ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.11, November 2010, pp.2838−2851
doi: 10.3724/SP.J.1001.2010.03699

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

癌症识别中一种基于组合 GCM 和 CCM 的分类算法*

卢新国[1+], 林亚平[1,2], 骆嘉伟[1], 李 丹[1]

[1](湖南大学 计算机与通信学院,湖南 长沙 410082)

[2](湖南大学 软件学院,湖南 长沙 410082)

Classification Algorithm Combined GCM with CCM in Cancer Recognition

LU Xin-Guo[1+], LIN Ya-Ping[1,2], LUO Jia-Wei[1], LI Dan[1]

[1](School of Computer and Communication, Hu'nan University, Changsha 410082, China)

[2](School of Software, Hu'nan University, Changsha 410082, China)

+ Corresponding author: E-mail: hnluxinguo@126.com

Lu XG, Lin YP, Luo JW, Li D. Classification algorithm combined GCM with CCM in cancer recognition. *Journal of Software*, 2010,21(11):2838−2851. http://www.jos.org.cn/1000-9825/3699.htm

Abstract: In this paper, two cancer recognition models, global component model (GCM) and cancer component model (CCM), are constructed. Due to the fact that GCM and CCM complement each other, a weighted voting strategy is applied, and an ensemble algorithm based on GCM and CCM for cancer recognition (EAGC) is proposed. Independent test experiments and cross validation experiments are conducted on Leukemia, Breast, Prostate, DLBCL, Colon, and Ovarian cancer dataset, respectively, and EAGC performed well on all datasets. The experimental results show that recognition, solution, and the generalization are strengthened by the combination of GCM and CCM.

Key words: gene expression profile; cancer recognition; global component model; cancer component model

摘 要: 根据基因表达谱数据的特点,提出了全局分量模型(global component model,简称 GCM)和癌症组分量模型(cancer component model,简称 CCM)两种癌症识别模型.结合 GCM 模型和 CCM 模型的互补性,利用基于权值的投票组合策略提出一种基于组合 GCM 和 CCM 的癌症分类算法(ensemble algorithm based on GCM and CCM for cancer recognition,简称 EAGC).在 Leukemia,Breast,Prostate,DLBCL,Colon,Ovarian 这 6 个数据集上进行了独立测试实验和交叉测试实验.实验结果表明,EAGC 有效地综合了 GCM 和 CCM 识别模型的解决方案,弥补了单个分类器的不足,具有较好的泛化性,在所有数据集上都取得较好的分类性能.

关键词: 基因表达谱;癌症识别;全局分量模型;癌症组分量模型

中图法分类号: TP181 **文献标识码**: A

近年来,基因微阵列(microarray)技术的发展使得研究人员可以在同一实验中获得成千上万个基因的表达

* Supported by the National Natural Science Foundation of China under Grant No.60873184 (国家自然科学基金); the National Science Foundation for Post-Doctoral Scientists of China under Grant No.20100471790 (国家博士后科学基金); the Hu'nan Provincial Natural Science Foundation of China under Grant No.07JJ5085 (湖南省自然科学基金)

Received 2008-10-14; Revised 2009-04-10; Accepted 2009-07-07

水平(expression level),从而产生了大规模基因表达谱数据(gene expression profile),对癌症诊断及治疗的研究具有非常重要的意义.在癌症的诊断和治疗过程中,对癌症的精确分类是提高诊断准确率和癌症治愈率至关重要的一个环节[1,2].但是,利用微阵列基因表达谱数据进行癌症分类是典型的高维、高噪和高冗余问题[3,4].

特征选择法是一种主要的基因表达谱数据预处理方法,如信噪比(signal to noise ratio,简称 SNR)、排秩法(rank)、信息增益(information gain)等[4−7].文献[8]分析和比较不同特征选择法在癌症分类中的特征基因选取情况发现,在相同数据集中,不同方法挑选出的特征基因明显不同,导致经过不同特征选择方法预处理之后的癌症识别效果也不相同.其主要原因是,不同的特征选择法基于不同的搜索机制和评价策略,挑选出来的特征基因偏向于致癌病理的一个方面或多个方面中的一部分,而不是全面地反映癌症病理因素.对于一种癌症识别分类器,如果选取合适的特征子集则会获得较好的分类结果,反之则分类结果不理想.这样就导致分类结果不稳定,缺乏泛化性.一种有效的解决方法是进行分类器组合[9].文献[8]采用多数投票法(majority voting)组合 4 种不同的分类器进行癌症识别.文献[10]提出一种基于装袋(bagging)的组合决策树的癌症分类算法.目前已有的方法都是首先采用不同的特征选择方法选择不同的基因子集,然后利用这些基因子集来训练分类器以进行分类器组合.这些方法的共同不足是:不同子集之间存在较多的重叠特征,导致分类器训练时输入较多的冗余信息;同时,没有充分考虑特征基因子集选取时的互补性以及分类器之间的差异性.具有互补性的分类器组合可以弥补单个分类器的缺点,同时也保持它的优点,有利于优化分类器的组合结果.

神经网络是一种有效的模式识别模型[11,12],但是由定量数据建立的单一神经网络模型往往缺乏泛化能力.结合组合分类算法的优点,本文提出了一种基于组合神经网络的癌症分类算法.首先,利用主分量分析法(principal component analysis,简称 PCA)选取基因特征空间中大于分量累积贡献率阈值的 r 个主要分量,利用这些主要分量训练识别癌症的神经网络模型以构造全局分量模型(global component model,简称 GCM);然后,针对每一种癌症类型抽取癌症组分量,利用癌症组分量训练识别癌症的神经网络模型以构造癌症组分量模型(cancer component model,简称 CCM);最后,利用基于权值的投票组合策略提出一种基于组合 GCM 和 CCM 的癌症分类算法(ensemble algorithm based on GCM and CCM for cancer recognition,简称 EAGC),并在 Leukemia,Breast,Prostate,DLBCL,Colon,Ovarian 这 6 个数据集上分别进行独立测试实验和交叉测试实验.由于在基因特征的抽取和癌症识别模型的构造上 GCM 和 CCM 都具有很强的互补性,EAGC 综合了 GCM 和 CCM 识别模型的解决方案,有效扩展了算法的解决方案,以弥补单个分类器的不足,提高整个系统的泛化能力.

本文第 1 节介绍背景知识.第 2 节构建全局分量模型(GCM)和癌症组分量模型(CCM),提出一种基于组合 GCM 和 CCM 的癌症分类算法(EAGC),并进行简要的算法分析.第 3 节在不同的数据集上进行独立测试实验和交叉测试实验,并分析实验结果.第 4 节给出结论.

1 背景知识

1.1 基因微阵列表达谱

基因表达谱是指利用 DNA 微阵列测定组织样本中基因的表达水平值,通常利用矩阵形式表示.假设 X 为一 $m\times n$(通常 $m>>n$)的基因表达矩阵,矩阵 X 的第 i 行是第 i 个基因在所有观测样本中的表达值,第 j 列是第 j 个样本中所有观测基因的表达值.矩阵 X 的元素 x_{ij} 表示第 i 个基因在第 j 个观测样本中的表达水平,也可以表示第 j 个样本下第 i 个观测基因表达水平.

1.2 BP网络

BP 网络是一种应用非常广泛的神经网络模型,在模式识别、智能控制和信号处理等领域都有大量的应用.BP 网络实际上就是多层感知器(multi-layer perceptron,简称 MLP).BP 网络是由输入层、输出层和若干隐层互相连接构成.BP 网络结构为:前后相邻层的任意两节点均连接,同层和非相邻层的节点均无任何耦合,从输入层开始逐层连接,到输出层连接结束.

