doi:10.3969/j.issn.1672-5565.2013.03.01

基于不同分类模型的基因芯片癌症诊断方法研究

孙远帅[1],陈 垚[1],玄 萍[2],江 弋[1*]
(1. 厦门大学 信息科学与技术学院,福建 厦门 361005;2. 黑龙江大学 计算机科学技术学院,黑龙江 哈尔滨 150080)

摘 要:基因芯片技术的发展为生物信息学带来了机遇,使在基因表达水平上进行癌症诊断成为可能。但基因芯片数据高维小样本的特征也使传统机器学习方法面临挑战。本文利用真实的基因表达数据,测试了目前主要的分类方法和降维方法在癌症诊断方面的效果,通过实验对比发现:基于线性核函数的支持向量机可以有效地分类肿瘤与非肿瘤的基因表达,从而为癌症诊断提供借鉴。

关键词:基因芯片; 癌症诊断; 分类; 主成份分析
中图分类号:Q343.1+5 **文献标识码**:A **文章编号**:1672-5565(2013)-03-161-06

Study of different classification models based-on microarray

SUN Yuan-shuai[1], CHEN Yao[1], XUAN Ping[2], JIANG Yi[1*]
(1. *School of Information Science and Technology*, *Xiamen University*, *Xiamen* 361005, *China*;
2. *Department of Computer Science and Technology*, *Heilongjiang University*, *Harbin* 150080, *China*)

Abstract: The development of microarray technology will bring opportunities to bioinformatics and makes it possible to diagnose cancer on the level of gene expression. But the high-dimensional characteristics and small number of samples in microarray data sets also challenges the traditional machine learning methods. In this paper, we compare the effect among the popular classification and dimensionality reduction methods in the diagnosis of cancer using the real gene expression data, the result demonstrates that SVM based on the linear kernel can better classify tumor and non-tumor gene expression, and thereby provide a reference for cancer diagonsis.

Keywords: Microarray; Cancer Diagnosis; Classification; Principal Component Analysis

基因芯片(Microarray)又称生物芯片,是基因研究领域发展最快的技术之一。它把DNA序列有规律地固定排列在细胞膜或硅片等支持物上,将数以万计的DNA探针添加到一个基片上来测量基因的活性。研究人员使用该技术测量基因表达的变化,研究疾病和环境如何影响细胞,进而解决一些如癌症和虫害等常见问题[1]。尽管基因芯片能够提供大量的生物信息,但是在临床诊断中,使用基因表达数据在分子水平上分类疾病仍是一项非常具有挑战性的研究课题。首先,需要从不同的试验中评估基因表达水平,以提取表达相关的基因(这一过程也称为基因提取),然后使用精确的容易解释的分类规则进行分类。基因芯片技术的发展为生物信息学带来了机遇,但同时也为传统机器学习方法带来了挑战[2]。研究人员通过基因芯片技术可以同时检测数以万计的基因表达情况,使基因表达水平上的癌症分类和检测成为可能。在最近的研究工作中,如RBF神经网络,MLP神经网络,贝叶斯,决策树和随机森林等分类方法已经被用来在基因表达数据中识别不同表达基因。一些研究人员对这些经典分类算法进行改进,并在此基础上提出新的算法,以提高对基因芯片数据分类的准确度。Li L等用遗传算法结合人工神经网络进行特征选择和分类,并取得了较好的实验效果[3]。Inza等则利用贝叶斯网络的性

收稿日期:2013-01-23;**修回日期**:2013-04-16.
资助项目:国家自然科学基金(61001013),黑龙江省教育厅科学研究项目(12521392),黑龙江省自然科学基金(F201119)。
作者简介:孙远帅,男,在读硕士,河南省濮阳人,主要研究方向为矩阵分解、推荐系统和基因表达数据分类。
*通讯作者:江弋,男,副教授,福建省福州人,主要研究方向为数据库、数据挖掘和生物信息学。

