基因表达数据分类算法及应用探讨

张 浩[1]，叶明全[1]，汪 楠[2]
（1.皖南医学院 计算机教研室，安徽 芜湖 241000；2.安庆职业技术学院 电子信息系，安徽 安庆 246003）

摘 要：通过机器学习方法辅助分析生物信息学中的数据，使用微阵列测试技术所获得的基因表达数据能够将任何给定条件下的基因表达模式表现出来，有利于研究人员更加深入地对众多生物过程的本质进行了解和掌握。文章对基因功能分类方法和基因表达数据的肿瘤分类进行了分析。对于基因表达数据的基因功能分类，按照功能类的隶属关系，提出基于功能树的优势因子决策和基于功能树的置信度调整准则，按照这两种标准进行基因功能树的基因功能分类算法改进。对于基因表达数据的肿瘤分类，将传统SVM算法和kNN算法两者进行结合，形成一种新型的分类算法，主要适用于肿瘤的分类。

关键词：基因表达数据；分类算法；应用；肿瘤分类；功能树
DOI:10.3969/j.issn.1674-5043.2014.03.013
中图分类号: TP301 文献标志码: A 文章编号: 1674-5043(2014)03-0055-05

在过去的几年中，计算机技术在医学和生命科学的各个领域中得到了广泛的发展和应用。生物信息学是指人们利用计算机技术，对生物信息进行分析、存储、检索等，从而产生一个新的知识范畴[1]。随着基因组和其他测序项目的快速发展，研究的主要内容由积累数据向如何解释数据方向发展。计算机科学也从生命系统中得到了启发，形成了较多的新概念，应用于各个相关领域中，使生物信息学得到了进一步的发展和进步。计算机学习方法能够较好地适用缺乏统一理论、含有噪声模式、数据量较大的领域，其基本思想是利用样本学习或者是建模、推理等在数据中进行自动学习，将数据、计算机、概率体系进行合理结合，有利于生物信息学和其他领域得到加大的发展[2]。

1 常用分类算法的性能

分类方法是一种具有监督功能的机器学习方式，可以利用事先已分类的样本集和某一算法之间建立一个参数预测模型，再通过这种模型对没有进行分类的对象进行分类和学习。正确的运用分类方法能够有效地对样本点特征进行选取，使分类的预测精度得到提高[3]。kNN、朴素贝叶斯和支持向量机是常用的分类算法，每种算法都具有不同的实现方式，且算法性能不同。本文对这3种算法在基因表达数据肿瘤分类中的性能进行分析，为基因表达数据的应用提供了依据。

1.1 kNN算法

机器学习中有一种学习方法是利用实例而开展的学习模式，该学习模式能够对训练样例进行简单的储存，并将这些实例中泛化的工作推迟到必需分类的新的实例中，当学习器遇到一个新的查询实例，则开始对这个新实例和已有的储存实例关系进行分析，按照此关系将一个目标函数值赋给新的实例以便计算。最邻近法(简称NN)是一种依靠实例而开展的学习方法，设定实例可以被作为欧式空间中的点进行标记。k近邻法(简称kNN)是最近邻法的一种。当k为1时，则为最邻近法，NN则是对最近点的重要性进行强调。kNN是从整体出发，是一种较为常用的方法，较NN的错误率低。kNN的算法主要是指系统在训练集中对一个需要识别的样本进行最近的k个紧邻寻找，查看k个紧邻中哪类属性较高，就将需要识别的样本划分到哪一类种。利用k紧邻分类器在分类样本中检索与需要识别样本最相近的类别，最终获得需要识别样本的种类信息。

收稿日期: 2014-03-26
作者简介: 张浩(1979-),男,安徽阜阳人,硕士,讲师,主要从事数据挖掘、数据库、网络安全等方面的研究.
基金项目: 安徽省高校自然科学研究重点项目(KJ2014A266).

