

针对动态非平衡数据集鲁棒的在线极端学习机

张 晶 冯 林

(大连理工大学电子信息与电气工程学部计算机科学与技术学院 辽宁大连 116024)

(大连理工大学创新实验学院 辽宁大连 116024)

(zhangjing_0412@163.com)

An Algorithm of Robust Online Extreme Learning Machine for Dynamic Imbalanced Datasets

Zhang Jing and Feng Lin

(School of Computer Science and Technology, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, Liaoning 116024)

(School of Innovation Experiment, Dalian University of Technology, Dalian, Liaoning 116024)

Abstract With the coming of big data age, dynamic data has gradually appeared in various application fields, such as safety monitoring, financial forecasting, and medical diagnostics. Although existing knowledge discovery and data mining techniques have shown great success in many real-world applications, dynamic data has the features of imbalance and instability of data classes, the dynamic change of data volume, which makes it difficult for the classification of dynamic data. To solve these problems, in this paper a robust weighed online sequential extreme learning machine algorithm (RWOSELM) based on the online sequential extreme learning machine algorithm (OSELM) is presented. RWOSELM generates the local dynamic weighted matrix with the help of cost sensitive learning theory, thereby it optimizes the empirical risk of the classification model. Meanwhile, RWOSELM takes the data distribution changes which are caused by temporal properties change of dynamic data into consideration, thus it introduces the forgetting factor to enhance the sensitivity of the classifier to the change of data distribution. The method is able to deal with the data with imbalanced class distribution, and maintains the good robust on dynamic data. This paper tests on 24 datasets with different distribution, and the results show that RWOSELM gets good results on imbalanced dynamic dataset.

Key words imbalanced dataset; extreme learning machine; online sequential extreme learning machine; cost sensitive learning; forgetting factor

摘 要 动态数据存在数据量动态改变,数据类别分布非平衡、不稳定等问题,这些问题成为分类的难点.针对该问题,通过对在线极端学习机模型进行拓展,提出鲁棒的权值在线极端学习机算法.为解决动态数据非平衡性,该算法借助代价敏感学习理论生成局部动态权值矩阵,从而优化分类模型产生的经验风险.同时,算法进一步考虑动态数据由于时序性质改变造成的数据分布变化,而引入遗忘因子增强分类器对数据分布变更的敏感性.算法在不同数据分布的24个非平衡动态数据集上测试,取得了较好的效果.

收稿日期:2014-03-06;修回日期:2014-09-25

基金项目:国家自然科学基金项目(61173163,51105052,61370200)

通信作者:冯 林(fenglin@dlut.edu.cn)

关键词 非平衡数据集; 极端学习机; 在线极端学习机; 代价敏感学习; 遗忘因子

中图法分类号 TP181

随着互联网技术的发展, 实际应用中数据产生量日益增大, 数据分布更多样化, 此时类别分布均匀、性质好的数据集十分稀有. 动态数据通常数据量不稳定, 且分布非平衡, 甚至相差悬殊. 针对出现的动态数据类别分布非平衡问题, 原有分类算法对少样本类别训练不足, 泛化性降低, 将对分类结果产生误导, 在应用领域中产生严重后果. 目前, 对非平衡数据集分类的研究日益增加, 主要集中在更改样本分布与改进分类算法 2 个方面.

1) 通过更改非平衡的数据分布来保证原有方法的分类效果. 早期的方法是随机重采样技术. 该方法需要获得不同类别数据量差异 N_{ex} , 从少量样本中随机选择 N_{ex} 个样本使得 2 类样本达到平衡, 该方式易于理解, 且可得到类别分布平衡的训练数据集, 后续采用传统分类方法. 但该方法会生成冗余数据、信息, 并在后续分类中出现过拟合^[1-3]. Chawla 等人^[4]提出用内插法合成少类别样本的过采样技术 (synthetic minority over-sampling technique, SMOTE), 但该方法依然存在以上问题. Han 等人^[5]与 He 等人^[6]均针对 SMOTE 算法内插数据时对类边界数据难以判断的问题提出自适应合成样本方法, 从而对 SMOTE 进行了优化. 楼晓俊等人^[7]通过剔除少类样本边界的噪音数据对 SMOTE 优化, 提高合成数据利用率. 2014 年 Barua 等人^[8]根据少类别到多类别的距离生成新样本权值, 避免了冗余样本的产生. 另一种更改数据分布的做法称为样本的欠采样技术, 数据量较多的样本数据包含更丰富的信息, 该技术将数量较多的数据进行约减后得到分布均衡的训练样本数据. 针对该问题的研究方法较多, 最简单的是随机欠采样技术, 但该技术在随机删除样本时删除了性质较好的样本. Fithian 等人^[9]通过保证特征空间中类别局部平衡的方法得到欠采样后的数据集. 数据清理技术在欠采样方法中被应用, 比较有代表性的是 Kubat 等人^[10]提出单侧选择方法去除类别边缘的有重叠的样本.

上面方法以得到分布性质更好的数据集为目的, 后续利用经典分类方法. 由于数据的分布可能是极不均衡的, 此时以上方法可能使得样本数变得极少, 而降低分类精度.

2) 对已有分类算法进行改进以适应非平衡数据集. 核学习方法成为解决非平衡数据问题的新方

法, Raskutti 等人^[11]、Akbani 等人^[12]、Wu 等人^[13]均利用核 (support vector machine, SVM) 解决该问题. Abe^[14]、Ertekin 等人^[15]、Provost 等人^[16-17]将主动学习法进行改进, 使算法能够很好地适应于非平衡数据集. Freund 等人^[18]和付忠良^[19]将集成学习法用于解决非平衡数据集分类问题, 提高了少量样本类别的分类准确率. Sun 等人^[20]将集成学习与代价函数相结合, 提出 AdaC1, AdaC2, AdaC3, 同时 McCarthy 等人^[21]将代价敏感度学习 (cost-sensitive learning, C-S) 与决策树方法相结合用于解决该问题, 达到更加准确的分类效果. 神经网络具有非线性等良好特性与代价函数等方法相结合解决非平衡数据集分类问题时效果显著^[22]. 但神经网络需要迭代求解, 计算速度慢, 且容易陷入局部极小值. 2006 年 Huang 等人^[23]提出极端学习机 (extreme learning machine, ELM) 算法, 该方法是对单隐层前馈神经网络 (single hidden layer feedforward neural networks, SLNFs) 的改进, 解决了传统神经网络计算速度慢的问题, 且解不会陷入局部极小. 为了拓展极端学习机的应用, 国内外学者对其进行了改进^[24-26]. Zong 等人^[27]将 ELM 算法加以改进用于解决非平衡数据集的分类问题. 以上算法在解决静态非平衡数据分类问题时取得了较好的效果, 但难以适应不断增加的数据量, 而随着大数据时代的到来动态数据分类问题将成为研究的热点. 为解决动态数据分类问题 Liang 等人^[28]在 ELM 的基础上推导了在线极端学习机 (online sequential extreme learning machine, OSELM), 当增加新的数据时, 该方法无需对所有已有的数据重新训练, 仅对新增加的数据训练得到新的分类器, 结合已有分类器, 是一种速度很快的增量式学习算法, 该算法及其改进算法在图像分类、预测等应用领域取得了很好效果^[29-30]. 但该方法仅针对动态数据, 没有考虑到数据分布的改变及非平衡性质.

本文给出了对动态非平衡数据分类的统一模型——鲁棒的权值在线极端学习机 (robust weight online extreme learning machine, RWOSELM), 并对文献^[28]提出的方法进行改进, 通过生成局部权值矩阵, 动态调整模型经验风险, 并利用遗忘因子使模型适应动态数据分布的改变, 达到提升分类效果

的目的. 经过 24 个数据集的实验, 表明本文的分类方法可以显著提高分类准确率.