能作为子集评价标准[4]。杨帆等从自适应 k 近邻的角度分析了随机森林的分类机理,分析其存在的信息损失,并提出了基于随机森林的潜在 k 近邻算法 RF-PN,充分利用了决策树上的 OOB 样本信息[5]。Paul 等采用遗传规划的多种规则集来进行分类,提出了多数投票遗传规划分类器[6];Ghorai 等集成了非平行平面近似分类器形成一个分类器集,并提出了基于最低平均近似的 NPPC 集决策组合,使分类器集的准确度优于单个分类器[7-8]使用极限学习机(Extreme Learning Machine)对基因芯片数据进行多类别分类,ELM 能够有效避免局部优化、过拟合和计算复杂度高等常见学习算法所面临的问题。除此之外,也有研究人员从基因芯片数据表示结构上着手,Benso 等提出了基于图理论的分类方法,用顶点表示基因,边表示基因表达关系,基于训练样本图和测试样本图的拓扑对比来分类[9]。针对数据高维小样本的特征,出现了诸如人工神经网络(Artificial Neural Networks,ANN)、K 近邻(K nearest neighbors,KNN)、线性判别(Linear Discriminant Analysis,LDA)和支持向量机(Support Vector Machine,SVM)等优秀算法。一些研究人员在这些算法基础上进行了改进。针对 SVM-RF(Support Vector Machine-Recursive Feature Elimination)的不稳定性,提出了两阶段的 SVM-RF 算法,该算法并拥有较好的准确度和较低的时间复杂度。[10]传统基于经验共同信息的基因选择由于样本的数量少而遭受稀疏问题。为了克服这个问题,提出了基于模型的方法来估计模型中类变量的熵[11]。针对线性判别分析(Linear Discriminant Analysis)不适用于基因芯片数据的小样本数据问题,提出了不相关线性判别方法(Uncorrelated Linear Discriminant Analysis)[12]。Au 等提出了一个属性聚类方法,根据基因的内部关系来聚类基因,然后对聚类后的数据进行分类[13]。Statnikov 等提出并构建了基因表达模型选择器(gene expression model selector,GEMS),该系统可以根据基因表达数据自动创建和评估诊断模型和生物标志发现[14-15]。

随着对癌症等复杂疾病研究的深入,研究人员发现,基因对复杂疾病的影响不是通过一个或几个基因完成的,而是通过多个基因组成的复杂调控网络共同作用造成的。因此,单纯根据基因表达情况进行降维可能会丢失调控信息,进而影响疾病诊断的准确率。对高维数据分类一直是机器学习领域研究的热点问题,在文本分类、图像分类领域提出的处理高维数据的分类方法均可以为基因表达数据分类提供指导和借鉴。本文利用真实的基因表达数据,测试了目前主要的分类方法和降维方法在癌症诊断方面的效果,通过实验对比发现:基于线性核函数的支持向量机可以有效地分类肿瘤与非肿瘤的基因表达,从而为癌症诊断提供借鉴。

1 方法

1.1 支持向量机

支持向量机(Support Vector Machine,SVM)广泛应用于数据分类、模式识别和机器学习之中,并且其使用也非常简单。SVM 的两个参数 cost(c)和 gamma(g)对于结果的正确率来说非常重要,选择错误的话将会导致分类结果非常差的情况。但是对于 SVM 方法及其参数的选择,目前国际上还没有形成一个统一的模式,最优 SVM 算法参数选择还只能是凭借经验、实验对比、大范围的搜寻或者利用软件包提供的交互检验功能进行寻优。

LibSVM 是台湾大学林智仁(Chih-Jen Lin)教授等开发的一个简单、易于使用和快速有效的 SVM 模式识别与回归的软件包,提供了关于 SVM 的预测、分类等功能。对于最优参数的选择,可以利用其提供的 grid. py 进行搜寻。