2 基于组合 GCM 和 CCM 的癌症识别(EAGC)

2.1 神经网络模型

根据基因表达谱数据的特点,本节将构建两种神经网络的癌症识别模型,并依据抽取的输入变量称之为全局分量模型(GCM)和癌症组分量模型(CCM).

2.1.1 全局分量模型(GCM)

对于基因表达矩阵 $X_{m\times n}$,不妨设 $X^T=(\bar{g}_1,\bar{g}_2,...,\bar{g}_m)$,利用主分量分析法(PCA)抽取基因表达谱中的 $r(r\leqslant m)$ 个隐含变量 $\bar{h}_i(1\leqslant i\leqslant r)$,用如下公式表示:

$$\begin{cases}\bar{h}_i^T=a_1\bar{g}_1+a_2\bar{g}_2+...+a_m\bar{g}_m=\bar{a}_i^TX^T\\ Var(\bar{h}_i)=\bar{a}_i^T\Sigma\bar{a}_i=\lambda_i\end{cases}\tag{1}$$

其中,$\bar{a}_i^T=(\bar{a}_1,\bar{a}_2,...,\bar{a}_m)$,$\Sigma=(\upsilon_{ij})_{m\times m}=(Cov(\bar{g}_i,\bar{g}_j))$,$Var$ 表示方差,Cov 表示协方差,$\bar{h}_i$ 表示第 i 个主分量即 PC_i,λ_i 是 Σ 的第 i 个特征值,$\bar{a}_i$ 是 λ_i 对应的特征向量,表示观察基因变量在 $\bar{h}_i$ 上的载荷.

定义 1(分量贡献系数). 在基因表达数据中,定义 $Var(\bar{s}_i)$ 为隐含分量的分量贡献系数.

定义 2(累积贡献率). 不妨设 $\lambda_1\geqslant\lambda_2\geqslant\ldots\geqslant\lambda_m$,给定 $r(1\leqslant r\leqslant m)$,分量 $\bar{h}_1,\bar{h}_2,...,\bar{h}_r$ 的累积贡献率为

$$CR=\frac{\sum_{i=1}^{r}Var(\bar{s}_i)}{\sum_{i=1}^{m}Var(\bar{s}_i)}.$$

定义 3(全局分量空间). 对于 $r\leqslant m$,设 $\min(\lambda_1,\lambda_2,\ldots,\lambda_r)\geqslant\max(\lambda_{r+1},\lambda_{r+2},\ldots,\lambda_m)$,则由 $\bar{a}_1,\bar{a}_2,...,\bar{a}_d$ 组成了基因表达数据的 r 维全局分量空间 ε_g, $\varepsilon_g=\{\bar{a}_1,\bar{a}_2,...,\bar{a}_d\}$.

假设基因表达谱中分为 k 个癌症类别,我们抽取 $CR\geqslant$阈值的 r 个主分量并利用 BP 网络构建识别癌症类型的全局分量模型(GCM),如图 1 所示.

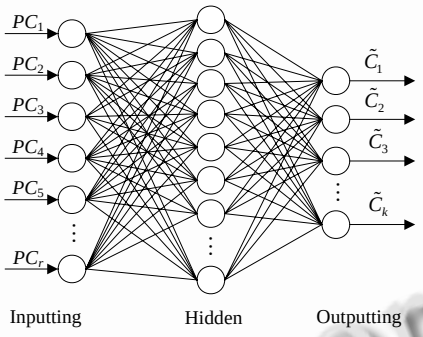


Fig.1 GCM for cancer recognition

图 1 癌症识别中的全局分量模型

GCM 包含输入层(I 层)、隐层(H 层)和输出层(O 层)等 3 层,I 层有 r 个神经元节点,O 层有 k 个神经元节点,设 H 层有 q 个神经元节点.全局分量模型用下面的数学公式描述:

$$\begin{cases}net_j=\begin{cases}\sum_i w_{ij}I_i-\theta_j, & \text{if 节点 } j \text{ in H层}\\ \sum_i w_{ij}H_i-\theta_j, & \text{if 节点 } j \text{ in O层}\end{cases}\\ out_j=f(net_j)\end{cases}\tag{2}$$

其中,out_j 是神经元节点 j 的输出,w_{ij} 是节点 i 到 j 的权值,I_i 是 I 层节点 i 的输入,H_i 是 H 层节点 i 的输出,θ_j 是节点 j 的激活阈值,节点的特性函数是 S 型函数 $f(x)=\frac{1}{1+e^{-x}}$. 权值和激活阈值的调节如公式(3)和公式(4)所示:

$$\begin{cases} w_{ij}(t+1)=w_{ij}(t)+\Delta w_{ij}=w_{ij}(t)-\eta\delta_j out_i \\ \delta_j=\begin{cases}-(\hat{O}_j-O_j)f'(net_k), & \text{if 节点 } j \text{ in O层} \\ f'(net_k)\sum_k \delta_k w_{jk}, & \text{if 节点 } j \text{ in H层}\end{cases}\end{cases} \tag{3}$$

$$\theta_j(t+1)=\theta_j(t)+\eta\delta_j \tag{4}$$

其中:w_{ij},out_i,θ_j 和 net_k 与上式相同;η是增益因子,$0<\eta\leqslant 1$;O_j 是 O 层节点 j 的输出,$\hat{O}_j$ 是对应的期望输出.对于癌症样本 $\bar{s}$,如果 $\bar{s}\in C_{i'}$,那么,

$$\hat{O}_j=\begin{cases}1, & \text{if } j=i' \\ 0, & \text{else}\end{cases}.$$

2.1.2 癌症组分量模型(CCM)

假设基因微阵列表达谱中第 i 类癌症样本集合是 $\tilde{C}_i(1\leqslant i\leqslant k)$,$\tilde{C}_i$ 的样本数目为 n_i,C_i 是 $\tilde{C}_i$ 的 $m\times n_i$ 基因表达矩阵.对于每一个$\tilde{C}_i(1\leqslant i\leqslant k)$,我们首先获取 $\tilde{C}_i$ 的最小扩展空间$\hat{\varepsilon}$,并将癌症样本在 $\hat{\varepsilon}$ 上映射抽取有价值的基因特征,称为癌症组分量(cancer component,简称 CC),然后利用 BP 网络构建识别癌症类型的癌症组分量模型(CCM).获取 $\tilde{C}_i$ 的最小扩展空间 $\hat{\varepsilon}$ 以及样本的 CC 分量如下所示:

设 $C_i^T=(\bar{g}_1,\bar{g}_2,...,\bar{g}_m)$,其中,$\bar{g}_j$ 是第 j 个基因在 $\tilde{C}_i$ 样本中的基因表达向量.C_i^T 的协方差矩阵为 $Cov(C_i^T)$,$Cov(C_i^T)$ 是半正定的 m 维方阵,可以进行如下矩阵分解:

$$Cov(C_i^T)=\sum\lambda_r\bar{p}_r\bar{p}_r^T=P\wedge P^T \tag{5}$$

其中:λ_r 是 $Cov(C_i^T)$ 特征值;$\wedge$是非负的对角矩阵,对角线上元素由$\lambda_r(1\leqslant r\leqslant m)$组成;$\bar{p}_r$ 是λ_r 对应的特征向量:

$$P=(\bar{p}_1,\bar{p}_2,...,\bar{p}_m)$$

定义 4(关联空间). 设癌症 $\tilde{C}_i$ 和表达矩阵 C_i,$\lambda_1,\lambda_2,\ldots,\lambda_m$ 是 $Cov(C_i^T)$ 的特征值,$\bar{p}_1,\bar{p}_2,...,\bar{p}_m$ 是$\lambda_1,\lambda_2,\ldots,\lambda_m$对应的特征向量.对于$d\leqslant m$,则由 $\bar{p}_1,\bar{p}_2,...,\bar{p}_d$ 组成了 $\tilde{C}_i$ 上秩为 d 的关联空间ε,$\varepsilon=\{\bar{p}_1,\bar{p}_2,...,\bar{p}_d\}$,$\bar{p}_i$ 为ε的第 i 维方向.λ_i 称为方向 $\bar{p}_i$ 的方向扩展系数,$P=(\bar{p}_1,\bar{p}_2,...,\bar{p}_d)$ 为 $\tilde{C}_i$ 的关联空间矩阵.

