文章编号: 1001-0920 (2016) 12-2147-08 DOI: 10.13195/j.kzyjc.2015.1257

面向贯序不均衡分类的粒度极限学习机

毛文涛[a,b], 田杨阳[a], 王金婉[a], 何 玲[a]

(河南师范大学 a. 计算机与信息工程学院, b. 河南省高校计算智能与数据挖掘工程技术研究中心, 河南 新乡 453007)

摘 要: 针对现有算法对贯序到达的密度型不均衡数据分类效果不佳的缺陷, 提出一种基于粒度划分的在线贯序极限学习机算法. 离线阶段, 根据数据分布特性对多类样本进行粒度划分, 用粒心代替原有样本, 建立初始模型; 在线阶段, 根据更新后的分布特性对多类边界数据进行二次粒度划分, 替换原有边界数据, 并动态更新网络权值. 理论分析证明该算法存在信息损失上界. 实验结果表明, 该算法能有效提高贯序不均衡数据上的整体泛化性能和分类效率.

关键词: 极限学习机; 粒度划分; 贯序不均衡数据; 欠取样

中图分类号: TP181 **文献标志码:** A

Granular extreme learning machine for sequential imbalanced data

MAO Wen-tao[a,b], *TIAN Yang-yang*[a], *WANG Jin-wan*[a], *HE Ling*[a]

(a. College of Computer and Information Engineering, b. Computational Intelligence and Data Mining Engineering Technology Research Center of Colleges and Universities in He'nan Province, He'nan Normal University, Xinxiang 453007, China. Correspondent: MAO Wen-tao, E-mail: maowt@qq.com)

Abstract: Aiming at the shortcomings of the present classification algorithms on density-based imbalanced data which are selected sequentially, an online sequential extreme learning machine based on granular division is proposed. In the offline stage, majority class samples are divided by using granularity according to the data distribution property, and the centre of granule is introduced for replacing the samples in this granule. In the online stage, the boundary majority samples are divided again by using granularity according to the new updated distribution property, and then are replaced by the new centre of granule to update network weight dynamically. Furthermore, a theoretical proof is given to testify the proposed algorithm had upper bound of information loss. The experimental results show that the proposed method can improve the total generalization performance and classification efficiency compared with some state-of-the-art algorithms.

Keywords: extreme learning machine; granular division; sequential imbalanced data; under-sampling

0 引 言

实际工程应用中, 在线贯序分类数据经常会出现类别严重不均衡的现象. 利用传统的分类算法解决此类问题时, 往往具有偏向性. 例如 1 000 个样本中有 990 个多类样本, 而少类样本仅有 10 个, 此时即使总体判别正确率达到 99%, 也极有可能出现多类样本全部判断正确, 而少类样本几乎全部误判的现象, 显然这样的分类结果对实际应用毫无价值. 大多数的实际在线预测问题中, 少类样本的错分代价通常远远大于多类样本, 对少类样本的识别往往更具有实际意义, 例如医疗诊断、天气预测等. 因此, 降低在线不均衡数据中少类样本的误判率具有极高的理论研究价值和工程意义[1].

目前针对不均衡数据分类的方法主要有两种: 1) 基于数据的策略, 主要通过欠采样和过采样进行预处理, 从而达到均衡样本, 提高分类精度的目的; 2) 基于算法的策略, 主要通过改进传统分类算法, 从而提高少类样本的分类精度和判别能力, 例如加权支持向量机、代价敏感学习等. 陶新民等[1-2]提出了基于 ODR 和 BSMOTE 相结合的不均衡数据采样方法, 通过逐级优化递减对多类样本进行欠采样, 同时利用 BSMOTE 算法[3]对边界少类样本进行过采样, 从而对

收稿日期: 2015-10-13; **修回日期:** 2016-01-04.

基金项目: 国家自然科学基金项目(U1204609); 中国博士后科学基金项目(2014M550508); 河南省高校科技创新人才计划项目(15HASTIT022); 河南师范大学优秀青年基金项目(14YQ007); 河南省高校青年骨干教师计划项目(2014GGJS-046).

作者简介: 毛文涛 (1980–), 男, 副教授, 博士, 从事机器学习、弱信号检测等研究; 何玲 (1990–), 女, 硕士生, 从事泛化性理论的研究.

样本数进行均衡. 王文剑等[4-5]提出了基于粒度划分的SVM, 根据粒度划分对多类样本进行欠采样以此来平衡样本. 上述研究方法虽然在一定程度上解决了类别的不均衡现象, 但通常存在两点问题: 1) 欠采样过程中未充分考虑样本的分布特性, 容易导致严重的信息损失; 2) 无法解决在线不均衡数据的分类问题. 目前在模式识别领域广泛使用的在线贯序极限学习机(OS-ELM)[6]是一种单隐层前馈神经网络, 尽管其具有速度快、泛化性好等优点, 但该模型在面对严重类别不均衡数据的分类问题时, 性能却表现不佳, 极易对少类样本产生误判.

由上述分析可知, 对于在线贯序不均衡数据而言, 有效分类的关键在于: 1) 采用合理的采样方法, 充分利用样本的分布特性, 提高均衡后样本数据的可靠度和可信度; 2) 降低多类样本欠采样后的信息损失, 使得其分类精度的下降可控. 但是, 当面对密度型数据时, 上述基于数据的策略的有效性会降低. 这里密度型数据是指不具有明显的分布规律的数据, 对其分布特性的提取通常存在困难. 粒度计算通过将复杂的问题进行抽象, 将其转化为相对于原信息数量较少的问题, 并保留原始信息大概的类别特征, 减少信息丢失造成的损失, 因此可应用于密度型数据的处理. 目前粒计算已经被引入基于支持向量机的不均衡数据分类中, 并在密度型数据上取得了较好的效果[4-5]. 但是, 这些方法最初并非针对在线贯序到达的不均衡数据分类, 对贯序数据的分布特性提取存在着一定的缺陷, 并且没有对欠采样过程中的信息损失给出分析.