LibLINEAR 也是台湾大学林智仁教授等开发的一个适用于大规模数据的线性分类库。对于高维的海量数据的分析,线性分类是最好的技术。LibLINEAR 支持 L2 常逻辑回归,L2 和 L1 无损线性分类并支持向量机。同时,他也继承了 LibSVM 的很多特性,例如:简单好用,文档全,开源等。LibLINEAR 对于大规模高维数据的训练也是十分有效的,耗时非常短且准确率很高。

1.2 随机森林

随机森林是一个含有多个决策树的集合,其采用多数投票法(针对分类问题)或平均(针对回归问题)得到最后的输出结果。RF 算法的具体描述如下:

算法 Random Forests

输入:1. 训练集 $S=\{(x_i,y_i),i=1,\cdots,n\}$,$(X,Y)\in R^d\times R$

2. 待测样本 $xt\in R^d$

~ For i = 1,…,Ntree

~ (1)对原始训练集 S 进行 Boostrap 抽样,生成训练集 Si

~ (2)使用 S_i 生成一棵不剪枝的树 h_i:

~ a. 从 d 个特征中随机选取 M_{try} 个特征

~ b. 在每个节点上从 Mtry 个特征依据 Gini 指标选取最优特征

~ c. 分裂直到树生长到最大

~ End

输出:1. 树的集合$\{h_i, i = 1, \cdots, N_{tree}\}$

2. 对待测样本 xt,决策树 hi 输出 $h_i(x_t)$

回归:$f(x_t) = \frac{1}{N_{tree}} \sum_{i=1}^{N_{tree}} h_i(x_t)$

分类:$f(x_t) = \text{majority vote}\{h_i(x_t)\}_{i=1}^{N_{tree}}$

1.3　主成分分析法

主成分分析(Principal Component Analysis, PCA)是一种掌握事物主要矛盾的统计分析方法,它可以从多元事物中解析出主要影响因素,揭示事物的本质,简化复杂的问题。计算主成分的目的是将高维数据投影到较低维的空间中,即实施对数据的降维。

对数据降维的意义是显而易见的。低维的数据更易发现数据内部的联系,从而能够提出更好的预测模型,同时也降低了预测过程中计算的复杂度。

进行 PCA 时,我们首先对数据进行中心化(Center)或标准化(Standardize),获得协方差矩阵(Covariance Matrix)或相关矩阵(Correlation Matrix),再对矩阵求其特征值和特征向量,由这些特征向量便可以得到由原始特征线性组合而成的新的特征。将原数据值带入到新的特征中获得新的特征的值,并可计算出其方向上的方差贡献率。当所选取的几个特征向量方向上的方差贡献率累积达到一定值时,便可以认为他们可以取代原特征代表数据的变化情况。通过这种方式,从而达到了对原始数据降维的作用。

对数据进行预处理时,选择中心化或者标准化数据对结果有着至关重要的影响。一般,数据的各维量纲不同,数据水平相差较多的情况下,先标准化数据获得相关矩阵再进行运算,而当其数据水平差异较小时,可以使用协方差矩阵进行运算。

2　实验与分析

本部分将通过 8 个公开的癌症数据集来比较 RF、LibSVM 和 LibLINEAR 的分类性能以及它们在 PCA 后的分类性能变化情况。这 8 个数据集均从网站 http://www.gems-system.org/获得。这些数据被国内外学者广泛引用,具有一定的标准性。数据集的具体信息如表 1 所示,其中包含了 1 个二分类数据集和 7 个多分类数据集,均属于公共卫生领域所发现的人类癌症肿瘤的诊断数据集。

表 1　数据集详细信息

Table 1　Detailed information of data set

Lung Cancer	4 种肺部癌症肿瘤	12 600	203	5	62
DLBCL	弥漫性大 B 细胞淋巴瘤和滤泡性淋巴瘤	5 469	77	2	71
Brain-Lumor1	5 种脑部肿瘤	5 920	90	5	66
Brain-Lumor2	4 种脑部肿瘤	10 369	50	4	207
Leukemia1	3 种急性白血病亚型	5 327	72	3	74
Leukemia 2	3 种白血病	11 225	72	3	156
9 Lumors	9 种人类肿瘤	5 726	60	9	95
11 Lumors	11 种人类肿瘤	12 533	174	11	72

