一种鲁棒非平衡极速学习机算法*

孟凡荣, 高春晓†, 刘 兵
(中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116)

摘 要: 极速学习机(ELM)算法只对平衡数据集分类较好,对于非平衡数据集,它通常偏向多数样本类,对于少数样本类性能较低。针对这一问题,提出了一种处理不平衡数据集分类的 ELM 模型(ELM-CIL),该模型按照代价敏感学习的原则为少数类样本赋予较大的惩罚系数,并引入模糊隶属度值减小了外围噪声点的影响。实验表明,提出的方法不仅对提高不平衡数据集中少数类的分类精度效果较明显,而且提高了对噪声的鲁棒性。

关键词: 极速学习机; 不平衡数据集; 基于核的可能性模糊 C-均值聚类; 神经网络
中图分类号: TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2014)04-0985-04
doi:10.3969/j.issn.1001-3695.2014.04.006

Robust extreme learning machine algorithm based on imbalanced datasets

MENG Fan-rong, GAO Chun-xiao†, LIU Bing
(*School of Computer Science & Technology, China University of Mining & Technology, Xuzhou Jiangsu* 221116, *China*)

Abstract: ELM algorithm can achieve better classification results for balanced datasets. For the imbalanced datasets, it is usually favor of the majority class and has lower class performance for a small number of samples. To solve this problem, this paper proposed an ELM model (ELM-CIL model) dealing with imbalanced dataset classification problems. The model was in accordance with the principles of cost-sensitive learning that gave a greater penalty coefficient for the minority class sample. At the same time, the introduction of fuzzy membership value reduced the impact of external noise points. The experiments show that the proposed method is more obvious to improve minority class classification accuracy and has better robustness to noise than traditional ELM algorithm.

Key words: extreme learning machine(ELM); imbalanced datasets; KPFCM; neural networks

极速学习机(ELM)算法是 Huang 等人[1,2]基于 SLFN 算法提出的,由于具有扎实的数学背景、学习速度快、能逼近复杂非线性函数以及高度泛化能力的优点,成为现在广泛使用的机器学习技术,已经成功应用到模式识别、数据挖掘等很多领域。不同于传统的采用基于梯度下降思想[3]的误差反向传播的神经网络算法[4,5],需要在训练过程中迭代多次确定所有参数最优解,导致算法计算量较大、学习速度较慢等;ELM 算法根据 SLFN 的学习能力只与隐藏层节点的数目有关,而与输入层的权值无关,所以,ELM 学习过程中不需要大量人为地设定网络参数,算法学习速度相比于传统的神经网络学习算法有显著提高[6]。同 SVM[7,8]相似,尽管 ELM 算法对于平衡数据集能够取得较好的分类结果,但对于非平衡数据集,它通常产生的是次优结果。例如一个训练非平衡数据集的 ELM 通常产生一个偏向多数样本类的模型,对于少数样本的类性能较低,但是这些少数类往往正是使用者所感兴趣的[9]。例如,在故障诊断应用中,故障状态样本的数量由于成本等原因通常远少于正常状态的样本,而使用者显然更关注故障状态的类别。如何有效地提高不平衡数据集中少数类的分类性能成为 ELM 亟待解决的问题。本文考虑了二分类中的不平衡问题,正类作为少数类,负类作为多数类。

由于 ELM 算法平等地考虑所有的训练样本,使得数据集中的每个正类样本和负类样本对分类面的形成产生同样的影响,而负类作为多数类所有样本累加,最终对分类面的形成产生较大的影响。针对这一问题,本文提出了一种处理不平衡数据集分类问题的 ELM 模型(ELM-CIL 模型),该模型按照代价敏感学习的原则为少数类样本赋予较大的惩罚系数,而为多数类样本赋予较小的惩罚系数;同时,算法中引入了模糊隶属度的概念,减小了外围噪声点对分类的影响。实验表明,本文提出的 ELM-CIL 方法相比于原 EIL 方法,对于提高不平衡数据集中少数类的分类精度方面效果较明显,同文献[10]中处理不平衡数据集的 FPSVM-CIL 方法相比,测试精度近似且略低于支持向量机,但在参数选择与学习速度上优于 SVM。

1 ELM 模型

已知训练样本集$\{(x_i,t_i)\}$,其中,$x_i=\{x_{i1},x_{i2},\cdots,x_{in}\}\in R^n$,$t=\{t_{i1},t_{i2},\cdots,t_{im}\}\in R^m$,含有 L 个隐藏层节点且激励函数为 $f(x)$。单输出节点情况下,ELM 的输出函数为

$$f_L(x)=\sum_{i=1}^{L}\beta_i h_i(x)=h(x)\beta \tag{1}$$

其中:$\beta=[\beta_1,\cdots,\beta_L]^T$ 表示 L 个隐藏层节点的隐藏层与输出节点之间的输出权值向量; $h(x)=[h_1(x),\cdots,h_L(x)]$表示关于输入 x 的隐藏层输出向量,即 $h(x)$将输入数据从 d 维的输入空间映射为 L 维的隐藏层特征空间。ELM 模型考虑了最小