为此, 本文同时从数据策略和算法策略入手, 提出一种基于动态粒度划分的在线贯序极限学习机(DG-OSELM). 该算法分为离线和在线两个阶段. 离线阶段, 为了提取样本分布特性, 对多类样本进行粒度划分, 利用粒中心实现对多类数据集的欠采样, 构建均衡样本集, 建立初始分类模型; 在线阶段, 对分布在分类边界的多类样本进行二次粒度划分和欠采样, 并动态更新分类模型参数. 同时, 在理论上证明了本文所提出方法具有信息损失上界, 表明了本文算法的合理性. 最后在两类不同分布特性的仿真数据集上验证了所提出算法的有效性.

1 在线贯序极限学习机

极限学习机(ELM)[7]是一种单隐层前馈神经网络. 该算法的输入层参数随机初始化, 并利用Moore-Penrose广义逆, 求得最小L_2范数形式的输出权重. 除隐神经元个数外其他各项参数均不可调, 所以整体算法结构简单, 学习速度快, 泛化能力良好[8]. 基于原始ELM算法, 在线贯序极限学习机(OS-ELM)是一种在线增量式的快速学习算法. 算法步骤可分为两个阶段[6]:

1) 初始化阶段. 从给定训练集$D=\{(x_i,t_i),i=1,2,\cdots,N\}$中选取部分数据集$D_0=\{(x_i,t_i),i=1,2,\cdots,N_0\}$, 其中$N_0\geqslant L$.

Step 1: 随机选取输入权值w_i和b_i, 其中$i=1,2,\cdots,L$, 计算隐层输出矩阵H_0;

Step 2: 计算初始输出权值$\beta^0=P_0{H_0}^{\mathrm{T}}T_0$, 其中$P_0=({H_0}^{\mathrm{T}}H_0)^{-1}$, $T_0=[t_1,t_2,\cdots,t_{N_0}]^{\mathrm{T}}$;

Step 3: 置$k=0$.

2) 序列学习阶段.

Step 4: 学习第个数据d_{k+1}=$(x_{N_0+k+1},t_{N_0+k+1})$;

Step 5: 令$T_{k+1}=[t_{N_0+k+1}]^{\mathrm{T}}$, 计算新学习数据的隐层输出矩阵

$$H_{k+1}=[g(w_1x_{N_0+k+1}+b_1)\cdots g(w_Lx_{N_0+k+1}+b_L)]_{1\times L};\quad(1)$$

Step 6: 计算输出权值

$$\beta^{k+1}=\beta^k+P_{k+1}{H_{k+1}}^{\mathrm{T}}(T_{k+1}-H_{k+1}\beta^k),\quad(2)$$

其中

$$P_{k+1}=P_k-P_kH_{k+1}^{\mathrm{T}}(I+H_{k+1}P_k{H_{k+1}}^{\mathrm{T}})^{-1}H_{k+1}P_k;$$

Step 7: 置$k=k+1$, 若$k<N$, 则返回Step 4, 否则算法结束.

2 基于动态粒度划分的不均衡在线贯序极限学习机算法

为叙述方便, 首先给出几个定义. 给定多类样本数据集$D=\{(x_i,t_i),i=1,2,...,N\}$, 其中$x_i$表示$m$维向量, $t_i=1$表示多类样本.

定义 1 粒的最大半径

$$R_j=\max_{x_{ji}\in X}(|C-X|),\quad(3)$$

即该粒内离该粒粒心最远的点到该粒心的距离. 其中: $C=(c_1,c_2,\cdots,c_{k_1})$, $X=(x_{j1},x_{j2},\cdots,x_{jN_j})$, $j=1,2,\cdots,k$, N_j的值为第j个粒内包含的总样本个数, c_j为粒心坐标, k为粒度划分所得粒个数的总数.

定义 2(粒的聚合度) 用来衡量该粒中样本的聚合程度, 因为粒的聚合度与粒内所含样本的多少成正比, 与粒内所有样本到粒心的距离和的大小成反比, 与粒的最大半径成反比, 即

$$\mathrm{Polymerize}(c_j)=\frac{\frac{N_j^2}{R_j}}{\sum_{i=1}^{N_j}|c_j-x_{ji}|}=\frac{N_j^2}{R_j\sum_{i=1}^{N_j}\sqrt{c_j^2-2x_{ji}c_j+x_{ji}^2}}.\quad(4)$$

x_{ji} 是以 c_j 为粒心的粒内的样本, 同定义 1 一样, N_j 的值为第 j 个粒内包含的总原始样本个数. 易知, 粒的聚合度越大, 粒中的样本越集中, 粒心越具有代表性.

定义 3 (粒的支持度)　粒内所含样本个数越多, 粒内样本占总体样本的比例越大, 越具有代表性, 即

$$\mathrm{Support}(c_j)=\frac{\mathrm{size}(c_j)}{N}=\frac{N_j}{N}. \tag{5}$$

其中: $\mathrm{size}(c_j)$ 为粒 c_j 内的样本规模, 即第 j 个粒内所含原样本个数 N_j, N 内为多类样本数据集 D 的大小, 支持度越大, 表明该粒包含样本越多.

2.1　初始离线阶段

离线阶段, 重构初始训练样本集 $S=\{(x_i,t_i)|i=1,2,\cdots,N_1\}$, 并建立初始模型.

首先采用粒度划分的思想削减多类样本, 对不均衡样本重构, 将初始训练样本集 S 分为多类样本集 Δ_1 和少类样本集 Δ_2.