1 鲁棒的权值在线极端学习机

1.1 OSELM 算法

ELM 模型是单隐层神经网络的改进算法. SLNFs 求解速度慢, 容易陷入局部极小值, 可描述如下: K 个不同的样本 $(\mathbf{x}_i, t_i)$, 其中 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K)^T \in \mathbb{R}^{D \times K}$, $\mathbf{t}_i = (t_{i1}, t_{i2}, \dots, t_{im})^T \in \mathbb{R}^m$, 网络中含有 $\tilde{N}$ 个隐层结点, 激活函数是 $g(\cdot)$, 可为 sigmoid 神经元, 或者 RBF 神经元等.

$$\sum_{i=1}^K \beta_i g(\mathbf{a}_i \mathbf{x}_j + b_j) = \boldsymbol{\varepsilon}_j, \quad (1)$$

其中, $j=1, \dots, K$, $\mathbf{a}_i = (a_{i1}, a_{i2}, \dots, a_{ik})$ 为连接 i 个隐层结点与输入结点的输入权值向量, $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \dots, \beta_{ik})^T$ 为连接第 i 个隐层结点和输出结点的输出权值向量, b_i 为 i 个隐层结点的偏置值. 令隐层结点为 $\tilde{N}$, 式(1)可写成矩阵形式: $\mathbf{H}\boldsymbol{\beta} = \mathbf{T}$, 其中网络隐层输出矩阵

$$\mathbf{H} = \begin{bmatrix} g(\mathbf{a}_1 \mathbf{x}_1 + b_1) & \cdots & g(\mathbf{a}_{\tilde{N}} \mathbf{x}_1 + b_{\tilde{N}}) \\ \vdots & & \vdots \\ g(\mathbf{a}_1 \mathbf{x}_K + b_1) & \cdots & g(\mathbf{a}_{\tilde{N}} \mathbf{x}_K + b_{\tilde{N}}) \end{bmatrix}_{K \times \tilde{N}},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1^T \\ \vdots \\ \boldsymbol{\beta}_{\tilde{N}}^T \end{bmatrix}_{\tilde{N} \times m}, \quad \mathbf{T} = \begin{bmatrix} t_1^T \\ \vdots \\ t_K^T \end{bmatrix}_{N \times m}.$$

传统 SLFNs 通过梯度下降等方式求解得到值 $\hat{\mathbf{a}}_i, \hat{\mathbf{b}}_i, \hat{\boldsymbol{\beta}}_i (i=1, \dots, \tilde{N})$, 使得:

$$\| \mathbf{H}(\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_{\tilde{N}}) \boldsymbol{\beta} - \mathbf{T} \| = \min_{\mathbf{a}_i, \mathbf{b}_i, \boldsymbol{\beta}} \| \mathbf{H}(\mathbf{a}_1, \dots, \mathbf{a}_{\tilde{N}}, b_1, \dots, b_{\tilde{N}}) \boldsymbol{\beta} - \mathbf{T} \|. \quad (2)$$

Huang 等人^[23]证明隐层权值的随机选取将不会影响求解结果, 此处随机选取输入权值 $\mathbf{a}_i$ 和隐层的偏置值 b_i . 由式(2)可知, SLFNs 网络的解可通过线性系统 $\mathbf{H}\boldsymbol{\beta} = \mathbf{T}$ 的最小二乘解得到. $\mathbf{H}\boldsymbol{\beta} = \mathbf{T}$ 解的形式可写为 $\hat{\boldsymbol{\beta}} = \mathbf{H}^+ \mathbf{T}$, 其中 $\mathbf{H}^+$ 为 $\mathbf{H}$ 的广义逆矩阵, 可表示为

$$\mathbf{H}^+ = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T. \quad (3)$$

将式(3)代入 $\hat{\boldsymbol{\beta}} = \mathbf{H}^T \mathbf{T}$, 得到:

$$\hat{\boldsymbol{\beta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{T}. \quad (4)$$

此处为了减小结构化风险增加正则化参数 C , C 的引入使得方程组的解不仅获得尽可能小的训练误

差, 而且使边缘距离最大化, 从而具有更好泛化性能. ELM 模型可以表示为

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \| \boldsymbol{\beta} \| + C \sum_{i=1}^K \boldsymbol{\varepsilon}_i^2, \\ \text{s. t.} \quad & \sum_{i=1}^{\tilde{N}} \beta_i g(\mathbf{a}_i \mathbf{x}_j + b_j) - t_j = \boldsymbol{\varepsilon}_j, \\ & j = 1, 2, \dots, K. \end{aligned} \quad (5)$$

从 ELM 的优化方程中可看到, 算法希望最小化结构风险, 相当于最大化样本到分类边界距离. 由图 1 可知, ELM 为了最小化结构风险, 分类边界明显偏向于少数类别样本, 从而产生少数样本类别泛化性降低的问题.

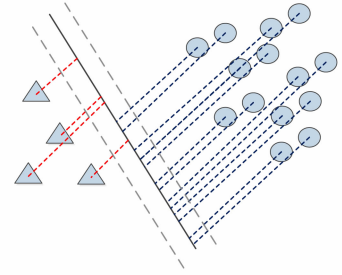


Fig. 1 The relationship between the classifier with samples.

图 1 分类器与样本之间的关系

针对 ELM 模型对非平衡数据集产生的问题, 文献[27]采用增加权值的方法, 针对不同类别数据调整其经验风险. 根据式(5)得到改进后的优化方程, $\mathbf{W}$ 为对角矩阵每一行对应一行训练数据:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \| \boldsymbol{\beta} \| + C \mathbf{W} \sum_{i=1}^K \boldsymbol{\varepsilon}_i^2, \\ \text{s. t.} \quad & \sum_{i=1}^{\tilde{N}} \beta_i g(\mathbf{a}_i \mathbf{x}_j + b_j) - t_j = \boldsymbol{\varepsilon}_j, \\ & j = 1, 2, \dots, K. \end{aligned} \quad (6)$$

由 KKT 条件, ELM 优化方程式(6)可以写成式(7):

$$L = \frac{1}{2} \| \boldsymbol{\beta} \|^2 + C \mathbf{W} \frac{1}{2} \sum_{i=1}^K \boldsymbol{\varepsilon}_i^2 - \sum_{i=1}^K \alpha_i (g(\mathbf{x}_i) \boldsymbol{\beta} - t_i + \boldsymbol{\varepsilon}_i), \quad (7)$$

该拉格朗日多项式中, α_i 为常数因子, 需求解 $(\boldsymbol{\beta}, \boldsymbol{\varepsilon}_i, \alpha_i)$ 参数值. 对式(7)求偏导数使其结果为 0, 有 $\frac{\partial L}{\partial \boldsymbol{\beta}} = 0, \frac{\partial L}{\partial \boldsymbol{\varepsilon}_i} = 0, \frac{\partial L}{\partial \alpha_i} = 0$. 权值矩阵为对角阵, 对应训练样本 $i=1, \dots, N$, 可得到: $\boldsymbol{\beta} = \mathbf{H}^T \boldsymbol{\alpha}, \alpha_i = C \mathbf{W} \boldsymbol{\varepsilon}_i$. 带入后可得到:

$$\hat{\beta} = \left(\frac{I}{C} + H^TWH\right)^{-1} H^TWT. \tag{8}$$

为解决在线问题,Liang 等人^[28]提出了 OSELM, OSELM 为 ELM 模型的在线求解方法,能快速处理在线计算问题. 对 $\Delta K (\Delta K \geq 1)$ 个新样本进行学习后将计算得到的旧样本与权值新样本进行矩阵计算,得到新的输出权值 $\beta_{K+\Delta K}$. 当 ΔK 个新样本到达时,隐层输出矩阵

$$H_{K+\Delta K} = \begin{bmatrix} g(x_1;a_1,b_1) & \cdots & g(x_1;a_{\tilde{N}},b_{\tilde{N}}) \\ \vdots & & \vdots \\ g(x_K;a_1,b_1) & \cdots & g(x_K;a_{\tilde{N}},b_{\tilde{N}}) \\ \cdots & & \cdots \\ g(x_{K+1};a_1,b_1) & \cdots & g(x_{K+1};a_{\tilde{N}},b_{\tilde{N}}) \\ \vdots & & \vdots \\ g(x_{K+\Delta K};a_1,b_1) & \cdots & g(x_{K+\Delta K};a_{\tilde{N}},b_{\tilde{N}}) \end{bmatrix} = \begin{bmatrix} g_1 \\ \vdots \\ g_K \\ \cdots \\ g_{K+1} \\ \vdots \\ g_{K+\Delta K} \end{bmatrix} = \begin{bmatrix} H_K \\ \cdots \\ H_{K+\Delta K} \end{bmatrix},$$

