基于近邻传播聚类的集成特征选择方法

孟 军 尉双云
(大连理工大学计算机科学与技术学院 大连 116024)

摘 要 针对高维数据中的类标记仅与少部分特征关联紧密的问题,提出了基于排序聚合和聚类分组的特征随机选择集成学习方法。采用排序聚合技术对特征进行过滤,选出与样本分类相关的特征,以bicor关联系数作为关联衡量标准,利用近邻传播聚类算法进行分组,使不同组的特征互不关联,然后从每个分组中随机选择一个特征生成特征子集,便可得到多个既存在差异性又具备区分能力的特征子集,最后分别在对应的特征子空间训练基分类器,采用多数投票进行融合集成。在7个基因表达数据集上的实验结果表明,提出的方法分类误差较低,分类性能稳定,可扩展性好。

关键词 分类,排序聚合,近邻传播聚类,集成特征选择
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.3.050

Affinity Propagation Clustering Based Ensemble Feature Selection Method

MENG Jun YU Shuang-yun
(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract Aiming at the problem that only a small part of features are associated with the sample classification in high-dimensional data containing thousands of features, a filtering and grouping based feature random selection ensemble learning method was proposed. Rank aggregation technique was used to select the relevant features, and we grouped them by affinity propagation clustering algorithm using bicor correlation coefficient as distance measure. The feature clusters were produced and the feature pairs from any two different clusters are not correlated. A feature from each cluster was selected randomly, and then a relevant and discriminative feature subspace was generated. In this way, many feature subspaces can be generated. Base classifiers were trained in the produced feature subspaces and fused together using a majority voting method. The experiments on 7 gene expression data sets show that the proposed method can effectively reduce the classification error. Meanwhile, it also has more stable performance, and good expansibility.

Keywords Classification, Rank aggregation, Affinity propagation clustering, Ensemble feature selection

1 引言

集成特征选择是一种基于特征选择的集成学习方法,通过搜索数据集的特征空间选出不同的特征子空间,在得到的多个特征子空间中分别构建基分类器进行集成,能够获得比Bagging和AdaBoost更好的分类性能[1]。自1998年Ho[2]受到随机划分(Stochastic Discrimination, SD)理论的启发提出了一种基于随机子空间(Random Subspace Method, RSM)的决策树构建集成方法以来,这类方法引起了人们极大的兴趣。代表性的方法有:Optiz[1]提出的基于遗传算法的集成特征选择方法和Breiman[3]提出的随机森林(Random Forest)算法。近年来,Bock[4]采用RSM和(或)Bagging构建特征子空间,产生多个不同的基于广义可加模型(Generalized Additive Models, GAMs)的基分类器,得到GAMbag、GAMrsm和GAMens集成学习。姚旭[5]等为保证所选的具有较大差异性的特征子集内部特征也具有较小的相关性,运用抽样技术结合基于正则化互信息的特征度量准则生成特征子集,并用基分类器对测试样本的分类结果差异度作为评价特征子集之间差异性大小的准则选出差异性大的特征子集,在其对应的特征空间上训练基分类器进行融合,在仿真数据和UCI数据集上获得了较高的分类精度。

集成特征选择方法降低了高维度对学习算法的影响,同时能够产生具有差异性的基分类器集合,进而形成有效的集成学习,适合于高维数据分类问题。Moon[6]采用随机方法划分特征空间,形成互斥特征子空间训练基分类器,提出了一种源于随机划分的集成分类(Classification by Ensembles from Random Partitions, CERP)方法,其适合高维数据分类,并用于疾病诊断。Liu[7]基于快速关联过滤(Fast Correlated-Based Filter, FCBF)提出了一种简单、高性能、便于实现的分组集成基因选择(Ensemble Gene Selection by Grouping, EGSG)方法,该方法运用近似Markov Blanket进行基因分组,使同组内的基因相互关联,在此基础上,从每个分组的前t个与类标记关联紧密的基因中运用随机方法选择一个构建基因子集以保证所选基因子集的优越性,在生成的特征子空间