收稿日期: 2013-06-04; **修回日期**: 2013-08-05 **基金项目**: 高等学校博士学科点专项科研基金资助项目(20110095110010);国家“863”计划资助项目(2012AA011004)

作者简介:孟凡荣(1962-),女,辽宁沈阳人,教授,博士,主要研究方向为数据库技术、数据挖掘;高春晓,女(通信作者),硕士,主要研究方向为数据挖掘(892464651@qq.com);刘兵,男,博士,主要研究方向为数据挖掘.

化训练误差以及决策函数输出权值的范数,如下所示:

$$\min \| H\beta - T \|^2 \text{ and } \| \beta \| \tag{2}$$

其中: H 表示隐藏层输出矩阵,表示为

$$H = \begin{bmatrix} h(x_1) \\ h(x_2) \\ \vdots \\ h(x_n) \end{bmatrix} = \begin{bmatrix} h_1(x_1) & \cdots & h_L(x_1) \\ h_1(x_2) & \cdots & h_L(x_2) \\ \vdots & \vdots & \vdots \\ h_1(x_n) & \cdots & h_L(x_n) \end{bmatrix} \tag{3}$$

类似于标准 SVM,最小化输出权值的范数 $\| \beta \|$ 可以看做在 ELM 特征空间中最大化两个类之间的分类间隔 $2/\| \beta \|$。该范数实际上控制了在 ELM 特征空间中函数的复杂程度。

根据统计学理论可知,实际风险包括经验风险和结构风险两种成分[11],一个具有较好泛化性能的模型应该能权衡这两种风险, 并取得最佳折中。ELM 同时考虑这两种风险因素,并通过参数 C 调节两种风险。因此,单输出的两类 ELM 模型为

$$\begin{aligned} &\min L_{\text{ELM}} = \frac{1}{2} \| \beta \|^2 + \frac{C}{2} \sum_{i=1}^{N} \varepsilon_i^2 \\ &\text{subject to: } h(x_i)\beta = t_i - \varepsilon_i \quad i = 1, \cdots, N \end{aligned} \tag{4}$$

式中: $\frac{1}{2} \| \beta \|^2$ 代表结构风险; $\frac{C}{2} \sum_{i=1}^{N} \varepsilon_i^2$ 代表经验风险。

基于 KKT 原理,训练 ELM 的过程等同于解决下面的对偶优化问题:

$$L_{\text{ELM}} = \frac{1}{2} \| \beta \|^2 + \frac{C}{2} \sum_{i=1}^{N} \varepsilon_i^2 - \sum_{i=1}^{N} \alpha_i (h(x_i)\beta - t_i + \varepsilon_i) \tag{5}$$

根据式(5)可以得到 KKT 约束条件:

$$\frac{\partial L_{\text{ELM}}}{\partial \beta} = 0 \rightarrow \beta = \sum_{i=1}^{N} \alpha_i h(x_i)^{\text{T}} = H^{\text{T}}\alpha \tag{6}$$

$$\frac{\partial L_{\text{ELM}}}{\partial \varepsilon_i} = 0 \rightarrow \alpha_i = C\varepsilon_i \quad i = 1, \cdots, N \tag{7}$$

$$\frac{\partial L_{\text{ELM}}}{\partial \alpha_i} = 0 \rightarrow h(x_i)\beta - t_i + \varepsilon_i = 0 \quad i = 1, \cdots, N \tag{8}$$

这里 $\alpha = [\alpha_1, \cdots, \alpha_N]^{\text{T}}$,其中,每个拉格朗日乘子 α_i 对应于第 i 个训练样本。当训练样本较少(即 $N < L$)时,将式(6)(7)代入式(8),由上面的公式可以推出

$$\beta = H^{\text{T}}(HH^{\text{T}} + \frac{I}{C})T \tag{9}$$

同理当训练样本较大(即 $N > L$) 时,可以推出

$$\beta = (H^{\text{T}}H + \frac{I}{C})H^{\text{T}}T \tag{10}$$

最终,ELM 分类器的输出函数(1)可以写成:

$$f(x) = h(x)\beta = h(x)H^{\text{T}}(HH^{\text{T}} + \frac{I}{C})T \quad \text{或}$$