其中, $g_{K+r} = (g(x_{K+r}; a_1, b_1), g(x_{K+r}; a_2, b_2), \cdots, g(x_{K+r}; a_{\tilde{N}}, b_{\tilde{N}}))^T (r=1, 2, \cdots, \Delta K)$ 是与第 r 个新样本对应的输入向量的隐层输出矩阵, $H_{\Delta K} = (g_{K+1}^T, g_{K+2}^T, \cdots, g_{K+\Delta K}^T)^T$. 输出矩阵 $T_{K+\Delta K} = (t_{K+1}, t_{K+2}, \cdots, t_{K+\Delta K})^T$, 由式(4)可以得到:

$$\beta_{K+\Delta K} = (H_{K+\Delta K}^T H_{K+\Delta K})^{-1} H_{K+\Delta K}^T T_{K+\Delta K}, \tag{9}$$

此处可得到:

$$H_{K+\Delta K}^T H_{K+\Delta K} = H_K^T H_K + H_{\Delta K}^T H_{\Delta K}, \tag{10}$$

将式(9)代入式(10),可得到新的隐层输出权值:

$$\beta_{K+\Delta K} = (H_{K+\Delta K}^T H_{K+\Delta K})^{-1} H_{K+\Delta K}^T T_{K+\Delta K} = (H_K^T H_K + H_{\Delta K}^T H_{\Delta K})^{-1} (H_K^T T_K + H_{\Delta K}^T T_{\Delta K}). \tag{11}$$

1.2 RWOSELM 算法

上述 OSELM 算法实现了在动态数据到来时,仅对新数据训练达到增量式学习的目的,但算法没有考虑到数据分布的动态改变与其非平衡性质. 如图 2 所示,算法进行增量式学习时,新数据分布有一定随机性,且总体数据分布也随着新数据的到来发

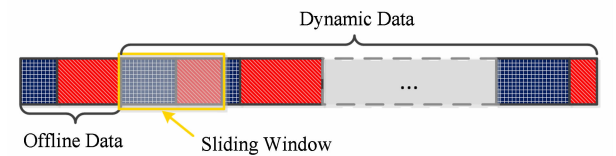


Fig. 2 Dynamic data distribution.
图 2 动态数据分布示意

生改变,需要对分类器进行调整以适应数据分布的动态变更.

首先,需要创建适合于动态非平衡数据集的增量式分类模型. 由式(11)知 $\beta_{K+\Delta K}$ 求解方式为:新数据到来,无需重新训练全部数据,而直接在原有模型上增量学习. 结合式(8)与式(11),为保证式(10)中隐层输出矩阵进行增量式学习时的可分解性,需要对式(8)中的权值矩阵 W 拆分. W 为对角阵,有 $W = \sqrt{W} \sqrt{W}$. 权值矩阵分解后得到隐层输出权值为

$$\hat{\beta} = \left(\frac{I}{C} + H^TWH\right)^{-1} H^TWT = \left(\frac{I}{C} + (\sqrt{W}H)^T \sqrt{W}H\right)^{-1} (\sqrt{W}H)^T \sqrt{W}T. \tag{12}$$

此时令 $A = \sqrt{W}H, B = \sqrt{W}T$, 式(12)可化为 $\hat{\beta} = (\frac{I}{C} + A^T A)^{-1} A^T B$. 根据 OSELM 的增量式学习算法,由式(11)可以得到:

$$\beta_{K+\Delta K} = \left(\frac{I}{C} + A_{K+\Delta K}^T A_{K+\Delta K}\right)^{-1} A_{K+\Delta K}^T B_{K+\Delta K}, \tag{13}$$

其中:

$$A_{K+\Delta K}^T A_{K+\Delta K} = A_K^T A_K + A_{\Delta K}^T A_{\Delta K}, \tag{14}$$

此处可令 $U_0 = \frac{I}{2C} + A_K^T A_K$, 因此可得到:

$$U_1 = \frac{I}{2C} + U_0 + A_{\Delta K}^T A_{\Delta K}, \tag{15}$$

有 $A_{K+\Delta K}^T B_{K+\Delta K} = A_K^T B_K + A_{\Delta K}^T B_{\Delta K}$, 根据式(14)和式(15),可以得到: $A_{K+\Delta K}^T B_{K+\Delta K} = K_0 K_0^{-1} A_0^T B_0 + A_1^T B_1 = K_0 \beta_0 + A_1^T B_1$, 又有 $K_0 = -\frac{I}{2C} + K_1 - A_{\Delta K}^T A_{\Delta K}$, 可得到:

$$A_{K+\Delta K}^T B_{K+\Delta K} = U_1 \beta_K - \frac{I}{2C} \beta_K - A_{\Delta K}^T A_{\Delta K} \beta_0 + A_{\Delta K}^T B_{\Delta K}. \tag{16}$$

又因为 $\beta_{\Delta K} = U_1^{-1} \begin{bmatrix} A_K \\ A_{\Delta K} \end{bmatrix}^T \begin{bmatrix} B_K \\ B_{\Delta K} \end{bmatrix}$, 根据式(13)和式(16),得到:

$$\beta_{K+\Delta K} = U_1^{-1} (U_1 \beta_K - \frac{I}{2C} \beta_K - A_{\Delta K}^T A_{\Delta K} \beta_K + A_{\Delta K}^T B_{\Delta K}) = \beta_K - U_1^{-1} (-\frac{I}{2C} \beta_K - A_{\Delta K}^T A_{\Delta K} \beta_K + A_{\Delta K}^T B_{\Delta K}). \tag{17}$$

当新数据动态到来时,由式(17)不仅可实现非平衡数据集的增量式学习,且在保证计算精度的前提下,提高了计算速度. 但如何动态地调整局部权值矩阵 W , 成为算法针对非平衡数据集分类效果的关键

键. 本文利用 Elkan^[31], Ting^[32] 提出的 C-S 学习方法, 动态生成局部权值矩阵. 如图 3 所示:

		True Class	
		P	N
Output	T	TP (True Positives) C_{00}	FP (False Positives) C_{01}
	F	FN (False Negatives) C_{10}	TN (True Negatives) C_{11}

Fig. 3 Data categories labeled.
图 3 数据类别标识

二分类问题中, P 与 N 均为样本的真实类别, 对应正类样本与负类样本. T 与 F 为分类后得到的样本类别, 对应正类样本与负类样本. C_{00}, C_{11} 为正确分类代价值, C_{10}, C_{01} 为错误分类代价值. 一般认为, 数据量少的类别被错误分类是难以接受的, 因此, $C_{10} > C_{01}$ 且 $C_{00} = C_{11}$, 有式(18):

$$p(j = 0 | \mathbf{x})C_{10} + p(j = 1 | \mathbf{x})C_{11} \geq p(j = 0 | \mathbf{x})C_{00} + p(j = 1 | \mathbf{x})C_{01}, \quad (18)$$