到稿日期:2014-03-28 返修日期:2014-06-23 本文受辽宁省自然科学基金项目(20130200029)资助。
孟 军(1964-),女,博士,副教授,主要研究方向为机器学习与数据挖掘,E-mail:mengjun@dlut.edu.cn;尉双云(1984-),男,硕士生,主要研究方向为机器学习与数据挖掘。

中训练基分类器进行集成，在癌症数据分类中获得了较高的分类准确度。

研究与实验表明，在保证特征子集多样性的同时，提升特征子集的区分能力能够改善集成学习算法整体的分类能力。针对目前多数集成特征选择方法集中在搜索整个特征空间构建特征子集，而高维数据中样本分类通常只与少部分特征相关联的问题，本文提出了一种基于近邻传播聚类的集成特征选择(Affinity Propagation Based Ensemble Feature Selection，APBEFS)方法。该方法中采用排序聚合技术(Rank Aggregation Techniques)[8]对特征进行过滤，选出对样本分类具有较强的影响能力的特征，对选出的特征以 bicor 相关系数[9]为特征关联度量标准，采用近邻传播聚类(Affinity Propagation Clustering，AP)[10]进行分组，然后从每个分组中随机选择一个特征构建特征子集，在保证生成的特征子集之间存在较大差异的同时，使得子集内的特征互不关联，从而提高了对应基分类器的整体分类性能，有效改善了集成学习的性能。

2 基本理论

2.1 排序聚合技术

大多数特征过滤方法可以视为排序问题。特征排序可描述为：对于数据集 $D=(X,Y)$，其中：样本观察值 $X=(x_{ij})_{j=1,2,\cdots,M}^{i=1,2,\cdots,N}$，$N$ 为样本个数，M 为特征个数，类标记 $Y=(y_1,y_2,\cdots,y_N)$，定义一个评分函数(Scoring Function) $S(x)$ 来衡量特征空间 $F=(f_1,f_2,\cdots,f_M)$ 中的特征在不同样本组的差异，然后根据评分估计值计算统计显著性(Statistical Significance)并排序，得到特征排序(Rankings) $R=(r_1,r_2,\cdots,r_M)$，$r_i(1\leqslant i\leqslant M)$ 表示特征 f_i 排序后所在的位置序号，由排序 R 可得到一个特征有序表(Ordered List) $L=(l_1,l_2,\cdots,l_M)$，$l_i(1\leqslant i\leqslant M)$ 表示位置 i 对应的特征序号，于是 $l_p=q\Leftrightarrow r_q=p(p,q\in[1,M])$，选择 *Top-K* 个特征作为特征子集。这类方法通常简单、快速、便于实现，因而被广泛运用到各种高维数据分析中。

虽然特征排序方法在大多数情况下能够获得令人满意的结果，然而数据的轻微扰动，或对同一个数据集采用不同的方法，特征排序选择的结果也不尽相同。排序聚合技术[8,11,12]的出现在一定程度上解决了这一问题，该技术采用集成学习的思想，通过执行多次特征排序，把各排序结果按照某一方式融合后选择特征子集，能够有效提高特征选择的稳定性。

特征排序聚合技术按照生成特征排序的方法划分，可分为基于排序标准(Different Ranking Criteria)的方法和基于数据扰动(Different Perturbed Versions of the Data Set)的方法两类[11]。基于排序标准的方法是对同一个数据集采用多个不同的排序方法对特征进行排序，然后把排序结果按照一定方式聚合起来，根据聚合结果最终形成一个优化的排序。如：令 $R^{(e)},R^{(f)},R^{(s)},R^{(m)},R^{(w)}$ 分别表示 eBayes、Fold-Change、SAM、maxT 和 Welch T-test 5 种排序方法对同一个数据集 D 的特征排序，采用平均聚合(Mean Aggregation)[11]方式，得到特征排序聚合观察值 $\bar{R}=(\bar{r}_1,\bar{r}_2,\cdots,\bar{r}_M)$，$\bar{r}_j=(r_j^{(e)}+r_j^{(f)}+r_j^{(s)}+r_j^{(m)}+r_j^{(w)})/5$，然后对 $\bar{R}$ 排序，得到优化后的特征有序表 $\hat{L}=(\hat{l}_1,\hat{l}_2,\cdots,\hat{l}_M)$，选择 *Top-K* 个特征作为最终的特征子集。这类方法排序结果通常与选择的多种排序方法有关，不同的方法组合产生的结果也不尽相同。