一次粒度划分: 将初始样本集 S 中的所有多类样本先划分为 k_1 个粒, k_1 根据离线阶段的样本不均衡比进行设置, 一般 k_1 的取值公式为

$$k_1=2.2\,\mathrm{size}(\Delta_2)\sim 1.8\,\mathrm{size}(\Delta_2), \tag{6}$$

其中 $\mathrm{size}(\Delta_2)$ 代表离线少类样本集的样本规模, 并设置聚合度的上下阈值为 $[\eta_1,\eta_2]$, 使得粒度划分尽可能地符合原始多类样本的整体分布. 根据样本集 S 中的少类样本与多类样本的数量比对多类样本粒划分的 k_1 进行设定, 均匀选取 k_1 个点作为初始粒心. 因为在离线阶段中只进行一次粒度划分, 根据文献 [9], 利用式 $c_j=\dfrac{1}{N_j}\displaystyle\sum_{i=1}^{N_j}x_{ji}$ 计算每次迭代的聚类中心, 直至所有样本与聚类中心满足条件

$$\min\sum_{j=1}^{k_1}\sum_{x\in c_i}\mathrm{dist}(c_j,x)^2, \tag{7}$$

最终得到 k_1 个聚类中心 $C=\{c_1,c_2,\cdots,c_{k_1}\}$. 根据式 (4) 和 (5) 计算 k_1 个粒的聚合度和支持度, 从而对 k_1 值进行相应调整. 选取 k_1 个粒的粒心代替这些粒内的众多原始多类样本参加分类, 从而完成一次样本削减过程. 通过一次粒度划分完成对多类样本的缩减将不均衡比降到 2:1 左右, 得到基本均衡的初始训练样本集 $S_0=\{(x_i,t_i)|i=1,2,\cdots,N_0\}$.

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 w_i 和偏置 b_i, $i=1,2,\cdots,L$. 计算隐层输出矩阵[6]

$$H_0=\begin{bmatrix} g(w_1x_1+b_1) & \cdots & g(w_Lx_1+b_L)\\ g(w_1x_2+b_1) & \cdots & g(w_Lx_2+b_L)\\ \vdots & \ddots & \vdots\\ g(w_1x_{N_0}+b_1) & \cdots & g(w_Lx_{N_0}+b_L)\end{bmatrix}. \tag{8}$$

$T_0=[t_1\quad t_2\quad\cdots\quad t_{N_0}]^{\mathrm{T}}$ 为输出向量, 输出权值为

$$\beta_0=H_0{}^{\dagger}T_0, \tag{9}$$

其中 $H_0{}^{\dagger}=(H_0{}^{\mathrm{T}}H_0)^{-1}H_0{}^{\mathrm{T}}$.

2.2　在线贯序学习阶段

设第 $k+1$ 步贯序到达的样本块为 $\Omega_{k+1}=\{(x_i,t_i)|i=N_1{+}1{+}k,\cdots,N_1{+}1{+}k{+}\mathrm{Block}\}$, 其中 Block 为第 $k+1$ 步的样本数. 根据 t_i 的值, 该样本块可分为多类样本块 Φ_d 和少类样本块 Φ_s.

根据整体样本分布特征, 利用离线阶段中粒度划分的方法对新到的样本块中的多类样本集 Φ_d 进行一次粒度划分. 此次令目标粒个数 k_1 等于二倍的新到样本块中少类样本的个数, 得到的 k_1 个聚类中心将作为粒度划分时的粒心, 这些初次粒度划分得到的粒心组成了临时多类样本 Φ_{nd}. 经过一次粒化后, 保证将 Ω_{k+1} 样本集的不均衡比例降到 2:1.

密度较大且离分界面较近的临时多类样本对最终的分类正确率影响相对较大, 所以针对边界的临时多类样本再次进行欠采样. 根据边界临时样本实际分布, 利用式 (9) 确定其边界大致范围, 有

$$\frac{1}{3}S_{\mathrm{all}}\leqslant S_b\leqslant\frac{1}{2}S_{\mathrm{all}}. \tag{10}$$

其中: S_{all} 为多类数据的总体分布面积, S_b 为多类样本靠近边界范围的面积. 对于分布较为规律、面积大致可求的数据而言, 根据面积公式 S_{all} 较为易求; 但对于分布不规律且分布不够严格的几何形状的数据而言, S_{all} 可根据二维图像分块估计面积求近似值, 此处对面积的求解要求并不精确, 主要是想得到分布在多类和少类数据大致分类面周围的多类样本数据, 该部分边界数据对分类面影响较大.

式 (10) 对边界值的选取给出了一个较为清晰的概念, 根据该公式可以确定多类样本边界大致范围, 使得多类边界数据在整体多类数据中的比重较为适中, 控制边界范围不至于过大, 从而影响整个样本集的重构.

根据选取的大致边界面积范围, 规定属于这一范围内的多类样本为边界临时多类样本, 而不属于边界范围内的多类样本为内部多类样本. 利用这一规则最终得到边界临时多类样本的数据集 Φ_{ndb1} 和内部多类样本的数据集 Φ_{ndn}. 根据现在的多类少类不均衡比和边界多类样本个数, 对 k_2 进行设置, 从而对边界临时多类样本再次粒划得到粒心, 合并初次粒划分后的内部多类样本集 Φ_{ndn} 和二次粒划分后的粒心, 得到最终的多类样本集 Φ_{ndb2}. 根据上述条件和最终类别比, 以及聚合度和支持度的阈值要求, 得到

$$k_2=\frac{1}{3}\mathrm{size}(\Phi_{ndb1})\sim\frac{1}{2}\mathrm{size}(\Phi_{ndb1}), \tag{11}$$

其中 $\mathrm{size}(\Phi_{ndb1})$ 为边界临时多类样本的数据集 Φ_{ndb1}

的样本规模. 再次根据 $\min\sum_{j=1}^{k_2}\sum_{x\in c_i}\mathrm{dist}(c_j,x)^2$, 代入 k_2 进行聚类计算, 得到 k_2 个聚类中心, 即粒中心, 使用新粒心代替初次划分后的边界粒心, 最终使得样本类别比例降低到 1.5:1 与 1.1:1 之间.

