

## 面向贯序不均衡数据的混合采样极限学习机

毛文涛<sup>1,2\*</sup>, 王金婉<sup>1</sup>, 何玲<sup>1</sup>, 袁培燕<sup>1,2</sup>

(1. 河南师范大学 计算机与信息工程学院, 河南 新乡 453007; 2. 智慧商务与物联网技术河南省工程实验室, 河南 新乡 453007)

(\*通信作者电子邮箱 maowt.mail@gmail.com)

**摘要:**针对现有机器学习算法难以有效提高贯序不均衡数据分类问题中少类样本分类精度的问题,提出一种基于混合采样策略的在线贯序极限学习机。该算法可在提高少类样本分类精度的前提下,减少多类样本的分类精度损失,主要包括离线和在线两个阶段:离线阶段采用均衡采样策略,利用主曲线分别构建多类和少类样本的可信区域,在不改变样本分布特性的前提下,利用可信区域扩充少类样本和削减多类样本,进而得到均衡的离线样本集,建立初始模型;在线阶段仅对贯序到达的多类数据进行欠采样,根据样本重要度挑选最具价值的多类样本,进而动态更新网络权值。通过理论分析证明所提算法在理论上存在损失信息上界。采用 UCI 标准数据集和实际的澳门空气污染预报数据进行仿真实验,结果表明,与现有在线贯序极限学习机(OS-ELM)、极限学习机(ELM)和元认知在线贯序极限学习机(MCOS-ELM)算法相比,所提算法对少类样本的预测精度更高,且数值稳定性良好。

**关键词:**极限学习机;在线贯序数据;不均衡分类;主曲线

**中图分类号:** TP181 **文献标志码:** A

### Hybrid sampling extreme learning machine for sequential imbalanced data

MAO Wentao<sup>1,2\*</sup>, WANG Jinwan<sup>1</sup>, HE Ling<sup>1</sup>, YUAN Peiyan<sup>1,2</sup>

(1. College of Computer and Information Engineering, Henan Normal University, Xinxiang Henan 453007, China;

2. Engineering Laboratory of Intellectual Business and Internet of Things Technologies, Henan Province, Xinxiang Henan 453007, China)

**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 hybrid sampling online extreme learning machine on sequential imbalanced data was proposed. This algorithm could improve the classification accuracy of minor class as well as reduce the loss of classification accuracy of major class, which contained two stages. In offline stage, the principal curve was introduced to model the confidence regions of minor class and major class respectively based on the strategy of balanced samples. Over-sampling of minority and under-sampling of majority was achieved based on confidence region. Then the initial model was established. In online stage, only the most valuable samples of major class were chosen according to the sample importance, and then the network weight was updated dynamically. The proposed algorithm had upper bound of the information loss through the theoretical proof. The experiment was taken on two UCI datasets and the real-world air pollutant forecasting dataset of Macao. The experimental results show that, compared with the existing methods such as Online Sequential Extreme Learning Machine (OS-ELM), Extreme Learning Machine (ELM) and Meta-Cognitive Online Sequential Extreme Learning Machine (MCOS-ELM), the proposed method has higher prediction precision and better numerical stability.

**Key words:** Extreme Learning Machine (ELM); online sequential data; imbalanced data classification; principal curve

## 0 引言

在实际工程问题中,在线贯序数据往往同时具有类别严重不均衡的特点。传统的分类算法通常是以提高样本的总体分类精度为目标,在解决此类问题时,往往会产生“虚假的”分类结果,即少类样本的分类精度远远低于多类样本。例如,100个样本中仅有10个少类样本,此时即使少类样本全部误判、同时多类样本全部分类正确,整体分类精度仍可达到90%,显然这一结果毫无意义。而实际的在线预测问题中,对

少类样本的识别通常更为重要,且少类样本的错分代价远大于多类样本,例如气象预测中将恶劣天气预判为良好、临床上将癌症患者诊断为正常。因此,提高在线贯序不均衡数据中少类样本的分类精度是目前的一个研究热点。

目前针对不均衡数据的分类方法主要包括基于数据和基于算法的策略。基于数据的策略强调通过欠采样和过采样来改善数据的类别不均衡程度。然而传统的采样方法易受随机性影响,稳定性较差。为解决此问题,张冬雪等<sup>[1]</sup>提出一种谱聚类欠采样方法,在核空间里使用谱聚类方法选择具有代

**收稿日期:** 2015-03-25; **修回日期:** 2015-05-12。 **基金项目:** 国家自然科学基金资助项目(U1204609); 中国博士后科学基金资助项目(2014M550508); 河南省高校科技创新人才资助计划项目(15HASTIT022); 河南省高校青年骨干教师资助计划项目(2014GGJS-046)。

**作者简介:** 毛文涛(1980-),男,河南新乡人,副教授,博士,CCF会员,主要研究方向:机器学习、弱信号检测; 王金婉(1991-),女,河南济源人,硕士研究生,CCF会员,主要研究方向:机器学习、模式识别; 何玲(1990-),女,河南鹤壁人,硕士研究生,主要研究方向:泛化性理论; 袁培燕(1978-),男,河南邓州人,副教授,主要研究方向:移动计算。