定义 5(最小扩展空间). 对于癌症 $\tilde{C}_i$ 和表达矩阵 C_i,假设 $\hat{\varepsilon}$ 是 $\tilde{C}_i$ 的 d 维关联空间,$\lambda_1,\lambda_2,\ldots,\lambda_d$ 是 $\hat{\varepsilon}$ 的方向扩展系数,当$\lambda_1,\lambda_2,\ldots,\lambda_d$ 满足 $\max(\lambda_1,\lambda_2,\ldots,\lambda_d)\leqslant\min(\lambda_d,\lambda_{d+1},\ldots,\lambda_m)$时,则称 $\hat{\varepsilon}$ 为 d 维最小扩展空间.

定义 6(癌症组分量). 对于癌症 $\tilde{C}_i$ 和 d 维最小扩展空间 $\hat{\varepsilon}$,设$\tilde{C}_i=\{\bar{s}_1,\bar{s}_2,...,\bar{s}_{n_i}\}$,$\bar{s}_l=(s_{l1},s_{l2},...,s_{lm})^T$,那么 $\bar{s}_l$ 在 $\bar{p}_j$ 上的癌症组分量为$CC_j=\bar{s}_l\cdot\bar{p}_j$.其中,$\bar{p}_j$ 为 $\hat{\varepsilon}$ 的第 j 维方向,$\bar{p}_j=(p_{j1},p_{j2},...,p_{jm})^T$,$\bar{s}_l\cdot\bar{p}_j=\sum_{k=1}^{m}s_{lk}p_{jk}$.

我们通过构造 $\tilde{C}_i$ 的最小扩展空间 $\hat{\varepsilon}$ 抽取样本的 d 个癌症组分量,然后利用 BP 网络构建识别癌症 $\tilde{C}_i$ 的癌症组分量模型(CCM),如图 2 所示.

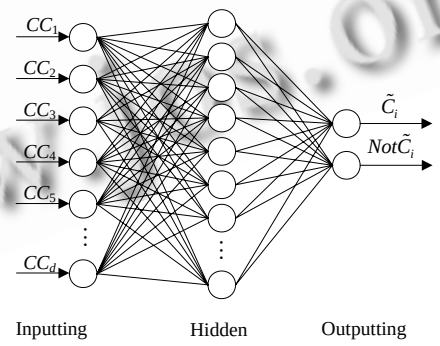


Fig.2 CCM for cancer recognition

图 2 癌症识别中的癌症组分量模型

CCM 包含输入层(I 层)、隐层(H 层)和输出层(O 层)等 3 层,I 层有 d 个神经元节点,O 层有 2 个神经元节点,设 H 层有 q'个神经元节点.CCM 模型中,神经元节点的输入、输出、转移函数、权值和激阈值调节同于 GCM 模型.对于癌症样本 $\bar{s}$ 的期望输出 $\hat{O}$ 通过下式给出:

$$\hat{O}=\begin{cases}[1\ 0]^T, & \text{if } \bar{s}\in\tilde{C}_i\\ [0\ 1]^T, & \text{else}\end{cases}.$$

2.2 基于组合GCM和CCM的癌症识别算法(EAGC)

GCM 模型利用 PCA 提取样本的主分量 $PC_j(1\leqslant j\leqslant r)$作为输入变量,经过隐层神经元和权值的作用,在输出层判别输入样本的癌症类别 $\tilde{C}_i(1\leqslant i\leqslant k)$.CCM 模型则利用癌症组内基因变量的相关性提取样本的癌症组分量 $CC_j(1\leqslant j\leqslant d)$,并输入 CCM 模型.在隐层神经元和权值的调节下,在输出层判别输入样本是否属于某种癌症类别(C_i 或 Not $\tilde{C}_i(1\leqslant i\leqslant k)$).GCM 模型和 CCM 模型在基因特征抽取和癌症识别模型的构造上具有很强的互补性.

本节提出一种基于组合 GCM 和 CCM 模型的癌症识别算法.首先,在训练阶段利用基因数据中的训练子集建立 GCM 模型和 $\tilde{C}_i$ 的 CCM 模型,然后在测试阶段分别利用 GCM 模型和 CCM 模型识别测试样本,并利用基于权值的投票组合策略以识别样本的癌症类型.

对于测试样本 $\bar{s}$,不妨设癌症 $\tilde{C}_i$ 的 CCM 模型的识别结果为 $R(\tilde{C}_i)^T=(r_{i1},r_{i2})$,GCM 模型的识别结果为 $R(\tilde{C})^T=(r_1,r_2,...,r_k)$,基于权值的投票组合策略描述如下:

$$\begin{cases}R(ensemble)^T=(r_1',r_2',...,r_k')\\ r_i'=\alpha r_{i1}+\beta r_i\\ \alpha+\beta=1\\ result=\tilde{C}_{i'},\ \text{if } r_{i'}'=\max(R(ensemble))\end{cases}\tag{6}$$

其中,$R(ensemble)$为 $R(\tilde{C}_i)$ 和 $R(\tilde{C})$ 的组合结果;α,β分别为 CCM 和 GCM 模型的权值;$result$ 为测试样本 $\bar{s}$ 的癌症类别.

EAGC 有效综合 GCM 和 CCM 模型的癌症识别结果,消除基因数据中内在的噪声和冗余对单个分类器的影响,优化分类器的癌症识别结果,提高 EAGC 的泛化能力.基于组合 GCM 和 CCM 络模型的癌症识别算法具体描述如下:

基于组合 GCM 和 CCM 模型的癌症识别算法(EAGC).

Inputing:训练集(training set)、测试集(test set).其中,训练集中有 k 种不同类型的癌症,第 i 类癌症样本集合是 $\tilde{C}_i,\tilde{C}_i$ 的表达矩阵 $C_i,\tilde{C}=\bigcup\tilde{C}_i$, q=10,$CR\geqslant$85%,d=15,η=0.5,α,β.

Begin

对 $\tilde{C}$ 的表达矩阵 X 进行 PCA 分解,获取全局分量空间 $\varepsilon_g=\{\bar{a}_1,\bar{a}_2,...,\bar{a}_r\}$;

给 GCM 模型的 w_{ij} 和 θ_j 赋随机初值;

训练 GCM 模型;

For i=1 to k

获取癌症 $\tilde{C}_i$ 表达谱的协方差矩阵 $Cov(C_i^T)$;

获取 $Cov(C_i^T)$ 的特征值 $\lambda_1,\lambda_2,\ldots,\lambda_m$,特征向量 $\bar{p}_1,\bar{p}_2,...,\bar{p}_m$;

选取 d 个最小的 $\lambda_1',\lambda_2',...,\lambda_d'$ 对应的 $\bar{p}_1',\bar{p}_2',...,\bar{p}_m'$ 构成 $\tilde{C}_i$ 的最小扩展空间 $\hat{\varepsilon}_i$;

给 CCM_{C_i} 模型的 w_{ij} 和 θ_j 赋随机初值;

训练 CCM_{C_i} 模型;