基于数据扰动的方法是采用 Bootstrap 或 Subsampling 对原数据集进行扰动，重复多次，对得到的多个扰动数据集选用一个排序方法进行特征排序，然后把生成的排序结果按照一定方式聚合起来，根据聚合结果最终形成一个优化的排序。令有序表 $R^{(1)},R^{(2)},R^{(3)},R^{(4)},R^{(5)}$ 分别表示采用 Subsampling 对数据集的 5 次扰动，利用 Welch T-test 方法排序的结果，采用简单平均聚合方式，得到特征排序聚合观察值 $\bar{R}=(\bar{r}_1,\bar{r}_2,\cdots,\bar{r}_M)$，$\bar{r}_j=(r_j^{(1)}+r_j^{(2)}+r_j^{(3)}+r_j^{(4)}+r_j^{(5)})/5$，然后对 $\bar{R}$ 排序，得到优化后的特征有序表 $\hat{L}=(\hat{l}_1,\hat{l}_2,\cdots,\hat{l}_M)$，选择 *Top-K* 个特征作为最终的特征子集。这类方法的排序结果通常与采取的扰动方法和次数有关，数据集扰动变化小，扰动次数多通常能够得到较稳定的排序。

多个特征排序的聚合方式主要有：平均(Mean)、中心值(Median)、分位点(Quantile)、马可夫链模型(Markov Chain Model)以及鲁棒排序(Robust Rank)等。Wald[12]通过在 11 个数据集上采用 5 种分类算法对聚合方式进行实验比对，结果表明平均聚合方式简单有效，而且计算代价相对较小，适用于高维数据处理。

2.2 近邻传播聚类算法

近邻传播聚类是 2007 年 Frey[10]在 Science 杂志上提出的一种聚类算法，该算法将所有的数据点都看作是潜在的类代表点(Exemplar)，通过数据点之间传递、更新信息，选出一个代表点的集合，最后将每个数据点归属到最近的代表点形成数据点划分。与传统的 K-means、K-center 方法比较，AP 算法具有 3 个优势[13]：(1)不需要预先指定类的个数，由算法自动形成；(2)能够产生更为稳定的、精确的聚类结果；(3)在达到相同聚类精度的前提下，AP 算法需要的时间更少。

AP 算法基于相似度矩阵。数据点之间的距离(通常采用欧氏距离的负值)组成相似度矩阵 $S_{N\times N}$，此矩阵可以是对称的，也可以是不对称的，对角线上的值 $s(k,k)$ 称为偏向参数 P(Preference)，它决定对应数据点 k 能否成为类代表点，该值越大，点 k 成为类代表点的可能性也就越大。通常所有数据点的 P 值均设置为相同的值，即所有的数据点具有相同的几率成为类代表点。P 的取值大小决定了算法产生的类簇个数的多少，取值较大时产生的类簇个数较多，较小时产生的类簇个数较少。

采用 AP 算法聚类时传递可信度(Responsibility，R)和可用度(Availability，A)两个重要信息。可信度 $r(i,k)$ 表示点 x_k 适合做点 x_i 类代表点的代表程度，可用度 $a(i,k)$ 则表示点 x_i 选择点 x_k 做类代表点的适合程度。迭代公式如下：

$$r(i,k)=s(i,k)-\max_{k'\neq k}\{a(i,k')+s(i,k')\} \tag{1}$$

$$a(i,k)=\begin{cases}\min\limits_{i'\neq i,k}\{0,r(k,k)+\sum\max\{0,r(i',k)\}, & i\neq k\\ \sum\limits_{i'\neq i,k}\max\{0,r(i',k)\}, & i=k\end{cases} \tag{2}$$

