利用协同分类方法识别癌症类型

卢新国 陈 东 杜家宜 周 娟
(湖南大学计算机与通信学院 长沙 410082)

摘 要 针对基因表达谱数据的特点提出了全局分量模型(GCM)和癌症组分量模型(CCM)两种癌症识别模型;通过基于权值的投票组合策略提出了一种基于GCM和CCM的协同分类方法(CAGC)来识别癌症类型。在Leukemia,Breast,Prostate,DLBCL,Colon,Ovarian等6个数据集上进行了测试实验,结果表明CAGC有效综合了GCM和CCM识别模型的解决方案,具有较好的泛化性。

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 TP18,TP391 **文献标识码** A

Using Co-classification Approach to Detect the Type of Cancer

LU Xin-guo CHEN Dong DU Jia-yi ZHOU Juan
(School of Computer and Communication, Hunan University, Changsha 410082, China)

Abstract Cancer recognition with gene expression profile was studied. Due to large redundant and noise information in the gene expression data and the sensitiveness to selected feature genes, the classification is lack of generalization capability. With studying the gene expression profiles, two cancer recognition models including global component model (GCM) and cancer component model (CCM) were constructed. And a weighted voting strategy was applied to propose an Co-classification Approach based on GCM and CCM for cancer recognition (CAGC). Test experiments were conducted on Leukemia, Breast, Prostate, DLBCL, Colon and Ovarian cancer dataset respectively, and great performance is acquired by CAGC on all datasets. The experiment results show that the recognition solution and the generalization are strengthened by combination of GCM and CCM.

Keywords Gene expression profile, Cancer recognition, Global component model, Cancer component model

近年来,研究人员获得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。如何利用基因表达谱数据对癌症进行精确分类,是目前基因诊断和治疗的一个研究热点[1,2]。

微阵列基因表达谱数据是一种典型的高维、高噪和高冗余数据[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, SNR)、排秩法(Rank)、信息增益(Information Gain)等[5-8]。文献[7]发现,在相同的数据集中,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亦不相同。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[7]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了较多的冗余信息;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, 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(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(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的协同分

到稿日期:2009-03-26 返修日期:2009-07-06 本文受国家自然科学基金项目(2007080504),湖南省自然科学基金(0002014014)资助。
卢新国(1979—),男,博士,讲师,主要研究方向为生物信息处理、数据挖掘等,E-mail:hnluxinguo@hotmail.com;陈 东(1969—),男,硕士生,主要研究方向为生物信息处理、机器学习等;杜家宜(1981—),男,硕士生,主要研究方向为生物信息处理等;周 娟(1981—),女,硕士生,主要研究方向为生物信息处理等。

类方法(CAGC),并在 Leukemia, Breast, Prostate, DLBCL, Colon, Ovarian 等 6 个数据集上分别进行交叉测试实验。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,因此 CAGC 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。

1 相关知识

1.1 基因微阵列表达谱

DNA 微阵列(Microarray)是在一定尺寸的基片(如硅片、玻璃、塑料等)表面固定一系列可寻址的识别分子的点阵,点阵中每一个点都可以视为一个传感器的探头。主要是通过 DNA-DNA 杂交反应或是蛋白质之间的特异性结合,同时观测数千甚至数万个基因。基因表达谱是指利用 DNA 微阵列测定的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,亦可以表示为第 j 个样本下第 i 个观测基因的表达水平。

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,其实质就是多层感知器 (Multi-Layer Perceptron, MLP),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 GCM 和 CCM 的协同分类(CAGC)

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(Global Component Model, GCM)和癌症组分量模型(Cancer Component Model, 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 $X^T=(\bar{g}_1,\bar{g}_2,\cdots,\bar{g}_m)$,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:

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

其中,$\bar{a}_i^T=(\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_m)$,$\sum=(\upsilon_{ij})_{m\times m}=(Cov(\bar{g}_i,\bar{g}_j))$,$Var$ 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 $\sum$ 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 $\bar{s}_i$ 的分量贡献系数。

定义 2(累积贡献率) 不妨设 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m$,给定 $r(1\leqslant r\leqslant m)$,分量 $\bar{h}_1,\bar{h}_2,\cdots,\bar{h}_r$ 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 $\min(\lambda_1,\lambda_2,\cdots,\lambda_r)\geqslant\max(\lambda_{r+1},\lambda_{r+2},\cdots,\lambda_m)$,则由 $\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_d$ 组成了基因表达数据的 r 维全局分量空间 ε_g,$\varepsilon_g=\{\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_d\}$。