其中, $\mathbf{x}$ 为样本数据, $j = 0, 1$ 为数据集类别. C-S 理论认为一个好的分类器, 应使得式(18)左右两侧代价值相等, 即需要找到分类算法, 使得:

$$\begin{aligned} p(j = 0 | \mathbf{x})C_{10} + p(j = 1 | \mathbf{x})C_{11} &= \\ p(j = 0 | \mathbf{x})C_{00} + p(j = 1 | \mathbf{x})C_{01}. \end{aligned} \quad (19)$$

本文解决二分类问题, 有 $p(j = 0 | \mathbf{x}) = 1 - p(j = 1 | \mathbf{x})$, 又 $C_{00} = C_{11} = 0$, 根据式(19), 可得到:

$$p(j = 0 | \mathbf{x}) = \frac{C_{01}}{C_{01} + C_{10}},$$

C_{10} 的意义是将类别 0 错分为类别 1 的代价, 而所有真值为 0 的样本均有可能分错, 因此, C_{10} 等于真值为 0 的样本数 E_P , 同理可得 C_{01} 等于真值为 1 的样本数 E_N . 最终得到将样本 $\mathbf{x}$ 作为类别 0 的可能性 $p(j = 0 | \mathbf{x})$. 结合 1.2 节, 可得到样本 $\mathbf{x}$ 对应权值:

$$w_{10} = \frac{E_N}{E_P + E_N}. \quad (20)$$

同时有 $w_{01} = 1 - w_{10}$, 即可得到权值矩阵 $\mathbf{W}$. 此外, 本文算法应用在动态非平衡数据的增量式学习中, 每一批次到来的样本个数均不相同, 本次到来的样本将会使所有样本的分布发生改变, 当前求解得到的仅为局部权值矩阵. 如图 4 所示, 图 4(a)为离线训练样本的分类效果; 图 4(b)中带条形阴影为新到来的数据, 如果只根据新数据调整分类模型, 忽略了总体数据分布, 分类边界将发生偏斜; 图 4(c)为分类边界正确的位置.

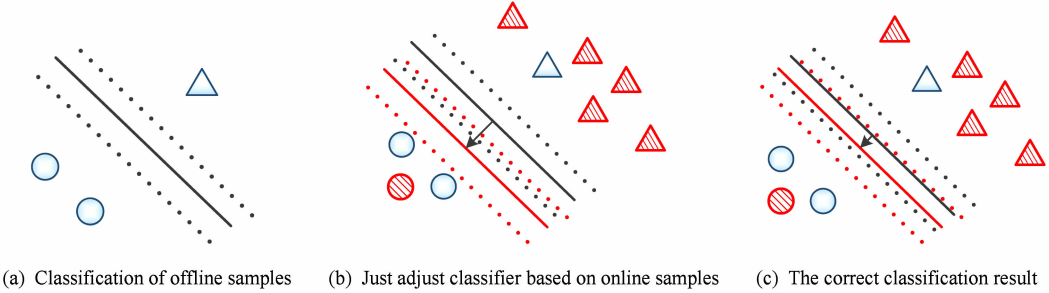


Fig. 4 The incremental data classification principle.
图 4 增量数据分类原理

针对该问题, 本文引入遗忘因子对前次分类器弱化, 根据新数据对总体数据分布改变的影响, 将前次分类器乘上遗忘因子 ρ , 实现对原有分类器的遗忘. 本文采用启发式方法选取 ρ 值, 此处有 $\rho = \left| \frac{S_P - E_P}{S_N - E_N} \right|$, 其中 S_P, S_N 分别为目前正、负类样本总数; 式(17)最终转化为

$$\begin{aligned} \beta_{K+\Delta K} &= \rho \times \beta_K - \mathbf{U}_1^{-1} \left[-\frac{\mathbf{I}}{2C} (\rho \times \beta_K) - \right. \\ &\quad \left. \mathbf{A}_{\Delta K}^T \mathbf{A}_{\Delta K} (\rho \times \beta_K) + \mathbf{A}_{\Delta K}^T \mathbf{B}_{\Delta K} \right]. \end{aligned} \quad (21)$$

RWOSELM 算法可被总结如下:

/ * 离线训练部分 * /

1) 随机指定网络输入权值 $\mathbf{a}_i$ 和偏置值 $b_i, i =$

- $1, \dots, \tilde{N}$, 设置正则化参数 C ;
- 2) 根据式(4)计算隐层输出矩阵 $\mathbf{H}$;
- 3) 根据式(20)计算离线数据对应权值矩阵 $\mathbf{W}$;
- 4) 根据上面得到的参数与式(12)计算离线数据, 输出权值 β ;
- / * 在线训练部分 * /
- 5) 根据式(20)计算新样本对应的权值矩阵 $\mathbf{W}$;
- 6) 计算前次分类器遗忘因子 ρ ;
- 7) 当到来新的数据时根据式(10)求解更新后的 β ;
- 8) 如果数据到来结束, 则利用式(21)求解结果; 否则转到步骤 4) 继续训练.

2 实验结果及分析

2.1 非平衡数据集评价标准

仅用分类率评价一个针对于非平衡集的分类方法不够全面,且该指标对数据分布变化并不敏感.目前已有很多有效的非平衡数据分类器评价指标,本文采用以下2种针对非平衡数据集较常用的评价标准来代替分类率.准确率(precision),召回率(recall)、 $G-mean$ 值均对数据类别分布改变敏感.准确率又称查准率,定义为式(22),其意义是在正类别中有多少是分类正确的;召回率,又称查全率,定义为式(23),其意义是在所有被分类为正类别的样本中,有多少是正确的; $G-mean$ 定义为式(24),即算法分类准确率的集合均值,该值对于数据分布的改变十分的敏感,当少数量分类正确个数为0时, $G-mean$ 的值将为0.

Precision = \frac{E_{TP}}{E_{TP} + E_{FP}}; (22)

Recall = \frac{E_{TP}}{E_{TP} + E_{FN}}; (23)

G-mean = \sqrt{\frac{E_{TP}}{E_{TP} + E_{FN}} \times \frac{E_{TN}}{E_{TN} + E_{FP}}}. (24)

另一种评价标准称为 ROC(receiver operating characteristics)曲线^[33-34],是一种对数据分布敏感的可视化评价标准,真阳性率和假阳性率,作为 ROC 曲线的纵轴与横轴.分别计算每一个样本的(false positive rate, $Rate_{TP}$)值和(true positive rate, $Rate_{FP}$)值,将其定义为式(25).在坐标空间中绘制对应的数据点,数据点的位置越靠近1则说明该分类器对不平衡数据的分类效果越好,ROC 曲线下方的面积定义为(area under an ROC curve, AUC),该指标作为评价分类器平均性能的标准, AUC 的值越接近于1,则分类效果越好,随机模型分类器的 AUC 结果为0.5.

Rate_{TP} = \frac{E_{TP}}{E_P}; Rate_{FP} = \frac{E_{FP}}{E_N}. (25)

2.2 数据集描述

本实验中采用文献[35]中非平衡数据集,为评价算法的有效性与普适性,选择分布尽可能多的情况,共3组24个数据集,所有数据集的维度不超过18维.离线训练样本为总样本量10%,在线学习每次到来的样本量为总样本量的10%,测试样本数为总样本量的10%.图5中可以看到数据集的全部数据量,以及数据类别的正负类别数量对比柱状图.

