

基于支持向量机的中国地鼠分类特征基因选取

杨俊丽¹, 刘田福²

(1. 山西医科大学 计算机教学部, 太原 030001; 2. 山西医科大学 实验动物中心, 太原 030001)

(hplkyjl@sohu.com)

摘要: 针对中国地鼠基因表达谱数据维数高和样本小的特点, 提出一种基于支持向量机(SVM)的分类特征基因选取方法。该方法利用改进的Fisher判别(FDR)基因特征计分准则剔除分类无关基因, 提出由空间距离和功能距离组成的新距离作为相似性度量的标准进行冗余基因的剔除, 采用SVM作为分类器检验特征基因的分类性能。实验结果表明, 该方法有效地剔除了分类无关基因和冗余基因, 选取的特征基因满足对中国地鼠正确分类的最小基因数。

关键词: 特征选取; 支持向量机; 分类器; 基因表达谱; 中国地鼠

中图分类号: TP391.4 **文献标志码:** A

Feature gene selection for Chinese hamster classification based on support vector machine

YANG Jun-li¹, LIU Tian-fu²

(1. Department of Computer Teaching, Shanxi Medical University, Taiyuan Shanxi 030001, China;

2. Laboratory Animal Center, Shanxi Medical University, Taiyuan Shanxi 030001, China)

Abstract: Concerning the gene expression profile of Chinese hamster feature, such as high-dimension and small sample, a method of feature selection for Chinese hamster classification based on Support Vector Machine (SVM) was proposed in this paper. The method used improved FDR gene feature score criterion to remove the genes irrelevant to the classification. A new distance composed by space distance and function distance was proposed as the criterion of comparability to remove redundant genes. A SVM was used as classifier to validate the classification performance of the feature genes selected. The experimental results show that this method effectively removes the irrelevant and redundant genes, and selected the feature genes that meet the needs of least feature genes which classify accurately on Chinese hamster.

Key words: feature selection; Support Vector Machine (SVM); classifier; gene expression profile; Chinese hamster

0 引言

中国地鼠因其染色体大、条数少、易于识别等特点^[1], 广泛应用于细胞遗传学、辐射遗传学、实验肿瘤和分子生物学等众多领域, 在医学和生物学实验研究中占有重要的地位。但由于中国地鼠的生物性状、基因组等基础资料报道甚少, 国内对于中国地鼠的分类学研究尚处在形态学分类阶段^[1]。随着基因表达谱技术的出现与不断发展, 利用基因序列中的基因表达谱数据建立分类模型, 已成为生物分类学研究的一种重要的分类方法。而分类特征基因的提取和选择方法又是建立分类模型的一个重要环节, 直接影响着分类器的设计和性能。因此, 如何选取生物序列中的特征基因, 成为特征基因提取与生物分类器研究的核心内容。目前, 常用的特征基因选取方法主要有因子分量分析、启发式搜索、支持向量机(Support Vector Machine, SVM)、线性判别分析等, 在实际应用中, 也常将多种方法结合起来使用^[2-6]。

中国地鼠的基因表达谱数据集具有高维数和小样本的特点, 而高维数及其所包含的高噪声和信息冗余等因素会降低分类器的分类性能。本文针对中国地鼠基因表达谱数据的特点, 设计了基于支持向量机的中国地鼠分类特征基因的选取方法。实验表明, 该方法有效地剔除了分类无关基因和冗余

基因, 选取的特征基因对中国地鼠的分类结果与传统的形态分类结果一致, 同时保证了对中国地鼠正确分类的最小基因数。

1 特征基因的预选

1.1 极端基因的过滤

极端基因是指偏离群体分布, 具有过大的变异性表达异常的基因^[2]。极端基因可以通过设置判别阈值进行识别, 判别阈值根据整个基因表达数据的分布百分位点或一定的标准差范围来确定。

1.2 冗余基因的预过滤

对于基因表达数据中的负值和极小值, 由于没有生物意义, 因此需要剔除。在计算基因表达数据的信号强度比率值时, 如果参考样本信号强度很小, 就可能造成单个异常大的峰数据, 当参考样本信号强度很大时, 又可能出现单个异常小的谷数据, 通常这些数据由噪声引起, 也需要剔除。最后就是对缺失数据的处理, 可将缺失数据项的行向量或列向量直接去掉。