假设基因表达谱中分为 k 个癌症类别 $C_1,C_2,\cdots,C_k$,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。GCM 包含输入层(I 层)、隐层(H 层)和输出层(O 层)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:

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

$$out_j=f(net_j)$$

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式(3)和式(4)所示:

$$w_{ij}(t+1)=w_{ij}(t)+\Delta w_{ij}=w_{ij}(t)-\eta\delta_j out_i$$

$$\delta_j=\begin{cases}-(\hat{O}_j-O_j)f'(net_k) & \text{if 节点 } j \text{ in O 层}\\ f'(net_k)\sum_k\delta_k w_{jk} & \text{if 节点 } j \text{ in H 层}\end{cases}\tag{3}$$

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

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:

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

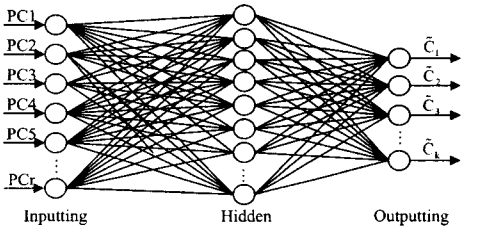

图 1 癌症识别中的全局分量模型

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 $\tilde{C}_i(1\leqslant i\leqslant k)$,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,称之为癌症组分量(Cancer Component, 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)。$\tilde{C}_i$ 的最小扩展空间 $\hat{\varepsilon}$ 以及样本的 CC 分量的获取如下所示。

设 $C_i^T=(\bar{g}_1,\bar{g}_2,\cdots,\bar{g}_m)$,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,$P=(\bar{p}_1,\bar{p}_2,\cdots,\bar{p}_m)$。

定义 4(关联空间) 设癌症 $\tilde{C}_i$ 和表达矩阵 C_i,$\lambda_1,\lambda_2,\cdots,\lambda_m$ 是 $Cov(C_i{}^T)$ 的特征值,$\bar{p}_1,\bar{p}_2,\cdots,\bar{p}_m$ 是 $\lambda_1,\lambda_2,\cdots,\lambda_m$ 对应的特征向量。对于 $d\leqslant m$,则由 $\bar{p}_1,\bar{p}_2,\cdots,\bar{p}_d$ 组成了 $\tilde{C}_i$ 上秩

为 d 的关联空间 ε，$\varepsilon=\{\bar{p}_1,\bar{p}_2,\cdots,\bar{p}_d\}$，$\bar{p}_i$ 为 ε 的第 i 维方向，λ_i 称为方向 $\bar{p}_i$ 的方向扩展系数，$P=(\bar{p}_1,\bar{p}_2,\cdots,\bar{p}_d)$ 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，$\lambda_1,\lambda_2,\cdots,\lambda_d$ 是 $\hat{\varepsilon}$ 的方向扩展系数，当 $\lambda_1,\lambda_2,\cdots,\lambda_d$ 满足 $\max(\lambda_1,\lambda_2,\cdots,\lambda_d)\leqslant\min(\lambda_d,\lambda_{d+1},\cdots,\lambda_m)$ 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 $\tilde{C}_i=\{\bar{s}_1,\bar{s}_2,\cdots,\bar{s}_{n_i}\}$，$\bar{s}_l=(s_{l1},s_{l2},\cdots,s_{lm})^T$，那么 $\bar{s}_l$ 在 $\bar{p}_j$ 上的癌症组分量 $CC_j=\bar{s}_l\cdot\bar{p}_j$，其中 $\bar{p}_j$ 为 $\hat{\varepsilon}$ 的第 j 维方向，$\bar{p}_j=(p_{j1},p_{j2},\cdots,p_{jm})^T$，$\bar{s}_l\cdot\bar{p}_j=\sum_{k=1}^{m}s_{lk}p_{jk}$。

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。CCM 包含输入层(I 层)、隐层(H 层)和输出层(O 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、权值和激活阈值调节同 GCM 模型。对于癌症样本 $\bar{s}$ 的期望输出 $\hat{O}$ 通过下式给出：

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

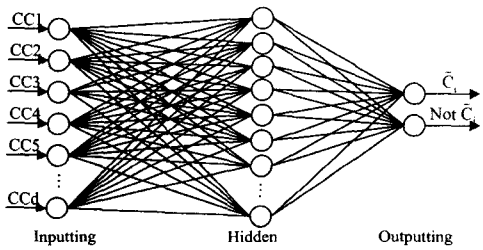

图 2 癌症识别中的癌症组分量模型

2.2 基于 GCM 和 CCM 的协同分类方法(CAGC)