算法迭代过程中，当有两个点或者多个点同时适合为同一簇类的聚类代表点的时候，算法就有可能出现振荡，无法收敛。针对这种情况，算法在迭代步骤中引入了阻尼因子(Damping Factor)λ，使得每一次迭代的 $r(i,k)$ 和 $a(i,k)$ 的值受上一次迭代值的约束，提高了算法的稳定性，计算公式如下：

$$R(t)=(1-\lambda)R(t)+\lambda R(t-1) \tag{3}$$

$$A(t)=(1-\lambda)A(t)+\lambda A(t-1) \tag{4}$$

AP 算法通过迭代更新每一个点的可信度和可用度值，

直到达到规定的迭代次数或迭代过程收敛，根据 $r(i,k)+a(i,k)$ 选出类代表点集，同时将其余的数据点分配给最近的类代表点，形成类簇。

3 基于近邻传播聚类的集成特征选择算法

3.1 基本思想

集成特征选择方法包含 3 个步骤：(1)生成特征子集；(2)训练基分类器；(3)融合预测结果。其中，生成特征子集是集成特征选择方法的基础，其产生过程可以看成是对特征空间搜索的过程。如果数据集的特征个数为 M，那么特征子集选择就是从 2^M-1 个特征组合中选出符合要求的特征子集，当 M 值较大时，搜索过程就会变得十分复杂。事实上，在高维数据中样本分类通常只与少部分特征有关联。采用排序聚合技术，剔除与分类无关的特征，选出较稳定、分类区分能力强的特征作为生成特征子集的基础，不仅能够缩减特征组合的搜索空间，提高算法整体运行效率，而且能够避免或减弱无关特征对后续分类的影响。

排序聚合技术选出的特征与样本分类关联紧密，但无法保证特征之间是互不关联的。一个好的特征子集满足两个条件：(1)特征和样本分类紧密相关；(2)特征之间应当互不关联。一些特征选择方法，如 HykGene[14] 和 mAP_KL[15]，在特征排序的基础上增加了特征关联分析，选取互不相关的特征作为特征子集，在一定程度上消除了子集特征之间的冗余，取得了较好的分类准确度。EGSG 方法首先根据特征之间关联度进行特征分组，使不同组的特征之间互不关联，而后从每组前 t 个与样本分类关联紧密的特征中随机选择一个构建特征子集，在排除大量无关特征的同时保证了子集特征之间互不关联，从而取得了较高的分类准确度。由此可见，生成特征子集时进行特征之间关联分析是十分必要的。

进行特征关联分析时，可采用的衡量特征之间关联的标准有很多，本文选择 bicor 来衡量特征之间的关联。bicor 基于 Tukey's biweight，是 Hardin[16] 于 2007 年提出的一种稳定的、高效的、衡量两个对象之间关联的度量。Song[9] 通过在 5 个基因表达数据集上分别采用互信息、Pearson 系数和 bicor 衡量基因之间的关联做了比较，结果表明 bicor 对含噪声的数据具有更好的鲁棒性，能够完全取代信息度量取得更准确的基因 GO 功能关联分析结果。对于两个数值向量 $X=(x_1,x_2,\cdots,x_M)$ 和 $Y=(y_1,y_2,\cdots,y_M)$，bicor 计算公式如下：

$$bicor(X,Y)=\sum_{i=1}^{M}\tilde{x}_i\tilde{y}_i \tag{5}$$

其中：

$$\tilde{x}_i=\frac{(x_i-med(X))w_i^{(x)}}{\sqrt{\sum_{k=1}^{M}[(x_k-med(X))w_k^{(x)}]^2}} \tag{6}$$

$$\tilde{y}_i=\frac{(y_i-med(Y))w_i^{(y)}}{\sqrt{\sum_{k=1}^{M}[(y_k-med(Y))w_k^{(y)}]^2}} \tag{7}$$

$$w_i^{(x)}=(1-u_i^2)^2I(1-|u_i|) \tag{8}$$

$$u_i=\frac{x_i-med(X)}{9mad(X)} \tag{9}$$

$$I(1-|u_i|)=\begin{cases}1, & 1-|u_i|>0\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

