第 57 卷 第 5 期 厦门大学学报(自然科学版) Vol. 57 No. 5
2018 年 9 月 Journal of Xiamen University (Natural Science) Sep. 2018
doi:10.6043/j.issn.0438-0479.201803022

基于支持向量机递归特征消除和特征聚类的致癌基因选择方法

叶小泉,吴云峰*
(厦门大学 信息科学与技术学院,福建省智慧城市感知与计算重点实验室,福建 厦门 361005)

摘要:癌症通常由基因发生突变引起,因此从大量基因中有效地识别出少量致癌基因具有重要意义.针对基因表达谱数据高维小样本的特点,将支持向量机递归特征消除(SVM-RFE)和特征聚类算法相结合,提出一种新的基因选择方法:K 类别 SVM-RFE(K-SVM-RFE).该算法通过特征排序算法去除大量无关基因,利用 K 均值聚类算法将相似基因聚为一类,并通过两次 SVM-RFE 算法精选致癌基因.随后将 K-SVM-RFE 算法应用于多个基因表达谱数据集,并对其中的关键参数设置进行了讨论.实验结果表明 K-SVM-RFE 算法所选基因较已有方法在分类准确率上有显著提高,特别是在选择少量致癌基因上效果提升更为明显.

关键词:基因表达谱;特征选择;K 均值聚类;支持向量机
中图分类号:TP 391.4 **文献标志码**:A **文章编号**:0438-0479(2018)05-0702-06

癌症通常缘于正常组织在物理或化学致癌物的作用下基因组发生突变,即基因表达水平的改变,使得许多生物过程失调[1].而基因表达信息可以通过基因芯片技术测得,基因芯片(通常也称为 DNA 微阵列或生物芯片)是附着于固体表面的微观 DNA 斑点的集合.在分子生物学领域,根据核苷酸分子在形成双链时遵循碱基互补原则,研究人员能够使用基因芯片测量大量基因的表达水平信息,从而得到基因表达谱.因此,若利用这些基因表达谱数据确定出与癌症有密切关系的基因,将对癌症的诊断和治疗发挥重要意义[2].

由于存在与测定相关的成本问题,基因表达谱数据具有高维小样本的特性.较高的维数是获得问题准确描述的有力保障,但它又难以避免地会引入大量冗余和与类别无关的噪声信息,这给传统的机器学习方法带来了挑战.因此,从成千上万个基因中判断出在不同疾病类别上具有差异性表达的少量致癌基因前,需要剔除掉大量无关基因,而特征选择是一种有效的手段.

在利用基因表达谱数据进行致癌基因选择的问题上,Golub 等[3]对急性白血病亚型识别和致病基因的判别进行了研究,用信噪比(SNR)指标来作为基因对样本类别的区分能力,其研究结果表明白血病亚型之间在基因表达上的差异可以通过一系列基因的表达水平检测来进行临床诊断,并可以由此指导后续治疗方案的制定.该方法运行速度较快,适用于高维数据,但由于其不能识别冗余基因,结果常常不尽人意.另外,Guyon 等[4]将支持向量机(SVM)与递归特征消除(RFE)相结合提出了 SVM-RFE 算法,该方法通过 SVM 每个维度权重的绝对值来度量对应特征的重要性,每次迭代删除权重排名靠后的一个特征,取得了良好的效果.但是它每次迭代只删除一个特征,在高维数据中仍耗时较长.因此 Ding 等[5]对它进行了改进,使得每次可以按比例删除特征,提高了计算速度,但同时也发现所选的特征对每次迭代删除的特征表现得十分敏感.此外 Yousef 等[6]提出了一种基于 SVM 的递归聚类特征消除(SVM-RCE)算法,该方法使用聚类方法对特征集进行聚类,随后利用 SVM