$$f(x) = h(x)\beta = h(x)(H^{\text{T}}H + \frac{I}{C})H^{\text{T}}T \tag{11}$$

2 类不平衡学习方法

近些年来,不平衡数据分类问题已成为数据挖掘和机器学习的研究热点,不平衡数据广泛存在于医疗诊断、文本分类、诈骗检测、市场行为分析、雷达图像监测等多个领域,具有极高的应用前景和现实意义。在两分类数据集中, 数量相当少的一类被称为少数类, 而另一类则被称为多数类,具有这样特征的两分类数据集则被称为是不平衡的。对于不平衡数据集,传统的分类方法结果明显偏向多数类,很多少数类样本被误分为多数类,导致多数类的分类结果远远优于少数类的分类结果。现在已有的 CIL 方法通常可以分为两类。

1)外部方法。其独立于学习算法,它在训练分类器之前进行数据集的预处理来平衡数据集,如过采样或欠采样等重采样方法属于这一类。随机欠采样即多数类的样本被随机删除,直到满足一个特定的类比[12]。随机过采样即少数类的样本被随机复制,直到满足一个特定的类比。SMOTE[13]是一种过采样方法,在该方法中,新生成的样本是在现有的少数类样本附近产生的,而不是直接复制它们;文献[9]中提出了一种基于遗传算法的抽样方法;文献[14]中提出基于聚类的抽样方法。

2)内部方法。通过改变学习算法来减少对类不平衡的敏感。在文献[15]中,Veropoulos 等人针对 SVM 提出了 DEC 方法,在该方法中,SVM 的目标函数如下:

$$\min\left(\frac{1}{2} w \cdot w + C_+ \sum_{[i|y_i=+1]}^{l_1} \delta_i + C_- \sum_{[i|y_i=-1]}^{l_2} \delta_i \right)$$

其中: C_+ 表示正类样本的惩罚系数, C_- 表示负类样本的惩罚系数,通过为正类样本分配一个比负类样本更高的惩罚系数,以减小类不平衡的影响。

zSVM 也是一种内部算法,在该方法中,由一个非平衡训练数据集生成一个 SVM 模型,然后对结果模型的决策边界进行修改来删除分类器偏向多数类(负)。改变后的 SVM 决策函数表示如下:

$$f(x) = \text{sgn}\{ z\sum_{i=1}^{l_1} \alpha_i^{+} y_i K(x_i, x) + \sum_{i=1}^{l_2} \alpha_i^{-} y_i K(x_i, x) + b \}$$

其中: a_i^{+} 是正类支持向量的系数, a_i^{-} 是负类支持向量的系数; l_1 和 l_2 分别代表正类和负类的训练样本数。在 zSVM 方法中, a_i^{+} 值随着所有支持向量乘以一个特别小的正数 z 而增大。

同时,一些学者将外部方法和内部方法相结合,Akbani 等人[16]将 Veropoulos 的不同惩罚因子方法同 SMOTE 相结合处理不平衡分类问题;Batuwita 等人[17]提出 FSVM-CIL 算法,该算法使用模糊 SVM 处理不平衡分类问题。

3 一种鲁棒非平衡的 ELM 算法

3.1 基于核的可能性模糊 C-均值聚类算法

本文模糊隶属度值采用文献[10]中提出的基于核的可能性模糊 C-均值聚类算法,该算法是将 FCM 算法和 PCM 结合得到的,其在高维特征空间中的目标函数如下:

$$\begin{aligned} J_{\text{KPFCM}} = &\sum_{i=1}^{C} \sum_{j=1}^{n} (au_{ij}^m + bt_{ij}^p) \| \phi(x_j) - v_i^{\phi} \|^2 + \\ &\frac{\beta}{m^2 c} \sum_{i=1}^{C} \sum_{j=1}^{n} (1 - t_{ij}^p) \end{aligned} \tag{12}$$

其中: $\phi(x_j)$ 为样本 x_j 在特征空间的映射; v_i^{ϕ} 为特征空间中第 i 类的中心; u_{ij} 是模糊隶属度,表示第 j 个样本属于第 i 类的相对程度; t_{ij} 是典型值,用来表示第 j 个样本属于第 i 类的绝对程度, m 和 p 分别为模糊隶属度和典型值的权重指数,这里要使式(12)取得最小值必须满足:

$$\begin{aligned} t_{ij} &= 1/(1 + bm^2 C \| \phi(x_j) - v_i^{\phi} \|^2/\beta) \\ u_{ij} &= 1/\sum_{k=1}^{C} \langle \| \phi(x_j) - v_i^{\phi} \|^2/\| \phi(x_j) - v_k^{\phi} \|^2 \rangle \\ v_i^{\phi} &= (\sum_{j=1}^{n} (au_{ij}^m + bt_{ij}^p)\phi(x_j))/(\sum_{j=1}^{n} (au_{ij}^m + bt_{ij}^p)) \\ \beta &= \sum_{j=1}^{n} \| \phi(x_j) - \overline{v^{\phi}} \|^2/n \end{aligned} \tag{13}$$