表性的多类样本,并和少类样本一起进行训练,提高分类性能。杨智明等<sup>[2]</sup>提出了一种自适应 SMOTE (Synthetic Minority Over-sampling TEchnique) 算法,根据样本集内部特征,自适应调整近邻选择策略,控制样本合成质量。上述研究虽然在一定程度上提高了少类样本的分类精度,但通常存在两类问题:1) 未考虑样本本身的分布特性,导致均衡后的样本数据缺乏可信度,这一点对于在线贯序数据尤为突出;2) 欠采样会使多类样本信息大量丢失,可能会导致多类样本分类精度的大幅度下降<sup>[3]</sup>。因此,并不能很好地应用于在线贯序不平衡数据分类问题。基于算法的策略强调在线学习算法对不平衡数据的学习速度和判别能力。其中,在线贯序极限学习机 (Online Sequential Extreme Learning Machine, OS-ELM)<sup>[4]</sup>作为一种单隐层前馈神经网络,具有极快的学习速度和良好的泛化能力,目前已在许多实际问题中得到了广泛应用。但 OS-ELM 在解决严重类别不平衡问题时,容易造成对少类样本的误判,从而限制了应用范围。

由上述分析可知,对于在线贯序不平衡数据而言,遵循样本的分布特性进行均衡采样,是提高少类样本分类精度的关键。为此,王金婉等<sup>[5]</sup>同时从数据策略和算法策略入手,提出了基于不平衡样本重构的加权在线贯序极限学习机,引入主曲线构建少类样本的可信区域并过采样,进而建立初始模型,根据训练误差为贯序到达的样本赋以相应大小的权重,进而动态更新网络模型。然而,该算法仅采用过采样对离线训练样本进行了重构,对在线样本的不均衡现象并未作相应处理,且该算法并未考虑在线样本与离线样本的关系,导致两个阶段孤立进行,也缺乏对样本重构合理性的理论分析。为此,本文提出一种基于混合采样策略的在线贯序极限学习机 (Hybrid sampling Online Sequential Extreme Learning Machine on imbalanced data, HOS-ELM),同时遵循样本的分布特性和考虑样本之间的联系,在提高少类样本分类精度的同时,尽可能减少多类样本的分类精度损失。在离线阶段引入主曲线分别构建多类样本和少类样本的可信区域,进而采用混合采样策略分别对多类样本欠采样和对少类样本过采样,得到符合样本分布特性的均衡样本集,并建立初始模型;为解决在线样本的不均衡现象,对多类样本欠采样,根据样本重要度指标筛选贯序数据中最具价值的多类样本,进而动态更新网络权重。同时受文献<sup>[6]</sup>的启发,在理论上给出了在线分类过程中舍弃样本的损失信息上界,从而证明了所提算法采用主曲线构建可信区域、在此基础上进行混合采样的合理性。最后在 UCI 标准数据和实际的澳门气象数据上验证了所提算法的有效性。

## 1 相关工作

### 1.1 在线贯序极限学习机

极限学习机 (Extreme Learning Machine, ELM)<sup>[7]</sup>是一种单隐层前馈神经网络。该算法随机挑选输入层参数,直接使用 Moore-Penrose 广义逆,即可求得最小 L2 范数的输出层权重。整个学习过程只有隐神经元个数可调,结构简单,具有非常快的学习速度和优秀的泛化能力。在线贯序极限学习机是在原始 ELM 算法的基础上提出的在线增量式快速学习算法。由文献<sup>[8]</sup>可知,算法步骤分为两个阶段<sup>[8]</sup>:

步骤 1 初始化阶段。

从给定训练集  $D = \{(\mathbf{x}_i, t_i), i = 1, 2, \dots, N\}$  中选取部分数据集  $D_0 = \{(\mathbf{x}_i, t_i), i = 1, 2, \dots, N_0\}$ , 其中  $N_0 \geq L_0$ 。

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

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

3) 置  $k = 0$ 。

步骤 2 序列学习阶段。

1) 学习第  $k+1$  个数据:  $d_{k+1} = (\mathbf{x}_{N_0+k+1}, t_{N_0+k+1})$ ;

2) 令  $\mathbf{T}_{k+1} = [t_{N_0+k+1}]^T$ , 计算新学习数据的隐层输出矩阵  $\mathbf{H}_{k+1}$ :

$$\mathbf{H}_{k+1} = \begin{bmatrix} g(\mathbf{w}_1 \cdot \mathbf{x}_{N_0+k+1} + b_1) & g(\mathbf{w}_2 \cdot \mathbf{x}_{N_0+k+1} + b_2) & \dots \\ g(\mathbf{w}_L \cdot \mathbf{x}_{N_0+k+1} + b_L) \end{bmatrix}_{1 \times L} \quad (1)$$

3) 计算输出权值  $\boldsymbol{\beta}^{k+1}$ :

$$\begin{cases} \mathbf{P}_{k+1} = \mathbf{P}_k - \mathbf{P}_k \mathbf{H}_{k+1}^T (\mathbf{I} + \mathbf{H}_{k+1} \mathbf{P}_k \mathbf{H}_{k+1}^T)^{-1} \mathbf{H}_{k+1} \mathbf{P}_k \\ \boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k + \mathbf{P}_{k+1} \mathbf{H}_{k+1}^T (\mathbf{T}_{k+1} - \mathbf{H}_{k+1} \boldsymbol{\beta}^k) \end{cases} \quad (2)$$

4) 置  $k = k+1$ , 若  $k > N$ , 则算法结束; 否则返回步骤 2 的 1)。

### 1.2 主曲线

主曲线是通过数据集“中间”的光滑无参数曲线,是线性主成分分析的非线性推广,也是嵌入高维数据的非欧空间的一维流形表述<sup>[9]</sup>。主曲线理论基于一定概率分布下曲线的“自相合”性,即曲线上任意点都是所有投影至该点的数据点的条件对偶。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”。主曲线的算法步骤<sup>[9]</sup>可概括为:

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  $\mathbf{x} \in \mathbf{R}^d$ , 求投影指标:

$$\lambda_{f^{(j)}}(\mathbf{x}) = \max_t \{t: \|\mathbf{x} - f(t)\| = \min_{\tau} \|\mathbf{x} - f(\tau)\|\} \quad (3)$$

