doi: 10.3969/j.issn.1005-3026.2017.06.008

基于加权极限学习机的肿瘤基因表达谱数据分类

姜琳颖, 余东海, 石 鑫
(东北大学 软件学院, 辽宁 沈阳 110169)

摘 要: 基因表达谱数据一般来源于临床试验,而在临床试验中,试验样本的类分布情况是不确定的,这就使得表达谱数据往往具有比较明显的不平衡性. 采用加权极限学习机来对不平衡基因表达谱数据进行分类,为了减少因为不平衡数据引起的分类误差,一个临时的权重被分配给每一个样本以增强少样本类的影响,同时减少多样本类的影响,进而提高肿瘤分类的准确率. 实验结果表明,所提方法能够提高少样本类的识别率,从而提高分类器的总体性能.

关 键 词: 基因;表达谱数据;加权极限学习机;不平衡性;肿瘤分类
中图分类号: TP 391.4 **文献标志码**: A **文章编号**: 1005-3026(2017)06-0798-06

Tumor Microarray Gene Expression Data Classification Based on Weighted Extreme Learning Machine

JIANG Lin-ying, *YU Dong-hai*, *SHI Xin*
(School of Software, Northeastern University, Shenyang 110169, China. Corresponding author: JIANG Lin-ying, E-mail: jiangly@swc.neu.edu.cn)

Abstract: With the development of gene microarray technology, gene expression profiling becomes a significant method for identifying different types of canners. Microarray gene expression data is from clinical trials in general, where the class distribution of samples is changeable, which makes the expression data have a chance to become more imbalanced. In this paper, the weighted extreme learning machine (WELM) was used to classify the imbalance microarray gene expressing data. In order to reduce classification error caused by the imbalance data, a weight was assigned to each sample in order to enhance the impact of minority class while reducing majority class's impact, and improve the accuracy of tumor classification. The experimental results show that the minority class recognition rate can be well improved by the proposed method, so as to improve the overall performance of classifiers.

Key words: gene; microarray expressing data; WELM; imbalance; tumor classification

基因芯片技术的出现,使得人们能够同时监控成百上千的基因表达水平. 但对于生物信息学来说,如何有效地利用这些基因表达数据来预测和诊断疾病也具有很大的挑战性. 肿瘤分类是其中一种典型的值得关注的应用,近年来已经有很多的研究证明了其可行性[1-5].

一般来说,基因表达谱数据来源于临床试验,而在临床试验中,试验样本的类分布情况是不确定的,这就使得表达谱数据往往具有比较明显的不平衡性. 如文献[1]中所用的结肠癌数据集(Colon),总共具有62个样本:其中40个为癌症样本,22个为正常样本. 而且数据类型的不平衡性不仅仅是表现在健康和病变样本之间,也有在不同肿瘤样本之间进行分类时出现的类型不平衡,例如文献[4]中的Lung-Brigham数据集,其中MPM样本31个,ADCA样本150个. 所以可以看出基因表达谱数据集中的类不平衡问题是普遍存在的. 而当要分类的数据具有复杂的类分布时,由于一般的学习算法都是默认地将数据假设为类平衡分布或者是相同的错分代价,而使得分

收稿日期: 2015-12-30
基金项目: 国家自然科学基金资助项目(61272176).
作者简介: 姜琳颖(1972-),女,山东龙口人,东北大学副教授.

类结果更偏向于多样本类别[6].