收稿日期:2018-03-08 录用日期:2017-07-16
基金项目:国家自然科学基金(61771331)
* 通信作者:yunfengwu@xmu.edu.cn
引文格式:叶小泉,吴云峰.基于支持向量机递归特征消除和特征聚类的致癌基因选择方法[J].厦门大学学报(自然科学版),2018,57(5):702-707.
Citation:YE X Q,WU Y F. Cancer gene selection algorithm based on support vector machine recursive feature elimination and feature clustering[J]. Xiamen Univ Nat Sci,2018,57(5):702-707.(in Chinese)

对各个特征类进行评分,最后迭代删除得分最低的那些特征类. 此类递归聚类特征选择算法能够有效去除大量无关特征,但最后剩下的部分特征之间存在相似性较高、容易导致特征冗余的问题. 因此,在特征排序和 SVM-RFE 算法的基础上,本研究将二者结合并引入聚类算法,提出一种新的、适用于基因表达谱数据的特征选择方法:K 类别 SVM-RFE(K-SVM-RFE).

1　相关工作介绍

在具有高维小样本特性的基因表达谱数据中,一个快速且有效获得致癌基因的方法是对特征排序. 因此,在 K-SVM-RFE 算法中,利用基于 SNR 的特征排序方法剔除大量无关基因,将剩余基因利用 K 均值算法聚成多个类别,并利用 SVM-RFE 算法精选致癌基因.

1.1　基于 SNR 的特征排序

SNR 通常用来表示电子信号中信号与噪声的比例,而在特征选择中,可以用 SNR 指标来度量特征的重要性,进而对特征排序. Golub 等[3]的研究表明基于 SNR 的特征排序方法是一个快速且有效的致癌基因判别方法. 基因 g_i 的 SNR 数值 R_{SN} 通过下式计算得到:

$$R_{\mathrm{SN}}(g_i)=\frac{|u_+(g_i)-u_-(g_i)|}{\sigma_+(g_i)+\sigma_-(g_i)}, \tag{1}$$

其中:$u_+(g_i)$ 和 $u_-(g_i)$ 分别表示第 i 个基因 g_i 在阴性类别和阳性类别的平均表达值;$\sigma_+(g_i)$ 和 $\sigma_-(g_i)$ 分别表示基因 g_i 在两个类别中表达水平的标准差.

用式(1)来衡量每个基因的重要性,值越大说明该基因越重要. 若某一基因在不同类别中的分布均值相等,那么它的 R_{SN} 等于零,则该基因便被认为是无关基因而剔除.

1.2　K 均值聚类算法

K 均值聚类算法[7]是最经典的聚类方法之一,它基于观测对象间的相似度将对象划分不同类别,使得类内具有较高的相似度,而类间的相似度较低. 对于给定的一组样本数据 $(x_1,x_2,\cdots,x_n)$,现要将其划分为 K 个子集合(类别),$S=\{S_1,S_2,\cdots,S_K\}$,K 均值的划分思想是:先从 n 个样本中随机选出 K 个样本作为初始聚类中心,随后将剩余样本分别划入与其距离最近的聚类中心的相应类别中,使得类内总距离达到最小,其目标函数可以表示为:

$$\underset{S}{\operatorname{argmin}}\sum_{i=1}^{K}\sum_{x\in S_i}|x-u_i|^2, \tag{2}$$

其中 u_i 表示集合 S_i 的聚类中心点. 所有样本的类别划分完毕后需要更新各个类别的中心点,第 $t+1$ 次的聚类中心通过下式计算:

$$u_i^{(t+1)}=\frac{1}{|S_i^{(t)}|}\sum_{x_j\in S_i^{(t)}}x_j, \tag{3}$$

随后对各个样本重新划分类别,重复以上过程直到中心值的变化可以忽略不计或者达到最大的迭代次数.