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

步骤 4 如果  $1 - \Delta(f^{(j+1)}) / \Delta(f^{(j)})$  小于预设阈值, 则停止 (其中  $\Delta(f^{(j)})$  表示点  $\mathbf{x}$  到曲线  $f$  的欧氏平方距离); 否则, 令  $j = j+1$ , 转步骤 2。

## 2 基于混合采样策略的在线贯序极限学习机

为提高不平衡在线贯序数据中少类样本的分类精度,同时尽可能减少多类样本的分类精度损失,本文同时遵循样本的分布特性和考虑样本本身的重要性,提出一种基于混合采样策略的在线贯序极限学习机,从离线和在线两个角度,对不平衡样本进行处理。

首先给出几个定义。给定多类样本集  $D = \{(\mathbf{x}_i, t_i), i = 1, 2, \dots, N\}$  和  $D_1 = \{(\mathbf{y}_i, t_i), i = 1, 2, \dots, N\}$ , 其中:  $\mathbf{x}_i$  和  $\mathbf{y}_i$  表示  $m$  维向量, 维数大小代表样本特征个数;  $t_i = 0$ , 表示多类样本。

定义 1 样本权重。构建基于多类样本集  $D$  的主曲线, 则  $D$  中每个样本  $\mathbf{x}_i$  对应的样本权重定义为:

$$w_i = 1 - f_i / \sum_{j=1}^N f_j \quad (4)$$

其中 $f_i$ 为样本点 $\mathbf{x}_i$ 到主曲线的投影距离。由式(4)可知,样本 $\mathbf{x}_i$ 到主曲线的距离越小,其对应的样本权重越大;反之 $f_i$ 越大, $w_i$ 越小。

**定义2** 自信息量。指某个样本在其所在样本集中所含的信息量大小。多类样本集 $D$ 内每个样本 $\mathbf{x}_i$ 的自信息量定义如下:

$$I(\mathbf{x}_i) = -\log(w_i) \quad (5)$$

**定义3** 相对信息量。指多类样本集 $D$ 中的样本 $\mathbf{x}_i$ 相对于多类样本集 $D_1$ 中的对应样本 $\mathbf{y}_i$ 所含的信息量大小。相对信息量在一定程度上可以反映出两个样本之间的关联性,其定义如下:

$$I(\mathbf{x}_i, \mathbf{y}_i) = -\log(w_i/w_i') \quad (6)$$

其中: $w_i, w_i'$ 分别为 $\mathbf{x}_i$ 和 $\mathbf{y}_i$ 在对应数据集中的样本权重。

**定义4** 样本重要度。结合定义2和定义3,得到样本 $\mathbf{x}_i$ 的样本重要度:

$$\text{Value}(\mathbf{x}_i) = I(\mathbf{x}_i)I(\mathbf{x}_i, \mathbf{y}_i) \quad (7)$$

$\text{Value}(\mathbf{x}_i)$ 值越大,表明对应样本 $(\mathbf{x}_i, t_i)$ 在多类数据集 $D_1$ 中的价值越大,即所包含的信息量越多,是最具价值样本。

## 2.1 初始离线阶段

离线阶段,在保证不改变样本分布特性的前提下,采用混合采样策略对不均衡样本重构,并建立初始模型。其基本思

$$\mathbf{H}_0 = \begin{bmatrix} g(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & g(\mathbf{w}_2 \cdot \mathbf{x}_1 + b_2) & \cdots & g(\mathbf{w}_L \cdot \mathbf{x}_1 + b_L) \\ g(\mathbf{w}_1 \cdot \mathbf{x}_2 + b_1) & g(\mathbf{w}_2 \cdot \mathbf{x}_2 + b_2) & \cdots & g(\mathbf{w}_L \cdot \mathbf{x}_2 + b_L) \\ \vdots & \vdots & \ddots & \vdots \\ g(\mathbf{w}_1 \cdot \mathbf{x}_{N_0} + b_1) & g(\mathbf{w}_2 \cdot \mathbf{x}_{N_0} + b_2) & \cdots & g(\mathbf{w}_L \cdot \mathbf{x}_{N_0} + b_L) \end{bmatrix}_{N_0 \times L}$$

输出向量为 $\mathbf{T}_0 = [t_1 \ t_2 \ \cdots \ t_{N_0}]^T$ ,则输出权值为:

$$\boldsymbol{\beta}_0 = \mathbf{H}_0^T \mathbf{T}_0 \quad (9)$$

式中:

$$\mathbf{H}_0^+ = (\mathbf{H}_0^T \mathbf{H}_0)^{-1} \mathbf{H}_0^T \quad (10)$$

令 $\mathbf{U}_0 = (\mathbf{H}_0^T \mathbf{H}_0)^{-1}$ ,则式(10)即为 $\mathbf{H}_0^+ = \mathbf{U}_0 \mathbf{H}_0^T$ 。

令 $k = 0$ 。

## 2.2 在线贯序阶段

设第 $k+1$ 步贯序到达的新样本块为 $\Omega_{k+1} = \{(\mathbf{x}_i, t_i)\}$ ,  $i = k + N_0 + 1, k + N_0 + 2, \dots, k + N_0 + \text{Block}$ ,其中 $\text{Block}$ 表示第 $k+1$ 步添加的数据个数。对新到样本块 $\Omega_{k+1}$ ,根据 $t_i$ 的值,将其分成多类和少类两部分,即 $\Phi_d$ 和 $\Phi_s$ 。

