计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
47(2): 328-335, 2010

基于相关性分析的微阵列数据集成分类研究

于化龙　顾国昌　刘海波　沈　晶　赵　靖
(哈尔滨工程大学计算机科学与技术学院　哈尔滨　150001)
(yuhualong@hrbeu.edu.cn)

Ensemble Classification of Microarray Data Based on Correlation Analysis

Yu Hualong, Gu Guochang, Liu Haibo, Shen Jing, and Zhao Jing
(*College of Computer Science and Technology, Harbin Engineering University, Harbin* 150001)

Abstract The tumor diagnosis method based on microarray data will be developed into a fast and effective molecular-level diagnosis method applied in clinic in the near future. However, it is a challenging task for traditional classification approaches due to the characteristics of high dimensionality and small samples for microarray data. Therefore, ensemble classification algorithms with better performance have attracted more researchers. A novel ensemble classification algorithm for microarray data based on correlation analysis is proposed in this paper to solve the problems of low classification accuracy and excessive computation for current ensemble classification algorithms. The proposed algorithm may extract some training subsets which have the most difference between each other by computing their correlation. Therefore, the proposed algorithm could effectively improve diversity among base classifiers. Support vector machine is selected as base classifier in this paper and the experiment results on leukemia dataset and colon tumor dataset show the effectiveness and feasibility of the proposed algorithm. Meanwhile, the performances of the proposed algorithm based on different parameters are tested and the results are helpful for selecting appropriate parameters.

Key words ensemble classification; microarray data; correlation analysis; feature selection; support vector machine

摘　要　基于微阵列数据的肿瘤诊断方法有望在不久的将来成为临床医学上一种快速且有效的分子层肿瘤诊断方法,但由于微阵列数据存在高维小样本的特点,因而对传统的分类方法提出了挑战,为此研究人员开始关注于性能更好的集成分类算法.针对现有的微阵列数据集成分类算法分类精度不高、计算量过大等问题,提出了一种基于相关性分析的微阵列数据集成分类算法.该算法可以通过计算训练子集间的相关性挑选出差异度最大的一组子集来进行训练,有效地增强了集成中的多样性.应用支持向量机作为基分类器,在急性白血病与结肠癌数据集上的实验结果表明了所提算法的有效性和可行性.同时,测试了算法在不同参数设置下的性能,测试结果为合理的参数设置提供了参考依据.

关键词　集成分类;微阵列数据;相关性分析;特征选择;支持向量机

中图法分类号　TP391.4; TP181

基因芯片又称为DNA微阵列,是一种新兴且意义重大的生物学技术.通过此技术,可以同时检测成千上万个基因在生物体内的活性,为从分子水平上对疾病尤其是对肿瘤进行诊断、分型以及致病机

收稿日期:2008-11-26;**修回日期**:2009-09-02
基金项目:国家自然科学基金项目(60873036);中国博士后科学基金资助项目(20060400809);黑龙江省青年科技基金项目(QC06C022)

理的研究提供了强有力的支持[1-3].然而,由于微阵列实验所检测的基因数量巨大,同时测试费用较高,因而造成了微阵列数据集高维小样本的特点,它的这一特点对传统的分类方法提出了挑战.为此,近年来研究人员开始更多地关注于分类精度更高、鲁棒性更强的集成分类方法.

鉴于集成分类方法在微阵列数据分类方面的优越性,目前已有很多相关的工作被报道.这些工作主要是围绕着如何增加集成中基分类器间的差异来进行展开的.Dettling[4]将两种经典的集成分类方法Bagging与Boosting相结合,提出了BagBoosting算法,它结合了以上两种方法的优点,分类性能与二者相比更优.然而不可忽视的是该方法需要集成大量的个体分类器,对计算时间与存储空间都是一个不小的挑战;文献[5]针对微阵列数据基因高度冗余的特点,提出了一种随机子空间集成分类算法,该算法与单分类器相比性能更优,但是由于基因选取的随机性,使基分类器的性能下降,因此需要集成的基分类器规模庞大,同样存在计算量与存储量过大的问题;文献[6]提出了一种个体差异最大化决策树集成分类算法:MDMT算法,该算法使用完全不同的基因来构造多个C4.5分类器,从而增加了个体间的差异,但由于个体分类器性能下降,使得最终的识别效果并不理想;Peng[7]提出的enSVM算法主要通过对各基分类器的判别空间进行相似性聚类来达到扩大差异的目的,然而由于数据集所含样本数规模的限制,使可靠的聚类难于实现;文献[8-9]均通过采用优化算法选取出的最优分类器集合来对样本进行集成分类,这类方法的分类性能较好,但缺点在于时间复杂度较高,且容易产生过适应的现象.