在kNN算法的基因分类中，使用多维向量表示对象，计算需识别对象和训练集中每个对象的距离主要有以下几种方法：

1）计算Minkowski距离。$L_k(x,y)=\left(\sum_{j=1}^{d}|x_j-y_j|^k\right)^{1/k}$，通常被称为$L_k$范数，欧氏距离是$L_2$范数，$L_1$范数经常被称为Manhattan距离或者是街区距离。

2）计算pearson相关系数。$Corr\_Coef(g_i,g_j)=\frac{\sum_{k=1}^{n}(x_{gik}-\bar{x}_{gi})(x_{gjk}-\bar{x}_{gj})}{\sqrt{\sum_{k=1}^{n}(x_{gik}-\bar{x}_{gi})^2\sum_{k=1}^{n}(x_{gjk}-\bar{x}_{gj})^2}}$，其中$x_{gik}$和$x_{gjk}$是基因$g_i,g_j$在训练集中第$k$个样本表达水平，$\bar{x}_{gi}$和$\bar{x}_{gj}$表示基因$g_i,g_j$在训练集所有样本中的平均表达水平。

3）计算向量间余弦。

假设g_i,g_j为两个基因向量，两者之间的相似度余弦定义为：$\cos(g_i,g_j)=\frac{g_i\cdot g_j}{\|g_i\|\cdot\|g_j\|}$，其中·表示两个向量间的点积，$\|v\|$表示这个向量的范数。

4）计算两者间的欧氏距离。如x为训练集合中点的输入向量，y为要分类点的输入向量，两者之间的欧式距离为：$\sum j(x_j-y_j)^2$。针对不同输入变量的相对重要性，缺少相应衡量方法的问题，尝试运用加权方法来处理，$\sum jw_j(x_j-y_j)^2$，w_j则是权。

1.2 朴素贝叶斯算法

朴素贝叶斯算法是一种概率方法，基于待考查的量按照某概率进行分布，按照已经观察到的数据和概率进行评判，进行最优的方案的设计。朴素贝叶斯算法是衡量多个假设的置信度提供定量的手段，朴素贝叶斯分类器能够与神经网络和决策树学习的性能相提并论，具有较高的实用性。

对朴素贝叶斯分类中的每个实例*x*可以利用属性的合取进行描述，从有限集合*v*中选取目标函数*f(x)*。将一系列关于目标函数的训练样例和新的实例$(a_1,a_2,a_3...a_n)$提供给学习器，进行新实例目标值得预测。

$$V_{MAP}=\underset{v_j\in V}{\operatorname{argmax}}\frac{p(a_1,a_2,a_3...a_n|v_j)p(v_j)}{p(a_1,a_2,a_3...a_n)}$$
$$=\underset{v_j\in V}{\operatorname{argmax}}\ p(a_1,a_2,a_3...a_n|v_j)p(v_j)$$

朴素贝叶斯分类器基于一个简单的假定：在实例目标值给定的情况下，观察到联合的$a_1,a_2,a_3...a_n$概率与每个单独属性的概率相乘的结果一致：$p(a_1,a_2,a_3...a_n|v_j)=\prod_i p(a_i|v_j)$通过换算可以得到朴素贝叶斯分类器所使用的方法为：$V_{NB}=\underset{v_j\in V}{\operatorname{argmax}}\ p(v_j)\prod_i p(a_i|v_j)$，其中朴素贝叶斯分类器输出目标值用$V_{NB}$表示。

朴素贝叶斯学习方法应对不同的$p(v_i)$、$p(a_i|v_j)$项进行估计，按照标准进行新实例的分类，只要满足需要的条件独立性能，则朴素贝叶斯分类V_{NB}和*MAP*分类结果相同。

1.3　支持向量机算法

支持向量机是一种新型的数据挖掘技术，是利用最优化方法进行机器学习问题解决的工具，主要对概率密度估计、回归函数估计、模式识别等问题进行解决。该算法通过结构风险最小化原则对二类问题进行解决。在一个向量空间中给予一组训练集，支持向量机通过选择最优决策超平面来对这两类进行区分。支持向量机具有在小样本情况下仍能够保持良好泛化性能的优点，对于确定学习机器推广性的界而言，处于类边界的少量样本具有决定性作用，被称为支持向量。