基于核的可能性模糊 C-均值(即 KPFCM)算法描述如下:

a)固定 c、m、p、a、b、σ。 $1 < c < n$, $1 < m < +\infty$, $1 < p < +\infty$,设置循环初始值 $r = 1$ 和最大循环数 $r_{\max}$,设置算法停止的阈值 ε。

b)计算训练样本集中每一类样本的类中心作为初始聚类中心 $V^{(0)}$,利用初始聚类中心得到初始 $U^{(0)}$ 和典型值矩阵 $T^{(0)}$,循环。

c)利用式(13)分别更新 $U^{(r)}$ 矩阵和 $T^{(r)}$ 矩阵,直到 $\| U^{(r)} - U^{(r-1)} \| < \varepsilon$ 或 $r > r_{\max}$。

3.2　基于 KPFCM 的非平衡 ELM 算法

从式(4)可以看出,ELM 为所有的正类样本和负类样本分配了相同的惩罚系数 C,由于负类样本数目远远多于正类样本,这会导致负类样本总体对分类的输出函数产生较大的影响。考虑到这一点,本文提出了一种改进的处理不平衡数据集分类问题的 ELM 算法,在该方法中,本文为正类样本和负类样本分配了不同的惩罚系数和,其中,正类样本的惩罚系数较大,负类样本的惩罚系数较小。

由于 KPFCM 算法对噪声具有鲁棒性,本文提出的基于核的可能性模糊 C-均值聚类的 ELM 算法也对噪声具有一定的鲁棒性。对于含有噪声的样本集,本文为新的 ELM 算法引入了一个模糊隶属度的概念,该模糊隶属度值采用 KPFCM 聚类方法求得。该聚类方法对于噪声数据具有较高的鲁棒性,本文引入该隶属度的 ELM 方法对于噪声的敏感性减小,则新的两类 ELM 模型(后面称 ELM-CIL)定义如下:

$$L_{\mathrm{ELM}} = \frac{1}{2}\|\beta\|^2 + \frac{C_+}{2}\sum_{i=1}^{N_+} m_i \varepsilon_i^2 + \frac{C_-}{2}\sum_{i=N_++1}^{N} m_i \varepsilon_i^2 \tag{14}$$

$$\text{subject to: } h(x_i)\beta = t_1 - \varepsilon_i \quad i = 1, \cdots, N$$

令 $C_+ = rC_-$,$C_- = C$,本文取 r 为正负类样本的比值,则式(14)可以变成下面的形式:

$$L_{\mathrm{ELM}} = \frac{1}{2}\|\beta\|^2 + r\frac{C}{2}\sum_{i=1}^{N_+} m_i \varepsilon_i^2 + \frac{C}{2}\sum_{i=N_++1}^{N} m_i \varepsilon_i^2 \tag{15}$$

$$\text{subject to: } h(x_i)\beta = t_i - \varepsilon_i \quad i = 1, \cdots, N$$

同理,可以得到 KKT 约束条件:

$$\frac{\partial L_{\mathrm{ELM}}}{\partial \beta} = 0 \rightarrow \beta = \sum_{i=1}^{N} \alpha_i h(x_i)^{\mathrm{T}} = H^{\mathrm{T}}\alpha \tag{16}$$

$$\frac{\partial L_{\mathrm{ELM}}}{\partial \varepsilon_i} = 0 \rightarrow \left\{ \begin{array}{ll} \alpha_i = rCm_i\varepsilon_i & i = 1, \cdots, N_+ \\ \alpha_i = Cm_i\varepsilon_i & i = N_+ + 1, \cdots, N \end{array} \right\} \tag{17}$$

$$\frac{\partial L_{\mathrm{ELM}}}{\partial \alpha_i} = 0 \rightarrow h(x_i)\beta - t_i + \varepsilon_i = 0 \quad i = 1, \cdots, N \tag{18}$$

将式(16)(17)带入式(18),最终,本文提出的 ELM-CIL 分类器的输出函数可以写成:

$$f(x) = h(x)\beta = h(x)H^{\mathrm{T}}\left(HH^{\mathrm{T}} + \frac{(MR)^{-1}}{C}\right)T \text{ 或}$$

$$f(x) = h(x)\beta = h(x)\left(H^{\mathrm{T}}H + \frac{(MR)^{-1}}{C}\right)H^{\mathrm{T}}T \tag{19}$$

$$M = \begin{bmatrix} m_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & m_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & m_{N-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & m_N \end{bmatrix}$$

$$R = \begin{bmatrix} r & 0 & 0 & \cdots & 0 & 0 \\ 0 & r & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