合并 Φ_{ndn}, Φ_{ndb2} 和 Φ_s 得到新的样本块

$$\Phi_{k+1}=\{(x_i,t_i)|i=N_1+k+1,N_1+k+2,\cdots,N_1+k+\mathrm{Block}-m\},$$

则新样本块对应的神经元矩阵

$$H_\Phi=[h_{N_1+k+1}\quad h_{N_1+k+2}\quad\cdots\quad h_{N_1+k+\mathrm{Block}-m}],$$

得到隐层输出矩阵为 $H_{k+1}=[H_k{}^{\mathrm{T}}H_\Phi{}^{\mathrm{T}}]^{\mathrm{T}}$, 输出向量为 $T_{k+1}=[T_k{}^{\mathrm{T}}T_\Phi{}^{\mathrm{T}}]^{\mathrm{T}}$. 更新网络权值[6], 得到

$$\beta_{k+1}=H_{k+1}{}^{\dagger}T_{k+1}. \tag{12}$$

其中

$$H_{k+1}{}^{\dagger}=(H_{k+1}{}^{\mathrm{T}}H_{k+1})^{-1}H_{k+1}{}^{\mathrm{T}},$$

$$H_{k+1}{}^{\mathrm{T}}H_{k+1}=[H_k{}^{\mathrm{T}}H_\Phi{}^{\mathrm{T}}][H_k{}^{\mathrm{T}}H_\Phi{}^{\mathrm{T}}]^{\mathrm{T}}=H_k{}^{\mathrm{T}}H_k+H_\Phi{}^{\mathrm{T}}H_\Phi.$$

H_{k+1} 可在 H_k 的基础上获得, 使得计算量大大降低.

算法整体流程如图 1 所示.

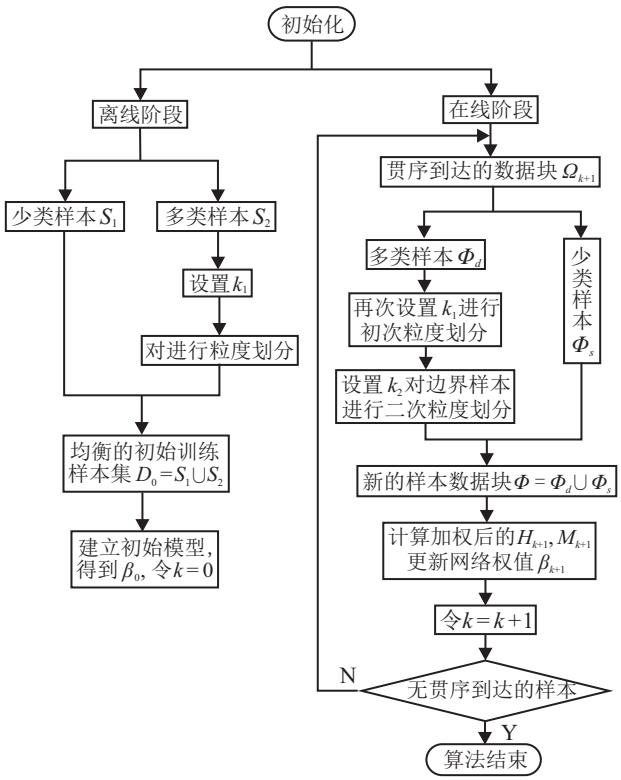


图 1 DG-OSELM 算法流程

3 损失信息上界分析

根据算法描述, 无论是在线阶段还是离线阶段, 均采用粒度划分对样本 $\Phi_d=\{(x_i,t_i),i=1,2,\cdots,M\}$ 进行欠采样, 得到 m 个最具价值的多类样本集合 Φ_d'. 本文利用信息熵[10]证明所提出算法欠采样过程中所损失的样本整体信息量和相对信息量存在上界, 从而证明该算法的合理性和有效性.

设每一次在线学习阶段在欠采样过程中的损失样本集合为 $E=\{(x_{ji},t_i),j=1,2,\cdots,m\}$, 其中 x_{ji} 是粒心为 c_j 的粒内包含的样本, c_j 最终代替该粒内所有样本作为最终的某个多类样本参加分类训练, 所以 c_j 内所有样本的样本权重和为 w_j, 则损失样本集 E 的总样本权重之和为

$$\sum_{j=1}^{m}w_j=\sum_{j=1}^{m}\Big(1-\frac{N_j^2}{R_j\sum\limits_{i=1}^{N_j}\sqrt{c_j^2-2x_{ji}c_j+x_{ji}^2}}\Big).$$

其中: N_j^2 为该粒中样本个数和的平方, $\sum\limits_{i=1}^{N_j}(c_j^2-2x_{ji}c_j+x_{ji}^2)^{1/2}$ 为该粒内所有样本到粒心的距离和. 根据已知聚类 k 值判断, 当粒内样本个数越多, 样本到粒心的距离和越小, 则该粒越紧密, 聚合度越大, 粒内的样本权重和越小. 从离线角度判断, 与贯序多类样本集 E 对应的离线多类损失样本集合为 E_1, 记为 $E_1=\{(y_{ij1},t_i),i=1,2,\cdots,m_0\}$, c_{j1} 内除粒心外各样本的样本权重为 w_{j1}, 则 E_1 的总体样本权重之和为

$$\sum_{j=1}^{m_0}w_j=\sum_{j=1}^{m_0}\Big(1-\frac{N_{j1}^2}{R_j\sum\limits_{i=1}^{N_j}\sqrt{c_{j1}^2-2x_{ji1}c_{j1}+x_{ji1}^2}}\Big).$$

定理 1 设 $H(E)$ 表示欠采样过程中的样本整体信息损失, 则存在

$$H(E)\leqslant\Big(m-\sum_{j=1}^{m}\frac{N_j^2}{R_j\sum\limits_{i=1}^{N_j}\sqrt{c_j^2-2x_{ji}c_j+x_{ji}^2}}\Big)\times\log\frac{m}{m-R_j\sum\limits_{j=1}^{m}\dfrac{N_j^2}{\sum\limits_{i=1}^{N_j}\sqrt{c_j^2-2x_{ji}c_j+x_{ji}^2}}},$$

且整体信息损失 $H(E)$ 的上界与样本集聚合度有关. 而样本聚合度由样本总体分布和聚类 k 值决定, 在该算法中单类样本块分布均匀, 且 k 值设置合理, 所以聚合度均较高.