$med(.)$ 表示取向量的中值，$mad(.)$ 表示取向量绝对中位差。

在 bicor 计算特征之间的关联之后，为强化强关联和弱化弱关联，因此特征之间的关联距离负值计算公式为：

$$s(i,j)=(\frac{bicor(f_i,f_j)+1}{2})^{\beta}-1 \tag{11}$$

β 值通常为一个常量，在本文实验中，当 $\beta=2$ 时，APBEFS 算法获得了较好的分类性能。

由式(11)可计算得到特征关联距离负矩阵 S，在矩阵 S 的基础上，用近邻传播聚类算法对特征分组得到 K 个分组(类簇)，从每个特征分组中随机选择一个特征生成一个大小为 K 的特征子集，重复 $nFit$ 次，从而得到 $nFit$ 个特征子集，而后分别在对应的特征子空间中，选择分类算法，训练基分类器。

基分类器生成之后，不同的融合方式会产生不同的集成学习结果。常用的融合方式有选择集成和融合集成两种。选择集成是选取使集成分类性能较优的基分类器子集对样本预测的结果按某一方式融合，如爬山法，这类方法通常能够获得较好的集成学习效果，但需要确定基分类器的评价标准和搜索融合方法，相对较复杂。融合集成是把所有的基分类器对样本的预测结果按某一方式融合，如多数投票方法，这类方法操作简单，有研究表明，在所有基分类器验证误差小于 0.5 时，基于多数投票方法的集成学习能够获得比单分类器更好、更稳定的分类性能。本文选择多数投票方法融合基分类器。

3.2 算法描述

基于近邻传播聚类的集成特征选择 APBEFS 算法，首先采用排序聚合技术对训练数据集的特征空间进行过滤，使得选出的特征具有较强的样本区分能力，然后以 bicor 相关系数为关联度量准则，利用近邻传播聚类算法对选出特征进行分组，使得相互关联的特征划分到同一个组内，接下来从每个分组中随机选择一个特征组合成特征子集，保证特征子集的差异性，重复选择多次，从而得到多个特征子集，在各自对应训练数据集的特征子空间中分别训练基分类器，并对投影到该特征子空间的待测样本进行预测，最后采用多数投票法融合并输出预测结果。

APBEFS 具体算法描述为：

输入：训练数据集 D，测试数据集 X，过滤后保留个数 Top-K，基分类器个数 nFit。

输出：测试数据集的样本类标记 $C=(c_1,c_2,\cdots,c_N)$。

方法：

(1)用排序聚合技术对训练数据集 D 的特征进行排序，选出与样本分类相关的、大小为 Top-K 的特征子集 SF；

(2)用式(11)计算集合 SF 中的特征之间的关联距离，得到特征关联距离负矩阵 S；

(3)基于矩阵 S，用近邻传播聚类算法对选出的 Top-K 个特征进行划分，得到特征簇集 $FG=(G_1,G_2,\cdots,G_K)$，其中，K=size(FG)；

(4)For j=1 to nFit

(5) 从每个分组 $G_i(i\in[1,K])$ 特征中随机选择一个特征，构成大小为 K 的特征子集 FS_j；

(6) 把训练数据映射到特征子集 FS_j 对应的特征子空间训练基分类器 cf_j；

(7) 把测试数据映射到特征子集 FS_j 对应的特征子空间，并用基分类器 cf_j 对测试样本类标记进行预测，得到预测结果 $pred_j$；

(8)End For

(9)N=length{sample:sample∈X}

(10)For i=1 to N

(11) 待测样本 i 的类标记 c_i 预测为 $Pred=\{pred_j:j\in[1,nFit]\}$ 中得票最多的类标记；

(12)End For

4 实验结果与分析

4.1 实验数据

本文选择来源于 Keng Ridge Bio-medical Dataset(http://datam.i2r.a-star.edu.sg/datasets/krbd/) 7 个常用的基因表达数据集，分别是：Breast Cancer、CNS、Conlon Cancer、Leukemia、Lung Cancer、Ovarian Cancer 和 Prostate Cancer。数据集的相关信息如表 1 所列。