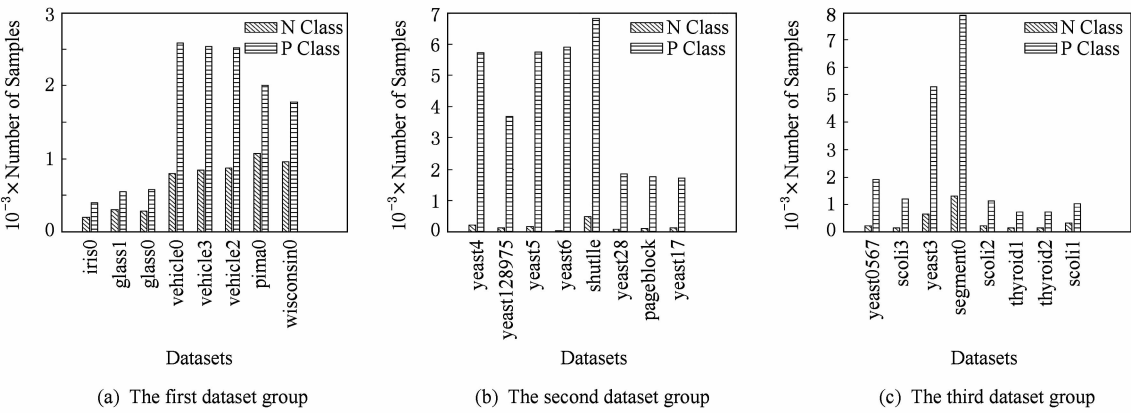


Fig. 5 Comparison of different classes of datasets of sample size.

图5 数据集不同类别样本量对比

2.3 实验结果

实验中选择隐层结点个数为50,文献[27]中指出sigmoid激活函数对隐层结点的变更不敏感,本文均采用sigmoid激活函数.正则化因子C值一般采用交叉验证或在固定取值范围内遍历的方式选取,本文采用后者,使C值在{2⁻¹⁸, 2⁻¹⁶, ..., 2⁴⁸, 2⁵⁰}内进行搜索.实验中对比了OSELM与RWOSELM

算法的准确率、召回率、 $G-mean$ 值与AUC值,并给出算法的ROC曲线图及分类效果的优劣对比. RWOSELM通过动态数据集分布生成局部权值矩阵,优化不同类别数据经验风险,并利用遗忘因子增强对动态数据分类的鲁棒性.从实验结果可知, RWOSELM算法在解决非平衡数据集的分类问题时对比原有OSELM算法综合指标有较大提升.本

① 实验中用到的数据库可从 <http://sci2s.ugr.es/keel/imbalanced.php#subA> 得到.

文算法用式(20)求解权值矩阵 $\mathbf{W}$ 与遗忘因子时通过不同类别样本数运算得到,对角阵与矩阵相乘部分增加 $2mn$ 次乘法,并且多了 $\mathbf{W}$ 与 ρ 的计算过程, RWOSELM 比 OSELM 算法的训练时间要长.

表 1 与图 6 为数据集组 1 中数据集 $E_P/E_N<9$

的实验对比结果.这 8 个数据集类别非平衡程度较低,2 个算法各项指标结果相近,有的数据集 RWOSELM 与 OSELM 结果相等甚至 OSELM 略高,如 iris0,wisconsin0 等数据集.根据本文算法, $\mathbf{W}$ 的取值与正负类别样本数密切相关,当样本均匀分

Table 1 Comparison Between ELM and RWOSELM on the First Group Dataset

表 1 数据集组 1 的分类结果对比

Dataset	Imbalance Ratio	C	Algorithm	Precision/%	Recall/%	G-mean	AUC
iris0	2	2^{-4}	OSELM	1	1	1	1
		2^{-2}	RWOSELM	1	1	1	1
glass1	1.82	2^{24}	OSELM	0.673 9	0.823 0	0.677 5	0.689 7
		2^{18}	RWOSELM	0.688 4	0.826 1	0.688 6	0.771 6
glass0	2.06	2^{36}	OSELM	0.652 8	0.930 7	0.720 3	0.741 1
		2^{22}	RWOSELM	0.666 7	0.932 0	0.727 3	0.844 8
vehicle0	3.25	2^{18}	OSELM	0.961 4	0.984 2	0.878 2	0.963 8
		2^{16}	RWOSELM	0.919 6	0.983 5	0.932 3	0.973 1
vehicle3	2.99	2^{-12}	OSELM	0.747 6	0.945 8	0.697 7	0.671 5
		2^{-4}	RWOSELM	0.945 8	0.929 4	0.956 3	0.861 6
vehicle2	2.88	2^{10}	OSELM	0.893 3	0.954 8	0.837 9	0.773 0
		2^{16}	RWOSELM	0.974 5	0.950 8	0.983 2	0.952 2
pima0	1.87	2^4	OSELM	0.766 6	0.831 0	0.672 7	0.873 9
		2^2	RWOSELM	0.861 4	0.882 2	0.861 8	0.887 9
wisconsin0	1.86	2^{38}	OSELM	0.997 7	0.991 1	0.978 9	0.987 7
		2^{34}	RWOSELM	0.982 0	0.990 9	0.993 4	0.997 7

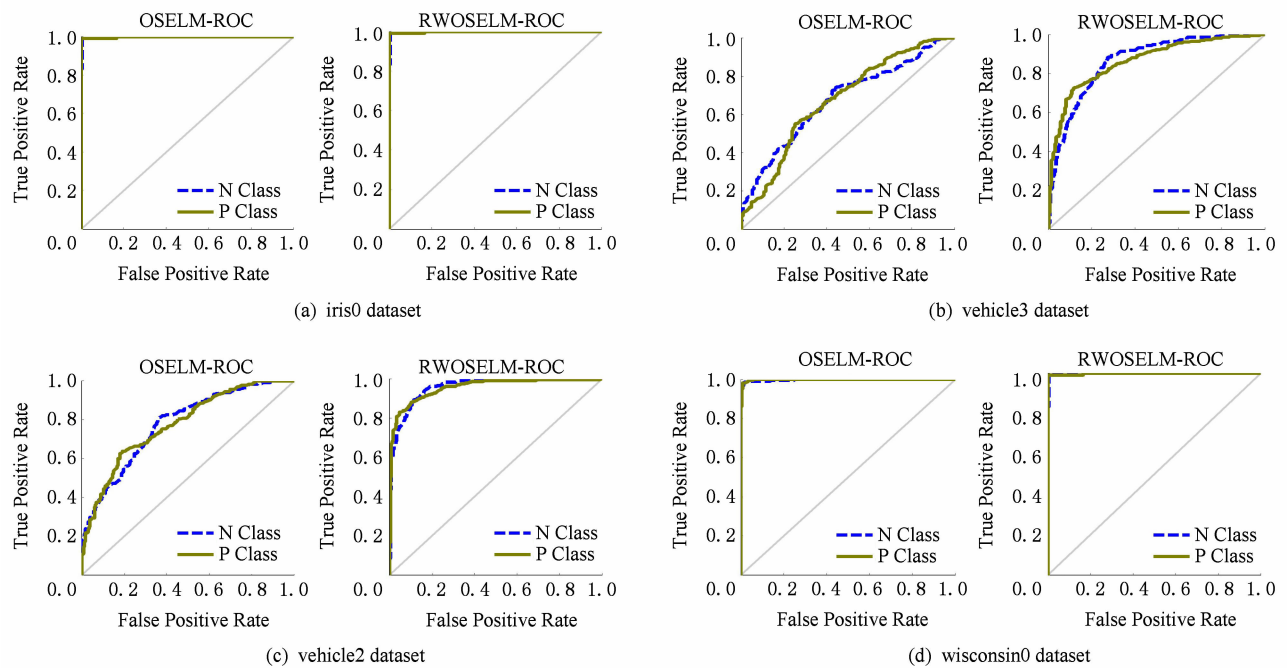


Fig. 6 Comparison of ROC curves of the first group dataset.

图 6 数据集组 1 的 ROC 曲线对比

布时正负类别样本权值相等, W 的作用仅类似于正则化参数 C , 用于增强算法的泛化性. 考虑文章篇幅, 本文给出数据集组 1 中 4 个数据集的 ROC 对比曲线图. 从图 6 中也能看到, 数据集的非平衡程度越低, 本文算法的优势越小.