对于不平衡数据分类问题,现如今主要采用的方法有:过采样、欠采样以及代价敏感方法等.过采样是指对少数样本类(少类)的样本重复采样以降低数据的不平衡性,如文献[7]中的少类样本随机重复采样(oversampling),文献[8]中采用 SMOTE 方法对少类样本进行过采样处理等,但过采样比较难以区分重复采样的是有益样本还是冗余样本,这就使得采样的结果得不到有效保证;欠采样是指对多数样本类(多类)的样本进行部分采样以使数据尽量达到平衡,如文献[9]中基于聚类融合欠抽样的改进 AdaBoost 分类算法,就是采用欠采样的方式来处理数据的不平衡问题. 欠采样会导致样本信息的丢失,基因表达谱数据的样本信息普遍较少,故不适合做欠采样处理,而且欠采样也具有与过采样相同的问题;难以区分冗余和有益样本数据;代价敏感学习是通过对少类样本赋予更多的错分代价方式来处理不平衡数据问题,如文献[10]中的代价敏感超网络算法,文献[6]中的加权极限学习机算法. 总之,代价敏感学习能够在保全所有样本信息的情况下,通过赋予一个代价敏感错分矩阵来处理不平衡数据问题,但是该方法的难点在于难以确定一个有效的代价敏感错分矩阵. 本文所提出的算法是在文献[6]提出的加权极限学习机基础上增加类别权重值的可调性,使其能够更好处理数据的不平衡性,并将其应用于基因表达谱数据.

1　方　　法

1.1　基因特征提取算法

每个基因表达谱样本都记录了组织细胞中所有可测基因的表达水平,这使得每个样本都具有很高的维度,但实际上只有其中的少数基因包含了样本的某种分类信息,与样本的类别有关,这些特殊的基因被称为分类特征基因. 而分类特征基因的选取是建立有效分类模型的关键之一. 目前,已经有很多有效的特征提取算法应用于基因表达谱数据. 例如文献[4]中的 ReliefF 算法,文献[11]提出的 GBC 算法,文献[12]中的分层特征提取算法等. ReliefF 算法原理较为简单,而且已在其他文献中证明其具有较好的性能,所以本文采用该算法对数据集进行特征提取.

ReliefF 是在 Relief 算法的基础上提出的利用特征之间的依赖强度来估计特征质量的算法. 该算法的基本思想在于调整权重矩阵 $\boldsymbol{W}=[w(1),w(2),\cdots,w(m)]$ 来获得特征之间更多的关联,从而用来更好区分不同类别.

算法随机选出一个样本 x,然后分别在其同类样本(h_j)和不同类样本(m_j)中分别选出 k 个最近邻域的样本对权值进行更新. 更新方程为

$$\boldsymbol{W}(f)=\boldsymbol{W}(f)-\frac{\sum_{j=1}^{K}\operatorname{dist}(f,x,h_j)}{t\cdot k}+\sum_{c\neq\operatorname{class}(x)}\frac{P(c)}{1-P(\operatorname{class}(x))}\cdot\frac{\sum_{j=1}^{K}\operatorname{dist}(f,x,m_j)}{t\cdot k}. \tag{1}$$

其中:$P(c)$表示类别 c 的先验概率;$P(\operatorname{class}(x))$表示包含 x 样本的类别的先验概率;$\operatorname{dist}(f,x,h_j)$和 $\operatorname{dist}(f,x,m_j)$分别表示基于 f 特征,样本 x 和同类样本 h_j及不同类样本 m_j 之间的距离.

在重复运算 t 次之后,算法最终可以得到每一个特征的相关系数权重值,然后利用该权重值,从原始样本 m 个基因中提取权重值排名靠前的 g 个基因作为特征基因($g<m$).

1.2　评价指标

通常情况下,总精度能够用来很好地评价一个分类器的性能. 然而对于不平衡数据,这种评价方法有可能会忽略少样本类的低识别率[6],而过多加大多样本类别带来的影响. 为了考虑到每一个类别的识别率,本文采用 G - mean 评价指标方法. 在计算每一个类别的准确率之后,然后取这些准确率的平方根. 二分类问题举例:

$$\text{G - mean}=\sqrt{\frac{\text{TP}}{\text{TP}+\text{TN}}\times\frac{\text{TN}}{\text{TN}+\text{FP}}}. \tag{2}$$