对于集合 $\Phi_d = \{(\mathbf{x}_i, t_i), i = 1, 2, \dots, M\}$ ,根据式(3)分别计算每个多类样本 $\mathbf{x}_i$ 到主曲线 $P_1$ 的投影距离 $d_i$  ( $i = 1, 2, \dots, M$ ),并根据式(4)计算其对应的样本权重 $w_i$ 。则根据定义2,集合 $\Phi_d$ 中各样本所含的自信息量大小即为 $I(\mathbf{x}_i) = -\log(w_i)$ 。

从初始阶段均衡的多类样本集 $S_1$ 中,随机选取 $M$ 个样本,得 $F = \{(\mathbf{y}_i, t_i), i = 1, 2, \dots, M\}$ 。同样的方法,计算投影距离 $d_i'$ ,并计算样本权重 $w_i'$ 。根据定义3,集合 $\Phi_d$ 和 $F$ 中对应样本的相对信息量为 $I(\mathbf{x}_i, \mathbf{y}_i) = -\log(w_i/w_i')$ 。

根据定义4,计算集合 $\Phi_d$ 中各样本的样本重要度 $\text{Value}(\mathbf{x}_i) = I(\mathbf{x}_i)I(\mathbf{x}_i, \mathbf{y}_i)$ 。

由此得到贯序到达样本块中多类样本集 $\Phi_d = \{(\mathbf{x}_i, t_i), i = 1, 2, \dots, M\}$ 中各样本对应的样本重要度。由定义可知, $\text{Value}(\mathbf{x}_i)$ 越大,表明样本 $(\mathbf{x}_i, t_i)$ 在 $\Phi_d$ 中的价值越大,在贯序学习中所起的作用越大。为均衡在线阶段样本的类别不均衡

想是采用主曲线分别构建多类和少类样本的可信区域,进而削减多类样本和扩充少类样本,得到均衡的训练样本集,最后建立初始模型。

对给定初始样本集 $D = \{(\mathbf{x}_i, t_i), i = 1, 2, \dots, N\}$ ,分别构建多类和少类样本的主曲线 $P_1$ 和 $P_2$ ,基于主曲线设定上下阈值 $\eta_1$ 和 $\eta_2$ ,进而得到以主曲线为中心的带状区域,即可信区域。对多类样本,选择可信区域内的样本点,得到多类样本集 $S_1$ ;对少类样本,在可信区域内随机插值,生成新的样本点,与原始少类样本集合并得到最终的少类样本集 $S_2$ 。如图1所示(以多类样本为例,其中实心点表示原始多类样本,空心点表示欠采样后的多类样本点,即 $S_1$ )。合并 $S_1$ 和 $S_2$ ,得到均衡的初始训练样本集 $D_0 = \{(\mathbf{x}_i, t_i), i = 1, 2, \dots, N_0\}$ 。

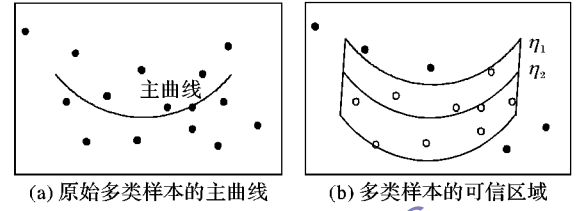


图1 可信区域形成过程

给定隐层激活函数 $g(\mathbf{x})$ 和隐层神经元个数 $L$ ,随机选取输入权值 $\mathbf{w}_i$ 和偏置 $b_i$  ( $i = 1, 2, \dots, L$ ),计算隐层输出矩阵<sup>[7]</sup>:

$$\mathbf{H}_0 = \begin{bmatrix} g(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & g(\mathbf{w}_2 \cdot \mathbf{x}_1 + b_2) & \cdots & g(\mathbf{w}_L \cdot \mathbf{x}_1 + b_L) \\ g(\mathbf{w}_1 \cdot \mathbf{x}_2 + b_1) & g(\mathbf{w}_2 \cdot \mathbf{x}_2 + b_2) & \cdots & g(\mathbf{w}_L \cdot \mathbf{x}_2 + b_L) \\ \vdots & \vdots & \ddots & \vdots \\ g(\mathbf{w}_1 \cdot \mathbf{x}_{N_0} + b_1) & g(\mathbf{w}_2 \cdot \mathbf{x}_{N_0} + b_2) & \cdots & g(\mathbf{w}_L \cdot \mathbf{x}_{N_0} + b_L) \end{bmatrix}_{N_0 \times L} \quad (8)$$

问题,同时避免欠采样过程中多类样本信息的大量丢失,仅选取最具价值多类样本进行训练,即从 $\Phi_d$ 中选取前 $m$ 个样本重要度最大的样本,得到集合 $\Phi_d'$ 。

合并集合 $\Phi_d'$ 和 $\Phi_s$ ,得到新样本块 $\Phi = \{(\mathbf{x}_i, t_i), i = 1, 2, \dots, m + \text{Block} - M\}$ 。则 $\Phi$ 对应的神经元矩阵为 $\mathbf{H}_\Phi = [\mathbf{h}_{k+N_0+1} \ \mathbf{h}_{k+N_0+2} \ \cdots \ \mathbf{h}_{k+N_0+m+\text{Block}-M}]$ 。此时,隐层输出矩阵为 $\mathbf{H}_{k+1} = [\mathbf{H}_k^T \ \mathbf{H}_\Phi^T]^T$ ,输出向量为 $\mathbf{T}_{k+1} = [\mathbf{T}_k^T \ \mathbf{T}_\Phi^T]^T$ 。更新网络权值:

$$\boldsymbol{\beta}_{k+1} = \mathbf{H}_{k+1}^+ \mathbf{T}_{k+1} \quad (11)$$

其中: $\mathbf{H}_{k+1}^+ = (\mathbf{H}_{k+1}^T \mathbf{H}_{k+1})^{-1} \mathbf{H}_{k+1}^T$ 。令 $\mathbf{U}_{k+1} = (\mathbf{H}_{k+1}^T \mathbf{H}_{k+1})^{-1}$ ,则有:

$$\mathbf{H}_{k+1}^+ = \mathbf{U}_{k+1} \mathbf{H}_{k+1}^T \quad (12)$$

因为:

$$\mathbf{H}_{k+1}^T \mathbf{H}_{k+1} = [\mathbf{H}_k^T \ \mathbf{H}_\Phi^T][\mathbf{H}_k^T \ \mathbf{H}_\Phi^T]^T = \mathbf{H}_k^T \mathbf{H}_k + \mathbf{H}_\Phi^T \mathbf{H}_\Phi \quad (13)$$