当数据非平衡程度增加时, 如数据集组 2 中所

给数据 $E_P/E_N > 9$. 表 2 给出该数据组的实验结果, 以及为考虑文章篇幅, 图 7 中给出 4 个数据集的 ROC 对比曲线图. 通过该结果可知, 本文给出的 RWOSELM 算法对比 OSELM 算法的准确率、召回率、 $G-mean$ 值及 AUC 值均有较大提高. 尤其对于非平衡度高的数据集, 如数据集 yeast6, bageblock1

Table 2 Comparison Between ELM and RWOSELM on the Second Group Dataset

表 2 数据集 2 分类结果							
Dataset	Imbalance Ratio	C	Algorithm	Precision/%	Recall/%	$G-mean$	AUC
yeast1	14.3	2^{44}	OSELM	0.8207	0.9882	0.7121	0.7614
		2^{20}	RWOSELM	0.8758	0.9890	0.7742	0.7972
yeast1-2-8-9_vs_7	30.57	2^{48}	OSELM	0.6532	0.9677	0.4718	0.6599
		2^{20}	RWOSELM	0.9498	0.9776	0.6532	0.7732
yeast5	32.73	2^2	OSELM	0.8604	0.8936	0.6328	0.7634
		2^2	RWOSELM	0.8624	0.9938	0.8428	0.9355
yeast6	41.4	2^2	OSELM	0.7336	0.9963	0.6712	0.6174
		2^2	RWOSELM	0.8682	0.9968	0.8721	0.9398
shuttle	13.87	2^2	OSELM	0.9837	1	0.8660	0.9919
		2^2	RWOSELM	1	1	1	0.9919
yeast-2_vs_28	23.1	2^6	OSELM	0.8723	0.9853	0.7347	0.8356
		2^2	RWOSELM	0.9242	0.9861	0.8308	0.8503
bageblock1	164	2^{38}	OSELM	0.7185	0.9755	0.6668	0.7568
		2^6	RWOSELM	1	0.9823	0.9911	0.9965
yeast-1_vs_7	14.3	2^6	OSELM	0.6057	0.6054	0.6819	0.7787
		2^2	RWOSELM	0.7865	0.7864	0.7702	0.8131

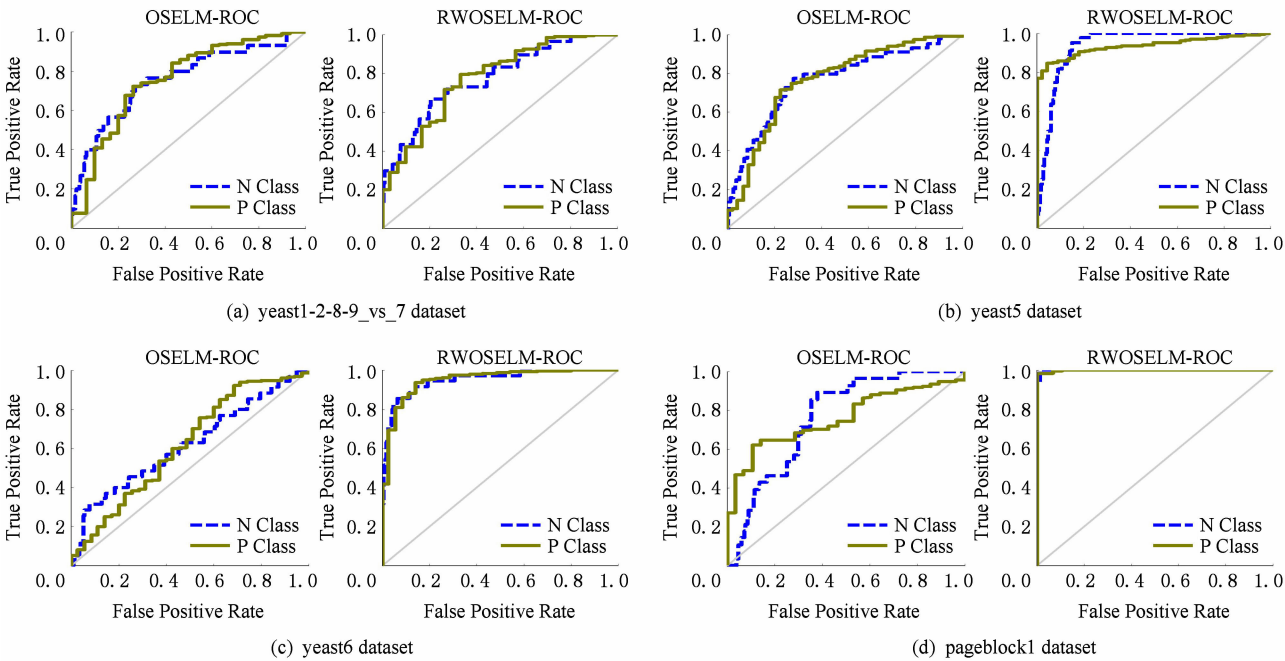


Fig. 7 Comparison of ROC curves of the second group dataset.

图 7 数据集组 2 算法 ROC 曲线对比图

等 RWOSELM 算法优势显著, 该算法动态生成局部权重矩阵, 受动态数据分布波动小, 同时通过遗忘因子适应全局数据分布的改变. 从 $G-mean$, AUC 值与 ROC 曲线的对比上看本文的算法对于非平衡数据集分类有更好的普适性.

数据集组 3 中数据非平衡程度较大, 且包含噪音数据, 该类数据为分类算法加大了难度. 从表 3 的实验结果, 以及为考虑文章篇幅, 图 8 中给出 4 个数据集的 ROC 对比图. 可看到本文提出的算法明显优于原算法, 提升了算法的分类效果.

Table 3 Comparison Between ELM and RWOSELM on the Third Group Dataset

表 3 数据集组 3 的分类结果

Dataset	Imbalance Ratio	Outlier/s	C	Algorithm	Precision/%	Recall/%	G-mean	AUC
yeast-0-5-6-7-9_vs_4	9.35	60	2^{-2}	OSELM	0.7317	0.9614	0.4643	0.8663
			2^{-8}	RWOSELM	0.8512	0.9667	0.5755	0.8991
ecoli3	8.6	40	2^6	OSELM	0.5949	0.9372	0.3856	0.6893
			2^{10}	RWOSELM	0.9037	0.9577	0.6509	0.8628
yeast3	8.1	100	2^8	OSELM	0.8335	0.9709	0.6005	0.7873
			2^2	RWOSELM	0.8630	0.9719	0.6734	0.8893
segment0	6.02	100	2^{-8}	OSELM	0.9975	0.9980	0.9914	0.5074
			2^6	RWOSELM	0.9717	0.9979	0.9226	0.9731
ecoli4	15.7	40	2^{-2}	OSELM	0.8765	0.8405	0.7325	0.7139
			2^{-8}	RWOSELM	0.9643	0.9673	0.9012	0.9660
new-thyroid1	4.84	20	2^{-4}	OSELM	0.5000	1	0.5292	0.8157
			2^{-16}	RWOSELM	0.7611	1	0.6699	0.9940
new-thyroid2	5.55	20	2^6	OSELM	0.9167	1	0.8367	0.8232
			2^{-2}	RWOSELM	0.9611	1	0.9129	0.9538
ecoli1	3.36	40	2^{-2}	OSELM	0.4865	0.9767	0.5000	0.8081
			2^{-16}	RWOSELM	0.6834	0.9833	0.6830	0.8983