Next

For each $\bar{t}$ in Test Set

获取 $\bar{t}$ 的主分量 $PC_j=\bar{t}\cdot\bar{a}_j$, 并输入 GCM,识别结果为 $R(\tilde{C})=(r_1,r_2,...,r_k)^T$;

For i=1 to k

获取 $\bar{t}$ 在 $\tilde{C}_i$ 中的癌症组分量 $CC_j=\bar{t}\cdot\bar{p}_j$,并输入 CCM_{C_i},识别结果为

$R(\tilde{C}_i)=(r_{i1},r_{i2})^T$

End

计算组合策略结果 $R(ensemble)=(r_1',r_2',...,r_k')^T$, 其中, $r_i'=\alpha r_{i1}+\beta r_i$;

识别 $\bar{t}\in\tilde{C}_{i'}$ if $r_{i'}'=\max(R(ensemble))$;

Next

End

2.3 讨论与分析

为了方便描述,设样本 $\bar{s}$ 的主分量为 $PC_j(\bar{s}_l)(1\leqslant j\leqslant r)$, 全局分量为 $CC_j(\bar{s}_l)(1\leqslant j\leqslant d)$.$M(\tilde{C}')$ 是一个虚拟样本, $M(\tilde{C}')=\frac{1}{|\tilde{C}'|}\sum_{\bar{s}_l\in\tilde{C}'}\bar{s}_l$. 其中, $\tilde{C}'\in\{\tilde{C},\tilde{C}_i\}$,$|\tilde{C}'|$ 是 $\tilde{C}'$ 的样本数量, $M(\tilde{C}')$ 表示 $\tilde{C}'$ 的中心.

定义 7(样本能量). 对于癌症数据集 $\tilde{C}$,$E(\tilde{C},\varepsilon_g)$ 表示 $\tilde{C}$ 中样本在全局分量空间 ε_g 的样本能量,

$$E(\tilde{C},\varepsilon_g)=\frac{1}{n}\sum_{\bar{s}_l\in\tilde{C}}\left\{\sum_{j=1}^{r}(PC_j(\bar{s}_l)-PC_j(M(\tilde{C})))^2)\right\}^2,$$

其中,n 是 $\tilde{C}$ 的样本数目.

定理 1. 对于癌症数据集 $\tilde{C}$,假设 ε_g 是 $\tilde{C}$ 的全局分量空间,$\varepsilon_g=\{\bar{a}_1,\bar{a}_2,...,\bar{a}_r\}$,$\lambda_j(1\leqslant j\leqslant r)$是 $\bar{a}_i$ 对应的特征值,则

$$E(\tilde{C},\varepsilon_g)=\sum_{j=1}^{r}\lambda_j.$$

证明:不妨设 $\tilde{C}=\{\bar{s}_1,\bar{s}_2,...,\bar{s}_n\}$,由 PCA 分析可知,

$$PC_j(\bar{s}_l)=\bar{s}_l\cdot\bar{a}_j,$$

$$E(\tilde{C},\varepsilon_g)=\frac{1}{n}\sum_{l=1}^{n}\sum_{j=1}^{r}(\bar{s}_l\cdot\bar{a}_j-M(\tilde{C})\cdot\bar{a}_j)^2=\frac{1}{n}\sum_{j=1}^{d}\sum_{l=1}^{n}(\bar{s}_l\cdot\bar{a}_j-M(\tilde{C})\cdot\bar{a}_j)^2.$$

由公式(1)可知, $\sum_{l=1}^{n}(\bar{s}_l\cdot\bar{a}_j-M(\tilde{C})\cdot\bar{a}_j)^2=n\lambda_j$. 因此, $E(\tilde{C},\varepsilon_g)=\sum_{j=1}^{r}\lambda_j$. □

从样本能量的定义可知, $E(\tilde{C},\varepsilon_g)$ 表示癌症集 $\tilde{C}$ 中样本和中心 $M(\tilde{C})$ 在 ε_g 上的距离,是 $\tilde{C}$ 中的癌症样本在 ε_g 上相似性的度量.$E(\tilde{C},\varepsilon_g)$ 反映了 $\tilde{C}$ 的子类 $\tilde{C}_i$ 中样本的类间相异性.样本能量越大,$\tilde{C}$ 中样本具有越大的活跃性,则样本之间的相似程度越小,子类 $\tilde{C}_i$ 的类间相异性越大.由于 $\min(\lambda_1,\lambda_2,\ldots,\lambda_r)\geqslant\max(\lambda_{r+1},\lambda_{r+2},\ldots,\lambda_m)$,因此,样本在 r 维的 ε_g 上具有最大的样本能量,则 $\tilde{C}$ 中样本具有最大的活跃性.

定义 8(组能量). 对于癌症 $\tilde{C}_i$,$E(\tilde{C}_i,\varepsilon_i)$ 表示 $\tilde{C}_i$ 在关联空间 ε_i 的组能量:

$$E(\tilde{C}_i,\varepsilon_i)=\frac{1}{n_i}\sum_{\bar{s}_l\in\tilde{C}_i}\left\{\sum_{j=1}^{d}(CC_j(\bar{s}_l)-CC_j(M(\tilde{C})))^2)\right\}^2,$$

其中,n_i 是 $\tilde{C}_i$ 的样本数目.

定理 2. 对于癌症样本集合 $\tilde{C}_i$ 和表达矩阵 C_i,假设 ε_i 是 $\tilde{C}_i$ 的关联空间,ε_i 的秩为 d,$\lambda_j(1\leqslant j\leqslant d)$是 ε_i 的方向扩展系数,则 $E(\tilde{C}_i,\varepsilon_i)=\sum_{j=1}^{d}\lambda_j$.

证明:对于癌症样本集合 $\tilde{C}_i$,设 $\tilde{C}_i=\{\bar{s}_1,\bar{s}_2,...,\bar{s}_{n_i}\}$,因为 $CC_j(\bar{s}_l)=\bar{s}_l\cdot\bar{p}_j$,所以,

$$E(\tilde{C}_i,\varepsilon_i)=\frac{1}{n_i}\sum_{l=1}^{n_i}\sum_{j=1}^{d}(\bar{s}_l\cdot\bar{p}_j-M(\tilde{C}_i)\cdot\bar{p}_j)^2=\frac{1}{n_i}\sum_{j=1}^{d}\sum_{l=1}^{n_i}(\bar{s}_l\cdot\bar{p}_j-M(\tilde{C}_i)\cdot\bar{p}_j)^2.$$

又设 $p_j=\overline{s}_l\cdot\overline{p}_j=\overline{p}_j^T\overline{s}_l$, 由于 $Var(p_j)=\overline{p}_j^T Cov(C_i^T)\overline{p}_j=\overline{p}_j^T\lambda_j\overline{p}_j=\lambda_j$, 所以 $\sum_{l=1}^{n_i}(\overline{s}_l\cdot\overline{p}_j-M(\tilde{C}_i)\cdot\overline{p}_j)^2=n_i\lambda_j$, 因此 $E(\tilde{C}_i,\varepsilon_i)=\sum_{j=1}^{d}\lambda_j$. 证毕. □

同理可知,组能量 $E(\tilde{C}_i,\varepsilon_i)$ 反映的是 $\tilde{C}_i$ 中的样本和 $\tilde{C}_i$ 的中心在 ε_i 上的距离,是 $\tilde{C}_i$ 中的癌症样本在 ε_i 上相似性的度量.组能量越小,组内样本的相似程度越大;反之则组内样本的相异程度越大.从最小扩展空间的定义可知,$\max(\lambda_1,\lambda_2,\ldots,\lambda_d)\leqslant\min(\lambda_d,\lambda_{d+1},\ldots,\lambda_m)$.因此,在 d 维最小扩展空间 $\hat{\varepsilon}_i$ 上,癌症 $\tilde{C}_i$ 中的样本具有最大的相似性.

