基于最小二乘支持向量机微阵列基因特征分类

高 振 斌
(西安财经大学统计学院数学与应用数学研究所 陕西 西安 710100)

摘 要 基因表达分析中的微阵列数据具有高维、高冗余的特点,给基因表达数据分类带来很大的困难。机器学习中的最小二乘支持向量机算法具有计算效率高的优势,从而为数据挖掘提供了一条有效途径。针对两类典型的癌症微阵列数据集(结肠癌集和白血病集),进行归一化预处理并且计算其相关系数矩阵;使用主成分分析法进行降维处理,得到用于特征选取和分类的信息基因集(各取 10 个基因);采用最小二乘支持向量机分类器对信息基因集进行分类。实验结果表明,该算法在两类癌症数据集上的留一交叉检验的准确率分别为 97.5% 和 100%,具有比其他分类器都高的测试准确率,为进一步医学临床应用提供可靠的诊断依据。

关键词 微阵列 特征分类 降维 最小二乘支持向量机

中图分类号 TP18 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.08.048

MICROARRAY GENE FEATURE CLASSIFICATION BASED ON LS-SVM

Gao Zhenbin
(*Institute of Mathematics and Applied Mathematics, School of Statistics, Xi'an University of Finance and Economics, Xi'an 710100, Shaanxi, China*)

Abstract Microarray data in gene expression analysis is characterized by high dimensionality and redundancy, which makes it difficult to classify gene expression data. The least-squares support vector machine (LS-SVM) algorithm in machine learning has the advantage of high computational efficiency, which provides an effective way for data mining. For two types of typical cancer microarray data sets (colon cancer set and leukemia set), we normalized the data and calculated the correlation coefficient matrix. The dimensionality reduction was carried out by principal component analysis, and the information gene sets (10 genes each) for feature selection and classification were obtained. Then, we used LS-SVM classifier to classify information gene sets. The experimental results show that the accuracy of this algorithm is 97.5% and 100% respectively, which is higher than other classifiers. It provides reliable diagnostic basis for further clinical application.

Keywords Microarray Feature classification Reducing dimension Least-square support vector machine (LS-SVM)

0 引 言

随着大规模基因表达谱技术的发展,人类各种组织的正常基因表达已经获得,各类病人的基因表达谱都有了参考的基准,因此基因表达数据的分析与建模已经成为生物信息学研究领域中的重要课题。

众多的研究者在此方向上进行了卓有成效的研究[1-4]。Chiaretti 等[5]对 T 细胞急性白血病的基因表达谱的分类进行了研究,并应用到临床治疗和预测之中;Sun 等[6]在肺癌临床治疗中通过对脱氧核糖核酸(DNA)微阵列数据特征分类从而作出预判;Devi 等[7]基于互信息选择信息基因,进而使用支持向量机(SVM)分类器对微阵列数据集进行分类评价;Wang 等[8]采用改进的偏最小二乘递归式特征消除(PLS-RFE)算法对多个微阵列数据集进行特征分类和选择,计算效率得到提高;Sharbaf 等[9]先采用 Fisher 指标进行滤波,降低数据集维数,然后将元胞学习自动机

收稿日期:2018-11-02。**高振斌**,副教授,主研领域:复杂系统的控制与优化。

(CLA)与蚁群算法(ACO)相结合,提高了基因特征分类精度;Khan 等[10]提出了一种新的自适应径向基核函数,并在非线性系统辨识、微阵列数据分类以及函数近似计算中做了仿真研究;Xiao 等[11]提出一种基于多模型集成的深度学习算法,并对三种癌症数据集进行验证;李颖新等[12]研究了急性白血病的分类信息基因选取,并以 SVM 作为分类器进行亚型识别;马煜等[13]将密度聚类与共享近邻法相结合,对微阵列数据进行聚类分析;韩利等[14]将粗糙集与 SVM 结合,通过粗糙集进行基因特征约简,然后用 SVM 进行数据分类;朱钦平等[15]提出了一种微阵列基因差异表达的多重假设检验方法,有效地减弱了数据噪声带来的假阳性结果;姚全珠等[16]研究了最小二乘支持向量机(LS-SVM)特征选择时参数优化算法;孙刚等[17]采用改进的 LASSO 算法对信息基因进行特征选择,剔除冗余基因;杨勤等[18]提出一种核最小二乘特征基因选择方法,对微阵列数据进行降维,然后用极限学习机进行训练和预测。