其中 TP,TN,FP,FN 分别表示真阳,真阴,假阳和假阴. 敏感性(sensitivity)和特异性(specificity)见式(3):

$$\text{sensitivity}=\frac{\text{TP}}{\text{TP}+\text{FN}},\ \text{specificity}=\frac{\text{TN}}{\text{TN}+\text{FP}}. \tag{3}$$

1.3　加权极限学习机

极限学习机(extreme learning machine)[13]是一种基于最小二乘,拥有快速训练和较好泛化能力的单隐层前馈神经网络算法(SLFNS). 它已经在许多肿瘤诊断应用中表现出很好的性能[14-16]. 加权极限学习机(weight extreme learning machine)是在 ELM 的基础上引入加权矩阵,对每一个样本进行加权,减少样本类间可能存在的不平衡性,从而提高样本总体的识别率.

对于包含 N 个样本的数据集$\{(x_i,t_i)\mid x_i\in$

$\mathbf{R}^n, t_i \in \mathbf{R}^o, i = 1, 2, \cdots, N\}$,其中 x_i 为单个样本,$\boldsymbol{t}_i$ 为目标向量,那么包含 n 个神经元结点的 SLFN,假设其激励函数为 $g(x)$,则其模型可表示为

$$f_n(x_i) = \sum_{k=1}^{n} \boldsymbol{\theta}_k g(\boldsymbol{w}_k \cdot x_i + b_k) = \boldsymbol{t}_i . \quad (4)$$

其中 $\boldsymbol{w}_k = [w_{k1}, \cdots, w_{km}]^{\mathrm{T}}$ 表示为输入结点到第 k 个隐藏层结点之间的权重值,$\boldsymbol{\theta}_k = [\theta_{k1}, \cdots, \theta_{kn}]$ 为链接第 k 个隐藏层结点到输出结点的输出权重,$\boldsymbol{t}_i = [t_1, \cdots, t_N]^{\mathrm{T}}$ 是目标向量,b_k 为第 k 个隐藏层结点的偏移值.

使 $\boldsymbol{H} = \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_n \cdot x_1 + b_n) \\ \vdots & & \vdots \\ g(w_1 \cdot x_n + b_1) & \cdots & g(w_n \cdot x_n + b_n) \end{bmatrix}$,

$\boldsymbol{\theta} = [\theta_1, \cdots, \theta_n]$ 和 $\boldsymbol{t}_i = [t_1, \cdots, t_N]$,则式(4)可简化为

$$\boldsymbol{H\theta} = \boldsymbol{T} . \quad (5)$$

则输出链接权重 $\boldsymbol{\theta}$ 可以通过求解线性系统的最小二乘解得到:

$$\boldsymbol{\theta} = \boldsymbol{H}^{+}\boldsymbol{T} . \quad (6)$$

其中,$\boldsymbol{H}^{+}$ 为隐藏层结点输出的 Moore - Penros 广义逆矩阵. 为了提高分类器的稳定性,本文采用正交投影分解法求解 $\boldsymbol{\theta}$. 然后对不同类别的样本进行不同的权值加权,则 WELM 算法求解隐藏层输出权重可以表示为

$$\boldsymbol{\theta} = \begin{cases} \boldsymbol{H}^{\mathrm{T}}(1/c + \boldsymbol{WHH}^{\mathrm{T}}) & N < L, \\ (1/c + \boldsymbol{H}^{\mathrm{T}}\boldsymbol{WH})\boldsymbol{H}^{\mathrm{T}}\boldsymbol{WT} & N \geqslant L . \end{cases} \quad (7)$$

其中 $\boldsymbol{W}$ 为一个对角矩阵,对角线的每一个元素代表相应样本的权重值;c 为正则化参数,在文献[6]中已证明 c 值较小时(小于 2^0)分类器的性能会很差,而当 c 取值大于 2^0 时,分类器的性能表现得很稳定,所以本文 c 值都取 2^{10}.