R 是一个由 N_+ 个 r 和 N_- 个 1 构成的对角矩阵。

4　实验

4.1　ELM-CIL 绩效评估

作为 ELM 训练环境,本文选择广泛使用的 LIBSVM 软件包,可以方便、快速处理矩阵运算。训练 ELM 分类器之前,每个数据集缩放在[-1, +1]中。在进行 SVM 分类研究过程中,多次实验结果表明当训练样本不平衡时,常用的性能测试方法(即求所有样本正确分类的比例)对于算法结果的绩效评估较为片面。因此本文使用几何平均敏感度 $GM = \sqrt{SE \times SP}$(其中 SE 表示正确分类的正类比例,SP 表示正确分类的负类比例)进行分类器的绩效评估。

本文中 ELM 学习过程中需要调整隐藏层节点数目以及惩罚系数,且当样本数目一定时,隐藏层节点数目增加到一定数目后趋于稳定。本文实验中隐藏层节点数目 L 统一设置为 L = 100(Waveform 除外),惩罚系数 C 在 $\{2^{-8}, 2^{-7}, \cdots, 2^8\}$ 区间选择。对于 SVM 训练过程中的参数选择,令 C 以及 γ 参数都从 $\{2^{-8}, 2^{-7}, \cdots, 2^8\}$ 区间选择。

4.2　人工数据集

本文随机生成 270 个正态分布的样本点,这些样本点被分为两类,两个类别的正态分布 $N(u, \Sigma)$ 分别为

类 1　N:([1;0],[0.2 0;0 0.3]),共 70 个点。

类 2　N:([2.5;0],[0.3 0;0 0.4]),共 200 个点。

为了验证本文提出的方法对噪声的鲁棒性,随机生成 100 个噪声数据,加入到正态分布的数据集中,图 1、2 分别是未加噪声和加噪声的样本集的二维分布。

本文分别采用 SVM、ELM 和 ELM- CIL 方法对两个数据集进行分类。这三个算法在无噪声数据集上的分类结果显示在表 1 中,加噪声数据集的分类结果在表 2 中。结果表明 ELM-CIL 方法比传统 SVM 和 ELM 方法具有较高的准确度值,尤其对于加噪声数据集 ELM-CIL 方法具有较高的鲁棒性。

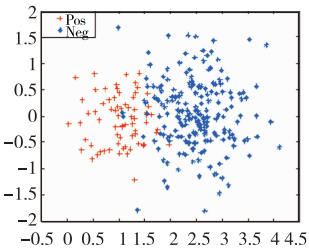

图 1　正常分布的数据集分布

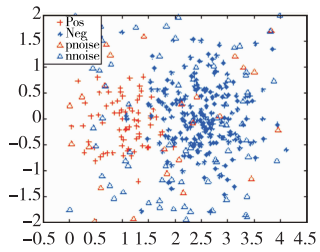

图 2　加噪声后的数据集分布

表 1　无噪声数据集的 SVM、ELM、ELM_CIL 分类结果

参数	normal SVM	ELM	ELM-CIL
accuracy	95.19	95.56	96.30
SE	90.00	85.71	97.14
SP	97.00	99.00	96.00
GM	93.43	92.12	96.57

表 2　加噪声数据集的 SVM、ELM、ELM_CIL 分类结果

参数	normal SVM	ELM	ELM-CIL
accuracy	82.97	83.78	85.14
SE	62.50	62.5	73.33
SP	92.80	94	90.80
GM	76.16	78.85	81.60

4.3　UCI 数据集对比实验

本文采用五个取自 UCI 机器学习知识库的实际非平衡数据集验证提出的 ELM-CIL 方法。表 3 总结了这些数据集中正负类数据比例的详细信息。其中包括了正类样本数 Pos、负类样本数 Neg、样本总数 total、正负类样本非平衡比例 ratio。对于多类数据集除了选为正类的样本,属于其他类的样本都作为负类数据集。

表 3　数据集信息

dataset	total	Pos	Neg	ratio
Pima-Indians	768	268	500	7 : 13
Transfusion	748	178	570	6 : 19
Ecoli	336	77	259	23 : 77
Haberman	306	81	225	13 : 37
Waveform	5 000	1 657	3 343	33 : 67