支持向量机方法是以线性可分情况下最优分类面为基础，设线性可分样本集为：$(x_1,y_1),(x_2,y_2),...,(x_m,y_m)$，$x\in \mathrm{R}^n, y\in\{+1,-1\}$，其中 x 为一个向量，由 n 个部分组成，该样本中的表达值由每部分最为一个基因进行表示，y 为类别标号，使用 ± 1 进行表示。

*n*维空间中，线性判断函数的一般形式为 $g(x)=w.x+b$，分类面方程为 $w.x+b=0$。

最优分类面是指当一个训练集中的向量可以被一个超平面无错误的进行线性分割，并且距离该超平面最近的向量间的距离最大。最优超平面的获得可以转化为：$\min(w)=\frac{1}{2}\|w\|^2$，满足约束条件：$y_i(w.x_i+b)\geqslant 1, i=1,2,3...,n$。当支持向量到超平面的距离为$1/\|w\|$时，支持向量间的间隔为$2/\|w\|$。

2　基因表达数据的功能分类及改进方法

2.1　基于功能树的优势因子决策算法及改进方法

受基因功能分类算法多分类、多标记特性的影响，为确保置信度能按照一定的设定进行基因归属于某种功能类的判定，在基因功能分类过程中使用经典的算法时，需要人为的对*K*个最大置信度功能作为判定的准则进行确定，为了避免人为武断因素的影响，提出来基于功能树的优势因子决策算法[4]。使用基于优势因子的判断方法标记判断多类别多标记分类能够有效地改善对*K*个最大置信度功能的判断方法的武断性影响。

设训练样本 T_i 在训练中所获得的最高置信度功能类是 $N_{\max}$，对应的置信度是 $Ci_{N_{\max}}$，则样本在其他任何类别 N_x，给定优势因子*y*，那么 $Ci_{N_x}\geqslant\gamma\times Ci_{N_{\max}}$，能够对 T_i 具有类别 N_x 的功能进行判定。

该种方法虽然具有高效性和实用性，但是在具有稀疏性和非平衡的分类中，每种类别的平均置信度水平都具有较大的差异性，若应用一致的优势因子则会出现类别间歧视的现象，较大的优势因子会降低召回率，较小的优势因子则会降低正确率。因此，为了避免这一现象，可以采用加权优势因子的判断方式。

设基因功能树中的节点为 $N(i,x)$，基因表达数据训练集为 Tr，最大置信度的节点为 $N_{\max}(BTree|N(i,x))$，对应置信度为 $C^A_{N_{\max}(BTree|N(i,x))}$，该节点的对应优势因子为 y，测试集为 Tst，测试基因 G 在节点 $N(i,x)$ 的置信度为 $C^A_{G,N(i,x)}>C^A_{N_{\max}(BTree|N(i,x))}\times\gamma\times\frac{C^{(A,avg)}_{Tst,N(i,x)}}{C^{(A,avg)}_{Tst,N_{\max}(BTree,N(i,x))}}$，其中基因功能子树的权重因子为 $\sigma=\frac{(C^{(A,avg)}_{Tst,N(i,x)}}{(C^{(A,avg)}_{Tst,N_{\max}(BTree,N(i,x))}}$。

2.2　基于功能树的置信度调整算法

1）利用功能隶属置信度调整标准进行算法改进。利用功能隶属置信度调整标准可以避免分类器在训练过程中因为训练数据覆盖不全所产生的个别功能丢失现象。利用基因功能树的逻辑关系，不仅可以提高召回率，还可以提高基因在某个节点功能上可能被忽略的置信度水平。

设某一功能树叶子节点为 $N(K,x)$，$K=N(BTree|N(K,x))$，$BTree|N(K,x)=[N(1,x_1),N(2,x_2),...,N(K,x_k)]$，