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。如图 3 所示，假设数据集合中存在 3 个癌症模式 m_1,m_2,m_3，则 $\not{m}_1$ 是不属于 m_1 的样本构成的模式(称之为非 m_1 模式)。相应地，$\not{m}_2,\not{m}_3$ 是非 m_2 模式和非 m_3 模式。不妨设 $\hat{\varepsilon}_{c_1}=\{x_1,y_1\}$，$\hat{\varepsilon}_{c_2}=\{x_2,y_2\}$，$\hat{\varepsilon}_{c_3}=\{x_3,y_3\}$，$\varepsilon_k=\{x,y\}$，在 x_1-y_1 中可以区分 m_1 和 $\not{m}_1$，在 x_2-y_2 中可以区分 m_2 和 $\not{m}_2$，在 x_3-y_3 中可以区分 m_3 和 $\not{m}_3$，在 $x-y$ 中可以区分 m_1,m_2 和 m_3。CCM 模型从癌症组内的基因特征中挖掘不同癌症模式 m_i，是一种具有局部相关性的癌症识别模型。GCM 模型从所有基因特征及特征组间的相异性发现各种癌症模式 m_i，是一种具有全局相关性的癌症识别模型。CCM 模型可以识别单个癌症模式，GCM 模型则可以同时识别多个癌症模式。从上述分析可知，GCM 和 CCM 是两种具有互补性的癌症识别模型。

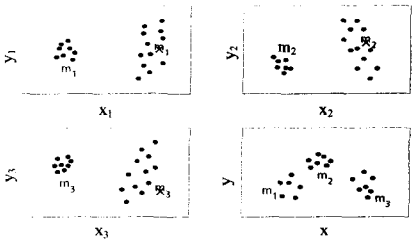

图 3 CCM 的癌症组分量和 GCM 的全局分量

本节提出一种基于 GCM 和 CCM 模型的协同分类方法。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来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 $R(\tilde{C})^T=(r_1,r_2,\cdots,r_k)$，基于权值的投票组合策略描述如下：

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

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。

CAGC 有效综合了 GCM 和 CCM 模型的癌症识别结果，消除了基因数据中内在的噪声和冗余对单个分类器的影响，优化了分类器的癌症识别结果，提高了 CAGC 的泛化能力。基于组合 GCM 和 CCM 模型的癌症识别算法具体描述如表 1 所列。

表 1 基于 GCM 和 CCM 模型的协同分类方法(CAGC)

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，$\tilde{C}=\sum\tilde{C}_i$，$q=10$，$CR\geqslant85\%$，$d=15$，$\eta=0.5$，$\alpha=0.4$，$\beta=0.6$。

Begin

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

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 $\lambda_1,\lambda_2,\cdots,\lambda_m$，特征向量 $\bar{p}_1,\bar{p}_2,\cdots,\bar{p}_m$；

选取 d 个最小的 $\lambda_1',\lambda_2',\cdots,\lambda_d'$ 对应的 $\bar{p}_1',\bar{p}_2',\cdots,\bar{p}_m'$ 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 $R(\tilde{C})=(r_1,r_2,\cdots,r_k)^T$；

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 $R(ensemble)=(r_1',r_2',\cdots,r_k')^T$，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

3 实验和分析

本节利用下面的6个基因表达谱数据集[10,14](如表2所列)来进行仿真实验,并分析基于组合GCM和CCM分类算法的癌症识别性能。在此,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。

表2 基因表达谱数据集

Dataset	Genes	Training samples	Test samples
ALL-AML Leukemia	7129	38(27:11)	34(20:14)
Breast Cancer	24481	78(34:44)	19(12:7)
Prostate Cancer	12600	102(52:50)	34(25:9)
DLBCL	4026	47(24:23)	0
Colon Tumor	2000	62(40:22)	0
Ovarian Cancer	15154	253(91:162)	0

3.1 数据集

3.1.1 急性白血病数据集(ALL-AML Leukemia)

急性白血病数据集[5]包含72例急性白血病样本,每个样本均含7129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,ALL),25例样本被诊断为急性骨髓白血病(Acute Myeloid Leukemia,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)。

3.1.2 乳腺癌数据集(Breast Cancer)

乳腺癌数据集[5]包含97例乳腺癌样本,每个样本均含24481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)。