Lung_Cancer 数据集中包含了 186 例肺癌病例样本和 17 例正常病例样本。在 186 例病例样本中包含 139 例肺腺癌、21 例鳞状细胞肺癌、20 例肺类癌和 6 例小细胞肺癌 SCLC。

DLBCL 数据集中包含了弥漫性大 B 细胞淋巴癌病例样本和正常样本共 77 个。弥漫性大 B 细胞淋巴癌是常见的承认淋巴癌,小于 50% 的患者有望治愈。预后模型是目前被广泛应用于预测 DLBCL 的治疗结果,该模型基于预处理过的特征,例如国际预后指数。然而治疗结果预测模型没有办法辨别不同治疗方法的分子学基础,也无法分辨具体的治疗目标。

Leukemia1 数据集中包含了 72 个 3 种急性白血病亚型的病例样本,包括急性髓系白血病(AML)、急性淋巴细胞白血病(ALL-B-CELL)B 细胞核急性淋巴细胞白血病 T 细胞(ALL-T-CELL)。

Leukemia2 数据集中包含了 72 个 3 种不同白血病的病例样本,包括急性髓细胞性白血病(AML)、急性淋巴细胞白血病(ALL)和混合系白血病(MLL)。

Brain_Tumor1 数据集中包含了 90 个 4 种不同脑部神经肿瘤和正常样本。4 种脑部神经肿瘤包含成神经管细胞瘤、原始神经外胚层肿瘤、恶性胶质瘤和非典型畸态瘤。

Brain_Tumor2 数据集中包含了 50 个 4 种不同的脑部神经肿瘤。其分别是典型性恶性胶质瘤、典型性少突神经胶质瘤、非典型恶性胶质瘤和非典型少突神经胶质瘤。

9_Tumors 数据集中包含了 60 个 9 种人类常见

的癌症肿瘤病例样本,包括肺癌肿瘤、乳腺癌肿瘤、结肠癌肿瘤、肾癌肿瘤、骨髓癌肿瘤、中枢神经肿瘤、前列腺肿瘤和卵巢肿瘤。

11_Tumors 数据集中包含了 174 个 11 种人类常见的癌症病例样本和基因数据,包括前列腺癌、膀胱/尿道癌(移行细胞癌和鳞状细胞癌)、浸润性乳腺导管癌、直肠癌、胃腺癌、透明肾细胞癌、肝癌、卵巢浆液性乳头状腺癌、胰腺癌和肺癌(肺腺癌和鳞状细胞癌)。

从表 1 中可以看出,这些数据包含的特征数(基因数)都在 5 237 ~ 12 600 之间,而样本数在 50 ~ 203 之间,特征数与样本数之比很大,属于典型的高维小样本数据。

实验均采用的是 5 折交叉验证法,将每个数据集随机划分成 5 份大小相同或相近的集合,每次实验依次取 1 折数据作为测试数据,余下的作为训练数据,记录 5 次实验的平均准确率作为 5 折交叉的结果。

为保证 RF 算法的结果是较有代表性的,其森林规模(trees)设置为 4 000,其他参数保持默认值不变。LibSVM 算法则根据 grid 函数算出最佳的 cost 和 gamma 参数值的组合并带入其中。LibLINEAR 的参数不做任何改变,保持默认值。

2.1 无降维的分类效果对比

对于原始数据不做任何特征选择等处理,保持其高维小样本特性,所实验得到的结果如下表 2 和图 1 所示。

表 2 三种算法在 8 个数据集上的分类效果对比表

Table 2 Comparative table of the classification results of the three algorithms on eight data sets