针对以上工作中所存在的问题,本文提出了一种基于相关性分析的微阵列数据集成分类算法.该算法首先通过样本重取样技术生成大量的训练子集,然后在不同的训练子集上交替使用几种不同的特征选择方法来构造特征子集,以确保各训练子集之间的差异最大化.接下来,通过判断各特征子集之间的相关性挑选出一些彼此间都存在着较大差异的个体,并训练相应的分类器.最后,采用这些分类器组成决策委员会对样本的类属作出判断.在两个数据集上的实验结果表明,与其他集成分类算法相比,本文所提算法可以通过集成更少的基分类器获得与之相当或更好的分类精度,且大大降低了计算的复杂度.

1 问题描述

设一个微阵列数据集由N个样本组成,每个样本包括M个基因,则数据集可以描述为如下形式:$S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_i,y_i),\cdots,(x_N,y_N)\}$,其中$x_i\in\mathbb{R}^M$表示第$i$个样本,$y_i\in\{0,1,\cdots,C-1\}$为样本$i$的类别标签,在本文中,只考虑$C=2$的情形.分类任务就是要在$S$上建立$x_i\rightarrow y_i$的精确映射函数$f$,使得在新样本$x'\in\mathbb{R}^M$出现时,可以通过$f$对其类别标签作出判断$f(x')=y',y'\in\{0,1\}$.

与单分类器不同,集成分类方法需要在S上构造多个不同的映射函数$f_1,f_2,\cdots,f_K$,其中K为基分类器的个数.在二分类问题中,样本x'的集成分类结果可以表示为

$$y'=\begin{cases}0, & \text{if}\left(\sum_{i=1}^{K}f_i(x')\right)\leqslant K/2,\\ 1, & \text{otherwise.}\end{cases}\tag{1}$$

即只有多数成员认定样本属于某一类时,集成分类器才会将该样本划归为此类.

Krogh和Vedelsby[10]指出:集成的泛化误差E等于集成中个体分类器的平均泛化误差$\bar{E}$与平均差异度$\bar{A}$之差.即

$$E=\bar{E}-\bar{A}.\tag{2}$$

从式(2)可以看出,要想提高集成分类器的性能,需要从两方面着手:一是要提高各基分类器的分类准确率;二是要尽量增加基分类器之间的差异度.针对第1个问题,可以通过选用效果较优的分类方法来加以保证.而对于后者,根据Zhou等人[11]的思想可知,如果采用某种策略从大量的基分类器中选出一些差异度较大的个体进行集成,预测的结果可能更加准确.本文的工作主要基于这一思想进行展开.

2 基于相关性分析的集成分类方法

2.1 特征选择

如上所述,微阵列数据不同于其他的数据载体,由于大量冗余基因与噪声基因的存在,使其具有高维小样本的特点.因此,如何从成千上万个基因中选取出与分类密切相关的基因就显得尤为重要[12].特征选择的优点在于不仅可以提高分类的精度,而且可以降低实验的成本[13]并且更好地从分子层上解

释致病的机理[3]. 所以,特征选择在基于微阵列数据的疾病分类问题中起着举足轻重的作用.

近年来,研究人员已经提出了大量的特征基因选择方法[2-3,9,14-15],主要可以分为以下两类:Filter方法(也称为基因排序方法)以及Wrapper方法(也称为基因子集选取方法)[16]. Filter方法会根据某种策略为每个基因的重要程度打分,然后根据分数由高到低进行排列,最后选取一些高分值的基因作为疾病相关基因. 与Filter方法不同,Wrapper方法将分类器嵌入到特征选择的过程中,根据分类精度的高低来评价所选取特征子集的优劣. 两种方法相比,Filter方法的速度更快,而Wrapper方法的分类效果更好. 考虑到时间复杂度与计算量等问题,本文采用Filter方法来选择疾病相关基因. 另外,为了增加各训练子集之间的差异度,采用了5种不同的Filter特征选择方法交替对不同的训练子集进行处理,这5种方法分别列举如下:

1) 信噪比(signal-noise ratio, SNR)[2]

$$SNR = \frac{|\mu_0(g) - \mu_1(g)|}{\sigma_0(g) + \sigma_1(g)}. \tag{3}$$

2) 皮尔森相关系数(Pearson correlation, PC)[9]

$$PC = \frac{\sum_{i=1}^{n}(ideal_i - \mu_{ideal})(g_i - \mu_g)}{\sqrt{\sum_{i=1}^{n}(ideal_i - \mu_{ideal})^2}\sqrt{\sum_{i=1}^{n}(g_i - \mu_g)^2}}. \tag{4}$$

3) 斯皮尔曼相关系数(Spearman correlation, SC)[9]

$$SC = 1 - \frac{6\sum_{i=1}^{n}(ideal_i - g_i)^2}{n(n^2-1)}. \tag{5}$$

4) 欧几里德距离(Euclidean distance, ED)[9]

$$ED = \sqrt{\sum_{i=1}^{n}(ideal_i - g_i)^2}. \tag{6}$$

5) 余弦相似度(cosine coefficient, CC)[9]

$$CC = \frac{\sum_{i=1}^{n} ideal_i \times g_i}{\sqrt{\sum_{i=1}^{n} ideal_i^2 \times \sum_{i=1}^{n} g_i^2}}. \tag{7}$$

其中,$\mu_i(g)$与$\sigma_i(g)$分别表示基因g在第i个类别样本上的平均表达值与标准差,n为样本数,g_i表示基因g在第i个样本上的表达值,而μ_g表示基因g在所有样本上的平均表达值,$ideal$为特征标签,它与类别信息高度相关,可以用来衡量基因与类别的相关程度. 如在二类问题中,可以用二进制字符串来表示$ideal$($ideal_i$为第i个样本所对应的二进制字符),在$ideal_1$中用1和0来分别标识类A和类B中的样本,而$ideal_2$则相反,用0标识类别A中的样本,1标识类B中的样本. 如:

$$ideal_1 = \{1,1,1,1,0,0,0,0,0,0\};$$
$$ideal_2 = \{0,0,0,0,1,1,1,1,1,1\}. \tag{8}$$

μ_{ideal}代表某个特征标签在所有样本上的平均值,如在式(8)中,$\mu_{ideal_1}=0.4$,而$\mu_{ideal_2}=0.6$. 分别将$ideal_1$与$ideal_2$代入式(4)~(7)中进行计算,在PC,SC与CC中保留二者中的较大值,而在ED中保留较小值作为基因重要程度的度量值. 最后,根据度量值对基因进行排序(其中在SNR,PC,SC与CC中按从大到小降序排列,而在ED中按升序排列),选取那些排在前面的基因作为特征基因.

2.2 相关性分析

如果两个训练子集之间存在较大的差异,那么用它们训练的分类器之间也必定有很大的不同. 因此,如何评判两个训练子集间的相关程度就成为挑选差异分类器的关键所在. 一个简单的设想就是通过两子集之间的特征基因重合率来量化二者之间的相关性. 但是,由于特征子集中可能含有很多的冗余基因,因此,这种简单的度量方式并不能真实地反映二者之间的关系. 为此,可以首先在原始训练样本集上对基因进行相似性聚类,用聚类的标号来替代基因的编号,这种做法的优点是使表达相似的基因可以归为同一类,以最大限度地降低冗余基因在相关性分析中所起到的不确定作用,从而更准确地评估二者之间的关系. 两特征子集之间的相关度具体定义如下:

定义1. 给定两特征子集$S_i=\{g_{i1},g_{i2},\cdots,g_{iN}\}$和$S_j=\{g_{j1},g_{j2},\cdots,g_{jN}\}$,其中$g$代表特征子集中相应位置基因的聚类编号,$N$为选取的特征基因的规模. 则二者之间的相关度可表示为$Corr(i,j)=(M_i+M_j)/2N$,其中M_i与M_j分别表示在S_j中可以找到相同聚类编号的S_i中的基因个数,以及在S_i中可以找到相同聚类编号的S_j中的基因个数(如当S_i与S_j中的基因聚类编号分别为$S_i=\{1,2,3\}$与$S_j=\{1,1,2\}$时,有$M_i=2$,而$M_j=3$).

定理1. 任意两特征子集S_i与S_j之间的相关度$Corr(i,j)\in[0,1]$.

证明. 根据M_i与M_j的定义可知:$0\leqslant M_i\leqslant N$,$0\leqslant M_j\leqslant N$,由此可推知$0\leqslant M_i+M_j\leqslant 2N$,又根据相关度定义$Corr(i,j)=(M_i+M_j)/2N$. 证毕.