文献[6]中只是简单考虑到用样本数量来确定权值,本文为了提高算法的分类精度以及使算法的适用性更广,引入一个调整参数. 初始权重值设定为该类样本的数量的倒数.

$$w_{ii} = \frac{1}{\#(\boldsymbol{t}_i)} . \quad (8)$$

其中$\#(\boldsymbol{t}_i)$为第 i 个训练样本对应类的样本总量. 假设 w_a 和 w_b 分别代表多类和少类的样本初始化权值,w_a' 和 w_b' 为本文所用权值.

$$\begin{aligned} w_a' &= w_a \cdot (1 - \alpha) \\ w_b' &= w_b . \end{aligned} \quad (9)$$

其中参数 α 的确定将在下文详细介绍. 本文所提出算法流程如下:

1) 将实验数据分为训练集(Tr)和测试集(Te),然后将训练集数据分为参数测试集(P - Tr)和参数寻优集(P - Op);

2) 使用参数测试集(P - Tr)建立训练分类器,初始化权值矩阵为式(9),参数 $\alpha = 0$,正则化参数 c 为 2^{10};

3) 使用参数寻优集(P - Op)测试步骤 2)建立的分类器,计算 α 取值在 0 到 1 之间变动时分类器 G - mean 值,并选取 G - mean 值最大时的 α 值;

4) 使用步骤 3)确定的参数 α 值和训练集(Tr)建立分类器,然后使用测试集(Te)测试该分类器性能.

2 实 验

为了验证本文提出算法的有效性,在实验中,将使用三个基因芯片数据集分别采用本文算法,以及 SVM,ELM 和文献[6]的算法建立分类器,然后用 G - mean 评价指标对每一个分类器的性能进行评价. 数据集 1 为急性白血病数据集,该数据集包含 2 096 个样本,54 675 个基因,14 个类别的数据,本文只讨论二分类问题,所以选取其中 AML 正常核形(351 个样本)和 AML 复杂异常核形(48 个样本)作为建立分类器的原始数据集. 数据集 2 是 Kent Ridge 生物医学数据库的肺癌数据集,该数据集包含 181 个样本,12 533 个基因,其中 31 个恶性胸膜间皮瘤(MPM)样本,150 个恶性腺瘤(ADCA)样本. 表 1 给出了这两个数据集的简要描述,可以看出以上两个实验数据集其数据都具有较大不平衡性. 为了验证本文算法对于平衡数据集也有较高的支持度,实验中还加入了第 3 个数据集(前列腺癌数据集),该数据集总共包含 136 个样本,每一样本有 12 600 个基因特征值. 其中癌症样本为 77 个,健康样本为 59 个. 前列腺癌数据集较急性白血病和肺癌数据集平衡性高很多.

表 1 实验数据集

Table 1 Description of experimental dataset

数据集	Class1	Class2	总样本数	基因数
急性白血病数据	351	48	399	54 675
肺癌数据	150	31	181	12 533
前列腺癌数据	77	59	136	12 600

在实验过程中,首先在总样本中随机选出2/3作为测试集(Tr),其余 1/3 作为训练集(Te). 本文提出算法为了确定参数 α,在测试集中再随机

选出 1/2 作为参数训练集(P – Tr),余下为参数寻优集(P – Op). 同时为了证实本文算法的性能,也使用 SVM,ELM 和文献[6]的算法分别对该数据进行分类,然后和本文算法进行对比. 这里假设少类为正标签,多类为负标签,则测试结果中的 sensitivity 表示的是少类的识别率,specificity 表示的是多类的识别率.

2.1　急性白血病数据集

由于急性白血病数据集每个样本有 54 675 个特征值,但并不是每个特征值都对分类具有帮助,且样本维数过高会降低算法的学习速度,所以需先将测试集数据采用 ReliefF 算法进行特征值提取. 图 1 为提取的特征值在 WELM 分类算法下的特异性值.