数据集名称	准确率		
	RandomForest	LibSVM	LibLINEAR
Lung-Cancer	89.655 2%	68.472 9%	95.073 9%
DLBCL	87.013 0%	75.324 7%	96.103 9%
Brain-Tumor1	82.222 2%	66.666 7%	86.666 7%
Brain-Tumor2	70.000 0%	39.000 0%	38.000 0%
Leukemia1	88.888 9%	52.777 8%	93.055 6%
Leukemia2	93.055 6%	38.888 9%	98.611 1%
9-Tumors	48.333 3%	13.333 3%	60.000 0%
11-Tumors	84.482 8%	16.666 7%	95.402 3%

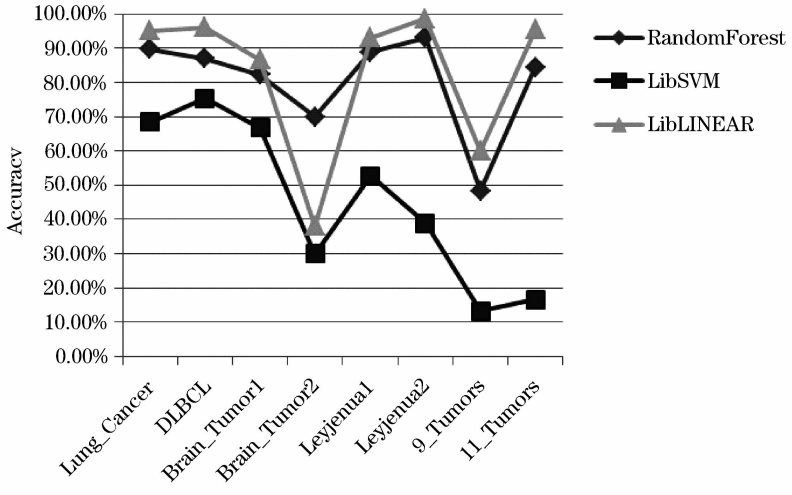

图 1 三种算法在 8 个数据集上的分类效果对比图

Fig. 1 Comparison chart of the classification results of the three algorithms on eight data sets

由以上图表可以看出,对于这种具有高维小样本特征的数据,LibLINEAR 算法拥有非常好的处理能力,基本上在各个数据集上的表现都好过其他两种算法。但其在 Brain_Tumor2 数据集上的表现却非常差,可见其在处理维数/样本数非常高的数据上也不是很理想,有待提升。

LibSVM 算法则表现出对这种类型的数据的处理能力的薄弱,模型的正确率一直在三种算法中排在最后,这点尤其在 11_Tumors 数据集上表现得更加明显。因此,处理这种具有高维小样本特征的数据,不适合直接使用 LibSVM 算法。

相较于前两者,RF 算法则表现得非常稳定,对于各个数据集,都能达到较高的准确率,没有出现对于个别数据集的“不适应”现象,由此可见其实际的应用价值之高。而实际上,RF 算法也是在各种医学临床等实验中最为常用的算法。

2.2 降维后的分类效果对比

对原数据进行 PCA 处理。由于原数据中特征与特征之间的差异水平非常大,在做 PCA 时采用了相关矩阵(correlation matrix)进行运算,将数据标准化,而累积贡献率设定为 0.95,即保持默认值不变。

实验得到的结果如表 3 和图 2 所示。

表 3　PCA 后三种算法在 8 个数据集上的分类效果对比表

Table 3　Comparative table of the classification results of the three algorithms on eight data sets after PCA

数据集名称	PCA 后特征数	PCA	准确率		
			Randomforest	FibSVM	LibLINEAR
Lung-Cancer	146	86	69.950 7%	94.088 7%	57.635 5%
DLBCL	56	98	80.519 0%	96.103 9%	66.233 8%
Brain-Tumor1	65	91	74.444 4%	86.666 7%	50.000 0%
Brain-Tumor2	36	288	66.000 0%	70.000 0%	46.000 0%
Leukemia1	59	90	66.666 7%	93.055 6%	63.888 9%
Leukemia2	53	212	79.166 7%	94.444 4%	93.055 6%
9-Tumors	52	110	53.333 0%	51.666 7%	46.666 7%
11-Tumors	135	93	84.482 8%	87.356 3%	85.057 5%