1.3　SVM-RFE 特征选择算法

SVM 是一种基于统计理论的分类方法,它利用核函数将普通低维空间中难以用一条直线分开的数据映射到一个较高维度的空间中,使其达到线性可分的目的. 在 SVM 超平面上的每个维度对应着输入数据集中的每个特征,因此可以把超平面上各个维度权重的绝对值看作该维度(或特征)的贡献(或重要性). 所以,权重的绝对值便可以用来对特征排序,从中选出关键特征. SVM-RFE 便是基于此思想的嵌入式特征选择方法,最初由 Guyon 等[4]提出,它是将 SVM 与 RFE 的后项搜索方法相结合的产物. SVM-RFE 的特征选择过程如下所示.

输入:训练数据集 E(n 个样本,m 个特征),类标签(n,1).

1) 初始化当前特征集合 E_{now} 为原始数据集,最优特征集合 E_{best} 为空,最优特征子集分类正确率 S_{best} 为 0.

2) 设置每次删除的特征数量比例 $p(0<p<1)$.

3) 重复以下步骤,直至当前特征集合 E_{now} 为空:

由 E_{now} 建立 SVM 模型,得到正确率评估值 S_{now};

按特征权重的绝对值 $|w|$ 降序排列 E_{now} 中的特征;

删除当前子集 E_{now} 中排名靠后的 $p\%$个特征;

若当前特征子集 E_{now} 的正确率 S_{now} 大于 S_{best}:$E_{\mathrm{best}}=E_{\mathrm{now}}$.

输出:最优特征子集 E_{best}.

SVM-RFE 算法用 SVM 超平面的每个维度的权重绝对值来代表相应特征的重要性,随后通过权重对特征按从大到小排列. 从降序排列的特征集合开始,每次删除排名最后的那个特征;随后继续使用 SVM 在剩余特征集合上训练分类器,再删除特征;如此多次重复进行直到该特征集合为空,或者达到了用户设定的特征数量为止. 由于其优异的性能表现,SVM-RFE 算法广泛用于图像处理,文本分析,生物信息处理等领域.

2 *K*-SVM-RFE 基因选择方法

特征排序算法(如基于 SNR 的特征排序算法)能够快速且有效地得到在不同类别中具有差异性表达的特征,特别是对于具有高维小样本特性的数据,特征排序算法可以迅速去除无关特征. 但是,在排名靠前的特征中,往往部分特征之间具有较高的相似性,造成了特征的冗余,这将会对少数关键特征的确定造成困扰,进而影响最终的分类性能.

因此,特征排序方法能够高效地去除无关特征,但是不能识别和去除冗余特征,它适用于关键基因的初步筛选. 基于此,本研究提出一种三阶段的基因选择方法 *K*-SVM-RFE. 首先,利用 SNR 指标计算各个基因的权重,并按权重降序排列基因,初步过滤掉大量权重值较低的基因;其次,为了去除冗余基因,将初步筛选后基因通过聚类算法聚成 k_1 个类别,并对各个类别利用 SVM-RFE 方法选出 k_2 个具有代表性的基因,组成新的基因集合 F;最后,再次利用 SVM-RFE 算法从 F 中选择出 k 个关键基因. 算法描述如下所示,流程如图 1 所示.

输入:原始数据集(n 个样本,m 个特征),类标签(n,1),选择基因数量 k.

1) 将原始数据预处理,处理结果记为 D.

2) 特征排序算法从 D 中筛选出 d 个基因,记为 f_1,其维度为(n,d).

3) 使用 K 均值聚类算法,将 f_1 按基因聚成个 k_1 个类别,且 $\sum_{i=1}^{k_1}|c_i|=d$,其中 c_i 表示第 i 个类别中的基因集合,而 $|c_i|$ 表示集合中基因的数量.

4) i 从 1 循环至 k_1,令 $f_2=f_2+\text{SVM-RFE}(c_i,k_2)$,其中 $\text{SVM-RFE}(c_i,k_2)$ 表示使用 SVM-RFE 算法从 c_i 中选择出 k_2 个关键基因.

5) 使用 SVM-RFE 算法从 f_2 中选择出 k 个关键基因.

输出:k 个关键基因.