定理 2. 特征子集间的相关度具备自反性与对称性,即 $Corr(i,i)=1$,且 $Corr(i,j)=Corr(j,i)$.

证明. 从定义 1 可知,S_i 中的所有基因都可以在 S_i 中找到相同的聚类编号,即其自身,故 $M_i=N$,可推知 $Corr(i,i)=(M_i+M_i)/2N=(N+N)/2N=1$,即可证自反性成立;同理,根据相关性定义可知 $Corr(i,j)=(M_i+M_j)/2N=(M_j+M_i)/2N=Corr(j,i)$,即可证对称性成立. 证毕.

由定理 1 与定理 2 可以看出,本文所定义的特征子集相关性度量方法不但可以将特征子集间的相互关系进行量化,限制于[0,1]范围内,而且可以保证特征子集与其自身具备最强的相关性,并且两特征子集间具备唯一的相关性度量值. 接下来需要解决的问题就是如何设立一个标准来评判两个子集之间的关系,究竟是强相关还是弱相关.

定义 2. 给定阈值 $T\in[0,1]$,当两个特征集合 S_i 与 S_j 之间的相关度 $Corr(i,j)<T$ 时,称集合 S_i 与 S_j 是基于阈值 T 的弱相关集合,反之,则称为是强相关集合.

定义 3. 给定阈值 $T\in[0,1]$以及由多个特征子集组成的集群 S,若对 $\forall S_i\in S,S_j\in S(S_i\neq S_j)$,有 $Corr(i,j)<T$,则称此集群为基于阈值 T 的弱相关集群,反之,则称为强相关集群.

根据以上定义可知,一个基于阈值 T 的弱相关集群可以保证在设定阈值 T 的前提下,使特征子集间的差异最大化. 本文的目标就是在大量的特征子集中选取出一些个体组成这样的一个子集群,从而增加基分类器之间的差异度,以提高分类的精度与鲁棒性. 具体算法描述如下:

算法 1. 基于相关性分析的差异特征子集选取算法.

输入:备选特征子集群 $S_{candidate}=\{S_1,S_2,\cdots,S_K\}$,其中 $S_i=\{g_{i1},g_{i2},\cdots,g_{iN}\}$,阈值 $T\in[0,1]$;

输出:选出的差异特征子集群 $S_{Base}=\{S'_1,S'_2,\cdots,S'_L\}$.

过程:

步骤 1. 依次计算 $S_{candidate}$ 中第 1 个子集与其他所有子集之间的相关性,若 $Corr\geqslant T$,则将后者从 $S_{candidate}$ 中剔除,否则保留;

步骤 2. 将 $S_{candidate}$ 中第 1 个子集转移至 S_{Base} 中;

步骤 3. 判断 $S_{candidate}$ 是否为空,若不为空,则返回步骤 1 继续执行,否则退出.

定理 3. 通过算法 1 所选取出的差异特征子集群 S_{Base} 为基于阈值 T 的弱相关集群.

证明. 通过算法 1 显然可知,S_{Base} 中任一子集均与其后选入的所有子集基于阈值 T 为弱相关. 另根据定理 2 中的对称性可推知,其同时也必与在其之前选入的子集弱相关,则可知 S_{Base} 中任意两子集是基于阈值 T 的弱相关集合. 因此,由定义 3 可推知,S_{Base} 为基于阈值 T 的弱相关集群. 证毕.

由定理 3 可以看出,算法 1 能够保证在给定阈值 T 的前提下,使选取的特征子集间的差异最大化.

2.3 基于相关性分析的集成分类算法

根据上面的定义与描述,给出了一个基本的基于相关性分析的集成分类方法模型,如图 1 所示. 从图 1 中可以看出,构造集成分类器的过程主要由样本重取样、特征选择、挑选差异特征子集、训练基分类器以及构造决策委员会等 5 个基本步骤组成. 具体算法描述如下:

算法 2. 基于相关性分析的集成分类算法.

输入:原始训练样本集 D、阈值 $T\in[0,1]$、特征基因数 N、备选子集数 K;

输出:决策委员会 F.