表 1 数据集的相关信息

数据集	特征数	类别	样本数
Breast Cancer	24481	Non-relapse	51
		Relapse	46
CNS	7129	Survivors	21
		Failures	39
Conlon Tumor	2000	Positive	22
		Negative	40
Leukemia	7129	ALL	47
		AML	25
Lung Cancer	12600	Normal	17
		Non-normal	186
Ovarian Cancer	15154	Normal	91
		Non-normal	162
Prostate Cancer	12600	Normal	59
		Nor-normal	77

所有数据集都是衡量基因在不同条件下表达水平的数值型数据。实验前，数据预处理方法如下：采用 R 语言包 limma 提供的分位点标准化(Quantile Normalization)方法协同基因之间的分布，然后采用零-均值规范化，即使每个基因向量的平均值为 0，方差为 1。实现 EGSG 方法需对数据进行离散化，数值范围$[-\infty,-0.5]$,$(-0.5,+0.5)$,$[+0.5,+\infty]$分别对应离散化值 0,1,2。

4.2 实验结果和分析

为验证 APBEFS 的有效性和可比性，在选取的数据集上进行实验，并与 EGSG、RSM 以及 Random Forest 进行对比。采用 5 折交叉验证(5-Fold Cross Validation，5-CV)的方法在数据集上做分类测试，用准确度(Accuracy，ACC)来衡量分类性能。准确度是一种常用的分类性能度量标准，基于 4 种简单的指标：真阳性(True Positives，TP)、假阳性(False Positives，FP)、真阴性(True Negatives，TN)和假阴性(False Negatives，FN)。其计算公式为：

$$ACC=(TP+TN)/(TP+TN+FP+FN) \quad (12)$$

实验分为两个部分：第一部分研究基分类器数与集成精度之间的关系；第二部分比较了在固定的基分类器数下各方法的分类准确度。实验中，对 APBEFS 方法，在特征排序时采用了基于数据扰动的 eBayes 特征排序聚合方法(eBayes based method)对基因进行排序，选择与样本分类关联紧密的前 *Top-K* 个基因用于后续分析，其中数据扰动时采用欠抽样的方法，共进行 25 次抽样，采用平均聚合的方式对 25 次排序结果进行了融合。APBEFS、EGSG 和 RSM 方法以 K 近邻(K Nearest Neighbors，KNN)和径向支持向量机(Radial Basis Function Support Vector Machine，RBFSVM)为基分类器，其中 KNN 来自 Weka(R 语言包 RWeka)，RBFSVM 来自 Chang 和 Lin[17] 编写的 LibSVM(R 语言包 e1071)。实验在 Windows XP 系统，2.19GHz 酷睿处理器和 2G 内存环境下进行，APBEFS、EGSG 采用 R 语言实现；RSM 和 Random Forest 采用 Weka 提供的方法。

4.2.1 基分类器数与分类精度之间的关系

集成分类的准确度通常与基分类器数 *nFit* 有关，由不同的基分类器个数获得的分类准确度也不尽相同。为了验证基分类器个数对 APBEFS 方法是否有影响，对基于 eBayes 排序聚合技术的 APBEFS 方法在 7 个基因表达数据集上进行了实验，采用在所有数据集上重复进行 10 次分类的准确度的平均值来衡量分类性能。实验中发现，当 $Top\text{-}K=500$，APBEFS 分类性能最好，此时 KNN 分类器参数 K 设置为 3，RBFSVM 分类器采用默认参数，基分类器数与分类精度之间关系如图 1 所示。由图 1 可知，当基分类器数 *nFit* 达到某一数值时，APBEFS 方法的分类性能趋于稳定，而且不同类型的方法分类性能趋于稳定需要的基分类器数不相同，总之，当基分类器数 $nFit=45$ 时，两种类型的 APBEFS 方法的分类精度趋于一个稳定值。