3.1.3 前列腺癌数据集(Prostate Cancer)

前列腺癌数据集[5]共有136例前列腺组织样本,每个样本均含12600个基因表达数据。其中75例为前列腺癌肿瘤样本(Prostate Tumor Sample,PTS),59例为正常前列腺组织样本(Normal Prostate Sample,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)。

3.1.4 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,GCB)淋巴瘤样本和活性型周围B细胞样(Activated Peripheral B-like,APB)淋巴瘤样本,每例样本均含4026个基因表达数据。

3.1.5 结肠癌数据集(Colon Tumor)

结肠癌数据集[7]共有62例结肠组织样本,其中包括40例结肠癌组织(Tumor Colon Tissue,TCT)和22例正常结肠组织(Normal Colon Tissue,NCT),每例样本均含2000个基因表达数据。

3.1.6 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,NOS)和151例卵巢癌组织样本(Ovarian Cancer Sample,OCS),每例样本均含15154个基因表达数据。

3.2 过滤噪声基因

利用Fayyad等[13]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。样本集 $\tilde{S}$ 有 k 个类别 $\tilde{C}_1,\tilde{C}_2,\cdots,\tilde{C}_k$,假设 $\tilde{S}$ 被分为两个子集 $\tilde{S}_1$ 和 $\tilde{S}_2$,$p(\tilde{C}_i,\tilde{S}_j)$ 为样本子集 $\tilde{S}_j$ 中样本出现在 $\tilde{C}_i$ 中的频率,则子集 $\tilde{S}_j(j=1,2)$ 的熵为:

$$Ent(\tilde{S}_j)=-\sum_{i=1}^{k}P(\tilde{C}_i,\tilde{S}_j)\log(P(\tilde{C}_i,\tilde{S}_j)) \tag{7}$$

假设由特征 A 在点 T 划分获取子集 $\tilde{S}_1$ 和 $\tilde{S}_2$,则用 $E(A,T;\tilde{S})$ 描述此划分的类别信息熵:

$$E(A,T;\tilde{S})=\frac{|\tilde{S}_1|}{|\tilde{S}|}Ent(\tilde{S}_1)+\frac{|\tilde{S}_2|}{|\tilde{S}|}Ent(\tilde{S}_2) \tag{8}$$

在特征 A 所有候选点中选取具有最小的类别信息熵作为划分点,然后在样本子集 $\tilde{S}_j(j=1,2)$ 上递归划分,并利用最小描述长度原则(Minimum Description Length Principle,MDL)作为停止条件,即满足:

$$Gain(A,T;\tilde{S})<\frac{\log_2(N-1)}{N}+\frac{\delta(A,T;\tilde{S})}{N} \tag{9}$$

其中,N 是集合 $\tilde{S}$ 中特征值的数目,$Gain(A,T;\tilde{S})=Ent(\tilde{S})-E(A,T;\tilde{S})$,$\delta(A,T;\tilde{S})=\log_2(3^k-2)-[k\cdot Ent(S)-k_1\cdot Ent(S_1)-k_2\cdot Ent(S_2)]$,$k_i$ 为 $\tilde{S}_i$ 中类别的数目。

噪声基因过滤如表3所列。

表3 噪声基因过滤

Dataset	Genes (Before filtering)	Genes (After filtering)
ALL-AML Leukemia	7129	866
Breast Cancer	24481	834
Prostate Cancer	12600	3071
DLBCL	4026	336
Colon Tumor	2000	135
Ovarian Cancer	15154	2945

3.3 结果与分析

3.3.1 性能评价

为了描述方便,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)指标来进行性能评价,Accu定义为:

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

其中,TP,FP,TN,FN 分类后所得到的样本数目如表4所列。

表4 混乱矩阵

Observed	Predicted	
	Positives	Negatives
Positives	TP	FN
Negatives	FP	TN

3.3.2 交叉测试实验

在所有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,LOOCV)和"5折交叉检验"(Five-Fold Cross Validation,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来训练GCM模型和CCM模型。GCM

模型和CCM模型的输入变量与独立测试实验相同，然后利用CAGC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*。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CAGC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5给出了交叉测试实验的分类准确度 *Accu*。从表5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上取得了最好的分类准确度(97.8%)，高于CAGC。然而在Prostate上，*Accu* 低于其它分类器。在所有数据集中，KNN则相对表现了较好的泛化能力。CAGC除了在Prostate上分类准确度略低于SVM外，在其它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。CAGC除了在DLBCL上分类准确度略低于SVM外，在其它数据集上都高于加权投票法、KNN和SVM。