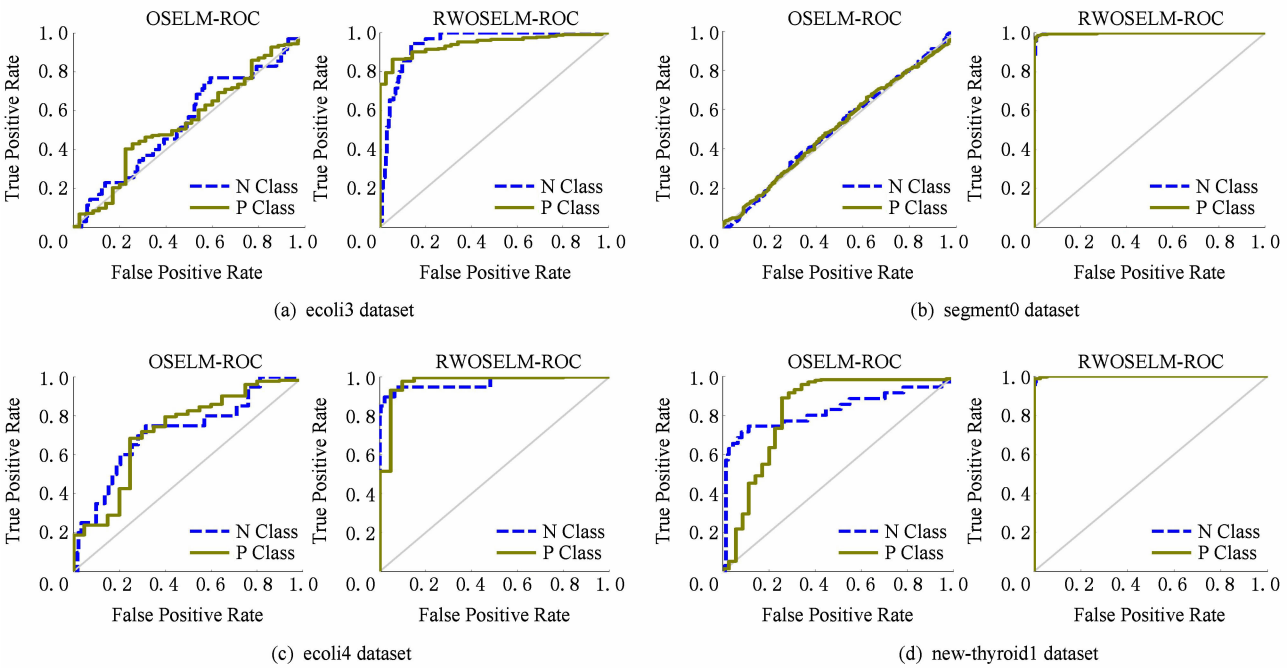


Fig. 8 Comparison of ROC curves of the third group dataset.

图 8 数据集组 3 算法 ROC 曲线对比图

本文给出了数据分批到来时算法 OSELM 和 RWOSELM 在 3 个数据集组上 $G-mean$ 值的对比效果,如图 9~11 所示. OSELM 与 RWOSELM 算

法的 $G-mean$ 值均随着数据的增加而上升并逐渐趋于稳定,可明显看到,RWOSELM 的分类效果要好于 OSELM.

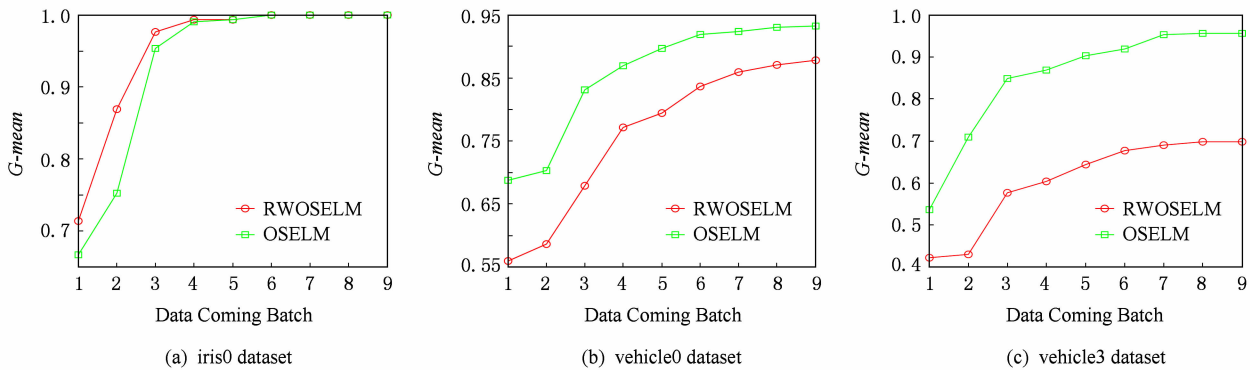


Fig. 9 According to the batch $G-mean$ values to compare three datasets of data in the first group.

图 9 数据集组 1 中 3 个数据集按批次的 $G-mean$ 值对比

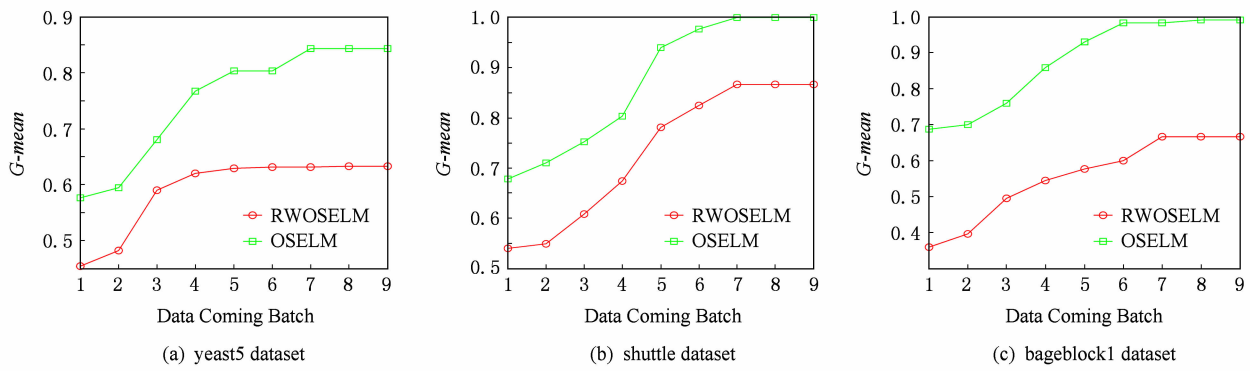


Fig. 10 According to the batch $G-mean$ values to compare three datasets of data in the second group.

图 10 数据集组 2 中 3 个数据集按批次的 $G-mean$ 值对比

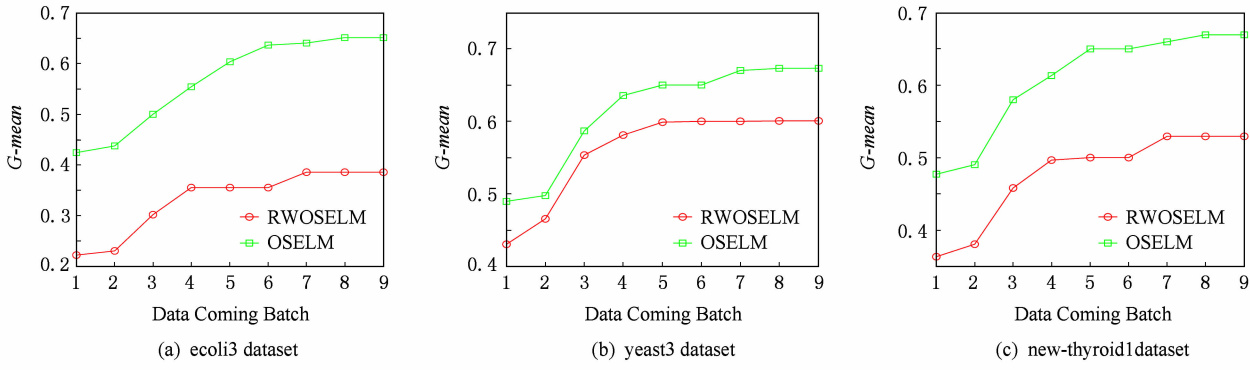


Fig. 11 According to the batch $G-mean$ values to compare three datasets of data in the third group.