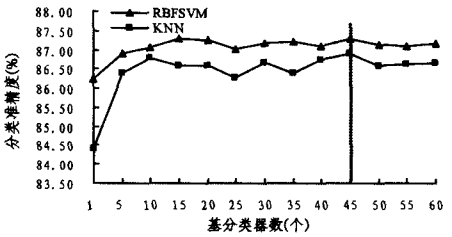

图 1 APBEFS 在不同数目基分类器下的分类准确度

4.2.2 固定基分类器个数下的分类准确度比较

表 2 列出了 EGSG、RSM、Random Forest 和 APBEFS 方法分别在 7 个基因表达数据上的 10 次运行的分类准确度的平均值。运行参数设置如下：所有分类器个数设置为 45，APBEFS 方法中 $Top\text{-}K=500$，EGSG 方法 $t=15$[7]，在实验中发现，当子空间大小设置为 100 时，RSM 方法分类性能较好。Random Forest 采用默认参数，基分类器 KNN 的 K 值设置为 3，RBFSVM 方法采用默认参数。

表 2 5 折交叉验证时的各方法分类准确率(%)

Datasets	KNN			RBFSVM			Random
	APBEFS	EGSG	RSM	APBEFS	EGSG	RSM	Forest
Breast Cancer	71.35	65.78	64.85	**70.41**	69.51	68.35	65.36
CNS	**65.29**	62.90	63.83	**65.29**	65.04	65.00	63.50
Conlon Tumor	85.22	82.42	81.13	86.00	83.32	**87.10**	83.55
Leukemia	98.32	97.63	92.64	**98.61**	97.21	84.44	93.33
Lung Cancer	98.38	**98.67**	98.23	98.33	98.43	97.09	97.00
Ovarian Cancer	98.05	99.09	94.17	99.29	**99.56**	94.17	95.71
Prostate Cancer	91.79	88.84	86.40	**93.24**	89.63	89.19	87.79
mean	86.92	85.05	83.03	**87.31**	86.10	83.62	83.75

由表 2 可知，APBEFS 和 EGSG 方法在 7 个数据集上的平均分类准确度明显优于 RSM 和 Random Forest 方法，APBEFS 方法优于 EGSG 方法约 1.5%。产生这种结果的原因主要有：(1)Random Forest 与 RSM 方法采用随机方法选择特征子空间，特征子空间内的特征不一定与样本分类关联紧密，而且特征之间有可能是相互关联的；而 EGSG 方法利用近似 Markov Blanket 对特征进行分组后从前 t 个特征中随机

(下转第 260 页)

046115

[10] MacQueen J. Some methods for classification and analysis of multivariate observations[C] //Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967: 281-297

[11] Park H S, Jun C H. A simple and fast algorithm for K-medoids clustering[J]. Expert Systems with Applications, 2009, 36(2): 3336-3341

[12] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976

[13] Ahn H J. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem[J]. Information Sciences, 2008, 178(1): 37-51

[14] Jeong B, Lee J, Cho H. Improving memory-based collaborative filtering via similarity updating and prediction modulation[J]. Information Sciences, 2010, 180(5): 602-612

[15] Bogdanova G, Georgieva T. Using error-correcting dependencies for collaborative filtering[J]. Data & Knowledge Engineering, 2008, 66(3): 402-413

[16] Resnick P, Iacovou N, Suchak M, et al. Grouplens: an open architecture for collaborative filtering of netnews[C] // Proceedings of the 1994 ACM on Computer Supported Cooperative Work. 1994: 175-186

[17] Sarwar B, Karypis G, Konstan J, et al. Recommender systems for large-scale e-commerce: scalable neighborhood formation using clustering[C] //Proceedings of the 5th International Conference on Computer and Information Technology. 2002