值得注意的是,*K*-SVM-RFE 方法中共涉及到 3 个关键参数,分别为 k,k_1 和 k_2. 其中,k 为最后 SVM-RFE 算法选择的基因个数,也即最终输出的基因数量;k_1 为聚类算法所聚的类数;k_2 为各个类别中使用 SVM-RFE 方法选择的基因数. k,k_1 和 k_2 均可通过用户设定,但为了保证最后一次的 SVM-RFE 方法能够选出足够的 k 个基因,应至少满足如下关系:

$$k_1\times k_2\geqslant k. \tag{4}$$

在本文中 3.2 节我们将进一步讨论这 3 个参数的设置关系,以使 *K*-SVM-RFE 算法所选择的特征达到最佳的分类效果.

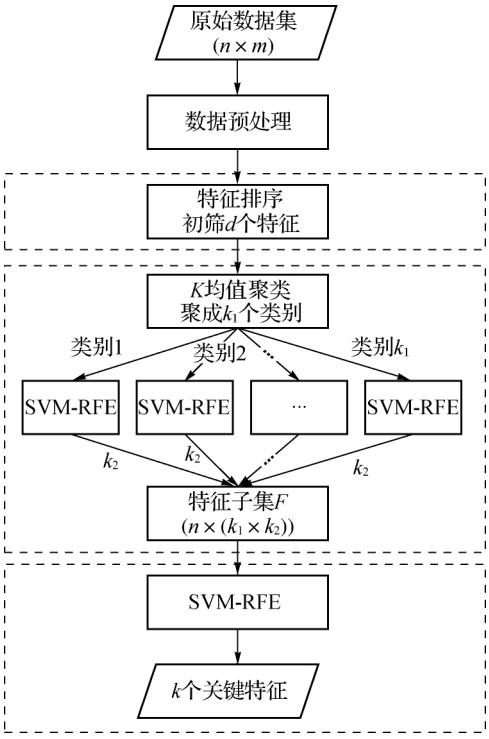


图 1 *K*-SVM-RFE 算法流程图

Fig. 1 Flowchart of the *K*-SVM-RFE algorithm

3 实验和结果分析

3.1 实验数据

实验主要以分类准确率来比较本研究所提出的 *K*-SVM-RFE 算法与基于 SNR 的特征排序算法以及 SVM-RFE 算法在分类上的性能差异. 为了验证 *K*-SVM-RFE 算法的有效性,本研究以 3 个公共的基因表达谱数据集作为实验对象,包括结肠癌基因表达谱数据集[8]、淋巴癌基因表达谱数据集[9]以及肺癌基因表达谱数据集[10]. 这些数据集均可以从生物识别研究计划的网站[11]下载得到,其数据构成如表 1 所示:

表 1 实验数据集

数据集序号	数据集名称	基因数量	样本数量(阳性/阴性)
1	结肠癌基因表达数据集	2 000	62(40/22)
2	淋巴癌基因表达数据集	7 129	77(58/19)
3	肺癌基因表达数据集	12 533	181(31/150)

在数据预处理阶段,由于原始数据集中存在着基因表达水平全为 0 的数据列,同时也存在着少量的基因有表达值,但基因信息为空白的数据列,因此在获得数据之后,本文中将这些全 0 列和信息不全的基因列作为问题数据剔除. 随后将数据离散化为 0,1,2 的整数,为下一步基因的分析研究做好准备工作. 对数据进行离散化处理,一方面是由于基因表达谱数据的数值表征基因的表达水平,相邻数据之间不具有连续性,另一方面数据离散化也可以看作是去噪的一个过程.