本文采用两种典型的肿瘤微阵列数据集(结肠癌数据集、白血病数据集),对数据进行归一化处理,计算其相关系数矩阵;使用主成分分析(PCA)法进行降维;使用 LS-SVM 对降维后的特征信息基因进行分类,并与其他几种分类方法进行了比较。

1　问题描述

假设微阵列特征分类问题可表示为集合 $O=(X,Y,F)$,其中,$X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 为样本集,共有 N 个样本;$Y=\{y_1,y_2,\cdots,y_N\}$ 为信息标签集;$F=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_N\}$ 为特征集;并且,$\boldsymbol{x}_k\in X$ 是一个包含 m 个元素(基因表达水平)的向量,可表示为 $\boldsymbol{x}_k=[x_{1k},x_{2k},\cdots,x_{mk}]^{\mathrm{T}}\in\boldsymbol{R}^m$;$y_k\in Y$ 是与 $\boldsymbol{x}_k$ 相对应的标量;假设为两分类(ω_1 和 ω_2)问题,则有:

$$y_k=\begin{cases}1 & \boldsymbol{x}_k\in\omega_1\\ -1 & \boldsymbol{x}_k\in\omega_2\end{cases}\quad k=1,2,\cdots,N \tag{1}$$

目的是寻找一组特征信息基因向量 $\boldsymbol{f}_k=[f_{1k},f_{2k},\cdots,f_{pk}]^{\mathrm{T}}\in F(p\leqslant m)$,使之能够精确区分样本的基因表达数据。假定所选的特征子集的数目 p 尽可能小。

2　SVM 和 LS-SVM

SVM 是一种基于统计学习理论,采用结构风险最小化原理的机器学习算法,可以有效地处理高维样本的分类问题,计算复杂度受样本维数的影响较小,适合处理小样本、高维数的基因表达数据的样本分类问题。

SVM 模型的目的是构造一个如下形式的最优分类函数:

$$f(x)=\mathrm{sgn}[\boldsymbol{w}^{\mathrm{T}}\varphi(x)+b] \tag{2}$$

式中:$\varphi(x)$ 为将输入数据映射到高维特征空间的非线性映射;$\boldsymbol{w}$ 为超平面权值系数向量;b 为偏置项。标准支持向量机分类问题可描述为如下优化问题:

$$\begin{aligned}&\min_{w,b,e}J(\boldsymbol{w},\boldsymbol{e})=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{1}{2}\gamma\sum_{k=1}^{N}e_i\\&\text{s.t.}\quad y_k[\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_k)+b]\geqslant 1-e_k\\&\qquad e_k\geqslant 0,k=1,2,\cdots,N\end{aligned} \tag{3}$$

式中:e_k 为误差;$\boldsymbol{w}=[w_1,w_2,\cdots,w_N]^{\mathrm{T}}$ 为权值系数向量;$\gamma>0$ 为惩罚系数,它控制对超出误差样本的惩罚程度。

LS-SVM 算法是在 SVM 的基础上通过最小二乘法利用误差平方和选择超平面,然后引进平方损失函数,将不等式约束转换为线性等式条件,将二次规划问题转化为线性求解问题。LS-SVM 分类问题可描述为求解下面的等式约束优化问题:

$$\begin{aligned}&\min_{w,b,e}J(\boldsymbol{w},\boldsymbol{e})=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{1}{2}\gamma\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}\\&\text{s.t.}\quad y_k[\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_k)+b]=1-e_k\\&\qquad k=1,2,\cdots,N\end{aligned} \tag{4}$$

式中:$\boldsymbol{e}=[e_1,e_2,\cdots,e_N]^{\mathrm{T}}$。

构造 Lagrange 函数如下:

$$\begin{aligned}L(\boldsymbol{w},b,\boldsymbol{e};\boldsymbol{\alpha})=J(\boldsymbol{w},\boldsymbol{e})-\sum_{k=1}^{N}\alpha_k\{y_k[\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_k)+\\ b]-1+e_k\}\end{aligned} \tag{5}$$