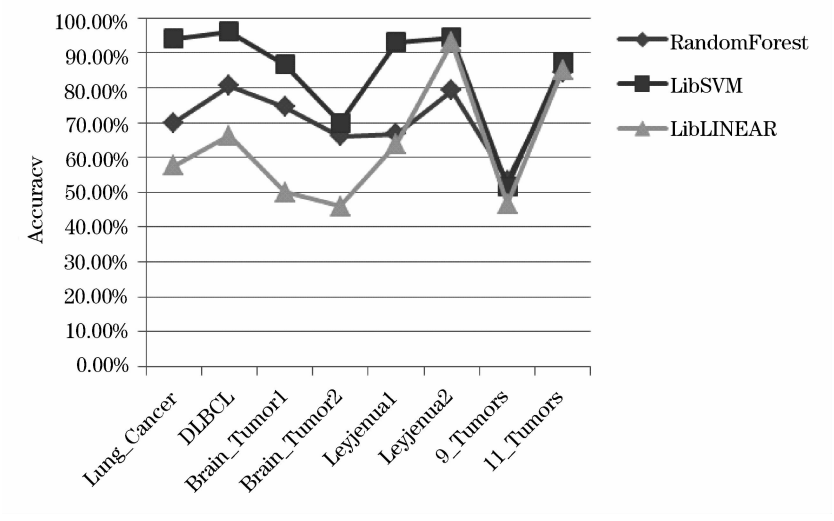

图 2　PCA 后三种算法在 8 个数据集上的分类效果对比图

Fig. 2　Comparison chart of the classification results of the three algorithms on eight data sets after PCA

由以上图表可以直观地看出,PCA 之后,数据的维度下降了非常多,压缩比均在 90 以上。而经过 PCA 之后,LibSVM 算法表现出了对这种低维样本的较强的处理能力,其在各个数据集上的表现都比较优秀,也没有出现明显的对数据集的“不适应”现象,比较稳定。

而 LibLINEAR 算法的表现则变得非常差,对于大部分数据集,其正确率都在三种算法中处于最劣的位置,可见其并不适合这种低维样本的处理。

RF 算法则依旧表现出非常好的稳定性,在处理 8 个数据集时,都保持着较高的准确率。这更加验证了其在实际诊断过程中应用的重要性。

3.3　综合分类效果对比

在使用了如上各种算法对数据进行处理之后,可以对各种方法的正确率进行横向对比,其统计如表 4。表中的黑体字表示在处理该数据集时的最高正确率。

表 4　各种方法的正确率对比表

Table 4　Comparison table of correct rate of the various methods

数据集名称	准确率(PCA 前)			准确率(PAC 后)		
	Randomforest	LibSVM	LibLINEAR	RandomForest	LibSVM	LibLNEAR
Lung-Cancer	89.655 2%	68.472 9%	**95.073 9%**	69.950 7%	94.088 7%	57.635 5%
DLBCL	87.013 0%	75.324 7%	**96.103 9%**	80.519 5%	**96.103 9%**	66.233 8%
Brain-Tumor1	82.222 2%	66.666 7%	**86.666 7%**	74.444 4%	**86.666 7%**	50.000 0%
Brain-Tumor2	**70.000 0%**	30.000 0%	38.000 0%	66.000 0%	**70.000 0%**	46.000 0%
Leukemial	88.888 9%	52.777 8%	**93.055 6%**	66.666 7%	**93.055 6%**	63.888 9%
Leukemia2	93.055 6%	38.888 9%	**98.611 1%**	79.166 7%	94.444 4%	93.055 6%
9-Tumors	48.333 3%	13.333 3%	**60.000 0%**	53.333 3%	51.666 7%	46.666 7%
11-Tumors	84.482 8%	16.667 0%	**95.402 3%**	84.482 8%	87.356 3%	85.057 5%