2 基于改进的FDR特征基因选择

中国地鼠基因表达谱数据的每个样本中都记录了所有可测基因的表达水平, 然而只有特征基因才包含样本的类别信

收稿日期: 2010-07-26。 基金项目: “十一五”国家科技支撑计划项目(2007BAK26B05); 山西省自然科学基金资助项目(20051087)。

作者简介: 杨俊丽(1978-), 女, 山西太原人, 讲师, 硕士研究生, 主要研究方向: 机器学习、医学数据整合、生物信息学; 刘田福(1954-), 男, 山西太原人, 教授, 主要研究方向: 实验动物学。

息,大部分与样本类别无关的基因称为“无关基因”或“噪声基因”^[3]。在衡量基因分类能力的问题上,Mika等人^[4]提出了Fisher判别(Fisher Discriminant Ratio, FDR)基因特征计分准则,即

$$FDR(g_i) = \frac{(\mu_i^+ - \mu_i^-)^2}{(\sigma_i^+)^2 + (\sigma_i^-)^2} \quad (1)$$

其中: μ_i^+ 表示基因 g_i 在正类样本中的表达水平均值, μ_i^- 表示基因 g_i 在负类样本中的表达水平均值, σ_i^+ 表示基因 g_i 在正类样本中的标准差, σ_i^- 表示基因 g_i 在负类样本中的标准差。由式(1)可知,如果基因 g_i 在正类和负类中表达水平均值相同或相近,则被作为噪声基因剔除;如果该基因在两个类中的表达水平标准差差异较大时,说明它在标准差很小的类别中具有近似一致性的基因表达,则该基因很可能是此类别的特征基因^[5]。因此,在衡量基因分类能力的问题上,还应该考虑基因表达水平分布方差不同对样本分类的贡献。为此本文将式(1)进行了修订,修订后的基因特征计分准则可表示为:

$$FDR(g_i) = \frac{1}{4} \frac{(\mu_i^+ - \mu_i^-)^2}{(\sigma_i^+)^2 + (\sigma_i^-)^2} + \frac{1}{2} \ln \frac{(\sigma_i^+)^2 + (\sigma_i^-)^2}{2\sigma_i^+ \sigma_i^-} \quad (2)$$

由式(2)知,修订后的基因特征计分准则由两部分组成:第一项体现了基因在两个类中分布均值的差异对样本分类的贡献;第二项体现了分布方差对样本分类的贡献。按照此计分准则对训练集中的每个基因进行计分,分值越大说明基因分类能力越强;然后按计算出的分值大小顺序对基因进行排序,并根据分类器的准确率选择前面一定数量的基因作为结果。

3 冗余基因的剔除

基因之间存在着调控和相互作用的关系,这在基因表达谱中反映为不同基因在表达水平上存在着一定程度的相关性^[6],即相似性。通过衡量基因之间的相似性,将相似基因中信息较少的基因去除,可有效地减少特征基因的数量。在实际应用中,常采用特征向量之间的距离作为相似性度量的标准。

本文在计算特征向量之间的距离时,将基因间的距离分为空间距离和功能距离两个部分,即

$$\delta_{ij} = \delta_{ij}^s + \delta_{ij}^f \quad (3)$$

其中: δ_{ij}^s 为空间距离, δ_{ij}^f 为功能距离。本文采用欧氏距离^[7]计算特征向量间的空间距离,欧氏距离表示为

$$\delta_{ij}^s = \sqrt{\sum_{k=1}^d (x_{ik} - x_{jk})^2} \quad (4)$$

功能距离包括减少具有相同功能基因间的距离和增加不同功能间的距离,即:

$$\delta_{ij}^f = \begin{cases} -F_i F_j^T, & i = j \\ 1 - F_i F_j^T, & \text{其他} \end{cases} \quad (5)$$

如果基因 i 具有功能 F 或者功能未知,则 F_i 取值为1;否则取值为0。如果特征向量间的距离小于给定的阈值,就认为它们是共表达的,阈值根据分类器的准确率来确定。