1)ELM 和 SVM 方法训练时间结果比较

用传统的 ELM 方法和 SVM 方法分别对上面提到的五个数据集进行分类,将各个数据集的训练时间统计在表 4 中。由于本文中 ELM 学习过程中只需要调整隐藏层节点数目以及惩罚系数,且当样本数目一定时,隐藏层节点数目增加到一定数目后趋于稳定,所以隐藏层节点设置较简单;SVM 方法需要调整核函数、惩罚系数等参数,学习过程需要迭代多次以寻找各个参数的最优值。所以,ELM 训练时间远远小于 SVM 的训练时间,两种算法训练时间结果如表 4 所示。

表 4 ELM 和 SVM 方法在各数据集上训练时间结果比较 /s

方法	Haberman	Ecoli	Pima	Transfusion	Waveform
SVM	21.17	8.35	162.6	105.79	5 812.04
ELM	1.06	1.28	8.14	17.73	960.56

2)传统 ELM 和本文提出的 ELM-CIL 方法结果比较

分别用传统的 ELM 方法和本文提出的 ELM-CIL 方法对各数据集进行分类,实验过程主要生成 ACCU、*SE*、*SP* 和 *GM* 四个结果。其中,ACCU 表示所有样本总体的分类正确率,*SE* 表示正确分类的正类样本比例,*SP* 表示正确分类的负类样本比例,*GM* 表示正负类样本的平均敏感度,将分类结果统计在表 5 中。从表 5 可以明显看出,ELM-CIL 分类器对正类的分类结果(即 *SE*)远远优于传统的 ELM 分类器,虽然总体分类精度(即 ACCU)略差于 ELM 方法,但其 *GM* 指数有明显的提高,所以本文提出的 ELM-CIL 方法大大提高了传统 ELM 算法处理不平衡数据集的能力,尤其提高了少数类的分类精度。结果表明 ELM-CIL 对于不平衡数据集敏感性降低。

表 5 传统 ELM 和本文提出的 ELM-CIL 方法分类结果

database	method	ACCU	*SE*	*SP*	*GM*
Haberman	ELM	78.10	32.10	94.67	55.13
	ELM-CIL	68.63	66.67	69.33	67.99
Ecoli	ELM	93.15	81.82	96.53	88.87
	ELM-CIL	88.10	97.40	85.33	91.16
Pima-Indians	ELM	73.96	43.66	90.2	62.75
	ELM-CIL	73.57	76.12	72.20	74.13
Transfusion	ELM	79.41	29.78	94.91	53.16
	ELM-CIL	64.30	61.75	72.47	66.90
Waveform	ELM	90.12	88.11	91.11	89.60
	ELM-CIL	90.16	94.09	88.21	91.10

3)本文提出的 ELM-CIL 和 FPSVM-CIL 方法结果比较

对上面提到的四个数据集分别统计其采用 SVM、文献[10]中提出的 FPSVM-CIL、ELM、ELM-CIL 方法获得的分类结果,从图 3 ~6 中可以看出,采用传统 SVM 及 ELM 方法比不平衡样本处理方法的样本总体分类比例(即 ACCU)略高,但是采用 ELM-CIL 和 FPSVM-CIL 方法正确分类的正类样本比例远远优于传统方法,即其 *SE* 指数以及 *GM* 指数比传统的 SVM 和 ELM 方法有明显的提高。

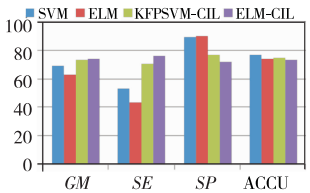

图 3 Pima-Indians 各分类方法结果比较

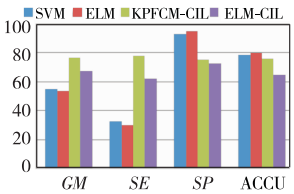

图 4 Transfusion 各分类方法结果比较

从四个数据集分类结果总体来看,ELM-CIL 总体测试精度略低于 FPSVM-CIL 方法,但由于 ELM-CIL 在学习速度上明显优于支持向量机,所以对于大规模样本集的不平衡数据分类来说,本文提出的算法具有很大的优势。

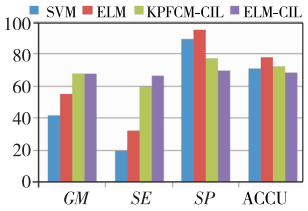

图 5 Haberman 各分类方法结果比较

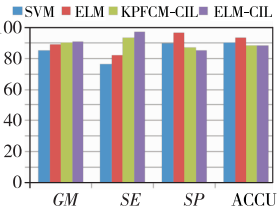

图 6 Ecoli 各分类方法结果比较

5 结束语

本文对传统的 ELM 方法进行了改进,提出了一种处理不平衡数据集分类问题的 ELM-CIL 方法。该方法相比于传统 ELM 方法,以略微降低多数类样本分类精度为代价,大大提高了少数类样本的分类精度,同时引入的模糊隶属度值减小了外围噪声点的影响。实验结果表明,本文提出的方法成为处理不平衡样本分类问题的一种有效方法,不同于 FPSVM-CIL 方法,ELM-CIL 学习过程中不需要调整大量参数,所以算法学习速度快,尤其适合处理大规模的不平衡数据集分类问题。

参考文献:

[1] HUANG Guang-bin, ZHU Qin-yu, SIEW C K. Extreme learning machine: theory and applications[J]. Neurocomputing, 2006, 70(1-3):489 -501.