证明 根据熵的定义, 可以得到

$$H(E)=-\sum_{j=1}^{M-m}w_j\log w_j.$$

根据最大熵原理, 当每个损失样本的权重 w_j 都取相同的值 $\Big(m-\sum\limits_{j=1}^{m}\Big(N_j^2\Big/R_j\sum\limits_{i=1}^{N_j}\sqrt{c_j^2-2x_{ji}c_j+x_{ji}^2}\Big)\times$

$\log\left(m\Big/\left(m-\sum_{j=1}^{m}\left(N_j^2\Big/R_j\sum_{i=1}^{N_j}\sqrt{c_j^2-2x_{ji}c_j+x_{ji}^2}\right)\right)\right)$

时, 整体信息损失 $H(\mathrm{E})$ 达到最大值, 则有

$$H(E)\leqslant -\left(m-\sum_{j=1}^{m}\frac{N_j^2}{R_j\sum\limits_{i=1}^{N_j}\sqrt{c_j^2-2x_{ji}c_j+x_{ji}^2}}\right)\times \log\frac{m}{m-\sum\limits_{j=1}^{m}\dfrac{N_j^2}{R_j\sum\limits_{i=1}^{N_j}\sqrt{c_j^2-2x_{ji}c_j+x_{ji}^2}}}. \tag{13}$$

由式(13)可知, $H(E)$ 上界仅与所有粒的 $N_j^2\Big/R_j\sum_{i=1}^{N_j}\sqrt{c_j^2-2x_{ji}c_j+x_{ji}^2}$ 有关, 且若该值越大, 则与之成反比的该值的信息上界越小.

定理 2　设 Λ 表示损失样本集合 E 对于离线阶段样本集 E_1 的整体相对信息损失, 且 $H(E_1)$ 表示 E_1 所含样本的整体信息量, 则有

$$\Lambda\leqslant \left(m_0-\sum_{j=1}^{m_0}\frac{N_{j1}^2}{R_j\sum\limits_{i=1}^{N_{j1}}\sqrt{c_{j1}^2-2x_{ji1}c_{j1}+x_{ji1}^2}}\right)\times \log\frac{m_0}{m_0-\sum\limits_{j=1}^{m_0}\dfrac{N_{j1}^2}{R_j\sum\limits_{i=1}^{N_{j1}}\sqrt{c_{j1}^2-2x_{ji1}c_{j1}+x_{ji1}^2}}}.$$

可以看出, 整体相对信息损失大小的上界仅与样本集 E_1 所得粒的聚合度 $k_{p1}^2\Big/R_j\sum_{i=1}^{k_{p1}}\sqrt{c_{j1}^2-2x_{ij1}c_{j1}+x_{ij1}^2}$ 有关, 且该值越大, 该信息上界越小.

证明　根据相对熵的定义, 有

$$\Lambda=\sum_{j=1}^{m}w_j\log\left(\frac{w_j{}'}{w_j}\right)=\sum_{j=1}^{m}(w_j\log w_j-w_j\log w_j{}')=-H(E)+H(E,E_1)\leqslant H(E)+H(E_1)-H(E)=H(E_1), \tag{14}$$

其中 $H(E,E_1)$ 表示 E 和 E_1 的叉熵. 由定理 1 易知

$$H(E_1)\leqslant -\left(m_0-\sum_{j=1}^{m_0}\frac{N_{j1}^2}{R_j\sum\limits_{i=1}^{N_{j1}}\sqrt{c_{j1}^2-2x_{ji1}c_{j1}+x_{ji1}^2}}\right)\times \log\frac{m_0}{m_0-\sum\limits_{j=1}^{m_0}\dfrac{N_{j1}^2}{R_j\sum\limits_{i=1}^{N_{j1}}\sqrt{c_{j1}^2-2x_{ji1}c_{j1}+x_{ji1}^2}}},$$

即

$$\Lambda\leqslant -\left(m_0-\sum_{j=1}^{m_0}\frac{N_{j1}^2}{R_j\sum\limits_{i=1}^{N_{j1}}\sqrt{c_{j1}^2-2x_{ji1}c_{j1}+x_{ji1}^2}}\right)\times \log\frac{m_0}{m_0-\sum\limits_{j=1}^{m_0}\dfrac{N_{j1}^2}{R_j\sum\limits_{i=1}^{N_{j1}}\sqrt{c_{j1}^2-2x_{ji1}c_{j1}+x_{ji1}^2}}}.$$

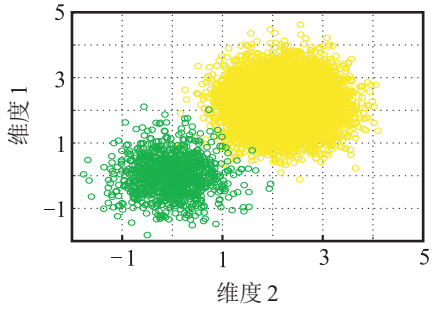
定理 1 和定理 2 从理论的角度出发, 证明了本文使用粒心代替原始样本的有效性和合理性. 考虑极端情况, 若部分粒的聚合度为 1, 则该粒相当于原始样本, 其对应的信息损失为 0, 说明该情况不影响样本的欠采样和整体信息损失. 利用熵值定理最终证明了本文算法根据粒度划分对多类样本进行欠采样的合理性.