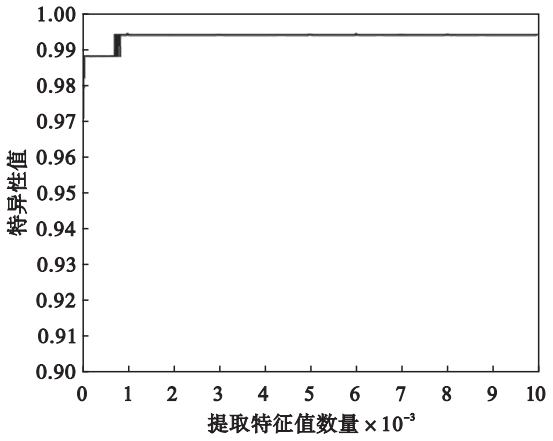

图 1　ReliefF 算法提取的特征值数量与 WELM 分类算法下的特异性关系图

Fig. 1　Classification specificity with the WELM and the number of genes selected by ReliefF

从图 1 可以看出,当特征值数量超过 1 000 之后,其特异性值较为稳定. 因此,可将测试集每一个样本均采用 ReliefF 算法提取 1 000 个特征值来建立分类器.

图 2 表示 WELM 在隐藏层神经元个数从0 ~ 2 000 下分类精度 G – mean 的数值变化.

从图 2 中可以看出当神经元个数超过 800 之后,G – mean 的数值已比较稳定,则隐藏层神经元个数可确定为 1 000.

从表 2 可以看出参数 α 取值越大(多类的权值越小),本文算法的敏感性就越高,相对的特异性变低,即对少类识别率更好,对多类识别率降低. 因此,可以认为能够取到一个适当的 α 值,使得少类识别率大幅提高,而多类识别率保持较高,从而使得 G – mean 值取得最大值.

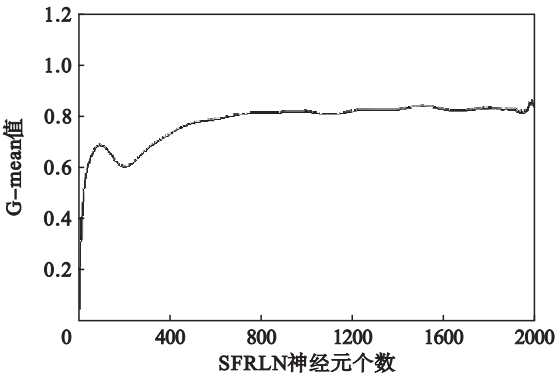

图 2　WELM 隐藏层不同数量节点下的 G – mean 变化值

Fig. 2　Classification G-mean with the WELM has different number of SFLN nodes

表 2　训练分类器中参数 α 取值对应下的分类精度

Table 2　Trainer classifier evaluation values when varying parameter α

α 值	Sen	Spe	G – mean
0	0. 625 0	**0. 991 4**	0. 787 2
0. 9	0. 625 0	0. 991 4	0. 787 2
0. 99	0. 750 0	0. 991 4	0. 826 3
0. 999	0. 937 5	0. 939 7	**0. 938 6**
0. 999 9	**1**	0. 362 1	0. 601 7

图 3,图 4 和图 5 分别表示 α 取值范围在 0. 99 到 0. 999 9 之间时 P – Op specificity 值、P – Op sensitivity 值和 P – Op G – mean 值. 从图中可以确定参数 α 的最佳值为 0. 997 ~ 0. 998.