即:

$$\mathbf{U}_{k+1}^{-1} = \mathbf{U}_k^{-1} + \mathbf{H}_\Phi^T \mathbf{H}_\Phi \quad (14)$$

对式(14)两端求逆,根据 Sherman-Morrison 矩阵求逆引理可得 $\mathbf{U}_{k+1}$ 的递推表达式<sup>[7]</sup>:

$$\mathbf{U}_{k+1} = (\mathbf{U}_k^{-1} + \mathbf{H}_\Phi^T \mathbf{H}_\Phi)^{-1} = \mathbf{U}_k - \frac{\mathbf{U}_k \mathbf{H}_\Phi^T \mathbf{H}_\Phi \mathbf{U}_k}{\mathbf{I} + \mathbf{H}_\Phi \mathbf{U}_k \mathbf{H}_\Phi^T} \quad (15)$$

因此, $\mathbf{U}_{k+1}$ 可以在 $\mathbf{U}_k$ 的基础上计算得到,从而大大简化了计算量。将式(15)代入式(12)即得到 $\mathbf{H}_{k+1}^+$ ,并根据式(11)更新网络权值得到 $\boldsymbol{\beta}_{k+1}$ 。

综合上述推导过程,HOS-ELM 算法流程如图2所示。

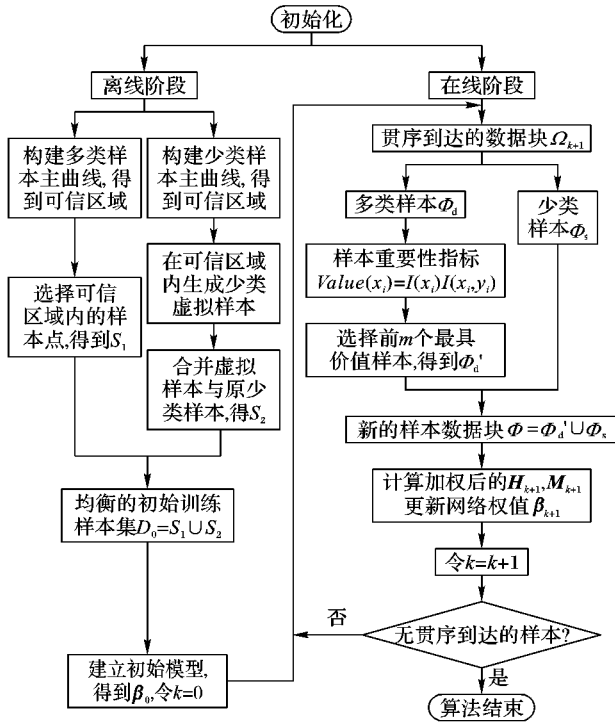


图2 HOS-ELM 算法流程

### 3 理论分析

由2.2节叙述可知,对贯序到达的不均衡数据块,根据样本重要度对多类样本  $\Phi_d = \{(x_i, t_i), i = 1, 2, \dots, M\}$  进行欠采样,得到  $m$  个最具价值多类样本  $\Phi_d'$ 。为说明该算法的合理性,本文从信息熵的角度证明本算法中欠采样过程的信息损失存在上界。

设每一次贯序学习阶段的损失样本集合为  $\Psi = \{(x_j, t_j), j = 1, 2, \dots, M - m\}$ , 其中  $x_j$  对应的样本权重为  $w_j$ , 则损失样本集  $\Psi$  的总体样本权重之和为  $\sum_{k=1}^{M-m} w_k = \sum_{k=1}^{M-m} (1 - d_k / \sum_{j=1}^M d_j)$ 。易知,  $\sum_{j=1}^M d_j$  表示多类样本集  $\Phi_d$  中所有样本到主曲线的投影距离之和,对给定的贯序数据块,  $\sum_{j=1}^M d_j$  为定值,因此,令  $\sum_{j=1}^M d_j = \Delta_1$ , 则  $\Psi$  的总体样本损失权重之和为  $\sum_{k=1}^{M-m} w_k = \sum_{k=1}^{M-m} (1 - d_k / \Delta_1) = (M - m) - \frac{1}{\Delta_1} \sum_{k=1}^{M-m} d_k$ 。又知,与损失样本集  $\Psi$  对应的离线多类样本点为  $F$  的子集,记为  $F_1 = \{(y_j, t_j), j = 1, 2, \dots, M - m\}$ , 其中:样本  $y_j$  的样本权重为  $w_j'$ , 则  $F_1$  的总体样本权重之和为  $\sum_{k=1}^{M-m} w_k' = \sum_{k=1}^{M-m} (1 - d_k' / \sum_{j=1}^M d_j')$ 。同上,  $\sum_{j=1}^M d_j'$  表示  $F$  中所有样本到主曲线的投影距离之和,因此为定值,令  $\sum_{j=1}^M d_j' = \Delta_2$ , 则  $F_1$  的总体样本权重之和为  $\sum_{k=1}^{M-m} w_k' = \sum_{k=1}^{M-m} (1 - d_k' / \Delta_2) = (M - m) - \frac{1}{\Delta_2} \sum_{k=1}^{M-m} d_k'$ 。由于熵可以表示一个数据集所包含的信息量,因此,本文从熵的角度给出损失样本集  $\Psi$  的整体信息量