当任意 $j\in[1,2,3,...,K]$，基因 G，某一指定 T，$C^{b\_avg}_{G,N(j,x_j)}=\frac{\sum_{i=j+1}^{j+T}C^{A}_{G,N(i,x_i)}}{T}$ 或者是 $C^{b\_avg}_{G,N(j,x_j)}=\frac{\sum_{i=j+1}^{K}C^{A}_{G,N(i,x_i)}}{K-N(BTree|N(j,x_j)}$。

当 $C^{A}_{G,N(i,x_i)}<C^{b\_avg}_{G,N(j,x_j)}$ 时，调整基因在功能节点 $N(j,x_j)$ 的置信度为 $C^{'A}_{G,N(j,x_j)}=C^{b\_avg}_{G,N(j,x_j)}$。

2）利用基本功能子树调整标准进行算法改进。利用基本功能子树调整标准可以避免分类器在训练过程中因为训练数据的稀疏性产生的个别功能突显现象。利用基因功能树的逻辑关系，能够降低基因在某个节点功能上可能突增置信度水平，提高精确度，使基因的平均置信度得到降低。

设某一功能树叶子节点为 $N(K,x)$，$K=N(BTree|N(K,x))$，$BTree|N(K,x)=[N(1,x_1),N(2,x_2),...,N(K,x_k)]$，

当任意 $j\in[1,2,3,...,K]$，基因 G，某一指定 T，$C^{f\_avg}_{G,N(j,x_j)}=\frac{\sum_{i=j-T}^{j-1}C^{A}_{G,N(i,x_i)}}{T}$ 或者是 $C^{f\_avg}_{G,N(j,x_j)}=\frac{\sum_{i=j-1}^{j-1}C^{A}_{G,N(i,x_i)}}{N(BTree|N(j,x_j))-1}$。

当 $C^{f\_avg}_{G,N(j,x_j)}<C^{A}_{G,N(j,x_j)}$ 时，调整基因在功能节点 $N(j,x_j)$ 的置信度为 $C^{f\_avg}_{G,N(j,x_j)}=C^{'A}_{G,N(j,x_j)}$

3 基于SVM-kNN算法的肿瘤分类

在使用SVM分类时，出错样本点几乎都分布在分界面周围[5]。分界面周围的样本通常为支持向量，将SVM看作每类只有一个代表点的1NN分类器，因为kNN将每类所有的支持向量都作为一个代表点，使分类器的分类准确率得到了有效的提高。x 表示需要识别的样本，对 x 和两类支持向量代表点 $x-$ 和 $x+$ 的距离距离差进行计算。当距离差大于给定的阈值时，表示 x 远离分界面，应使用SVM分类方法，反之则使用kNN分类方法进行计算。

SVM-kNN算法：

对相应的支持向量、系数、常数 b 进行求值，设测试集为 T，支持向量集为 Tsy，个数为 k。

① 当 $T\neq\varnothing$ 时，$x\in T$，当 $T\equiv\varnothing$ 时，停止；

② 对公式 $g(x)=\sum_i y_i a_i K(x_i,x)-b$ 进行计算；

③ 当 $|g(x)|>\varepsilon$ 时，直接计算 $f(x)=\mathrm{sgn}(g(x))$ 作为输出；当 $|g(x)|<\varepsilon$ 时，使用kNN算法，传递参数 x、Tsy、k，返回结果作为输出；

④ 当 $T\leftarrow T-\{x\}$ 时，返回步骤①。

在步骤③中可将支持向量集 Tsy 作为分类算法的代表点集合，在于进行测试样本计算和每个支持向量的距离时所使用的kNN算法与通常使用的不同，步骤③中的支持向量距离是在特征空间进行计算的，这时所使用的距离公式为 $d(x,x_i)\|\phi(x)-\phi(x_i)\|^2=k(x,x)-2k(x,x_i)+k(x_i,x_i),x_i\in T_{sv}$，分类阈值 e 一般设为1，当 ε 为0时，KSVM则是SVM算法。