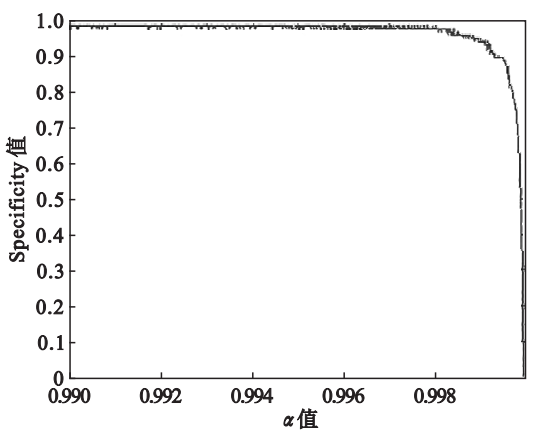

图 3　改变参数 α 下的 WELM 的 specificity 值

Fig. 3　Classification specificity with the WELM when varying the parameter α

通过 P – Op 确定参数 α 取值之后,则将全体训练集用于算法训练,然后用 Te 测试集数据来测试算法的性能,为了支持实验结果的可信性,实验中使用了 3 次 3 折交叉验证法,然后求实验结果的平均值. 从表 3 中可以看出几个算法的总精度都能达到很高的程度,但是少类识别率和 G – mean值是本文所提出的算法最好,文献[6]

算法其次.

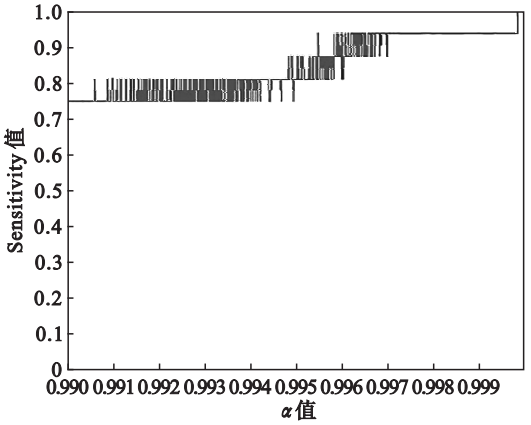

图 4 改变参数 α 下的 WELM 的 sensitivity 值

Fig. 4 Classification sensitivity with the WELM when varying the parameter α

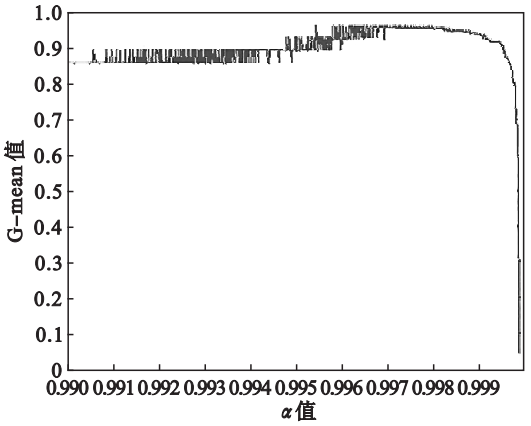

图 5 改变参数 α 下的 WELM 的 G – mean 值

Fig. 5 Classification G-mean with the WELM when varying the parameter α

表 3 急性白血病数据集的测试结果

Table 3 Testing results of acute leukemia dataset

分类算法	Sen	Spe	Acc	G – mean
SVM	0. 625 0	1	0. 954 5	0. 790 6
ELM	0. 437 5	0. 982 8	0. 916 7	0. 655 7
文献[6]	0. 750 0	0. 965 5	0. 939 4	0. 851 0
本文算法	**0. 875 0**	0. 931 0	0. 924 2	**0. 902 6**

本文算法需要确定参数 α,耗费的时间较其他算法有一定增加,但本文算法应用场景并不属于实时性场景,所以算法用时间上的消耗来换取分类准确度的提高是可以接受的.

2.2 肺癌数据集

肺癌数据集共包含 181 个样本,12 533 个特征值,其中 31 个恶性胸膜间皮瘤(MPM)样本,150 个恶性腺瘤(ADCA)样本.

从表 4 中的值可以看出,SVM,ELM,文献[6]的算法都不能使少类识别率达到 1,而本文的算法通过调整参数 α 达到该效果.