上界和整体相对信息量上界。

**定理1** 令  $H(\Psi)$  表示欠采样过程中的整体信息损失, 则  $H(\Psi) \leq ((M - m) - \frac{1}{\Delta_1} \sum_{k=1}^{M-m} d_k) \log \frac{M - m}{(M - m) - \frac{1}{\Delta_1} \sum_{k=1}^{M-m} d_k}$ ;

且整体信息损失大小  $H(\Psi)$  的上界仅与样本集  $\Psi$  中所有样本点到主曲线的投影距离之和  $\sum_{k=1}^{M-m} d_k$  有关。

**证明** 根据熵的定义有  $H(\Psi) = - \sum_{i=1}^{M-m} w_i \log(w_i)$ 。根据最大熵原理,当每一个  $w_i$  都取相同的值

$$\left( (M - m) - \frac{1}{\Delta_1} \sum_{k=1}^{M-m} d_k \right) / (M - m)$$

时,  $H(\Psi)$  达到最大值。则:

$$H(\Psi) \leq - \sum_{i=1}^{M-m} \frac{(M - m) - \frac{1}{\Delta_1} \sum_{k=1}^{M-m} d_k}{M - m} \log \frac{(M - m) - \frac{1}{\Delta_1} \sum_{k=1}^{M-m} d_k}{M - m} = \left( (M - m) - \frac{1}{\Delta_1} \sum_{k=1}^{M-m} d_k \right) \log \frac{M - m}{(M - m) - \frac{1}{\Delta_1} \sum_{k=1}^{M-m} d_k} \quad (16)$$

由式(16),  $H(\Psi)$  上界仅和  $\sum_{k=1}^{M-m} d_k$  有关,  $\sum_{k=1}^{M-m} d_k$  越大, 该界越小。

**定理2** 令  $R$  表示  $\Psi$  对于离线阶段样本集  $F_1$  的整体相对信息损失,  $H(F_1)$  表示  $F_1$  所含的整体信息, 则  $R \leq ((M - m) - \frac{1}{\Delta_2} \sum_{k=1}^{M-m} d_k') \log \frac{M - m}{(M - m) - \frac{1}{\Delta_2} \sum_{k=1}^{M-m} d_k'}$ ; 且整体相对信息损失大的上界仅与样本集  $F_1$  中所有样本到主曲线的投影距离之和  $\sum_{k=1}^{M-m} d_k'$  有关。

**证明** 根据相对熵的定义, 有:

$$R = \sum_{i=1}^{M-m} w_i \log(w_i / w_i') = \sum_{i=1}^{M-m} (w_i \log(w_i) - w_i \log(w_i')) = -H(\psi) + H(\psi, F_1) \leq H(\psi) + H(F_1) - H(\psi) = H(F_1) \quad (17)$$

其中  $H(\psi, F_1)$  表示  $\psi$  和  $F_1$  的叉熵。根据定理1得:

$$H(F_1) \leq - \sum_{i=1}^{M-m} \frac{(M - m) - \frac{1}{\Delta_2} \sum_{k=1}^{M-m} d_k'}{M - m} \log \frac{(M - m) - \frac{1}{\Delta_2} \sum_{k=1}^{M-m} d_k'}{M - m} = \left( (M - m) - \frac{1}{\Delta_2} \sum_{k=1}^{M-m} d_k' \right) \log \frac{M - m}{(M - m) - \frac{1}{\Delta_2} \sum_{k=1}^{M-m} d_k'} \quad (18)$$

即:  $R \leq \left( (M - m) - \frac{1}{\Delta_2} \sum_{k=1}^{M-m} d_k' \right) \log \frac{M - m}{(M - m) - \frac{1}{\Delta_2} \sum_{k=1}^{M-m} d_k'}$ 。



定理1和定理2在理论上证明了根据样本点到主曲线的投影距离挑选最具价值样本的有效性。考虑极端情况,若部分样本到主曲线的投影距离(即 $\sum_{k=1}^{M-m} d_k$ 和 $\sum_{k=1}^{M-m} d_k'$ )趋近于无穷大,则对应的信息损失上界趋近于无穷小,这意味着欠采样过程中删除该部分样本对整体信息可忽略不计。这证明了本算法采用主曲线构建可信区域,在此基础上进行混合采样的合理性。

#### 4 仿真实验

为进一步验证所提算法的有效性,本文采用UCI标准数据集和实际气象数据进行仿真实验,分别将ELM、OS-ELM和针对在线不均衡问题的元认知在线序列极限学习机(Meta-Cognitive Online Sequential Extreme Learning Machine, MCOS-ELM)<sup>[10]</sup>与本文算法HOS-ELM进行对比。由于ELM本身的随机性可能导致实验结果的不稳定,本文实验结果均为重复50次的平均值,以尽可能减少误差的影响。在训练前,所有样本被线性归一化到 $[-1,1]$ 。

##### 4.1 UCI标准数据集

限于篇幅,此处选择两个UCI标准数据集Pima和Abalone<sup>[11]</sup>进行仿真实验。两个数据集的统计结果如表1所示。

表1 Pima和Abalone数据集

数据集	属性数	离线训练样本数	在线训练样本数	测试样本数
Pima	8	300	200	100
Abalone	8	700	400	337

离线阶段,利用Relief特征选择法对Pima和Abalone均提取第8个属性作为主特征,分别构建多类和少类样本的主曲线。根据2.1节所述算法,对离线样本重构,得到均衡的离线训练样本集,见表2。