4 结 语

通过机器辅助学习方法对生物信息学中的数据进行分析是当今重要的一项研究内容，使用微阵列测试技术所获得的基因表达数据能够将任何给定条件下的基因表达模式表现出来，有利于研究人员更加深入地对众多生物过程的本质进行认识[6]。

从基因功能分类方面出发，由于功能数据集中的一个基因可能划分在几个功能类中，同时这些功能类又可能共同归属在一个分支中，使得基因表达数据的功能分类不同于传统的分类。本文按照功能类的隶属关系，提出了基于功能树优势因子决策和基于功能树置信度调整两个标准，按照这两个标准，进行

了基于功能树的基因功能分类算法改进。通过计算表明，调整基于功能树的置信度后，能够有效地提高SVM平均分类准确率。和传统的算法相比，该算法更适合应用在表达数据较为稀疏但具有较多隶属关系的基因功能判断中。

从基因表达数据的肿瘤分类方面出发，SVM-kNN算法是将kNN算法和SVM算法进行结合的一种新型的、较为通用的分类算法，在肿瘤分类中进行应用不但可以提高SVM分类器的准确率，而且在一定程度上不受核函数参数选择的影响，具有稳健性等优点。

参考文献：

[1] 岳峰,孙亮,王宽全,等.基因表达数据的聚类分析研究进展[J].自动化学报,2008,34(2):113-120.

[2] 蔡立军,沈小乔,林亚平,等.一种改进的基因表达数据分类方法[J].湖南大学学报:自然科学版,2007,34(3):79-82.

[3] 单连峰,李明,张惠丹,等.GLRT和LS_SVM应用于基因表达数据分类[J].数学的实践与认识,2010,09:82-86.

[4] 蔡显圣,杜芳.非负矩阵分解算法在胃癌基因表达数据分类中的应用[J].中国医疗设备,2011(04):28-32.

[5] Austin H,Yang Chenyin.The improvement of breast cancer prognosis accuracy from integrated gene expression and clinical data[J].Expert Systems with Application,2012,39(5):4785-4795.

[6] 詹超.支持向量机在基因表达数据分类中的研究[D].武汉:武汉理工大学论文,2006.

[7] 王丽美.微阵列基因表达数据双聚类的多目标优化算法研究[D].福州:福建农林大学论文,2013.

Algorithm and Application Approach of Gene Expression Data Classification

ZHANG Hao[1], YE Ming-quan[1], WANG Nan[2]

(1.Wannan Medical College, Wuhu 241002, China;

2.Anqing Vocational and Technical College, Anqing 246003, China)

Abstract: When machine learning method is used to help the analysis of bioinformatics data, the gene expression data obtained through microarray testing techniques can show the gene expression patterns under any given conditions, which contributes to the further understanding and mastering of the nature of massive biological processes for researchers. This article has analyzed the classification of gene functions and the tumor classification of gene expression data. For the former, according to the subjection relationship of function classes, dominate factor decision algorithm based on gene function tree and confidence adjustment algorithm based on gene function tree have been put forward to improve the gene function classification algorithm based on function tree. For the latter, the traditional SVM algorithm and kNN algorithm are combined to form a new classification type mainly applied to the classification of tumors.

Key words: gene expression data; classification algorithms; application; Tumor Classification; function tree

---

(上接第54页)

Study and Comparison of IETM Standards

CHEN Yao-pei[1], OUYANG Qing[1], SHI Guan-yu[1], ZHAO Han-bin[2]

(1.Naval Engineering University, Wuhan 430033, China;

2.The Fourth Academy of China Aerospace Science & Industry Corp, Wuhan 430040, China)

Abstract: This paper introduces the contents and features of some Interactive Electronic Technical Manual (IETM) standards, especially the development, advantages and disadvantages of the S1000D, which is widely used in recent years, as well as the main contents and change details of the latest Version S1000D4.1. By summarizing the characteristics and differences between the standards, and the domestic IETM research status, the paper presents issues for attention and the direction for future development.

Key words: IETM; S1000D; data module; CSDB; standard