从上述分析可知,GCM 模型利用主分量 PC_j 作为基因特征来训练分类器,从癌症数据集的整体性来分析癌症数据以识别各类癌症 $\tilde{C}_i$.CCM 模型则利用癌症组的癌症分量 CC_j 作为基因特征来训练分类器,从癌症组的局部性来分析癌症数据以识别癌症组 $\tilde{C}_i$.这两类癌症识别模型从高维的癌症数据的不同特性来分析癌症数据,基于组合 GCM 和 CCM 的解决方案,利用 GCM 模型和 CCM 模型的互补性,融合它们的识别结果以达到最优的识别结果.

定义 9(样本组能量). 对于数据集 $\tilde{C},\tilde{C}=\bigcup_i\tilde{C}_i,E(\tilde{C}_i,\varepsilon_i)$ 为 $\tilde{C}_i$ 在 ε_i 上的组能量, $E(\tilde{C},\varepsilon)$ 是样本组能量:

$$E(\tilde{C},\varepsilon)=\frac{1}{k}\sum_{i=1}^{k}E(\tilde{C}_i,\varepsilon_i)=\frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{l=1}^{n_i}\sum_{j=1}^{d}(\overline{s}_l\cdot\overline{p}_j-M(\tilde{C}_i)\cdot\overline{p}_j)^2=\frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{d}\sum_{l=1}^{n_i}(\overline{s}_l\cdot\overline{p}_j-M(\tilde{C}_i)\cdot\overline{p}_j)^2,$$

其中, $\varepsilon=\bigcup_i\varepsilon_i$.

由定义 9 可知,样本组能量 $E(\tilde{C},\varepsilon)$ 体现 $\tilde{C}$ 中的样本到各自所属类别 $\tilde{C}_i$ 的中心在 ε_i 的距离,体现 $\tilde{C}$ 中癌症样本在 ε 上相似度.设 CCM 模型的和 GCM 模型的权值分别为 α 和 β,不妨设 $\alpha+\beta=1$.由定义 8 可知, $E(\tilde{C},\varepsilon_g)$ 反映了 $\tilde{C}$ 中样本在 ε_g 的相异度,则令

$$\frac{\beta}{\alpha}=\left\lceil\frac{E(\tilde{C},\varepsilon_g)}{E(\tilde{C},\varepsilon)}\right\rceil \tag{7}$$

3 实验和分析

本节我们利用下面的 6 个基因表达谱数据集[10,13]来进行仿真实验.在此,我们将患者样本的癌症测试实验分为独立测试实验和交叉测试实验.在具有独立测试子集的前 3 个数据集上分别进行独立测试实验和交叉测试实验,在没有独立测试子集的后 3 个数据集上只进行交叉测试实验.

3.1 数据集

1. 急性白血病数据集(ALL-AML leukemia)

急性白血病数据集包含 72 例急性白血病样本,每个样本均含 7 129 个基因表达数据.其中,47 例样本被诊断为急性淋巴白血病(acute lymphoblastic leukemia,简称 ALL),25 例样本被诊断为急性骨髓白血病(acute myeloid leukemia,简称 AML).该数据集分为训练子集和测试子集,训练子集中包含 38 例训练样本(27 例 ALL+11 例 AML),测试子集中包含 34 例测试样本(20 例 ALL+14 例 AML).

2. 乳腺癌数据集(breast cancer)

乳腺癌数据集包含 97 例乳腺癌样本,每个样本均含 24 481 个基因表达数据.乳腺癌数据集记录了经过初次治疗超过 5 年时间后癌症患者的复发情况.在 46 例样本中癌症细胞发生转移(metastases)即癌症复发(relapse),51 例样本中癌症没有复发(non-relapse).该数据集分为训练子集和测试子集,训练子集中包含 78 例训练样本(34 例 relapse+44 例 non-relapse),测试子集中包含 19 例测试样本(12 例 relapse+7 例 non-relapse).

3. 前列腺癌数据集(prostate cancer)

前列腺癌数据集共有 136 例前列腺组织样本,每个样本均含 12 600 个基因表达数据.其中,75 例为前列腺癌

肿瘤样本(prostate tumor sample,简称 PTS),59 例样本正常前列腺组织(normal prostate sample,简称 NPS).该数据集分为训练子集和测试子集,训练子集中包含 102 例训练样本(52 例 PTS+50 例 NPS),测试子集中包含 34 例测试样本(25 例 PTS+9 例 NPS).

4. 弥漫性大 B 细胞淋巴瘤数据集(DLBCL)

弥漫性大 B 细胞淋巴瘤数据集共有 47 例弥漫性大 B 细胞淋巴瘤样本,其中包括 47 例胚中心 B 细胞样(germinal center B-like,简称 GCB)淋巴瘤样本和活性型周围 B 细胞样(activated peripheral B-like,简称 APB)淋巴瘤样本,每例样本均含 4 026 个基因的表达数据.

5. 结肠癌数据集(colon tumor)

结肠癌数据集共有 62 例结肠组织样本,其中包括 40 例结肠癌组织(tumor colon tissue,简称 TCT)和 22 例正常结肠组织(normal colon tissue,简称 NCT),每例样本均含 2 000 个基因的表达数据.

6. 卵巢癌(ovarian cancer)

卵巢癌数据集共有 253 例卵巢组织样本,其中包括 91 例正常卵巢组织样本(normal ovarian sample,简称 NOS)和 151 例卵巢癌组织样本(ovarian cancer sample,简称 OCS),每例样本均含 15 154 个基因的表达数据.

3.2 过滤噪声基因

利用 Fayyad 等人[14]提出的基于启发式熵最小化的离散方法(discretization)来过滤噪声基因,结果见表 1.

Table 1 Gene selection with noise reduction

表 1 噪声基因过滤

Dataset	Genes (before filtering)	Genes (after filtering)
ALL-AML leukemia	7 129	866
Breast cancer	24 481	834
Prostate cancer	12 600	3 071
DLBCL	4 026	336
Colon tumor	2 000	135
Ovarian cancer	15 154	2 945

3.3 结果与分析

3.3.1 性能评价

为了方便描述,将以上每个数据集划分为正例样本(positives)和负例样本(negatives).正例样本分别为 ALL,Relapse,PTS,GCB,TCT,NOS 样本,负例样本分别为 AML,Non-Relapse,NPS,APB,NCT,OCS 样本.利用准确度(*Accu*)、正确率(*Prec*)、灵敏度(*Sn*)和明确性(*Sp*)这 4 个指标来进行性能评价.*Accu*,*Prec*,*Sn*,*Sp* 定义如下:

$$Accu=\frac{TP+TN}{TP+FP+FN+TN},$$

$$Prec=\frac{TP}{TP+FP},$$

$$Sn=\frac{TP}{TP+FN},$$

$$Sp=\frac{TN}{TN+FP}.$$

其中,*TP*,*FP*,*TN*,*FN* 分类后所得到的样本数目见表 2.