4　仿真实验

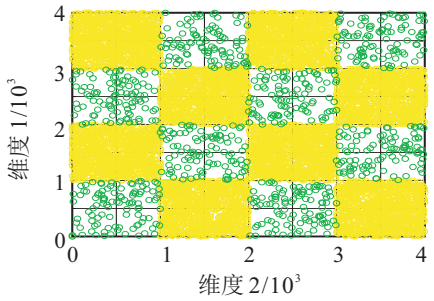
本文采用棋盘数据和标准正态分布数据集进行仿真实验. 分别采用 MC-OSELM、LS-SVM、OS-ELM 与本文所提出算法进行对比. 其中, MC-OSELM[11] 是针对在线不均衡问题的元认知在线序列极限学习机算法, 该算法在在线过程中对少类样本进行复制并加入训练样本集中, 从而降低样本集的不均衡程度. LS-SVM 是最小二乘支持向量机算法. 所有样本均需提前进行线性归一化, 所有计算结果均为 30 次重复实验的均值.

4.1　数据预处理

分别构造正态分布数据集和棋盘分布数据集. 正态分布数据集中少类样本与多类样本的中心分别为 $(0,0)$ 和 $(2.2,2.2)$, 以 $\begin{bmatrix}2.2 & 0\\ 0 & 2.2\end{bmatrix}$ 为协方差矩阵, 且样本规模分别为 1 000 与 10 000, 样本比为 1:10. 随机挑选 70% 的数据构成训练样本, 剩余 30% 为测试集, 如图 2(a) 所示. 对于棋盘数据集, 正负类数据分布面积一样, 分别占棋盘的 8 格, 棋盘内随机生成训练数据, 且样本规模分别为 100×8 与 $1\,000\times 8$, 样本比仍为 1:10, 测试数据生成方法相同且原始样本个数比也为 1:10, 如图 2(b) 所示.



(a) 正态分布数据

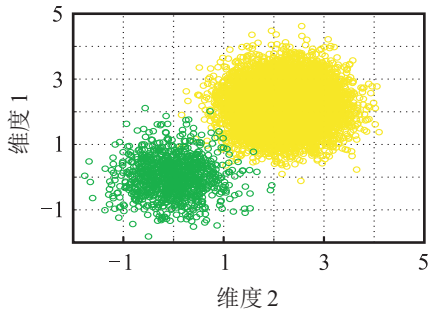


(b) 棋盘状数据

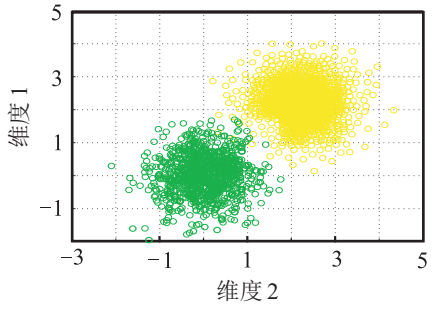
图 2 构造的仿真数据

4.2 实验结果

离线阶段, 采用粒度划分分别对正态分布样本集和棋盘状训练样本集进行样本重构, 图 3 和图 4 分别为两个数据集经过两次粒度划分前后样本变化情况. 由图 3 和图 4 可以看出, 经过两次粒度划分之后, 样本集已基本均衡. 表 1 给出了粒度划分前后的样本变化情况.



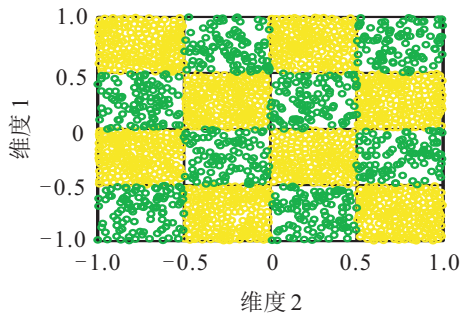
(a) 一次粒度划分结果



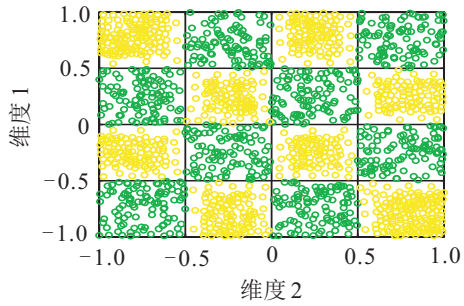
(b) 二次粒度划分结果

图 3 正态分布数据两次粒度划分变化情况

利用处理后的均衡样本集建立初始模型. 设定隐层激活函数为“Sigmoid”, 隐节点个数分别设置为 600 和 700. 4 种模型在两个数据集上的平均性能比较



(a) 一次粒度划分结果



(b) 二次粒度划分结果

图 4 棋盘数据两次粒度划分变化情况

表 1 均衡离线数据前后的样本数变化

数据集	属性	未处理前		处理后	
		多类	少类	多类	少类
棋盘数据	2	7 997	807	840	807
正态分布数据	2	10 000	1 000	1 057	1 000

表 2 正态分布数据集实验结果

算法	测试时间/s	少类测试精度/%	多类测试精度/%	总体测试精度/%	G-mean
DG-OSELM	0.189 9	98.86	97.60	97.71	0.982 3
OS-ELM	0.197 6	73.29	89.88	88.39	0.793 4
MC-OSELM	0.281 4	82.36	99.14	96.35	0.903 6
LS-SVM	42.738	67.53	99.92	96.97	0.819 7