4 基于SVM的特征基因分类性能检验

本文采用支持向量机(SVM)作为分类器检验特征基因的分类能力。SVM是建立在统计学习理论基础上的的一种机器学习算法^[8],具有很强的泛化能力。SVM的优点是能够处理高维数据,分类精度高,且抗噪能力强^[9]。因此,SVM在基因功能预测和基因分类方面非常有效。设训练样本个数为 n ,训练样本形式为 $\{(x_1, s_1), (x_2, s_2), \dots, (x_n, s_n)\}$,对于两类问题 $s_i \in \{1, -1\}$, $x_i \in \{0, 1\}$ 。对于多类问题可转化为两类问题处理。SVM的判别函数^[10]表示为:

$$f(x) = \text{sgn} \left[\sum_{i=1}^N a_i s_i K(x, x_i) + b \right] \quad (6)$$

其中: N 为支持向量的个数, $K(x, x_i)$ 为核函数。本文采用的核函数为径向基核函数(Radial Basis Function, RBF)^[11]。

$$K(x, x_i) = \exp \left(- \frac{\|x - x_i\|^2}{\sigma^2} \right) \quad (7)$$

式(6)可表示为:

$$f(x) = \text{sgn} \left[\sum_{i=1}^N a_i s_i \exp \left(- \frac{\|x - x_i\|^2}{\sigma^2} \right) + b \right] \quad (8)$$

由于训练集中没有互相矛盾的样本点,因此该判别函数对训练集的准确率可达到100%,据此来确定错误惩罚常数 C 和核参数 σ^2 。

5 仿真实验

5.1 实验数据描述

实验数据为山西医科大学实验动物中心饲养的中国地鼠近交系,它是我国目前唯一庞大的中国地鼠群体,已被英国收入“实验动物国际索引”^[12]。“山医群体近交系中国地鼠”^[1]分为A、E两家系,从A家系中随机抽取28个样本,其中包括18个训练样本和10个测试样本;从E家系中随机抽取22个样本,其中包括15个训练样本和7个测试样本。整个数据集的结构如表1所示。

表1 训练样本和测试样本数

类别	训练样本数	测试样本数
A家系	18	10
E家系	15	7

5.2 实验结果与分析

对训练集样本进行特征基因预选后得到892个基因,利用改进的FDR计分准则计算其分类信息分值,具体分布情况如图1所示。

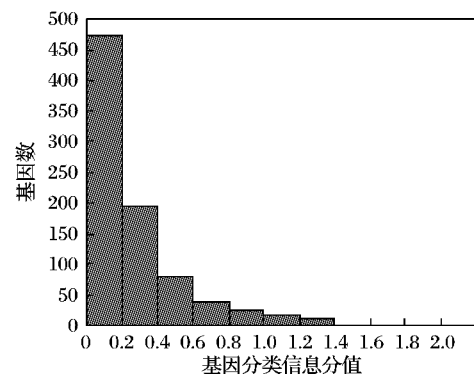


图1 基因分类信息分值分布情况

由图 1 可知,分值越大,基因数量越小。按基因的分值分别选取分值高的前 25 个、50 个和 100 个基因作为特征基因,然后将这些基因表达谱数据送入 SVM 分类器,进行分类能力的检验,实验结果如表 2 所示。

表 2 SVM 分类器分类结果

特征基因数	C	σ^2	准确率/%
100	1000	0.02	100.0
50	500	0.02	100.0
25	500	0.02	98.6

由表 2 可知,随着选取的特征基因数的减少,样本的准确率也随之下降。为了选择具有最小基因数并保持最高分类准确率的特征基因集,取 $C = 500$, $\sigma^2 = 0.02$, 选分值高的前 50 个基因作为特征基因,此时的分类准确率已达到 100%, 则该 50 个基因已经具备了完整的分类信息。接下来对这 50 个特征基因中可能存在的冗余基因进行剔除。本文采用空间距离和功能距离组成的新距离作为相似性度量的标准,当距离小于给定的阈值时,就认为它们是共表达的。图 2 给出了不同阈值对选择基因分类性能的影响。