从上表中可以看出,在处理具有高维小样本特征的数据时,若不考虑对原数据进行特征选择等处理而直接使用,LibLINEAR 具有非常高的正确率,但是其稳定性却有待考证,故更推荐使用 RF 算法,在保持较高正确率的同时,也非常稳定,避免做出错误的判断。而先对原数据进行 PCA 处理,则使用 LibSVM 算法为最佳,稳定性高且正确率高。

但是,对于高维的数据降维花费大量的时间,采集数据时,也难以按照降维后的新特征进行采集,并且,从含有几百、几千或者更多变量的线性组合中发现规律并不是一件简单的事,故其实用价值并没有用原数据直接运算来得高。

3 总结与展望

本文综合对比了 RF、LibSVM 和 LibLINEAR 三种不同的算法在 PCA 前后对于癌症诊断模型正确率的变化情况,总结了各种方法的优劣。通过它们对 8 个数据集的不同表现情况,发现 RF 算法在直接处理具有高维小样本特征的数据时,具有较高的实用价值,而 LibSVM 算法在处理降维后的数据时具有更好的效果。但是考虑到,RF 算法无法提供较高的正确率,以及对高维数据的降维意义不显著的情况,下一步工作应该考虑改进 LibLINEAR 算法,使其稳定性有所提升,对于较高的数据维度与数据实例比的数据能有更好的处理能力。

参考文献(References)

[1] 郭祖茂, 邹权, 李文滨, 韩英鹏. 生物信息学中的学习问题[J]. 山东大学学报(工学版), 2009, 39(3): 1-6.

[2] 邹权, 郭祖茂, 刘扬, 王峻. 类别不平衡的分类方法及在生物信息学中的应用[J]. 计算机研究与发展, 2010, 47(8): 1407-1414.

[3] Li Leping, Weinberg C, Darden T, Pedersen L. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method[J], Bioinformatics, 2001, 17(12): 1131-1142.

[4] Inza I, Larranage P, Etxeberria R, Sierra B. Feature Subset Selection by Bayesian network-based optimization[J]. Artificial Intelligence, 2000, 123(2): 157-184.

[5] 杨帆, 林琛, 周绮凤, 符长虹, 罗林开. 基于随机森林的潜在 k 近邻算法及其在基因表达数据分类中的应用[J]. 系统工程理论与实践, 2012, 32(4): 815-825.

[6] Paul T, Iba H. Prediction of Cancer Class with Majority Voting Genetic Programming Classifier Using Gene Expression Data[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2009, 6(2): 353-367.

[7] Ghorai S, Mukherjee A, Sengupta S, Dutta P. Cancer Classification from Gene Expression Data by NPPC Ensemble[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2011, 3(8): 659-671.

[8] Zhang Runxuan, Huang Guang-Bin, Sundararajan N, Saratchandran P. Multicategory Classification Using An Extreme Learning Machine for Microarray Gene Expression Cancer Diagnosis [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2007, 4(3): 485-495.

[9] Benso A, Carlo S, Politano G. A cDNA Microarray Gene Expression Data Classifier for Clinical Diagnostics Based on Graph Theory [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2011, 8 (3): 577-591.

[10] Tang Yuchun, Zhang Yanqing, Huang zhen. Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2007, 4(3): 365-381.

[11] Zhu Shenghuo, Wang Dingding, Yu Kai, Li Tao, Gong Yihong. Feature Selection for Gene Expression Using Model-Based Entropy[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2010, 7 (1): 25-36.

[12] Ye Jieping, Li Tao, Xiong Tao, Janardan R. Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2004, 1(4): 181-190.

[13] Au W, Chan K, Wong A, Wang Yang. Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2005, 2(2): 83-101.

[14] Statnikov A, Tsamardinos I, Dosbayev Y, Aliferis C. GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data[J]. International Journal of Medical Informatics, 2005, 74(5): 491-503.

[15] Statnikov A, Aliferis C, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis[J]. Bioinformatics, 2005, 21 (5): 631-643.