3.2　参数分析

K-SVM-RFE 算法中共涉及到 4 个参数,分别为待选择特征的数量 k,初步筛选特征数量 d,K 均值聚类算法所聚的类数 k_1 和在各个类别中使用 SVM-RFE 算法选择的基因数 k_2. 其中初步筛选特征的作用是首先去除大量无关的噪声特征,降低下一过程的计算复杂度,因此 d 的选择对实验结果影响不大,它满足远小于初始特征数量且稍大于待选特征数量即可. 因此本研究在 d 取 600 时进一步探究 k 与 k_1 和 k_2 之间的设置关系. 本实验以结肠癌基因表达谱数据集为实验对象,以 K 最近邻(KNN)作为分类器,设置不同的参数,采用五折交叉验证的方式重复实验 10 次,取分类准确率的平均值作为最终的结果,实验结果如表 2 所示. 由第 2 节知,k_1 与 k_2 需要满足式(4),所以表中不满足此条件的实验设为空.

表 2　不同参数下所选特征的分类准确率

Tab. 2　Classification accuracy of selected feature with different parameters　　%

k	k_2	分类准确率			
		$k_1=5$	$k_1=10$	$k_1=15$	$k_1=20$
5	1	85.4	84.9	**91.0**	89.9
	2	88.8	87.7	89.6	89.3
	3	90.8	88.5	87.5	87.9
10	1		87.8	88.8	91.9
	2	89.5	92.4	**93.7**	91.1
	3	92.7	91.8	92.2	93.5
15	1			88.3	92.0
	2	88.3	92.8	92.8	93.2
	3	92.3	91.6	**95.4**	94.2
20	2		93.9	94.2	91.9
	3	90.6	93.3	94.5	**96.9**
	4	93.3	95.0	96.1	96.2

在表 2 中,加粗的数据为所选特征数量 k 条件下的最佳分类准确率结果. 可以看出,当 k 取 15 和 20 时,分类准确率均在 k_1 与 k 相等,k_2 取 3 时达到最大值,此时有 $k_1\times k_2=3k$;当 k 取 5 和 10 时,虽然最大准确率不在 $k_1=k$ 条件下,但是依然满足 $k_1\times k_2=3k$ 的关系,且如果取 $k_1=k$,$k_2=3$,其结果也依然较好.

因此,设置聚类算法所聚的类数与要选择的特征数量相等,即 $k_1=k$ 且 $k_2=3$ 时,K-SVM-RFE 算法所选特征能够得到较好的分类性能.