表 4 肺癌数据集的测试结果

Table 4 Testing results of lung cancer dataset

分类算法	sen	spe	acc	G – mean
SVM	0. 953 3	1	0. 990 3	0. 976 4
ELM	0. 933 3	0. 986 7	0. 977 8	0. 959 6
文献[6]	0. 933 3	0. 986 7	0. 977 8	0. 959 6
本文算法	**1**	0. 986 7	0. 988 9	**0. 993 3**

2.3 前列腺癌数据集

前列腺癌数据集总共包含 136 个样本,每一样本有 12 600 个基因特征值. 其中癌症样本为 77 个,健康样本为 59 个. 前列腺癌数据集较急性白血病和肺癌数据集平衡性高很多,选择该数据集是为了验证本文算法对于平衡数据集也有较高的支持度. 该样本的数据分配方法与前两个数据集相同.

表 5 前列腺癌数据集的测试结果

Table 5 Testing results of acute prostate cancer dataset

分类算法	sen	spe	acc	G – mean
SVM	0. 631 6	0. 807 7	0. 733 3	0. 711 9
ELM	0. 789 5	0. 641 0	0. 703 7	0. 711 2
文献[6]	0. 789 5	0. 653 8	0. 711 1	0. 718 3
本文算法	**0. 952 4**	**0. 896 6**	**0. 920 0**	**0. 924 0**

从表 5 可以看出,本文算法在总体的分类准确度远远高于其他算法,证明本文算法对于平衡数据集也具有较好的性能.

3 结 语

WELM 使用加权的方式处理数据集,可以提高不平衡数据集中少类样本的识别率. 本文在 WELM 的基础上引入参数 α,通过调整参数 α 的值,来进一步提高少类的识别率,从而提高分类器整体性能. 虽然改进的算法分类在效率上有些降低,但是算法本身具有比其他算法更好的性能,这在肿瘤识别和分类问题上是更重要的. 同时本文算法能够在不平衡数据和平衡性较高的数据上都保持良好的分类性能,而且对于不平衡数据集的分类,本文算法明显优于 SVM,ELM 和文献[6]的算法.

参考文献:

[1] Lee K, Man Z H, Wang D H, et al. Classification of microarray datasets using finite impulse response extreme learning machine for cancer diagnosis[J]. *IEEE Industrial*

Electronics Society,2011,22(5):2347 – 2352

[2] Zong W W, Huang G B, Chen Y Q. Weighted extreme learning machine for imbalance learning [J]. *Neurocomputing*,2013,101(3):229 – 242.

[3] Su Y R, Wang R J, Li C X, et al. A dynamic subspace learning method for tumor classification using microarray gene expression data[C]//7th International Conference on Natural Computation. Shanghai,2011:396 – 400.

[4] Zhang X, Guan N Y, Jia Z L, et al. Semi-supervised projective non-negative matrix factorization for cancer classification[J]. *Plos One*,2015,10(9):e0138814.

[5] García V,Sánchez J S. Mapping microarray gene expression data into dissimilarity spaces for tumor classification[J]. *Information Sciences*,2015,294:362 – 375

[6] Zheng C H, Ng T Y, Zhang L, et al. Tumor classification based on non-negative matrix factorization using gene expression data[J]. *IEEE Transactions on Nanobioscience*, 2011,10(2):86 – 93.

[7] Bolón-Canedo V, Sánchez-Marono N, Alonso-Betanzos A. Distributed feature selection:an application to microarray data classification [J]. *Applied Soft Computing*, 2015, 30: 136 – 150.

[8] 李克文,杨磊,刘文英,等. 基于 RSBoost 算法的不平衡数据分类方法[J]. 计算机科学,2015,42(9):249 – 252.
(Li Ke-wen, Yang Lei, Liu Wen-ying, et al. Classification method of imbalanced data based on RSBoost[J]. *Computer Science*,2015,42(9):249 – 252.)