过程:

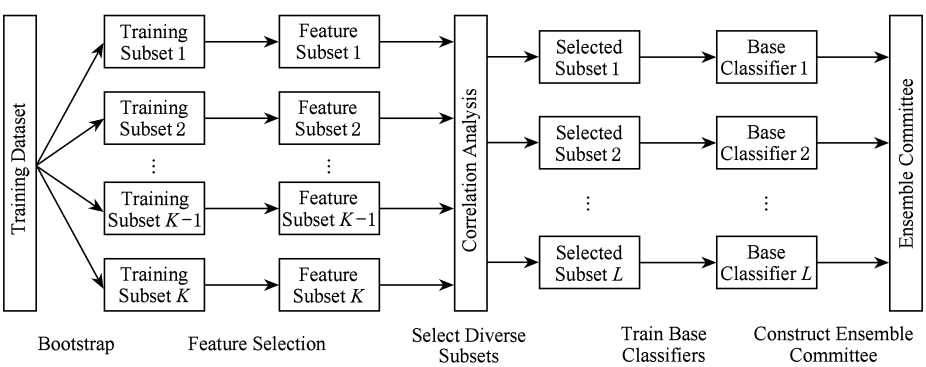

Fig. 1 Model of ensemble classification based on correlation analysis.

图 1 基于相关性分析的集成分类方法模型

步骤 1. 在原始训练样本集 D 上采用 Bootstrap 技术生成 K 个训练子集,每个训练子集与原始训练样本集所含样本数相同;

步骤 2. 在 K 个训练子集上交替使用 5 种特征选择方法生成 K 个特征子集,其中在每个子集上选出 N 个特征基因,形成备选特征子集群 $S_{\text{candidate}}=\{S_1,S_2,\cdots,S_K\}$;

步骤 3. 调用算法 1,得到基于阈值 T 的弱相关特征子集群 $S_{\text{Base}}=\{S'_1,S'_2,\cdots,S'_L\}$;

步骤 4. 使用选出的 L 个差异子集分别独立训练个体分类器 $f_1,f_2,\cdots,f_L$;

步骤 5. 采用步骤 4 中得到的 L 个个体分类器组成决策委员会 $F=\{f_1,f_2,\cdots,f_L\}$,应用多数投票法(见式(1))对未来无标签样本的类属作出判断.

3 实验结果与讨论

3.1 数据集

为便于对实验结果与同类的工作进行比较,我们采用两种研究得比较充分的肿瘤微阵列数据集,一种是急性白血病数据集(leukemia dataset),另一种是结肠癌数据集(colon tumor dataset).

急性白血病数据集由 72 个样本组成,每个样本均含有 7129 个基因的表达数据.其中 47 个样本被诊断为急性淋巴细胞白血病(acute lymphoblastic leukemia,ALL),另外 25 个被诊断为急性骨髓性白血病(acute myeloid leukemia,AML),详细描述可参考文献[2].

结肠癌基因表达数据集包括 62 个样本,其中的 40 个为结肠癌组织样本(标记为 negative),另外 22 个样本为正常组织样本(标记为 positive),每个样本由 2000 个基因组成.关于此数据集的具体描述详见文献[1].这两个数据集的原始数据均可以在 http://sdmc.lit.org.sg/GEDatasets/Datasets 上下载得到.

3.2 分类性能比较

在分类时,我们采用支持向量机(support vector machine, SVM)作为基分类器,这主要出自以下考虑:SVM 具有很坚实的理论基础与较强的泛化能力,并且特别适合于高维小样本分类问题.在本文实验中,我们采用的是基于径向基核函数的 SVM 分类器,宽度为 5,控制因子为 500.另外,为了与前人的工作进行比较,本文采用"留一交叉检验法"(leave-one-out cross validation,LOOCV)作为分类精度的评价方法.本文算法在 Matlab7.0 环境下实现,其中支持向量机的实现采用的是 Gunn 开发的支持向量机工具箱[17],而基因的聚类则采用的是分层聚类方法[18].

首先将本文算法与基于各种特征选择方法的单 SVM 分类器以及 Bagging 集成分类算法进行了比较.比较结果如表 1 所示,其中基因的聚类数以及备选训练子集数均设为 100,本文算法中的阈值设为 0.8,选取的特征基因个数分别为 50,100 与 200,括号中的数字代表在 LOOCV 中使用本文算法所构造的基分类器的平均数.