式中:$\alpha_k\geqslant 0(k=1,2\cdots,N)$ 为 Lagrange 乘子。对上式进行优化,即求 L 对 $\boldsymbol{w}$、b、e_k、α_k 的偏导数为零,经过化简,可得到如下的线性方程组:

$$\begin{bmatrix}\boldsymbol{I}&0&0&-\boldsymbol{Z}^{\mathrm{T}}\\0&0&0&-\boldsymbol{y}^{\mathrm{T}}\\0&0&\gamma\boldsymbol{I}&-1_N\\\boldsymbol{Z}&\boldsymbol{y}&1_N^{\mathrm{T}}&0\end{bmatrix}\begin{bmatrix}\boldsymbol{w}\\b\\\boldsymbol{e}\\\boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\0\\0\\1_N\end{bmatrix} \tag{6}$$

式中:

$\boldsymbol{Z}=[\varphi(\boldsymbol{x}_1)y_1,\varphi(\boldsymbol{x}_2)y_2,\cdots,\varphi(\boldsymbol{x}_N)y_N]^{\mathrm{T}}$,

$\boldsymbol{y}=[y_1,y_2,\cdots,y_N]^{\mathrm{T}}$,$1_N=[1,1,\cdots,1]^{\mathrm{T}}$,

$\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}}$,$\boldsymbol{I}$ 为单位矩阵。消除变量 $\boldsymbol{w}$、$\boldsymbol{e}$,再利用 Mercer 条件:

$$\begin{aligned}\Omega_{sl}=y_sy_l\varphi^{\mathrm{T}}(\boldsymbol{x}_s)\varphi(\boldsymbol{x}_l)=y_sy_lK(\boldsymbol{x}_s,\boldsymbol{x}_l)\\ s,l=1,2,\cdots,N\end{aligned} \tag{7}$$

可得矩阵方程:

$$\begin{bmatrix} 0 & -\boldsymbol{y}^{\mathrm{T}} \\ \boldsymbol{y} & \boldsymbol{\Omega}+\gamma^{-1}\boldsymbol{I} \end{bmatrix}\begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix}=\begin{bmatrix} 0 \\ 1_N \end{bmatrix} \tag{8}$$

式中:$\boldsymbol{\Omega}=[\Omega_{sl}]_{N\times N}$。假设 $\boldsymbol{A}=\boldsymbol{\Omega}+\gamma^{-1}\boldsymbol{I}$,由于 $\boldsymbol{A}$ 为对称半正定矩阵,因而 $\boldsymbol{A}^{-1}$ 存在,上式有解。得到 LS-SVM 分类器为:

$$f(x)=\mathrm{sgn}[\alpha_k y_k K(x,\boldsymbol{x}_k)+b] \quad k=1,2,\cdots,N \tag{9}$$

式中:α_k、b 为式(8)的解。

取径向基核函数为:

$$K(x,\boldsymbol{x}_k)=\exp\left\{-\frac{\|x-\boldsymbol{x}_k\|^2}{\sigma^2}\right\} \quad k=1,2,\cdots,N \tag{10}$$

3 算法实现

3.1 数据预处理

通常,原始数据集在特征选择之前应该被标准化。对于微阵列数据上一个基因中每个样本的表达值,减去该基因所有样本的平均值,再除以该基因所有样本的标准差。经过标准化之后,一个基因在所有样本上的表达值满足均值为 0,标准差为 1 的正态分布。

针对微阵列样本集 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$,且 $\boldsymbol{x}_k=[x_{1k},x_{2k},\cdots,x_{mk}]^{\mathrm{T}}$,数据归一化计算如下:

$$z_{ik}=\frac{x_{ik}-\mu_i}{\sigma_i} \quad i=1,2,\cdots,m,k=1,2,\cdots,N \tag{11}$$

式中:μ_i和 σ_i分别是为第 i 个基因表达值的均值和标准差。

3.2 相关系数矩阵

对归一化数据求相关系数为:

$$r_{ij}=\frac{\sum_{k=1}^{N} z_{ik}z_{jk}}{\sqrt{\sum_{k=1}^{N} z_{ik}^2\sum_{k=1}^{N} z_{jk}^2}} \quad i,j=1,2\cdots,m \tag{12}$$

3.3 提取主成分分量

主成分分析法的基本思想是在保留尽可能多的原始信息的前提下达到降维的目的。

求解特征方程:$|\lambda\boldsymbol{I}-\boldsymbol{R}|=0$,其中,$\boldsymbol{R}=[r_{ij}]\in\boldsymbol{R}^{m\times m}$为相关系数矩阵;求出 m 个特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m\geqslant 0$。确定主成分数量 $p(<m)$的值,使信息的利用率达到 85% 以上。