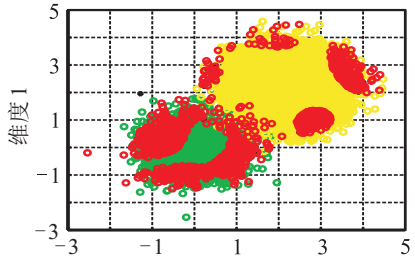
表 3 棋盘数据集实验结果

算法	测试时间/s	少类测试精度/%	多类测试精度/%	总体测试精度/%	G-mean
DG-OSELM	0.588 1	98.09	95.62	95.85	96.85
OS-ELM	0.283 4	59.01	99.56	92.79	76.64
MC-OSELM	0.351 5	60.98	99.81	96.27	77.99
LS-SVM	34.43	67.94	99.34	96.45	82.13

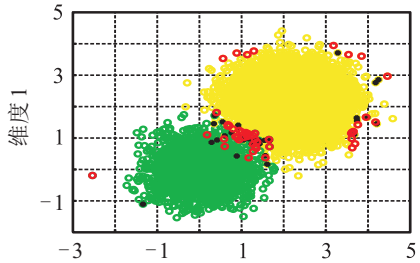
情况如表 2 和表 3 所示. 由表 2 和表 3 可以看出, 虽然 DG-OSELM 的总体测试精度并不总是最高, 但在少类上的测试精度却明显高于其他 3 种算法, 这表明本文所提出算法对少类样本的识别能力最好, 且所提出算法的 G-mean 值在两个数据集上分别比其他 3 种算法提高了 18.89%, 7.87%, 16.26% 和 20.21%, 18.86%, 14.72%. 由于所提出算法采用粒度划分的思想对不均衡样本重构, 避免了在对多类样本进行欠采样时样本信息的大量丢失, 不难发现, DG-OSELM 对多类样本的测试精度均达到 95% 以上, 进一步表明, 所提出算法可在保证多类分类精度的前提下, 增进少类的分类

正确率. 同时, DG-OSELM 的测试时间明显小于 LS-SVM, 且与 OS-ELM 和 MC-OSELM 在同一个数量级, 实时性较强.

为更直观地展示算法对比结果, 图 5 和图 6 给出了 DG-OSELM 和经典 OS-ELM 在正态分布数据和棋盘数据上的分类效果, 其中 "○" 与 "●" 表示被错误分类的样本点. 由图 5 和图 6 可以看出, 粒度划分前模型少类分类正确率较低, 极易出现异常分类情况, 而 DG-OSELM 分类效果明显优于粒度划分前的分类效果, 这表明本文算法通过粒度划分可充分考虑到样本总体分布特性, 使得 DG-OSELM 整体信息损失较少且数值稳定性良好.

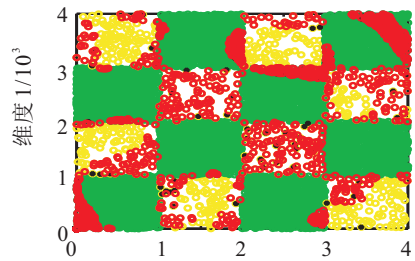


(a) OS-ELM 分类效果

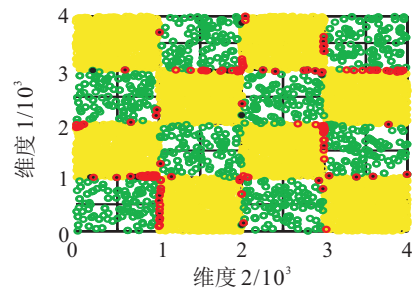


(b) DG-OSELM 分类效果

图 5　正态分布数据分类效果



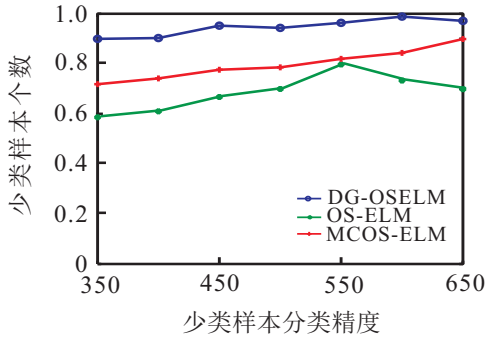
(a) OS-ELM 分类效果



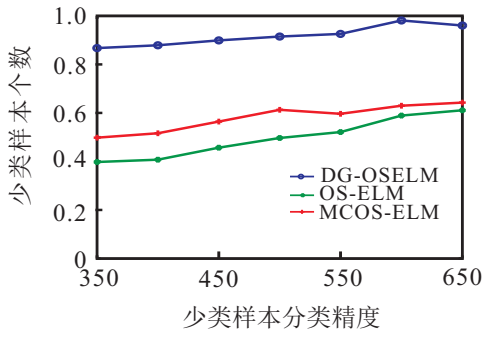
(b) DG-OSELM 分类效果

图 6　棋盘数据分类效果

调节隐节点个数, 观察少类精度变化, 如图 7 所示, 其中每个节点值取运行 15 次的平均值. 从图 7 中观察可得 GOS-ELM 相比于其他 3 种算法具有较强的稳定性.