表2 均衡离线数据前后的样本数

数据集	未处理前		处理后	
	多类	少类	多类	少类
Pima	250	50	169	168
Abalone	638	62	371	367

给定隐层激活函数为径向基函数(Radial Basis Function, RBF),隐层节点分别为30、45,运行50次取均值,四种模型的性能如表3、4所示。

从表3和表4可以看出,尽管HOS-ELM的总体训练精度和总体测试精度均低于OS-ELM和ELM,但HOS-ELM的少类训练精度和少类测试精度均明显高于两种经典算法,表明本文算法HOS-ELM对提高少类样本分类精度是有效的。不难发现,MCOS-ELM对少类样本的分类精度也优于OS-ELM和ELM,但由于其并未考虑样本的分布特性和样本重要度,在多数样本的测试精度上,MCOS-ELM低于HOS-ELM,表明HOS-ELM可在一定程度上缓解多数样本分类精度的损失;且在两个标准数据集上,本文算法HOS-ELM对少类样本的分类精度均高于MCOS-ELM,进一步表明HOS-ELM算法可有效提高少类样本的分类效果。

表3 四种模型在Pima标准数据集上的性能

算法	训练时间/s	测试时间/s	训练精度/%			测试精度/%		
			少类	多类	总体	少类	多类	总体
HOS-ELM	1.708 2	0.001 6	76.77	85.80	82.42	52.94	91.69	85.10
OS-ELM	0.159 1	0.059 9	34.58	97.82	87.32	26.47	96.51	84.60
ELM	0.006 2	0.012 5	32.65	97.43	86.68	24.12	96.51	84.20
MCOS-ELM	0.035 4	0.037 2	68.85	85.20	79.33	47.06	87.59	83.40

表4 四种模型在Abalone标准数据集上的性能

算法	训练时间/s	测试时间/s	训练精度/%			测试精度/%		
			少类	多类	总体	少类	多类	总体
HOS-ELM	3.270 8	0.018 7	89.57	97.48	94.65	93.10	94.83	94.27
OS-ELM	0.192 4	0.020 8	60.07	98.40	94.88	75.86	98.27	96.34
ELM	0.041 6	0.003 1	58.75	98.43	94.79	72.41	98.38	96.14
MCOS-ELM	0.064 2	0.066 5	88.78	96.47	95.16	90.80	93.02	92.66

为进一步证明HOS-ELM的数值稳定性,描绘随隐节点数变化,四种算法的少类测试精度变化曲线,如图3所示。其中,每个节点值均为运行10次所得结果的平均值。

从图3可以看出,随隐节点数变化,HOS-ELM的少类测试精度均明显优于其他三种算法,且曲线较为平滑,进一步证明了HOS-ELM对少类样本的识别能力更强,且数值稳定性良好。

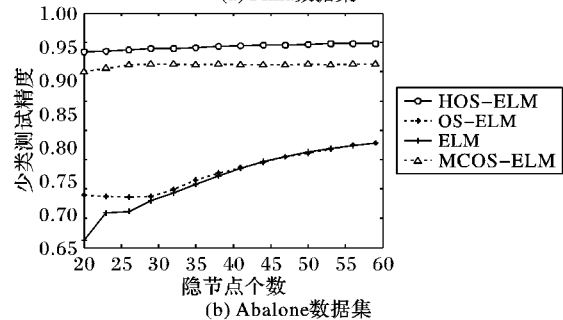
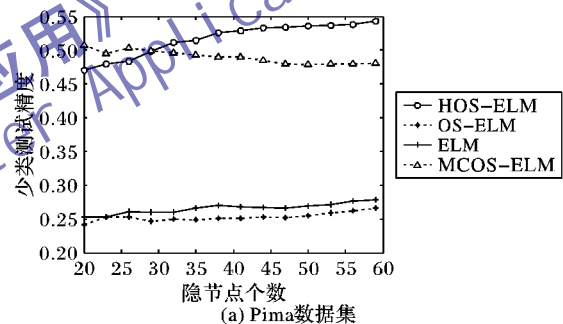


图3 随隐节点数变化少类测试精度的变化

##### 4.2 澳门气象数据

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大于空气严重污染的天数,因此是一种典型的在线贯序不均衡分类问题。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<sup>[12]</sup>进行仿真实验,分别采用ELM、OS-ELM和MCOS-ELM<sup>[9]</sup>与本文算法HOS-ELM进行对比。

###### 4.2.1 数据处理

给定训练数据集 $D = (x, t)$ ,  $x$ 表示输入变量,即当天的PM10、SO<sub>2</sub>、NO<sub>2</sub>、O<sub>3</sub>的浓度值,即 $x = (d(\text{PM10}), d(\text{SO}_2), d(\text{NO}_2), d(\text{O}_3))$ ;  $t$ 是输出向量即第二天的PM10的值,即 $t = d + 1(\text{PM10})$ 。

#### 4.2.2 实验结果

为验证 HOS-ELM 的有效性,利用 2011 年到 2013 年澳门氹仔岛格兰德气象站收集的序列数据进行实验。其中,2011 年的数据作为初始离线训练样本,2012 年数据作为在线训练样本,2013 年的数据作为测试样本。

对 2011 年初始离线样本,利用 Relief 特征选择法选择第一个属性值,即  $d(\text{PM}_{10})$  作为主特征,分别构建多类样本和少类样本的主曲线,并基于主曲线生成可信区域,如图 4 所示。