[2] HUANG Guang-bin, CHEN Lei. Enhanced random search based incremental extreme learning machine[J]. Neurocomputing, 2008, 71(16-18):3060-3068.

[3] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagation errors[J]. Nature, 1986, 323: 533-536.

[4] RUMELHART D E, McCLELLAND J L. Parallel distributed processing[M]. Cambridge: MIT Press, 1986: 125 -187.

[5] HUANG Guang-bin, CHEN Lei, SIEW C K. Universal approximation using incremental constructive feedforward networks with random hidden nodes[J]. IEEE Trans on Neural Networks, 2006, 17(4): 879-892.

[6] LIANG Nan-ying, HUANG Guang-bin, 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.

[7] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.

[8] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 6(1):32-42.

[9] ZOU Shu-xue, HUANG Yan-xin, WANG Yan, *et al*. SVM learning from imbalanced data by GA sampling for protein domain prediction [C]//Proc of the 9th International Conference for Young Computer Scientists. 2008:982-987.

[10] MENG Fan-rong, GAO Chun-xiao, LIU Bing. Fuzzy possibilistic support vector machines for class imbalance learning[J]. Journal of Convergence Information Technology, 2013, 8(3):692-701.

[11] HAYKIN S. Neural networks: a comprehensive foundation[M]. New Jersey: Prentice Hall, 1999.

(下转第 1004 页)

为了更客观地定量分析本文算法的解卷积效果和性能,采用相关系数的绝对值 ζ_{ik} 和重构信噪比 S/N 两个指标来对算法进行评价。相关系数的绝对值 ζ_{ik} 定义为[17]

$$\zeta = \zeta(\boldsymbol{y},\boldsymbol{s}) = \left| \frac{\sum_{t=1}^{q} y(t)s(t)}{\sqrt{\sum_{t=1}^{q} y^2(t) \sum_{t=1}^{q} s^2(t)}} \right| \tag{14}$$

式中:$\boldsymbol{s}$ 为原始信号,$\boldsymbol{y}$ 为解卷积信号。当 $y(t) = \lambda s(t)$ 时,$\zeta = 1$。ζ 的值越接近 1,表示解卷积信号与原始信号的相似程度越高。

重构信噪比 S/N 定义为[18]

$$S/N = -10 \lg \frac{\| \boldsymbol{y} - \boldsymbol{s} \|^2}{\| \boldsymbol{s} \|^2} \tag{15}$$

S/N 的值越大,表示解卷积信号与原始信号的误差越小,解卷积效果越好。本文算法中粒子群进化代数均设定为 100 代。表 1 所示为本文算法针对不同原始信号进行解卷积的性能,表中数据为 20 次独立仿真实验的统计平均值。

表 1 算法解卷积性能

原始信号	文献[16]算法相关系数绝对值	文献[16]算法重构信噪比	本文算法相关系数绝对值			本文算法重构信噪比		
			最大值	最小值	平均值	最大值	最小值	平均值
signal 1	0.9652	9.4229	0.9859	0.9652	0.9662	15.0666	9.4182	9.7050
signal 2	0.9804	13.1479	0.9954	0.9803	0.9816	18.5248	13.1436	13.5155
signal 3	0.9686	10.9716	0.9941	0.9679	0.9699	16.3731	10.8910	11.2393
signal 4	0.9776	12.0302	0.9798	0.9771	0.9777	12.5175	12.0236	12.0626

由于文献[16]的方法采用的是特征值分解的解析方法,所以每次实验中解卷积的结果完全相同。而本文算法采用进化计算的方法实现信号的盲解卷积,每次实验的结果具有随机特性。因此,实验中针对多次仿真过程统计了算法性能的最大值、最小值和平均值,从而更加客观地评价本文算法的性能。

通过观察表 1 中的数据可知,本文算法的盲解卷积效果较好,解卷积后的信号与原始信号的相关系数绝对值均高于 0.95,重构信噪比达到或接近 10 dB,且其平均性能优于文献[16]的方法。

4 结束语

本文在信号时间可预测性度量的理论基础上,提出了一种基于粒子群优化的盲解卷积算法。算法利用了信号的时间结构信息,采用粒子群优化算法对代价函数进行求解,实现了对声音信号的成功盲解卷积。由于本文算法仅利用卷积信号构成的协方差矩阵进行优化求解得到解卷滤波器,而无须对所有卷积信号样本点进行反复迭代计算。因此,算法在保证高解卷积精度的同时计算量很低。计算机仿真结果也表明,本文算法中粒子群进化收敛的速度很快,且算法对不同声音信号均具有良好的解卷积效果。

参考文献:

[1] GAZZAH H. SOS-based blind channel equalization with quadratic complexity[J]. IEEE Trans on Signal Processing, 2011, 59(2): 837-841.