[18] Herlocker J, Konstan J A, Borchers A, et al. An algorithmic framework for performing collaborative filtering[C] // Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieve. 1999: 230-237

(上接第 244 页)

选择特征构建特征子空间，APBEFS 方法先采用排序聚合技术过滤剔除样本分类不相关特征，后采用基于 bicor 的特征关联距离度量的近邻传播聚类算法对特征进行分组后，从不同组随机选择一个特征构建特征子空间，二者提高了子空间特征的质量，降低特征之间关联性，从而获得了较好的分类性能；(2)特征排序聚合方法，尤其是基于数据扰动的方法，能够获得更为稳定的、准确的特征排序，而且选择前 500 个与样本分类关联紧密的特征用于后续分析是合适的。

结束语 本文提出了一种基于特征排序聚合和聚类的集成特征选择方法。该方法能够与一般的特征过滤方法相结合，具有更广阔的运用前景，同时采用排序聚合技术，降低了不同特征选择方法或数据扰动带来的影响，以 bicor 相关系数为关联准则利用近邻传播聚类对特征进行分组，并在此基础上采用随机方法构建特征子空间，从而获得了更优的、存在差异性的基分类器，最后采用多数投票法进行基分类器融合，获得了比单分类器更好更稳定的分类性能。对 7 个常用的基因表达数据的实验结果表明，本文提出的方法能够获得准确度较高的分类结果，同时分类性能稳定，具有较好的可扩展性。

参 考 文 献

[1] Opitz D W. Feature selection for ensembles[C]//Proceedings of 16th National Conference on Artificial Intelligence (AAAI-99). Orlando, FL, USA, 1999: 379-384

[2] Ho T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844

[3] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32

[4] De Bock K W, Coussement K, Van den Poel D. Ensemble classification based on generalized additive models[J]. Computational Statistics & Data Analysis, 2010, 54(6): 1535-1546

[5] 姚旭，王晓丹，张玉玺，等. 基于正则化互信息和差异度的集成特征选择[J]. 计算机科学，2013，40(6)：225-228

[6] Moon H, Ahn H, Kodell RL, et al. Ensemble methods for classification of patients for personalized medicine with high-dimensional data[J]. Artificial Intelligence in Medicine, 2007, 41(3): 197-207

[7] Liu Hua-wen, Liu Lei, Zhang Hui-jie. Ensemble gene selection by grouping for microarray data classification[J]. Journal of Biomedical Informatics, 2010, 43(1): 81-87

[8] Wald R, Khoshgoftaar T M, Dittman D. Mean aggregation versus robust rank aggregation for ensemble gene selection[C]// 2012 11th International Conference on Machine Learning and Applications (ICMLA). Boca Raton, FL, USA, 2012: 63-69

[9] Lin Song, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices[J]. BMC Bioinformatics, 2012, 13(1): 328

[10] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976

[11] Boulesteix A L, Slawski M. Stability and aggregation of ranked gene lists[J]. Briefings in Bioinformatics, 2009, 10(5): 556-568

[12] Wald R, Khoshgoftaar T M, Dittman D, et al. An extensive comparison of feature ranking aggregation techniques in bioinformatics[C]//2012 IEEE 13th International Conference on Information Reuse and Integration (IRI). Las Vegas, NV, USA, 2012: 377-384

[13] Wang Chang-dong, Lai Jian-huang, Suen C, et al. Multi-Exemplar Affinity Propagation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(9): 2223-2237

[14] Wang Yu-hang, 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

[15] Sakellariou A, Sanoudou D, Spyrou G. Combining multiple hypothesis testing and affinity propagation clustering leads to accurate, robust and sample size independent classification on gene expression data[J]. BMC Bioinforma-tics, 2012, 13(1): 270

[16] Hardin J, Mitani A, Hicks L, et al. A robust measure of correlation between two genes on a microarray[J]. BMC Bioinformatics, 2007, 8(1): 220

[17] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27