实现微阵列数据的降维和特征分类步骤如下:

Step 1 数据预处理。对微阵列数据集进行归一化处理。

Step 2 提取主成分分量。计算相关系数矩阵;采用 PCA 法,对所选择的基因样本数据进行降维处理,得到样本的特征信息基因集。

Step 3 分类模型训练。对特征提取后的信息基因数据分别采用 LS-SVM 等分类器进行训练,得到分类模型。

Step 4 测试分类模型。将测试样本代入分类模型中,分别采用留一检测法和独立测试法评估各种分类器的性能。

4 仿真实验

4.1 实验数据及开发环境

实验采用两个公开的微阵列数据集来评估本文算法的性能。数据集的详细描述见表 1。结肠癌数据集包括 62 个样本,且分成两类:正常样本和结肠癌样本。其中,正常样本 22 个,结肠癌样本 40 个和 2 000 个基因[1];白血病数据集包括 128 个样本,分属于两类不同类型的肿瘤:T 细胞 ALL(共 33 例)样本和 B 细胞 ALL(共 95 例)样本和 12 625 个基因[5,19]。

表 1 实验数据及描述

数据集	基因数量	样本数量	类别数
结肠癌	2 000	62(22/40)	2
白血病	12 625	128(33/95)	2

本文的实验环境:Intel CPU 2.53 GHz 处理器,2 GB 内存的 PC 机,Windows XP 操作系统,MATLAB 2014b 开发环境。

4.2 实验结果分析

实验 1 结肠癌数据集分类

针对结肠癌数据集,实验首先经过数据预处理,然后,将正常样本和肿瘤样本按接近 2∶1 的比例随机地分配到训练集和测试集中。训练集有 40 个样本(其中正常样本 14,肿瘤样本 26),测试集有 22 个样本(正常样本 8 个,肿瘤样本 14 个)。

然后通过 PCA 降维方法,提取主成分前十的特征信息基因如表 2 所示。

表 2 结肠癌数据集中选取的特征基因

序号	基因名称	序号	基因名称
1	X53799	6	R80427
2	M29273	7	X75208
3	U21914	8	D29808
4	L00352	9	M59807
5	X90858	10	D13627

分别采用 LS-SVM 等分类器对选取的特征基因进行分类。各分类器分类准确率结果见表 3。图 1、图 2 为 LS-SVM 分类器训练模型准确率(100%)以及独立测试实验准确率(68.18%)结果图。

表 3　结肠癌数据集选取的特征基因集实验结果

分类器	基因数	留一交叉检验准确率	独立测试准确率
PNN	10	0.625	0.682
RBF	10	0.675	0.682
BP	10	0.650	0.636
SVM RBF_ker	10	0.675	0.727
SVM Lin_ker	10	0.600	0.636
SVM poly_ker	10	0.625	0.727
LS-SVM RBF_ker	10	0.975	0.682

注:PNN 表示概率神经网络;RBF 为径向基神经网络;BP 为反向传播神经网络;RBF_ker、Lin_ker、poly_ker 分别为径向基核函数、线性核函数和多项式核函数