Table 2 Confusion matrix

表 2 混乱矩阵

Observed	Predicted	
	Positives	Negatives
Positives	*TP*	*FN*
Negatives	*FP*	*TN*

3.3.2 独立测试实验

我们首先过滤 Leukemia,Breast 和 Prostate 数据集中的噪声基因,将 Leukemia,Breast 过滤后的基因分别作为各数据集的特征基因,在 Prostate 数据集上挑选累积 $Gain(A,T;\tilde{S})$ 最大的前 1 000 个作为 Prostate 的特征基因.然后,利用每个数据集的训练子集和特征基因来训练 Leukemia,Breast,Prostate 的 GCM 模型.GCM 模型的主分量设为 15 个,并分别训练正例样本子集和负例样本子集的 CCM 模型,CCM 模型的癌症组分量也设为 15 个.最后,利用 EAGC 算法给测试数据子集分类并计算性能评价指标 *Accu*,*Prec*,*Sn* 和 *Sp*.上述分类实验重复 10 次,计算平均性能评价指标,并与 Golub 提出的加权投票法(weighted voting)[6],SVM(support vector machine)和 KNN(*k*-nearest neighbor)所获得的分类结果进行了比较.其中,SVM 采用径向基函数(radial basis function,简称 RBF)作为核函数,KNN 相似性度量函数采用 Pearson 相关系数.在加权投票法中利用 50 个特征基因,在 SVM 和 KNN 中利用 SNR(信噪比)选取 50 个特征基因.加权投票法、SVM 和 KNN 的分类实验同样重复 10 次,计算平均性能评价指标.表 3 给出了独立测试实验的分类准确度(*Accu*),图 3 给出了独立测试实验的分类精确度(*Prec*),图 4 给出了独立测试实验的灵敏度(*Sn*)和明确性(*Sp*).

Table 3 Independent test results (*Accu* %)

表 3 独立测试实验结果(*Accu* %)

	Leukemia	Breast	Prostate
Weighted voting	85.3	78.9	67.6
SVM	94.1	94.7	75.6
KNN (*k*=5)	91.1	84.2	73.5
EAGC	97.1	94.7	91.4

由表 3 可以看出,在独立测试实验中,相对于加权投票法和 KNN,SVM 取得了较好的分类准确度,在 3 个数据集上都优于其他两种分类器,并在 Breast 数据集上具有与 EAGC 相同的准确度.并且,加权投票法、SVM 和 KNN 在 Prostate 数据集上都没有取得较好的分类效果,缺乏泛化性.然而,EAGC 结合了具有互补性能的 GCM 模型和 CCM 模型,有效综合了 GCM 和 CCM 的解决方案,弥补了单个分类器的不足,在所有数据集上都取了较高的分类准确度.

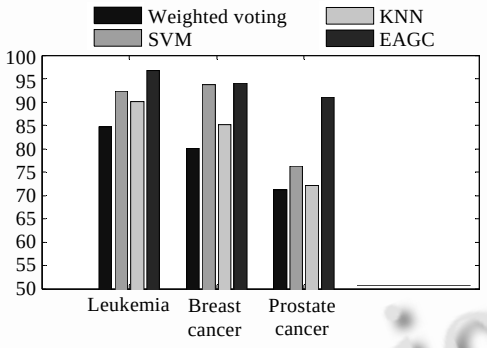


Fig.3 Independent test results (*Prec* %)

图 3 独立测试实验结果(*Prec* %)

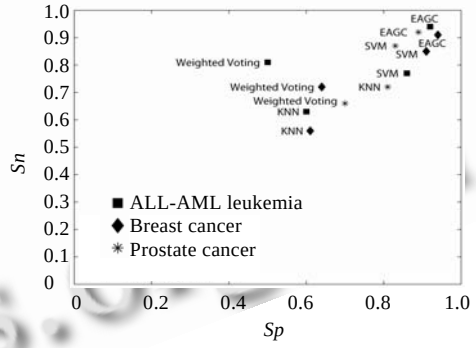


Fig.4 Independent test results (*Sn* and *Sp*)

图 4 独立测试实验结果(*Sn* 和 *Sp*)

从图 3 可以看出,在所有数据集中,SVM 都取得了优于 KNN 和 Leukemia 的精确度,甚至在 Breast 数据集超过 Weighted Voting 和 KNN 的分类精确度 14%和 10%,取得了和 EAGC 相近的分类性能.但在 Prostate 数据集中的的精确度只有 75.5%,与 Weighted Voting 和 KNN 性能相当,远小于 EAGC 的 91.2%.EAGC 在 3 个数据集上都取得了较好的分类精度.

从图 4 可以看出,在 Leukemia 数据集中,Weighted Voting 获得了较高的灵敏度,但明确性偏低.相对于 KNN,Weighted Voting 的灵敏度高于 KNN,但明确性远低于 KNN.在其他两个数据集中,SVM 的明确性都高于

Weighted Voting 和 KNN.在所有数据集中,Weighted Voting 方法的明确性偏低,KNN 的灵敏度偏低,SVM 则取得了较好的性能.EAGC 取得了明显优于 Weighted Voting,SVM 和 KNN 的效果.

3.3.3 交叉测试实验

在所有的 6 个数据集上进行交叉测试实验,包括留一交叉检验(leave-one-out cross validation,简称 LOOCV)和 5 折交叉检验(five-fold cross validation,简称 FFCV).在 LOOCV 中,每次从数据集中挑选一个不同的样本作为测试样本,其余样本作为训练数据集训练 GCM 模型和 CCM 模型,然后利用 EAGC 识别测试样本.重复该过程,直到每一个样本作为测试样本时为止.统计所有被正确识别的样本,并计算性能评价指标 *Accu*, *Prec*,*Sn* 和 *Sp*.上述分类实验重复 10 次,计算平均性能评价指标.在 FFCV 中,将数据集平均分为成 5 部分,每次挑选不同的一部分作为测试样本,其余样本作为训练数据集训练 GCM 模型和 CCM 模型,然后利用 EAGC 识别测试样本.重复分类过程 5 次,直到每部分样本作为测试样本时为止.统计所有被正确识别的样本,并计算性能评价指标 *Accu*,*Prec*,*Sn* 和 *Sp*.上述分类实验重复 10 次,计算平均性能评价指标,并与加权投票法、SVM 和 KNN 进行比较.其中,加权投票法、SVM 和 KNN 的分类参数设置与独立测试实验相同,分类实验同样重复 10 次,并计算平均性能评价指标.表 4 给出了实验的分类准确度 *Accu*,图 5 和图 6 给出了实验的灵敏度(*Sn*)和明确性(*Sp*).

Table 4 Cross validation test results (*Accu* %)

表 4 交叉测试实验结果(*Accu* %)

		Leukemia	Breast	Prostate	DLBCL	Colon	Ovarian
LOOCV	Weighted voting	90.3	77.3	70.4	88.6	93.5	63.5
	SVM	95.3	84.6	68.9	97.8	91.9	82.4
	KNN (*k*=5)	86.1	84.2	80.1	92.8	83.9	73.3
	EAGC	99.3	97.9	91.4	93.4	96.8	92.6
FFCV	Weighted voting	88.4	72.1	82.9	80.4	90.6	71.4
	SVM	92.1	94.4	80.5	96.3	93.2	84.4
	KNN (*k*=5)	84.1	80.4	80.6	86.3	84.2	67.9
	EAGC	96.2	93.6	94.3	95.5	97.3	96.3