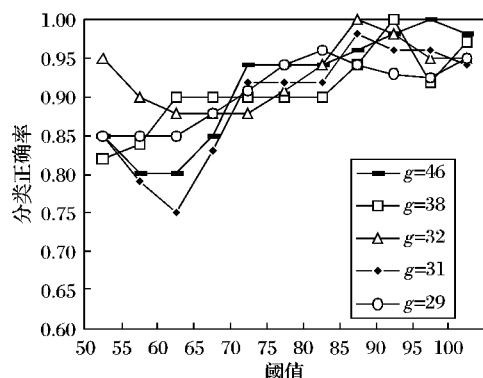


图 2 阈值对基因选择的影响

由图 2 可知,对于不同的基因数目 g 在阈值 ($\delta \in (0.8, 0.9)$) 的范围内基本都达到最高的准确率,这说明阈值过高或过低都不能得到最好的基因选择结果。当阈值 $\delta = 0.85$ 时,得到对原始样本分类准确率保持 100% 的最小基因数,因此,去冗余分析后最终得到 32 个分类特征基因。

6 结语

针对中国地鼠基因表达谱数据的特点,设计了一种分类特征基因选取的方法。该方法首先进行了特征基因的预选;然后利用改进的 FDR 基因特征计分准则对特征基因进行初选;最后采用由空间距离和功能距离组成的新距离作为相似性度量的标准进行冗余基因剔除。本文在特征基因选取的各阶段,均采用支持向量机作为分类器来检验选取的特征基因的分类能力,并以能正确分类作为标准选取最小特征基因数。实验表明,该方法选取的特征基因对中国地鼠的分类正确率达到 100%,并满足了对中国地鼠正确分类的最小基因数。

参考文献:

[1] 宋国华, 岳文斌, 刘田福. 中国地鼠线粒体 Cyt b 基因测序及其分子进化[J]. 中国实验动物学报, 2008, 16(2): 142-147.

- [2] VALENTINI G, DIETTERICH T G. Bias-variance analysis of support vector machines for the development of SVM based ensemble methods[J]. Journal of Machine Learning Research, 2004, 5(12): 725-775.
- [3] 李颖新, 阮晓钢. 基于支持向量机的肿瘤亚型分类特征基因选取[J]. 计算机研究与发展, 2005, 42(10): 1796-1801.
- [4] MIKA S, RATSCH G, WESTON J, et al. Fisher discriminant analysis with kernels[C]// IEEE Signal Processing Society Workshop of Neural Networks for Signal Processing IX. Washington, DC: IEEE Computer Society, 1999: 41-48.
- [5] 李泽, 包雷, 黄英武, 等. 基于基因表达谱的肿瘤分型和特征基因的选取[J]. 生物物理学报, 2002, 18(4): 413-417.
- [6] 黄德双. 基因表达谱数据挖掘方法研究[M]. 北京: 科学出版社, 2009.
- [7] RUIZ A, LOPEZ-de-TERUEL P E. Nonlinear kernel-based statistical pattern analysis[J]. IEEE Transactions on Neural Networks, 2001, 12(1): 16-32.
- [8] VAPNIK V N. Statistical learning theory[M]. New York: Wiley Interscience, 1998.
- [9] SOLLICH P. Bayesian methods for support vector machines: Evidence and predictive class probabilities[J]. Machine Learning, 2002, 46(1-3): 21-52.
- [10] KEERTHI S S, SHEVADE S K, BHATTACHARYYA C. A fast iterative nearest point algorithm for support vector machine classifier design[J]. IEEE Transactions on Neural Networks, 2000, 11(1): 124-136.
- [11] WILLIAMSON R C, SMOLA A J, SCHOLKOPF B. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators[J]. IEEE Transactions on Information Theory, 2001, 47(6): 2516-2532.
- [12] 宋国华, 刘田福, 赵嘉慧, 等. 近交系中国地鼠山医群体遗传结构的随机扩增多态分析[J]. 山西医科大学学报, 2005, 36(3): 270-274.

征订启事

《计算机应用》以促进计算机开发应用、创新为目标,以介绍最新应用技术为重点,注重学术水平高、指导性强、技术内容丰富的文章。现审稿周期为 2 个月,发表周期为 6 个月。欢迎投稿,欢迎订阅。全国各地邮局均可订阅,也可直接从编辑部订阅。

邮发代号: 62-110

定价: 28 元/册,全年 336 元/12 期。

通信地址: 成都市武侯区 237 信箱

《计算机应用》编辑部

邮政编码: 610041 联系人: 雍平

电话: 028-85224283-803

传真: 028-85222239-816

作者订刊请登录我刊网站查看作者优惠订刊活动:

http://www.computerapplications.com.cn

《计算机应用》编辑部

2011 年 1 月