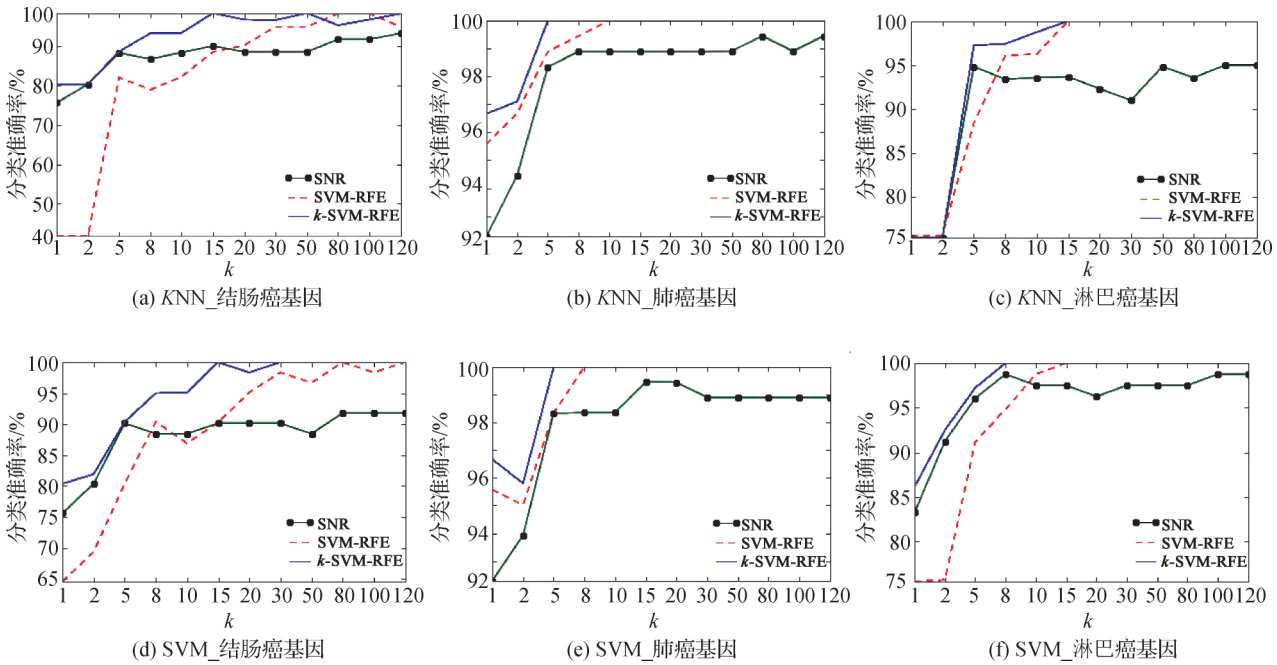
3.3　分类准确率的分析

为了分析比较不同特征数量对特征评价的准确性,实验分别测试重要特征数量为 1,2,5,8,10,15,20,30,50,80,100,120 时的分类性能. 实验中涉及到的一些参数包括:基于 SNR 的特征排序方法初步筛选出 $d=600$ 个重要基因,k,k_1 与 k_2 的取值根据 3.2 节取 $k_1=k$,$k_2=3$;SVM-RFE 算法每次迭代删除的特征比例设为 0.1,其他参数保持默认. 另外,在分类结果验证上,特征选择算法选出的关键基因分别作用于 KNN 和以径向基为核函数的 SVM 这 2 个分类器. 其中 KNN 分类器原理简单,易于理解与实现,而 SVM 分类器在解决小样本、非线性及高维模式识别中表现出许多特有的优势,将 K-SVM-RFE 算法同时作用于这 2 个分类器,可以验证 K-SVM-RFE 算法所选特征在不同分类器上的适用情况. 实验采用五折交叉验证的方式,取 5 次结果的平均值作为最终实验的准确率,实验结果如图 2 所示.

从图 2 中可以看出,K-SVM-RFE 算法在 2 种不同的分类器(KNN 和 SVM)下、3 个不同的数据集和多个不同的关键基因数量上均展现出了比 SVM-RFE 算法和基于 SNR 的特征排序方法更好的分类准确率. 首先,随着提取关键特征数量的递减,K-SVM-RFE 算法与经典的 SVM-RFE 算法的分类准确率在逐步拉开差距,K-SVM-RFE 算法在分类表现上较 SVM-RFE 算法有较大提升,表明 K-SVM-RFE 算法在提取少量关键基因上的有效性. 其次,在所有的结果中,基于 SNR 的特征排序方法所选择特征的分类准确率均不能达到 100%,表明了该过滤式特征选择方法不能去除冗余特征的局限性,而 K-SVM-RFE 算法能够进一步去除冗余特征,达到了特征精选的效果.

另外,对比相同数据集不同分类器条件下的结果,可以发现,以 SVM 作为分类器的分类结果总体都好于 KNN 分类器的结果. 特别是淋巴癌基因表达谱数据集上,SVM 的分类准确率在特征数量为 8 时达到 100%,而 KNN 分类器则在特征数量为 15 时分类准

确率才达到 100%. 产生这样的差异一方面是因为 *K*-SVM-RFE 算法基于 SVM 学习,所以用 SVM 进行分类可取得较好的结果;另一方面也是因为 SVM 在做分类器时它的惩罚因子的值主要是由样本的数量而不是特征数量决定的,因此在各种数据集上应用此模型都会有比较稳定的分类性能.