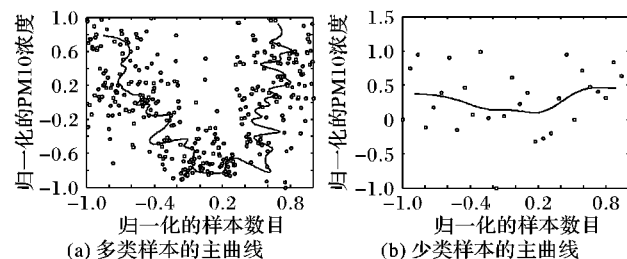


图 4 构建离线阶段多类和少类样本的主曲线

在可信区域内,分别对多类样本进行筛选和扩充少类样本,得到均衡的初始训练样本集。均衡样本前后 2011 年的样本数情况如表 5。

表 5 均衡样本前后 2011 年样本数变化情况

类别	均衡前		均衡后	
	样本数量	占比/%	样本数量	占比/%
少类样本	30	8.22	151	46.60
多类样本	335	91.78	173	53.40

利用处理后的均衡离线样本集建立初始模型。给定隐层激活函数为 RBF 核函数,对 2012 的数据进行在线贯序训练。

设定隐节点个数为 30,表 6 给出了四种模型的性能比较,所有数据为运行 50 次的平均值。

由表 6 可知,在处理实际不均匀问题时,HOS-ELM 有效地提高了少类样本的训练精度和测试精度,分别比其他三种算法提高了 66.28%,65.63%,5.51% 和 37.83%,46.13%,7.63%。且与 MCOS-ELM 相比,HOS-ELM 对多类样本的分类精度损失较低,有力地表明了在处理不均匀问题时,根据样本重要度挑选最具价值样本的重要性。虽然 HOS-ELM 总体的测试精度略低于两种经典算法,但有效提高了少类样本的分类精度,大大降低了少类样本错分的代价,对实际气象预测更具有实际的应用价值。

表 6 四种模型对澳门气象数据预测精度平均值比较

算法	训练时间/s	测试时间/s	训练精度/%			测试精度/%		
			少类	多类	总体	少类	多类	总体
HOS-ELM	7.6596	0.0107	86.94	95.58	92.55	83.71	86.58	84.99
OS-ELM	0.2558	0.0016	20.66	99.28	92.71	45.88	98.15	90.82
ELM	0.0125	0.0125	21.31	99.10	92.60	41.18	98.34	90.33
MCOS-ELM	0.5076	0.0596	81.43	94.47	91.33	76.08	84.89	82.64

## 5 结语

本文提出了一种基于混合采样策略的在线极限学习机,通过引入主曲线提取样本的分布特性,并根据样本重要度,挑选在线贯序数据中最具价值的多类样本,保证在减少多类样本分类精度损失的前提下,提高少类样本的分类精度,并从理

论上证明了该算法的合理性,对解决实际气象问题具有重要的理论和工程意义。继续完善和改进算法性能,是我们下一步研究的方向。

#### 参考文献:

- [1] ZHANG D, TAO X. Unbalanced data classification under-sampling algorithm based on SVM for research and application [D]. Harbin: Harbin Engineering University, 2013: 18-34. (张冬雪,陶新民.基于欠采样不平衡数据 SVM 算法与应用[D].哈尔滨:哈尔滨工程大学,2013:18-34.)
- [2] YANG Z, QIAO L, PENG X. Research on data mining method for imbalanced dataset based on improved SMOTE [J]. Acta Electronica Sinica, 2007, 35(B12): 22-26. (杨智明,乔立岩,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(B12):22-26.)
- [3] ZENG Z, WU Q, LIAO B, et al. A classification method for imbalance data set based on kernel SMOTE [J]. Acta Electronica Sinica, 2009, 37(11): 2489-2495. (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.)
- [4] LIANG N, HUANG G. A fast and accurate online sequential learning algorithm for feedforward networks [J]. IEEE Transactions on Neural Networks, 2006, 17(6): 1411-1423.
- [5] WANG J, MAO W, HE L, et al. Weighted online sequential extreme learning machine based on imbalanced samples-reconstruction [J]. Journal of Computer Applications, 2015, 35(6): 1605-1610. (王金婉,毛文涛,何玲,等.基于不平衡样本重构的加权在线贯序极限学习机[J].计算机应用,2015,35(6):1605-1610.)
- [6] YUAN P, MA H, FU H. Hotspot-entropy based data forwarding in opportunistic social networks [J]. Pervasive and Mobile Computing, 2015, 16, Part A: 136-154.
- [7] HUANG G, 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.
- [8] YANG L, ZHANG R. Online sequential ELM algorithm and its improvement [J]. Journal of Northwest University: Natural Science Edition, 2012, 42(6): 885-896. (杨乐,张瑞.在线序列 ELM 算法及其发展[J].西北大学学报:自然科学版,2012,42(6):885-896.)
- [9] ZHANG J, WANG J. An overview of principal curves [J]. Chinese Journal of Computers, 2003, 26(2): 129-146. (张军平,王钰.主曲线研究综述[J].计算机学报,2003,26(2):129-146.)
- [10] VONG C-M, IP W-F, WONG P-K, et al. Prediction minority class for suspended particulate matters level by extreme learning machine [J]. Neurocomputing, 2014, 128: 136-144.
- [11] NEWMAN D J, HETTICH S, BLAKE C L, et al. UCI repository of machine learning databases. Irvine: University of California, Department of Information and Computer Science [DB/OL]. [2015-02-06]. <http://archive.ics.uci.edu/ml/datasets.html?format=&task=cla&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table>.
- [12] SMG E-publication download page [DB/OL]. [2012-03-16]. [http://www.smg.gov.mo/www/ccaa/pdf/e\\_pdf\\_download.php](http://www.smg.gov.mo/www/ccaa/pdf/e_pdf_download.php).