[2] YU Cheng-pu, ZHANG Ci-sheng, XIE Li-hua. A blind deconvolution approach to ultrasound imaging[J]. IEEE Trans on Ultrasonics, Ferroelectrics and Frequency Control, 2012, 59(2): 271-280.

[3] WU Guo-rong, LIAO Wei, STRAMAGLIA S, *et al*. A blind deconvolution approach to recover effective connectivity brain networks from resting state fMRI data[J]. Medical Image Analysis, 2013, 17(3): 365-374.

[4] TAKAHATA A K, NADALIN E Z, FERRARI R, *et al*. Unsupervised processing of geophysical signals: a review of some key aspects of blind deconvolution and blind source separation[J]. IEEE Signal Processing Magazine, 2012, 29(4): 27-35.

[5] CADZOW J A, LI X K. Blind deconvolution[J]. Digital Signal Processing, 1995, 5(1): 3-20.

[6] FIORI S. Blind deconvolution by simple adaptive activation function neuron[J]. Neurocomputing, 2002, 48(1-4): 763-778.

[7] 段海滨, 张祥银, 徐春芳. 仿生智能计算[M]. 北京: 科学出版社, 2011.

[8] GORRIZ J M, PUNTONET C G, ROJAS F, *et al*. Optimizing blind source separation with guided genetic algorithms[J]. Neurocomputing, 2006, 69(13-15): 1442-1457.

[9] HSIEH S T, SUN T Y, LIN Chun-ling, *et al*. Effective learning rate adjustment of blind source separation based on an improved particle swarm optimizer[J]. IEEE Trans on Evolutionary Computation, 2008, 12(2): 242-251.

[10] 陈雷, 张立毅, 郭艳菊,等. 基于细菌群体趋药性的有序盲信号分离算法[J]. 通信学报, 2011, 32(4): 77-85.

[11] 陈雷, 张立毅, 郭艳菊,等. 基于粒子群优化的有序盲信号分离算法[J]. 天津大学学报, 2011, 44(2): 174-179.

[12] SUN T Y, LIU Chan-cheng, TSAI S J, *et al*. Cluster guide particle swarm optimization(CGPSO) for underdetermined blind source separation with advanced conditions[J]. IEEE Trans on Evolutionary Computation, 2011, 15(6): 798-811.

[13] 陈雷, 张立毅, 郭艳菊,等. 基于智能单粒子与信号变化的盲源分离算法[J]. 电路与系统学报, 2012, 17(4): 89-94.

[14] KENNEDY J, EBERHART R C. Particle swarm optimization[C]// Proc of IEEE International Conference on Neural Networks. 1995: 1942-1948.

[15] 胡旺, 李志蜀. 一种更简化而高效的粒子群优化算法[J]. 软件学报, 2007, 18(4): 861-868.

[16] STONE J V. Blind deconvolution using temporal predictability[J]. Neurocomputing, 2002, 49(1-4): 79-86.

[17] 谢胜利, 何昭水, 高鹰. 信号处理的自适应理论[M]. 北京: 科学出版社, 2006.

[18] BOFILL P, ZIBULEVSKY M. Underdetermined blind source separation using sparse representations[J]. Signal Processing, 2001, 81(11): 2353-2362.

(上接第 988 页)

[12] WEISS G. Mining with rarity: a unifying framework[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1):7-19.

[13] CHAWLA N, BOWYER K, HALL L, *et al*. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[14] JO T, JAPKOWICZ N. Class imbalances versus small disjuncts[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1):40-49.

[15] VEROPOULOS K, CAMPBELL C, CRISTIZNINI N. Controlling the sensitivity of support vector machines[C]//Proc of International Joint Conference on AI. 1999:55-60.

[16] AKBANI R, KWEK S, JAPKOWICZ N. Applying support vector machines to imbalanced datasets[C]//Proc of the 15th European Conference on Machine Learning. Berlin: Springer, 2004:39-50.

[17] BATUWITA R, PALADE V. FSVM-CIL: fuzzy support vector machines for class imbalance learning[J]. IEEE Trans on Fuzzy Systems, 2010, 18(3):558-571.