[9] 张枭山,罗强. 一种基于聚类融合欠抽样的不平衡数据分类方法[J]. 计算机科学,2015,42(B11):63 – 66.
(Zhang Xiao-shan,Luo Qiang. Unbalanced data classification algorithm based on clustering ensemble under-sampling[J]. *Computer Science*,2015,42(B11):63 – 66.)

[10] 郑燕,王杨,郝青峰,等. 用于不平衡数据分类的代价敏感超网络算法[J]. 计算机应用,2014,34(5):1336 – 1340.
(Zheng Yan, Wang Yang, Hao Qing-feng, et al. Cost-sensitive hypernetworks for imbalanced data classification[J]. *Journal of Computer Applications*,2014,34(5):1336 – 1340.)

[11] Alshamlan H M,Badr G H,Alohali Y A. Genetic bee colony (GBC) algorithm: a new gene selection method for microarray cancer classification[J]. *Computational Biology & Chemistry*,2015,56:49 – 60.

[12] Nguyen T,Khosravi A,Creighton D,et al. Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification[J]. *Plos One*,2015,10(3):e0120364.

[13] Huang G B,Zhu Q Y,Siew C K. Extreme learning machine: theory and applications [J]. *Neurocomputing*, 2006, 70: 489 – 501.

[14] Sánchez M J, Cruz R M, Fernández N F, et al. On the suitability of extreme learning machine for gene classification using feature selection [C]// International Conference on Intelligent Systems Design & Applications. Sanya,2010:507 – 512.

[15] Baboo S S,Sasikala S. Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis[J]. *Communication Control and Computing Technologies*,2010,4(3):748 – 757.

[16] Bharathi A, Natarajan A M. Microarray gene expression cancer diagnosis using machine learning algorithms[C]// International Conference on Signal & Image Processing. Quebec,2010:275 – 280.

(上接第 792 页)

[2] 屈乐乐,方广有,杨天虹. 频率步进探地雷达系统小型化的设计与实现[J]. 仪器仪表学报,2010,31(3):699 – 703.
(Qu Le-le, Fang Guang-you, Yang Tian-hong. Design and implementation of miniature stepped-frequency ground penetrating radar system[J]. *Chinese Journal of Scientific Instrument*,2010,31(3):699 – 703.)

[3] Nicolaescu I. Improvement of stepped-frequency continuous wave ground-penetrating radar cross-range resolution [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2013,51(1):85 – 92.

[4] Donoho D. Compressed sensing[J]. *IEEE Transactions on Information Theory*,2006,52(4):1289 – 1306.

[5] Yang J,Jin T,Huang X,et al. Sparse MIMO array forward-looking GPR imaging based on compressed sensing in clutter environment [J]. *IEEE Transactions on Geoscience and Remote Sensing*,2014,52(7):4480 – 4494.

[6] Krueger K R,McClellan J H,Scott W R. Efficient algorithm design for GPR imaging of landmines[J]. *IEEE Transactions on Geoscience and Remote Sensing*,2015,53(7):4010 – 4021.

[7] Ambrosanio M, Pascazio V. A compressive-sensing-based approach for the detection and characterization of buried objects[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*,2015,8(7):3386 – 3395.

[8] Qu L L,Yang T H. Investigation of air/ground reflection and antenna beam width for compressive sensing SFCW GPR migration imaging[J]. *IEEE Transactions on Geoscience and Remote Sensing*,2012,50(8):3143 – 3149.

[9] Wei X,Zhang Y. Interference removal for autofocusing of GPR data from RC bridge decks [J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*,2015,8(3):1145 – 1151.

[10] Tropp J A, Gilbert A C. Signal recovery from random measurements via orthogonal matching pursuit [J]. *IEEE Transactions on Information Theory*,2007,53(12):4655 – 4666.

[11] Solimene R,Cuccaro A,Aversano A et al. Ground clutter removal in GPR surveys[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*,2014,7(3):792 – 1151.