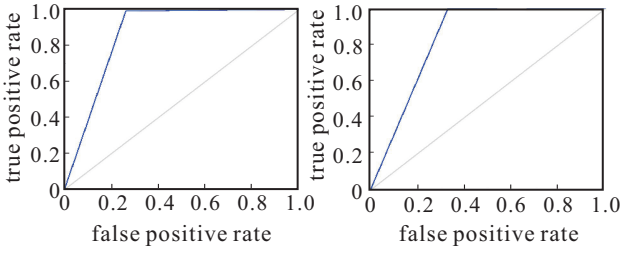
(a) 正则化数据集



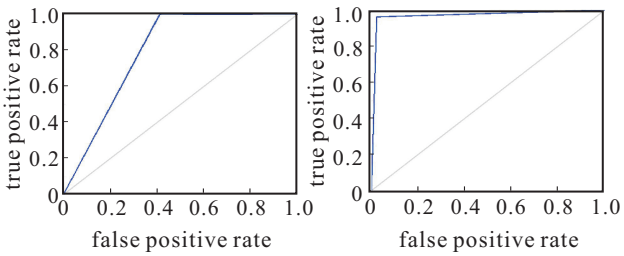
(b) 棋盘状数据集

图 7　不同隐节点上少类测试精度的变化

为了更好地体现该算法的优越性, 采用 ROC 曲线对样本的分类效果进行更直观的体现. 图 8 和图 9 所示为 4 种模型在棋盘数据和正态分布数据上的 ROC 曲线. 图 8 和图 9 中曲线下方面积越大, 说明分类器效果越显著. 可以看出, DG-OSELM 的分类效果表现得相对优越, 该算法少类样本的错分率明显较低, 识别能力更强, 对实际情况更有实用价值.



(a) SVM　(b) OS-ELM



(c) MC-OSELM　(d) DG-OSELM

图 8　4 种算法在棋盘数据上的 ROC 曲线

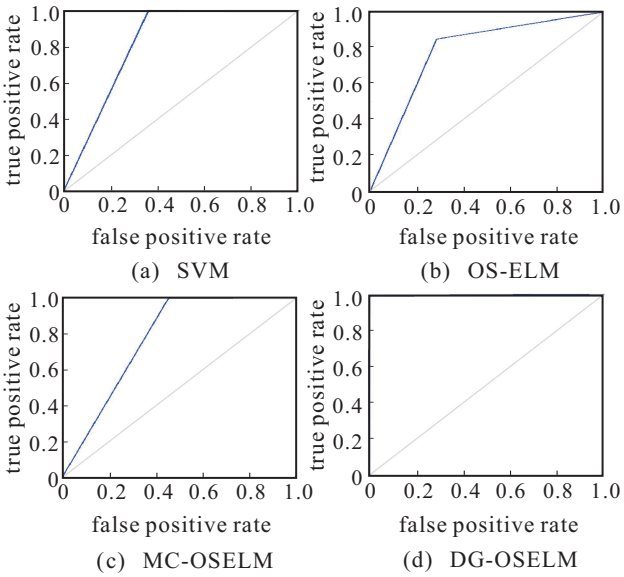


图 9　4 种算法在正态分布数据上的 ROC 曲线

5 结　论

本文提出了一种基于粒度划分的不均衡贯序极限学习机算法. 在离线阶段, 该算法根据多类样本分部特征和聚类特性进行粒度划分, 利用粒心代替原始样本, 得到最终的均衡样本; 在在线阶段, 该算法对边界多类样本进行二次动态粒度划分和欠采样, 提高了少类分类精度, 增强了模型对少类样本的判别能力. 为证明算法的合理性, 本文给出了算法欠采样过程的两种信息损失上界. 实验结果表明, 该算法可有效提高不均衡数据的整体泛化效果和分类效率, 对解决大规模的不均衡分类问题具有重要的意义.

参考文献(References)

[1] 陶新民, 童智靖, 刘玉. 基于ODR和BSMOTE结合的不均衡数据SVM分类算法[J]. 控制与决策, 2011, 26(10): 1535-1541.
(Tao X M, Tong Z J, L Y. SVM classifier for unbalanced data based on combination of ODR and BSMOTE[J]. Control and Decision, 2011, 26(10): 1535-1541.)

[2] 陶新民, 张冬雪, 郝思媛. 基于谱聚类欠抽样的不均衡数据SVM分类算法[J]. 控制与决策, 2012, 27(12): 1761-1775.
(Tao X M, Zhang D X, Hao S Y. SVM classifier for unbalanced data based on spectrum cluster-based under-sampling approaches[J]. Control and Decision, 2012, 27(12): 1761-1775.)

[3] 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, 12(A): 22-26.
(Yang Z M, Qiao L Y, Peng X Y. Research on datamining method for imbalanced dataset based on improved SMOTE[J]. Acta Electronica Sinica, 2007, 12(A): 22-26.)

[4] 郭虎升, 王文剑. 基于粒度偏移因子的支持向量机学习方法[J]. 计算机研究与发展, 2013, 50(11): 2315-2324.
(Guo H S, Wang W J. A support vector machine learning method based on granule shift parameter[J]. J of Computer Research and Development, 2013, 50(11): 2315-2324.)

[5] 程凤伟, 王文剑, 郭虎升. 动态粒度SVM学习算法[J]. 模式识别与人工智能, 2014, 27(4): 372-377.
(Cheng F W, Wang W J, Guo H S. Dynamic granular support vector machine learning algorithm[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(4): 372-377.)

[6] Liang N, Huang G. A fast accurate online sequential learning algorithm for feedforword networks[J]. IEEE Trans on Neural Networks, 2006, 17(6): 1411-1423.

[7] Huang G, Zhou H, Ding X, et al. Extreme learning machine for regression and multiclass[J]. IEEE Trans on Systems, Man, and Cybernetics-Part B: Cybernetics, 2012, 42(2): 513-529.

[8] 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.
(Yang L, Zhang R. Online sequential ELM algorithm and its improvement[J]. J of Northwest University: Natural Science Edition, 2012, 42(6): 885-896.)

[9] 蒋帅, K-均值聚类算法研究[D]. 西安: 陕西师范大学计算机科学学院, 2010.
(Jang S. K-means algorithm[D]. Xi'an: Shaanxi Normal University, School of Computer Science, 2010.)

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

[11] Vong C, Fip W, Wong P, et al. Prediction minority class for suspended particulate matters level by extreme learning machine[J]. Neurocomputing, 2014, 128: 136-144.

(责任编辑：孙艺红)