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 2012, 42(2): 513-529

[10] Liang N Y, Huang G B. A fast and accurate online sequential learning algorithm for feedforward networks[J]. IEEE Trans Neural Networks, 2006, 17: 1411-1423

[11] Yuan P, Ma H, Fu H. Hotspot-entropy based data forwarding in opportunistic social networks[J]. Pervasive and Mobile Computing, 2015(1), 16(A): 136-154

[12] Li Hao. Soft Sensing and its Applied Research Based on Principal Curves[D]. Hangzhou: Zhejiang University, 2013(in Chinese)
李浩. 基于主曲线的软测量及其应用研究[D]. 杭州：浙江大学，2013

[13] Nele V, Enislay R, Chris Cornelis, et al. Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection[J]. Applied Soft Computing Journal, 2014, 22: 511-517

[14] SMG. E-publication Download Page [OL]. http://www.smg.gov.mo/www/ccaa/pdf/e_pdf_download.php

[15] Newman D J, Hettich S, Blake C L, et al. UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science[OL]. http://www.ics.uci.edu/mlearn/ML Repository.html