(a) *K*NN_结肠癌基因 (b) *K*NN_肺癌基因 (c) *K*NN_淋巴癌基因

(d) SVM_结肠癌基因 (e) SVM_肺癌基因 (f) SVM_淋巴癌基因

图 2 不同分类器(*K*NN、SVM)在不同基因(结肠癌、肺癌、淋巴癌基因)表达谱数据集下 3 种特征排序方法的分类正确率与 k 的变化关系图

Fig. 2 Classification accurate rates of different classifiers (*K*NN, SVM) with respect to k on different genes (colon, lung, andlymphoma gene) expression datasets solved by three feature sorting methods

4 结 论

本研究将聚类算法与 SVM-RFE 方法相结合,提出了一种新的面向基因表达谱数据的特征选择方法 *K*-SVM-RFE,以多个基因表达谱数据为实验对象,并通过 2 个分类器分别验证所选基因的分类效果. 研究结果表明了 *K*-SVM-RFE 算法在致癌基因识别上的有效性,特别是在精选少量致癌基因上,性能更佳.

在取得上述成果的同时,本研究还有许多有待进一步研究的地方. 如本文中实验数据均只有 2 个类别,对于多类别数据的分类性能还有待进一步研究;SVM-RFE 和其他聚类算法的结合效果以及 k_1 和 k_2 2 个参数的最佳设置,也有待进一步探讨.

参考文献:

[1] 张东. 基因表达谱的复杂网络研究[J]. 计算机工程应用技术,2011,7(7):1671-1674.

[2] 李颖新,李建更,阮晓钢. 肿瘤基因表达谱分类特征基因选取问题及分析方法研究[J]. 计算机学报,2006,29(2):324-330.

[3] GOLUB TR, SLONIM DK, TAMAYO P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286(5439): 531-537.

[4] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for caner classification using support vector machines [J]. Machine Learning, 2000, 46(13): 389-422.

[5] DING Y, WILKINS D. Improving the performance of SVM-RFE to select genes in microarray data[J]. BMC Bioinformatics. 2006, 7(2): S12.

[6] Yousef M, Jung S, Showe L C, et al. Recursive cluster elimination(RCE) for classification and feature selection from gene expression data[J]. BMC Bioinformatics, 2007, 8(1): 144.

[7] LLOYD S. Least squares quantization in PCM[J]. IEEE Transactions on Information Theory, 1982, 28(2): 129-137.

[8] BANDRES E, MALUMBRES R, CUBEDO E, et al. A gene signature of 8 genes could identify the risk of recurrence and progression in Dukes' B colon cancer patients [J]. Oncology Reports, 2007, 17(5): 1089-1094.

[9] ROSENWALD A, WRIGHT G, WIESTNER A, et al. The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma[J]. Cancer Cell, 2003, 3(2): 185-197.

[10] GORDON G J, JENSEN R V, HSIAO L L, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma[J]. Cancer Research, 2002, 62(17): 4963-4967.

[11] RICHARD SIMON. BRB-array tools data archive for human cancer gene expression[DB/OL]. [2018-01-23]. https://linus.nci.nih.gov/~brb/DataArchive_New.html.

Cancer Gene Selection Algorithm Based on Support Vector Machine Recursive Feature Elimination and Feature Clustering

YE Xiaoquan, WU Yunfeng*

(Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, Xiamen 361005, China)

Abstract: Cancer is usually caused by mutations in genes. It is significant to effectively identify a small number of pathogenic genes from numerous genes. Based on characteristics of gene expression profile data, a novel algorithm (K-SVM-RFE) of gene selection is proposed by combining SVM-RFE with feature clustering algorithm. First, irrelevant genes were removed by feature ranking algorithm. Then, these genes were clustered by K-means and the SVM-RFE algorithm was applied twice to select key genes. We conducted experiments on some real-world data sets and discussed the parameter settings in our method. Results show that, compared with the existing methods, genes selected by the K-SVM-RFE algorithm have significantly improved the classification accuracy, especially in selecting a few key genes.

Key words: gene expression profile; feature selection; K-means; support vector machine