Table 1 Comparison of Classification Performance for Various Approaches

表 1 各种方法的分类性能比较

Dataset	Number of Feature Genes	Classification Accuracy/%						
		SNR	*PC*	*SC*	*CC*	*ED*	Bagging	Our Method
Leukemia Dataset	50	95.83	94.44	90.28	87.50	90.28	**97.22**	**97.22**(38.47)
	100	94.44	94.44	91.67	90.28	93.06	95.83	**97.22**(24.53)
	200	**95.83**	94.44	90.28	90.28	91.67	**95.83**	**95.83**(8.73)
Colon Tumor Dataset	50	77.42	82.26	70.97	74.19	70.97	85.48	**87.10**(26.97)
	100	75.81	75.81	77.42	74.19	77.42	85.48	**88.71**(13.92)
	200	80.65	80.65	72.58	82.26	72.58	**87.10**	**87.10**(5.52)

从表 1 中可以看出,无论选取多少特征基因,本文算法的分类精度都是最高的.其中在 3 组实验中与 Bagging 的分类精度相同,但需要注意的是,本文算法需要集成的基分类器数量与 Bagging 相比要少得多.如当选用 200 个特征基因时,Bagging 算法与本文算法在两个数据集上都可以获得 95.83% 和 87.10% 的分类准确率,但是 Bagging 算法需要集成 100 个基分类器,而本文算法只需要分别集成 8.73 与 5.52 个.因此,本文算法在计算复杂度与存储开销方面与 Bagging 算法相比更小.

3.3 参数对分类性能的影响

本文也测试了参数改变对算法的影响.参数分

别包括基因的聚类数(50,100,200,500,1000,2000)、阈值 T(0.5,0.6,0.7,0.8,0.9,1.0)以及备选训练子集的规模(25,50,75,100,125,150,175,200),而特征基因的个数恒设为 100.当一种参数发生改变时,其他参数均按照第 3.2 节中的参数进行设置.实验结果如表 2、表 3(其中括号中的数字代表在 LOOCV 中本文方法所构造的基分类器的平均数)及图 2 所示:

Table 2 Effect of Classification Accuracy by Number of Clusters for Genes

表 2 基因的聚类数对分类精度的影响 %

Dataset	Number of Clusters for Genes					
	50	100	200	500	1000	2000
Leukemia Dataset	97.22(5.73)	97.22(24.53)	98.61(49.75)	97.22(62.17)	95.83(70.83)	95.83(84.17)
Colon Tumor Dataset	83.87(5.03)	88.71(13.92)	87.10(30.32)	83.87(76.39)	85.48(97.16)	85.48(98.97)

Table 3 Effect of Classification Accuracy by Threshold T

表 3 阈值 T 对分类精度的影响 %

Dataset	Threshold T					
	0.5	0.6	0.7	0.8	0.9	1.0
Leukemia dataset	93.06(1.39)	95.83(3.28)	97.22(6.42)	97.22(24.53)	95.83(67.72)	95.83(100)
Colon tumor dataset	80.65(1.65)	82.26(3.13)	87.10(5.69)	88.71(13.92)	87.10(53.90)	85.48(100)

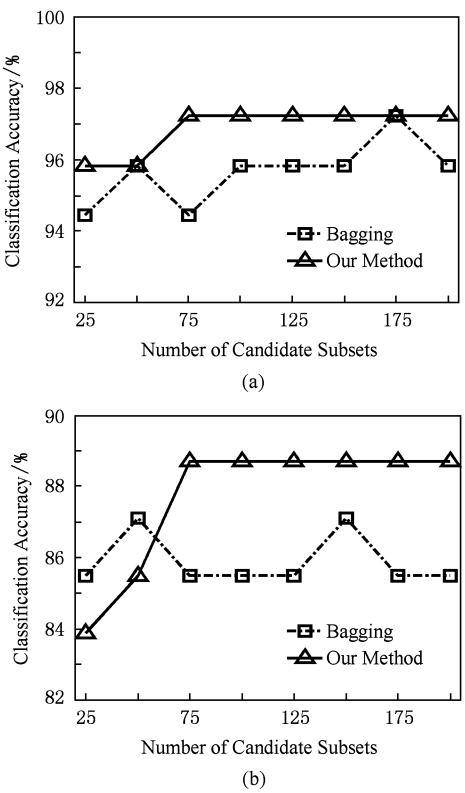

Fig. 2 Effect of classification accuracy by the size of candidate training subsets. (a) Leukemia dataset and (b) Colon tumor dataset.