图 11 数据集组 3 中 3 个数据集按批次的 $G-mean$ 值对比

3 结 论

动态非平衡数据集的分类在各个领域均有着广泛的应用,如金融、医疗以及网络入侵检测等,但动

态数据存在数据量波动,类别分布非平衡等问题,使得很多学习算法对该类数据不能很好适应,出现学习不充分、过拟合降低分类效果. 本文提出一种适合于该类数据的分类方法,该方法利用代价敏感学习理论,借助在线极端学习机的增量式学习方法,该方

法对动态数据的非平衡与时序性敏感,在分类中得到了满意的效果.但在实际应用中无标签样本越来越多,如何对半监督的非平衡数据集有效分类将成为一个值得研究的问题.同时,针对在线分类问题,实际应用中存在离线数据小样本问题,如何在在线学习中通过参数调整的方式使模型获得较好的效果且避免过拟合问题,也是一个有价值的研究问题.

参 考 文 献

- [1] Holte R C, Acker L, Porter B W. Concept learning and the problem of small disjuncts [C] //Proc of the 11th Int Joint Conf on Artificial Intelligence. San Francisco: Morgan Kaufmann, 1989: 813-818
- [2] Mease D, Wyner A J, Buja A. Boosted classification trees and class probability/quantile estimation [J]. The Journal of Machine Learning Research, 2007, 8: 409-439
- [3] Drummond C, Holte R C. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling [C] //Proc of the Learning from Imbalanced Datasets II, ICML. Menlo Park, CA: AAAI, 2003: 11-19
- [4] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357
- [5] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning [G] //Advances in Intelligent Computing. Berlin: Springer, 2005: 878-887
- [6] He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning [C] //Proc of IEEE World Congress on Computational Intelligence. Piscataway, NJ: IEEE, 2008: 1322-1328
- [7] Lou Xiaojun, Sun Yuxuan, Liu Haitao. Clustering boundary over-sampling classification method for imbalanced data set [J]. Journal of Zhejiang University: Engineering Science, 2013, 47(6): 944-950 (in Chinese)
(楼晓俊, 孙雨轩, 刘海涛. 聚类边界过采样不平衡数据分类方法[J]. 浙江大学学报: 工学版, 2013, 47(6): 944-950)
- [8] Barua S, Islam M, Yao X, et al. MWMOTE—Majority weighted minority oversampling technique for imbalanced data set learning [J]. IEEE Trans on Knowledge and Data Engineering, 2014, 26(2): 405-425
- [9] Fithian W, Hastie T. Local case-control sampling: Efficient subsampling in imbalanced data sets [OL]. 2013 [2013-02-25]. <http://arxiv.org/abs/1306.3706>
- [10] Kubat M, Matwin S. Addressing the curse of imbalanced training sets: One-sided selection [C] //Proc of the 14th Int Conf on Machine Learning. Menlo Park, CA: AAAI, 1997: 179-186
- [11] Raskutti B, Kowalczyk A. Extreme re-balancing for SVMs: A case study [J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 60-69
- [12] Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced datasets [C] //Proc of European Conf on Machine Learning. Berlin: Springer, 2004: 39-50
- [13] Wu G, Chang E Y. Class-boundary alignment for imbalanced dataset learning [C] //Proc of ICML. Menlo Park, CA: AAAI, 2003: 49-56
- [14] Abe N. Sampling approaches to learning from imbalanced datasets: Active learning, cost sensitive learning and beyond [C/OL] //Proc of ICML-KDD. Menlo Park, CA: AAAI, 2003 [2013-02-25]. http://www.site.uottawa.ca/~nat/Workshop2003/ICML03Workshop_Abe.ppt
- [15] Ertekin S, Huang J, Bottou L, et al. Learning on the border: Active learning in imbalanced data classification [C] //Proc of the 16th ACM Conf on Information and Knowledge Management. New York: ACM, 2007: 127-136
- [16] Ertekin S, Huang J, Giles C L. Active learning for class imbalance problem [C] //Proc of the 30th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2007: 823-824
- [17] Provost F. Machine learning from imbalanced data sets 101 [C] //Proc of the AAAI'2000 Workshop on Imbalanced Data Sets. Piscataway, NJ: IEEE, 2000: 1-3
- [18] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting [C] //Proc of Computational Learning Theory. Berlin: Springer, 1995: 23-37
- [19] Fu Zhongliang. Real adaBoost algorithm for multi-class and imbalanced classification problems [J]. Journal of Computer Research and Development, 2012, 48(12): 2326-2333 (in Chinese)
(付忠良. 不平衡多分类问题的连续 AdaBoost 算法研究[J]. 计算机研究与发展, 2012, 48(12): 2326-2333)
- [20] 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
- [21] McCarthy K, Zabar B, Weiss G. Does cost-sensitive learning beat sampling for classifying rare classes? [C] //Proc of the 1st Int Workshop on Utility-based Data Mining. New York: ACM, 2005: 69-77
- [22] Kukar M, Kononenko I. Cost-sensitive learning with neural networks [C] //Proc of the 13th European Conf on Artificial Intelligence (ECAI-98). New York: John Wiley & Sons, 1998: 445-449
- [23] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: Theory and applications [J]. Neurocomputing, 2006, 70: 489-501
- [24] Feng Lin, Liu Shenglan, Zhang Jing. Robust activation function of extreme learning machine and linear dimensionality reduction in high-dimensional data [J]. Journal of Computer Research and Development, 2014, 51(6): 1131-1140 (in Chinese)

- (冯林, 刘胜蓝, 张晶. 高维数据中鲁棒激活函数的极端学习机及线性降维[J]. 计算机研究与发展, 2014, 51(6): 1131-1140)
- [25] Deng Wanyu, Zheng Qinghua, Chen Lin. Regularized extreme learning machine [C] //Computational Intelligence and Data Mining. Los Alamitos, CA: IEEE Computer Society, 2009: 389-395
- [26] Horata P, Chiewchanwattana S, Sunat K. Robust extreme learning machine [J]. Neurocomputing, 2013, 102: 31-44
- [27] Zong W, Huang G B, Chen Y. Weighted extreme learning machine for imbalance learning [J]. Neurocomputing, 2013, 101: 229-242
- [28] Liang N Y, Huang G B, Saratchandran P, et al. A fast and accurate online sequential learning algorithm for feedforward networks [J]. IEEE Trans on Neural Networks, 2006, 17 (6): 1411-1423
- [29] Li B, Wang J, Li Y, et al. An improved on-line sequential learning algorithm for extreme learning machine [G] // LNCS 4491: Advances in Neural Networks-ISNN 2007. Berlin: Springer, 2007: 1087-1093
- [30] Liang N Y, Huang G B, Saratchandran P, et al. A fast and accurate online sequential learning algorithm for feedforward networks [J]. IEEE Trans on Neural Networks, 2006, 17 (6): 1411-1423
- [31] Elkan C. The foundations of cost-sensitive learning [C] // Proc of the 17th Int Joint Conf on Artificial Intelligence International. San Francisco, CA: Morgan Kaufmann, 2001: 973-978
- [32] Ting K M. An instance-weighting method to induce cost-sensitive trees [J]. IEEE Trans on Knowledge and Data Engineering, 2002, 14(3): 659-665
- [33] Fawcett T. ROC graphs: Notes and practical considerations for researchers [R]. Palo Alto, CA: HP Labs, 2003
- [34] Fawcett T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874
- [35] Fernández A, García S, et al. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets [J]. Fuzzy Sets and Systems, 2008, 159(18): 2378-2398



Zhang Jing, born in 1984. PhD candidate in the School of Computer Science and Technology, Dalian University of Technology, China. Her main research interests include Data mining.



Feng Lin, born in 1976. Received his PhD degree in mechanical design and theory from Dalian University of Technology, China, in 2004. Professor in the School of Innovation Experiment, Dalian University of Technology, China. His main research interests include intelligent image processing, robotics, data mining, and embedded systems.