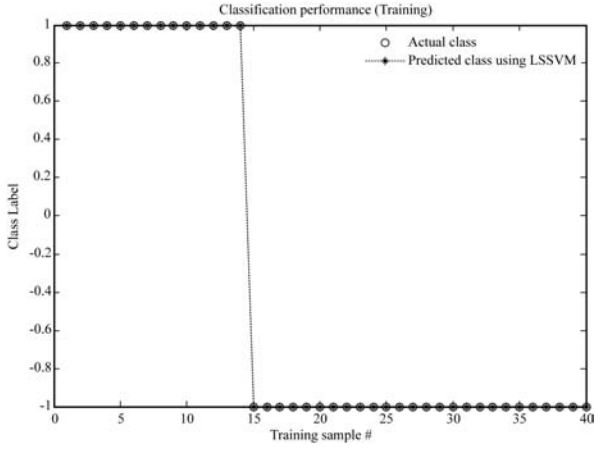

图 1　结肠癌数据 LS-SVM 训练模型准确率

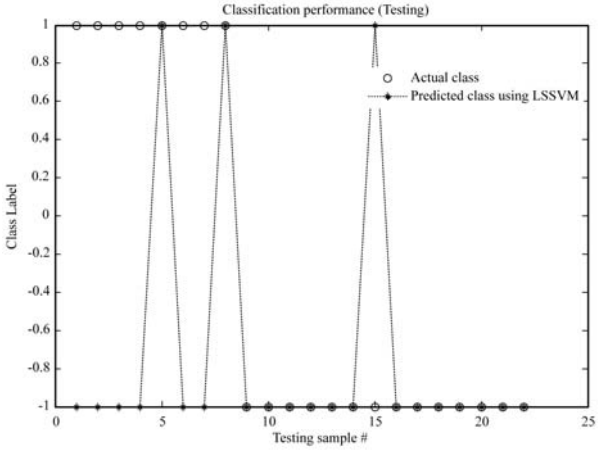

图 2　结肠癌数据 LS-SVM 独立测试准确率

实验 2　白血病数据集分类

白血病数据集经过数据预处理,由 PCA 法求得主成分前十的特征信息基因如表 4 所示。

表 4　白血病数据集中选取的特征基因

序号	基因名称	序号	基因名称
1	X110_at	6	X1221_at
2	X1098_at	7	X1210_s_at
3	X1189_at	8	X1086_at
4	X1106_at	9	X1173_g_at
5	X1178_at	10	X1199_at

将数据集中的两类样本分配到训练集和测试集中。训练集有 65 个样本(T 细胞样本有 48 个,B 细胞样本有 17 个),测试集有 63 个样本(T 细胞样本有 47 个,B 细胞样本有 16 个)。

对选取的特征基因进行分类,分别采用留一交叉检验和独立测试实验,结果见表 5。图 3、图 4 分别为 LSSVM 分类器留一交叉检验准确率(100%)和独立测试实验准确率(93.65%)结果图。

表 5　白血病数据集选取的特征基因集实验结果

分类器	基因数	留一交叉检验准确率	独立测试准确率
PNN	10	0.969	0.746
RBF	10	0.892	0.825
BP	10	0.738	0.746
SVM RBF_ker	10	0.738	0.825
SVM Lin_ker	10	0.738	0.746
LS-SVM RBF_ker	10	1.000	0.937

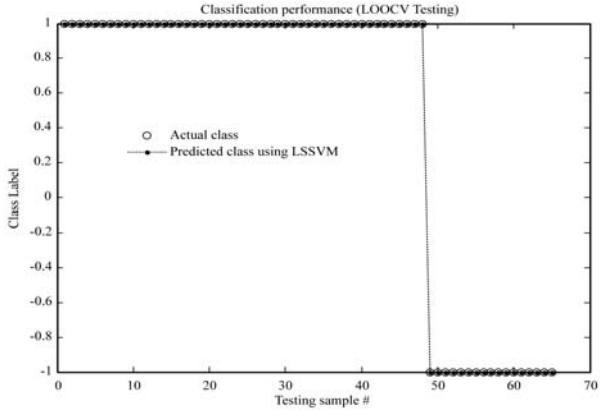

图 3　白血病数据 LS-SVM 留一交叉检验测试准确率

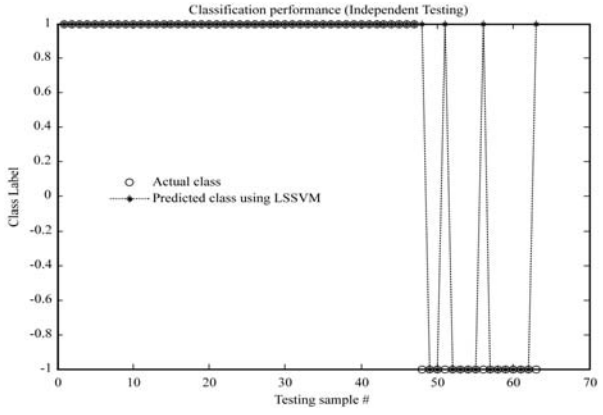

图 4　白血病数据集 LS-SVM 独立测试准确率

从表3、表5中看出,对于两个数据集的留一交叉检验结果,LS-SVM 分类器的准确率最高,分别为97.5%和100%,其次是 PNN 分类器和 RBF 分类器;独立测试实验结果中,白血病特征基因集的 LS-SVM 分类器的准确率最高,为93.65%,而结肠癌数据集 LS-SVM 分类结果与其他分类器的结果差别不大。