表5 交叉测试实验结果(Accu%)

		ALL-AML Leukemia	Breast Cancer	Prostate Cancer	DLBCL	Colon Tumor	Ovarian Cancer
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
	CAGC	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
	CAGC	96.2	93.6	94.3	95.5	97.3	96.3

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)，利用基于权值的投票组合策略提出了一种协同分类方法(CAGC)。在6个数据集上分别进行了交叉测试实验。CAGC有效综合了GCM和CCM识别模型的解决方案，在所有数据集上都取得了很好的分类性能。

参考文献

[1] Kuramochi M, Karypis G. Gene classification using expression profiles: a feasibility study[J]. International Journal on Artificial Intelligence Tools, 2005, 14(4): 641-660

[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[J]. 计算机学报，2004，27(5)：675-682

[3] Lu X G, Lin Y P, Wang H J, et al. A Novel Relative Space Based Gene Feature Extraction and Cancer Recognition[C]// Proc. of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007), LNAI 4426. 2007: 712-719

[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[J]. 计算机学报，2006，26(2)：324-330

[5] Shipp M A, Ross K N, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning[J]. Nature Medicine, 2002, 8(1): 68-74

[6] 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

[7] Veer L J V T, Dai H, Vijver M J V D, et al. Gene expression profiling predicts clinical outcome of breast cancer[J]. Nature, 2002, 415(6871): 530-536

[8] Cho S B, Won H H. Machine Learning in DNA Microarray Analysis for Cancer Classification[C]// Proc. of Bioinformatics 2003 First Asia-Pacific Bioinformatics Conference (APBC 2003). 2003: 189-198

[9] 王正群，陈世福，陈兆乾. 优化分类型神经网络线性集成[J]. 软件学报，2005，16(11)：1902-1908

[10] Tan A C, Gilbert D. Ensemble machine learning on gene expression data for cancer classification[J]. Applied Bioinfomatics, 2003, 2(3 Suppl): S75-S83

[11] Khan J, Wei J S, Ringne M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks[J]. Nature Medicine, 2001, 7(6): 673-679

[12] O'Neill M C, Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect[J]. BMC Bioinformatics, 2003, 4: 13

[13] Fayyad U, Irani K. Multi-interval discretization of continuous-valued attributes for classification learning[C]// Proceedings of the 13th International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, 1993: 1022-1029

[14] Liu B, Cui Q H, Jiang T Z, et al. A combinational feature selection and ensemble neural network method for classification of gene expression data[J]. BMC Bioinformatics, 2004, 5: 136

(上接第228页)

结束语 本文讨论了广义量子粒子模型的理论与性能，并设计了切实可行的将广义量子粒子模型应用于数据自组织聚类的并行算法。所提出的算法收敛速度快，并且在对噪声的不敏感性、聚类数据的强鲁棒性等方面都具有优势，仿真试验表明了所提算法的优越性。接下来要进一步充分挖掘和利用广义量子粒子模型的并行机制，研究其中参数设置的合理性，使得量子计算的优越性能够得到充分的利用。

参考文献

[1] Nielsen M A, Chuang I L. Quantum Computation and Quantum Information[M]. London: Cambridge University Press, 2000

[2] 帅典勋，冯翔，赵宏彬，等. 广义细胞自动机的结构及其硬件实现[J]. 计算机学报，2004，27(11)：1441-1450

[3] 刘燕，帅典勋，柴震川. 基于群体智能的网络信息自组织方式[J]. 计算机科学，2003，30(6)：84-87，125

[4] Shuai D X, Xue F L. Generalized particle model used for data clustering[J]. International Journal of Pattern Recognition, 2006, 20(7): 1001-1028

[5] Ramos R V, Souza R F. Calculation of the quantum entanglement measure of bipartite states, based on relative entropy, using genetic algorithms[J]. Journal of Computational Physics, 2002, 175(2): 576

[6] Grover L K. Quantum mechanics helps in searching for a needle in a haystack[J]. Physical Review Letters, 1997, 79(2): 325

[7] Feng X, Lau F C M, Shuai D X. The coordination generalized particle model-An envolutionary approach to multi-sensor fusion [J]. Information fusion, 2008, 9(4): 450-464

[8] Aczel A D. Entanglement: the greastest Mystery in physics[M]. New York: John Wiley and Sons Ltd, 2002