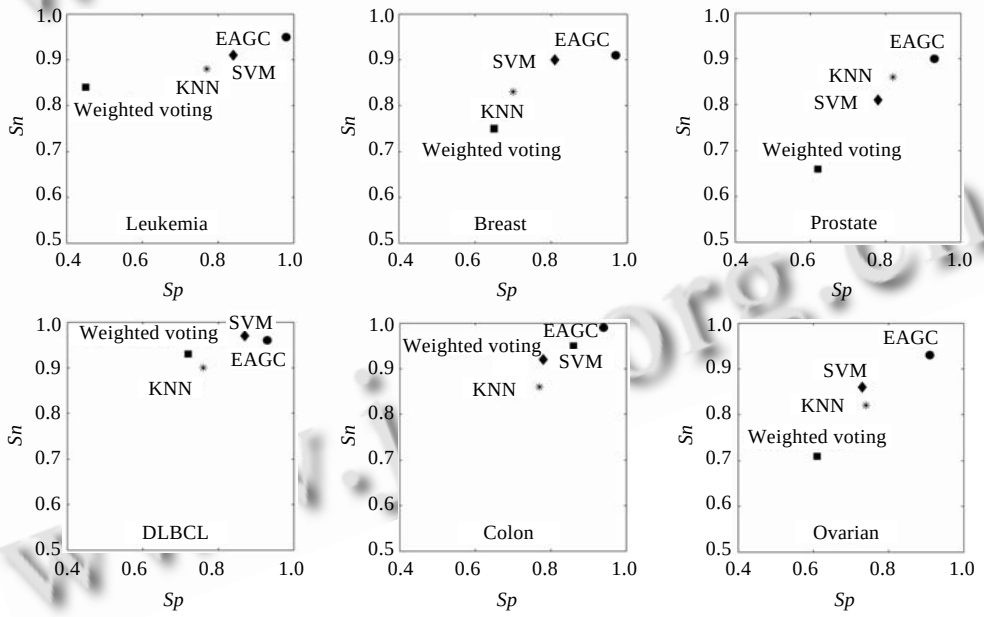


Fig.5 LOOCV test results (*Sn* and *Sp*)

图 5 LOOCV 交叉测试实验结果(*Sn* 和 *Sp*)

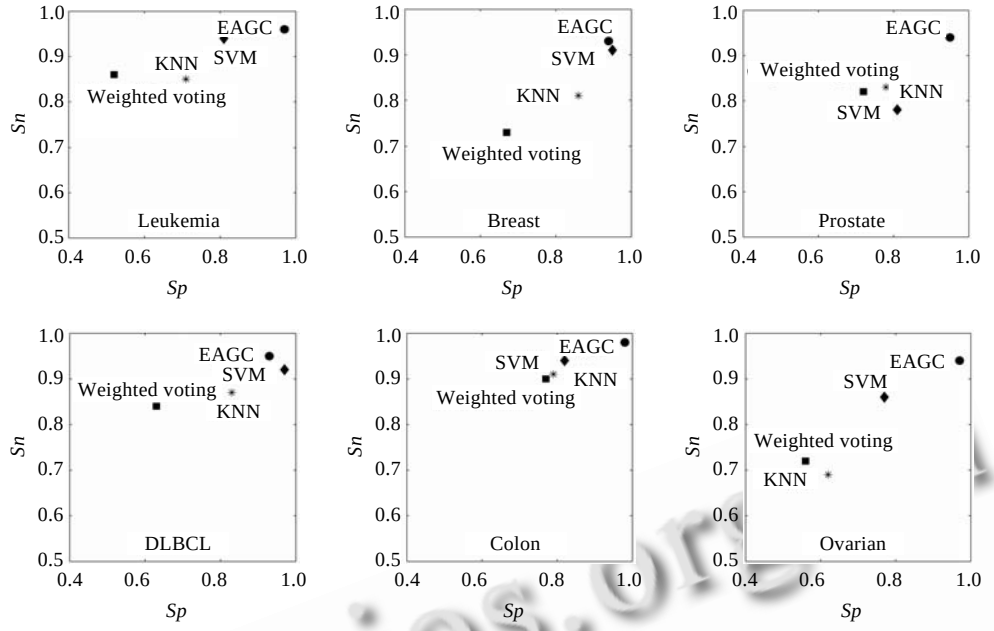


Fig.6 FFCV test results (*Sn* and *Sp*)

图 6 FFCV 交叉测试实验结果(*Sn* 和 *Sp*)

从表 4 可以看出,在 LOOCV 中,相对于加权投票法和 KNN,SVM 同样取得了较好的分类准确度,并且在 DLBCL 上取得最好的分类准确度 97.8%,高于 EAGC.然而在 Prostate 上,*Accu* 低于其他分类器.在所有数据集中,KNN 则相对表现出较好的泛化能力.EAGC 除了在 Prostate 上分类准确度略低于 SVM 之外,在其他数据集上都高于加权投票法、KNN 和 SVM.在 FFCV 中,SVM 同样取得了较好的分类准确度,加权投票法次之.EAGC 除了在 DLBCL 上分类准确度略低于 SVM 之外,在其他数据集上都高于加权投票法、KNN 和 SVM.

从图 5 和图 6 可以看出,在 LOOCV 实验中,Weighted Voting 在 Colon 数据集上取得了较好的分类效果,但在其他实验中的效率低于其他方法.除了在 FFCV 实验中的 Ovarian 数据集上,KNN 的分类效果明显不及 SVM 之外,在其他实验中则与 SVM 相当,并在 Prostate 的 LOOCV 和 FFCV 都优于 SVM.说明在不同的数据集上的分类效果由于方法不同存在较大差别,在癌症样本检测中分类方法缺乏泛化性.但是,EAGC 则表现出较好的泛化能力,在所有的实验中都表现了很好的分类性能.

3.3.4 特征值数和权值

首先讨论特征值数目的选取对不同算法的影响.对于 Leukemia,Breast,Prostate 数据集,分别将特征数目 *h* 选定从 1 递增到 100,分析不同分类方法准确度(*Accu*)的情况.即在构建的 GCM 和 CCM 模型,模型的分量分别设为 *h* 个;同样,在 Weighted Voting,SVM 和 KNN 方法中选取 *h* 个特征基因,训练分类器;其余的参数设置与第 2 节的独立测试实验相同.实验重复 10 次,计算平均分类准确度.图 7~图 9 给出了不同算法在 Leukemia,Breast,Prostate 数据集上的分类准确度,图中横坐标是基因特征数量.从图中可以看出,对于 Weighted Voting,SVM 和 KNN 方法,在特征基因数目 50 左右,可以获得一个较高的分类准确度.在 3 个数据集中,SVM 取得了更好的分类精度.但在 Prostate 数据集上,KNN 可以取得与 SVM 接近的分类精度.在所有数据集中,EAGC 算法则在特征基因数量为 15 左右时,可以获得很好的分类性能.根据经验值,Weighted Voting,SVM 和 KNN 方法中选取 50 个特征基因,在 EAGC 中选取 15 个特征基因.从图中同样可以看出,SVM 在 Prostate 数据集上的分类准确度不高(76.4%),在另外两个数据集上取得较好的性能,EAGC 则具有很好的泛化性能.