图 2 备选训练子集规模对分类精度的影响.(a) 急性白血病数据集;(b) 结肠癌数据集

从表 2 与表 3 中可以看出,当这两个参数过小或过大时,分类精度都不高.这主要是因为:当参数过小时,对特征子集间的差异要求过于严格,导致只有很少的基分类器可以被构造并参与最后的决策,因而降低了分类的准确率;而当参数过大时,则会导致大量的冗余分类器参与最后的决策,同样会降低分类的精确度.从表 2 的实验数据中可以看出,当将基因聚为 100~200 类时,分类精度可以达到最高.而从表 3 中可以看出,在将基因聚为 100 类的前提下,0.7~0.8 的阈值即可以保证得到最高的分类精度.图 2 则展示了备选训练子集数对分类性能的影响.从图 2 中可以看出,当备选训练子集数过少时,本文算法的分类精度较低,甚至要低于 Bagging 集成算法.这主要是因为可供挑选差异个体的备选子集规模过小,导致只能构造极少量的基分类器进行集成,从而降低了分类的准确率.当备选训练子集数达到一定规模后,数量上的增长并不会对分类的精度产生何种影响.如从图 2 中可以看出,当提供的备选训练子集数多于 75 个时,无论在急性白血病数据集还是结肠癌数据集上均可以得到最高的分类精度.由以上实验可见,合理的参数设置将有助于本文算法发挥出更大的功效.

3.4 与前人工作的比较

为了体现本文算法的优势,我们将其与前人的一些工作在分类精度与集成规模等两方面进行了比较,详细比较结果如表 4 所示:

Table 4 Related Works on the Two Datasets

表 4 两个数据集上的相关工作 %

Method	Classification Accuracy (Size of ensemble)	
	Leukemia Dataset	Colon Tumor Dataset
Our method	97.22(24.53)	88.71(13.92)
Bagging+SVM	97.22(100)	87.10(100)
Boosting[4]	94.33(100)	80.86(100)
BagBoosting[4]	95.92(100)	83.9(100)
Random subspace[5]		82.26(N/A)
MDMT[6]	97.50(25)	85.80(25)
enSVM[7]	97.22(25)	88.71(25)
EA+Majority voting[8]		88.87(42)
EDA+Majority voting[9]	100.00(20)	95.16(20)

从表 4 中可以看出,本文算法可以取得与其他集成分类算法相当或更优的分类精度,且集成规模与其他算法相比更小,这就意味着更小的计算复杂度与存储空间. 只有文献[9]的分类精度明显高于本文算法,但需要注意的是,它采用了优化算法来选取最优的分类器集合,需要大量的迭代,这也就表明该算法具有更高的计算复杂度,因此本文方法与其相比仍然具有一定的优势.

4 结 论

针对现有的微阵列数据集成分类算法分类精度不高、计算量过大等问题,本文提出了一种基于相关性分析的集成分类算法. 该算法的特点在于可以有效地挑选出那些具有最大差异度的基分类器,从而增强集成的多样性,以达到提高分类性能的目的. 经实验证明该算法与前人的工作相比,不但可以保持或提高分类的正确率,而且可以大幅度降低计算的复杂度与存储开销. 本文算法在其他的高维小样本分类问题中是否同样有效,有待在未来的工作中进行验证.

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 array [J]. Proc of the National Academy of Sciences, 1999, 96(12): 6745-6750

[2] 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(5439): 531-537

[3] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines [J]. Machine Learning, 2002, 46(1): 389-422

[4] Dettling M. Bagboosting for tumor classification with gene expression data [J]. Bioinformatics, 2004, 20(18): 3583-3593

[5] Bertoni A, Folgieri R, Valentini G. Bio-molecular cancer prediction with random subspace ensembles of support vector machines [J]. Neurocomputing, 2005, 63(1): 535-539

[6] Hu H, Li J Y, Wang H, et al. A maximally diversified multiple decision tree algorithm for microarray data classification [C] //Proc of 2006 Workshop on Intelligent Systems for Bioinformatics. Hobart, Australia: CRPIT, 2007: 35-38

[7] Peng Y H. A novel ensemble machine learning for robust microarray data classification [J]. Computers in Biology and Medicine, 2006, 36(6): 553-573

[8] Kim K J, Cho S B. An evolutionary algorithm approach to optimal ensemble classifiers for DNA microarray data analysis [J]. IEEE Trans on Evolutionary Computation, 2008, 12(3): 377-388

[9] Chen Y H, Zhao Y O. A novel ensemble of classifiers for microarray data classification [J]. Applied Soft Computing, 2008, 8(4): 1664-1669