5 结 语

微阵列数据对疾病的诊断有很重要的参考价值,但是,微阵列数据的高维和高冗余给进一步挖掘其中蕴含的知识带来极大困难,其中一个关键任务就是信息基因的选择。LS-SVM 分类器将 SVM 优化问题的不等式约束转换为线性等式条件,将二次规划问题转化为线性求解问题,避免了求解耗时,提高了运行效率。本文基于 LS-SVM 分类器对两类癌症微阵列数据集的基因分别进行提取和分类。首先,对微阵列数据进行归一化预处理,计算其相关系数矩阵,并运用 PCA 法进行降维。提取特征信息基因集(各取 10 个基因),运用不同的分类器(包括 LS-SVM、PNN、RBF、BP 及 SVM)进行分类。从留一交叉检验和独立测试两种实验结果可以看出,运用 LS-SVM 分类器,结肠癌集准确率分别达到97.5%和68.2%;白血病集准确率分别达到100%和93.7%,从而证明了本文提出的算法比运用其他分类器计算准确率相对较高,能够为医学临床实践提供较为可靠的判断依据。

参 考 文 献

[1] Alon U, Barkai N, Notterman D A, et al. Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays [J]. Proceedings of the National Academy of Sciences of the United States of America, 1999, 96(12):6745 - 6750.

[2] Furey T S, Cristianini N, Duffy N, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data[J]. Bioinformatics, 2000, 16(10):906 - 914.

[3] Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring [J]. Science, 1999, 286: 531 - 537.

[4] Guyon I, Weston J, Bamhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46:389 - 422.

[5] Chiaretti S, Li X C, Gentleman R, et al. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival[J]. Blood, 2004, 103(7):2771 - 2778.

[6] Sun Z F, Yang P. Gene expression profiling on lung cancer outcome prediction: present clinical value and future premise [J]. Cancer Epidemiology, Biomarkers & Prevention, 2006, 15(11):2063 - 2068.

[7] Devi A V C, Devaraj D, Venkatesulu M. Gene Expression data classification using support vector machine and mutual information-based gene selection [J]. Procedia Computer Science, 2015, 47: 13 - 21.

[8] Wang A G, An N, Chen G L, et al. Improving PLS-RFE based gene selection for microarray[J]. Computers in Biology and Medicine, 2015, 62:14 - 24.

[9] Sharbaf F V, Mosafer S, Moattar M H. A hybrid gene selection approach for microarray data classification using cellular learning automata and colony optimization [J]. Genomics, 2016, 107: 231 - 238.

[10] Khan S, Naseem I, Togneri R, et al. A Novel Adaptive Kernel for the RBF Neural Networks[J]. Circuits, Systems, and Signal Processing, 2017, 36(4):1639 - 1653.

[11] Xiao Y W, Wu J, Lin Z L, et al. A deep learning-based multi-model ensemble method for cancer prediction [J]. Computer Methods and Programs in Biomedicine, 2018, 153: 1 - 9.

[12] 李颖新,刘全金,阮晓钢. 急性白血病的基因表达谱分析与亚型分类特征的鉴别[J]. 中国生物医学工程学报, 2005,24(2):240 - 244.

[13] 马煜,陈莉,欧立奇. 基因微阵列数据的聚类分析算法研究[J]. 计算机工程与应用,2006(5):176 - 178.

[14] 韩利,祁云嵩,王俊. 基于粗糙集的支持向量机微阵列数据分类方法[J]. 科学技术与工程,2009,9(1):152 - 155.

[15] 朱钦平, 胡晓涵, 祁云嵩. 基于非线性回归分析的差异基因选择方法[J]. 科学技术与工程, 2010, 10(27): 6675 - 6678.

[16] 姚全珠,蔡婕. 基于 PSO 的 LS-SVM 特征选择与参数优化算法[J]. 计算机工程与应用,2010,46(1):134 - 136, 229.

[17] 孙刚,张靖. 面向高维微阵列数据的混合特征选择算法[J]. 小型微型计算机系统,2015,36(6):1209 - 1213.

[18] 杨勤,董洪伟,薛燕娜. 核多元基因选择和极限学习机在微阵列分析中的应用[J]. 传感器与微系统,2016,35(5): 146 - 148,153.

[19] Torgo L. 数据挖掘与 R 语言[M]. 李洪成,陈道轮,吴立明,译. 北京:机械工业出版社,2016.