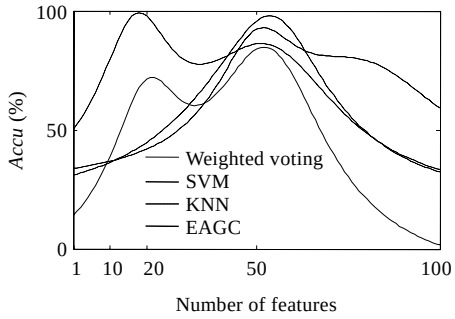


Fig.7 Different methods' accuracy with different gene features in Leukemia

图 7 Leukemia 中算法在不同特征数下的准确度

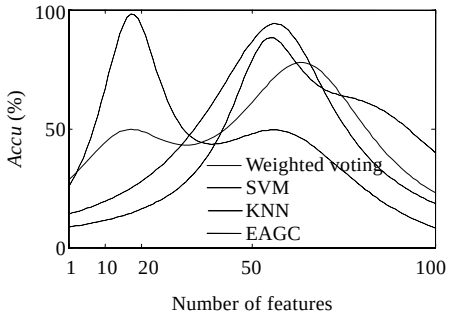


Fig.8 Different methods' accuracy with different gene features in Breast

图 8 Breast 中算法在不同特征数下的准确度

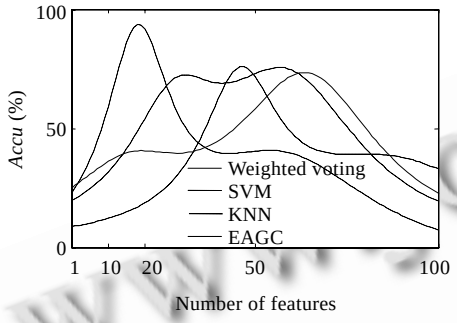


Fig.9 Different methods' accuracy with different gene features in Prostate

图 9 Prostate 中算法在不同特征数下的准确度

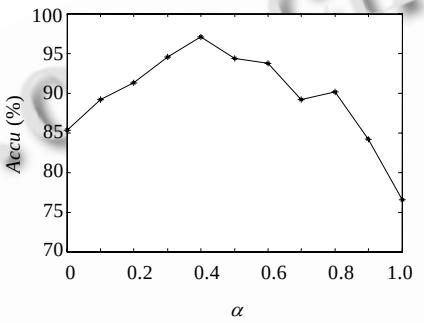


Fig.10 EAGC accuracy with different α in Leukemia

图 10 Leukemia 中 EAGC 在不同α下的准确度

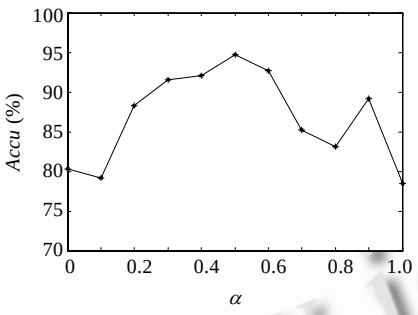


Fig.11 EAGC accuracy with different α in Breast

图 11 Breast 中 EAGC 在不同α下的准确度

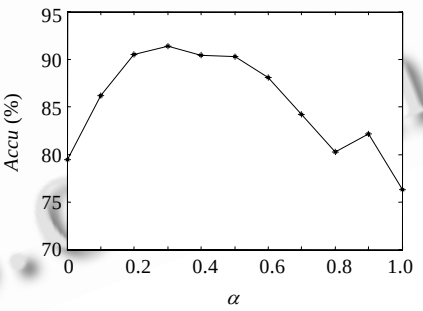


Fig.12 EAGC accuracy with different α in Prostate

图 12 Prostate 中 EAGC 在不同α下的准确度

其次讨论在 EAGC 算法中,基于权值的投票组合策略对 EAGC 算法的影响.在组合策略中,CCM 模型赋予的权值为α,GCM 模型赋予的权值为β.在 Leukemia,Breast,Prostate 数据集上,分别利用 EAGC 算法进行癌症识别.从 0 到 1 改变 CCM 模型权值α,其余参数设置与第 2 节的独立测试实验相同.实验重复 10 次,计算平均分类准确度.图 10~图 12 给出了 EAGC 算法在 Leukemia,Breast,Prostate 数据集上的分类准确度,图中横坐标是α值.从图中可得,对于 Leukemia,Breast,Prostate 数据集,当α值分别为 0.4,0.5 和 0.3,时,EAGC 可以获得最优分类效果.

同时,由前述讨论设 d=15,则由公式(7)可计算α和β.对于 Leukemia,Breast,Prostate 数据集,α值分别为 0.4,0.5 和 0.2,β值分别为 0.6,0.5 和 0.8.α值基本一致,在 Prostate 数据集上,α值存在较小的偏差.通过α值的组合策略没有获得最优分类准确度,但是也可以获得很好的分类准确度;在另外两个数据集上,都取得最优分类准确度.

4 结束语

针对基因表达谱数据特点提出了两种癌症识别模型(GCM 模型和 CCM 模型),并结合 GCM 模型和 CCM 模型的互补性,利用基于权值的投票组合策略提出一种组合分类算法(EAGC).相对于传统算法,EAGC 有效地综合了 GCM 和 CCM 识别模型的解决方案,弥补了单个分类器的不足,扩展了 EAGC 的解决方案,在所有数据集上都取了很好的分类性能.

References:

[1] Kuramochi M, Karypis G. Gene classification using expression profiles: A feasibility study. Int'l Journal on Artificial Intelligence Tools, 2005,14(4):641−660.

[2] Li X, Zhang TW, Guo Z. An novel ensemble method of feature gene selection based on recursive partition-tree. Chinese Journal of Computers, 2004,27(5):675−682 (in Chinese with English abstract).

[3] Lu XG, Lin YP, Wang HJ, Zhou SW, Li XL. A novel relative space based gene feature extraction and cancer recognition. In: Zhou ZH, Li H, Yang Q, eds. Proc. of the 11th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD 2007). Berlin: Springer-Verlag, 2007. 712−719.

[4] Li YX, Li JG, Ruan XG. Study of informative gene selection for tissue classification based on tumor gene expression profiles. Chinese Journal of Computers, 2006,26(2):324−330 (in Chinese with English abstract).

[5] Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine, 2002,8(1):68−74. [doi: 10.1038/nm0102-68]

[6] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasernbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 1999,286(5439):531−537. [doi: 10.1126/science.286.5439.531]

[7] Veer LJVT, Dai H, Vijver MJVD, He YD, Hart AA, Mao M, Peterse HL, Kooy KVD, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002,415(6871):530−536. [doi: 10.1038/415530a]

[8] Cho SB, Won HH. Machine learning in DNA microarray analysis for cancer classification. In: Chen YPP, ed. Proc. of the 1st Asia-Pacific Bioinformatics Conf. (APBC 2003). Adelaide: Australian Computer Society, 2003. 189−198.

[9] Wang ZQ, Chenh SF, Chen ZQ. An optimized neural network linear ensemble for classification. Journal of Software, 2005,16(11): 1902−1908 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/1902.htm [doi: 10.1360/jos161902]

[10] Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Applied Bioinfomatics, 2003, 2(Suppl.):S75−S83.

[11] Khan J, Wei JS, Ringne M, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 2001,7(6):673−679. [doi: 10.1038/89044]

[12] O'Neill MC, Song L. Neural network analysis of lymphoma microarray data: Prognosis and diagnosis near-perfect. BMC Bioinformatics, 2003,4:13. [doi: 10.1186/1471-2105-4-13]

[13] Liu B, Cui QH, Jiang TZ, Ma SD. A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC Bioinformatics, 2004,5:136. [doi: 10.1186/1471-2105-5-136]

[14] Fayyad U, Irani K. Multi-Interval discretization of continuous-valued attributes for classification learning. In: Bajcsy R, ed. Proc. of the 13th Int'l Joint Conf. on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1993. 1022−1029.

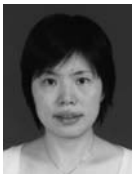
附中文参考文献:

[2] 李霞,张田文,郭政.一种基于递归分类树的集成特征基因选择方法.计算机学报,2004,27(5):675−682.

[4] 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.计算机学报,2006,26(2):324−330.

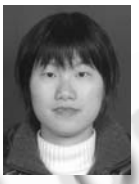
[9] 王正群,陈世福,陈兆乾.优化分类型神经网络线性集成.软件学报,2005,16(11):1902−1908. http://www.jos.org.cn/1000-9825/16/1902.htm [doi: 10.1360/jos161902]

卢新国(1979－),男,湖南汨罗人,博士,讲师,主要研究领域为数据挖掘与机器学习,生物信息学.



骆嘉伟(1964－),女,博士,教授,博士生导师,主要研究领域为数据挖掘与机器学习,生物信息学.

林亚平(1955－),男,博士,教授,博士生导师,主要研究领域为计算机网络,模式识别,数据挖掘与机器学习,生物信息学.



李丹(1987－),女,硕士生,主要研究领域为生物信息学.