[10] Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning [G] //Tesauro G, Touretzky D S, Leen T K. Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1995: 231-238

[11] Zhou Z H, Wu J X, Tang W. Ensembling neural networks: Many could be better than all [J]. Artificial Intelligence, 2002, 137(1): 239-263

[12] Xing E P, Jordan M I, Karp R M. Feature selection for high-dimensional genomic microarray data [C] //Proc of the 18th Int Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2001: 601-608

[13] Roth F P. Bringing out the best features of expression data [J]. Genome Research, 2001, 11(11): 1801-1802

[14] Li Yingxin, Ruan Xiaogang. Feature selection for cancer classification based on support vector machine [J]. Journal of Computer Research and Development, 2005, 42(10): 1796-1801 (in Chinese)
(李颖新, 阮晓钢. 基于支持向量机的肿瘤分类特征基因选取[J]. 计算机研究与发展, 2005, 42(10): 1796-1801)

[15] Wang Shulin, Wang Ji, Chen Huowang, et al. Heuristic breadth-first search algorithm for informative gene selection based on gene expression profiles [J]. Chinese Journal of Computers, 2008, 31(4): 636-649 (in Chinese)
(王树林, 王戟, 陈火旺, 等. 肿瘤信息基因启发式宽度优先搜索算法研究[J]. 计算机学报, 2008, 31(4): 636-649)

[16] Inza I, Larranaga P, Blanco R, et al. Filter versus wrapper gene selection approaches in DNA microarray domains [J]. Artificial Intelligence in Medicine, 2004, 31(2): 91-103

[17] Gunn S R. Support vector machines for classification and regression [EB/OL]. (1998-05-10) [2008-11-23]. http://users.ecs.soton.ac.uk/srg/publications/pdf/SVM.pdf

[18] Wang Y H, Makedon F S, Ford J C, et al. HykGene: A hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data [J]. Bioinformatics, 2005, 21(8): 1530-1537

Yu Hualong, born in 1982. Received his master degree in computer science from Harbin Engineering University in 2008. He is currently a PhD candidate of Harbin Engineering University. Student member of China Computer Federation. His current research interests mainly include bioinformatics, pattern recognition and machine learning.

于化龙,1982 年生,博士研究生,中国计算机学会学生会员,主要研究方向为生物信息学、模式识别与机器学习.

Gu Guochang, born in 1946. Professor and PhD supervisor of the College of Computer Science and Technology, Harbin Engineering University, Harbin, China. His main research interests include pattern recognition, image processing, machine learning and bioinformatics.

顾国昌,1946 年生,教授,博士生导师,主要研究方向为模式识别、图像处理、机器学习与生物信息学.

Liu Haibo, born in 1976. PhD and associate professor. His main research interests include image processing and pattern recognition.

刘海波,1976 年生,博士,副教授,主要研究方向为图像处理与模式识别.

Shen Jing, born in 1969. PhD and associate professor. Her main research interests include image processing and data mining.

沈　晶,1969 年生,博士,副教授,主要研究方向为图像处理与数据挖掘.

Zhao Jing, born in 1972. PhD and associate professor. Her main research interests include software testing, software reliability evaluation and machine learning.

赵　靖,1972 年生,博士,副教授,主要研究方向为软件测试、软件可靠性评估与机器学习.

Research Background

This work is partially supported by National Natural Science Foundation of China under grant No. 60873036, China Postdoctoral Science Foundation Funded Project under grant No. 20060400809 and the Science and Technology Special Foundation for Young Researchers of Heilongjiang Province of China under grant No. QC06C022.

Microarray is a novel biology technology and it provides the ability to measure the expression levels of thousands of genes simultaneously in a single experiment and makes it possible to provide diagnosis for disease, especially for tumor, at molecular level. However, it also presents challenge for traditional data analysis methods and pattern classification methods because there are a large number of gene expression values per experiment, and a relatively small number of experiments. Therefore, researchers have been more interested in ensemble classification algorithm with better performance. In this paper, we propose a novel ensemble classification algorithm of microarray data based on correlation analysis to solve the problems of low classification accuracy and excessive computation of the current ensemble classification algorithms, and apply the algorithm to Leukemia microarray dataset and colon tumor dataset to validate its feasibility and effectiveness. This work is expected to help medicine researchers and doctors to design an accurate, fast and low storage tumor clinical diagnostic system based on microarray data in the near future.