一种基于互信息的模糊粗糙分类特征基因快速选取方法

徐菲菲[1] 魏 莱[2] 杜海洲[1] 王文欢[3]
(上海电力学院计算机与信息工程学院 上海 200090)[1] (上海海事大学信息工程学院 上海 201303)[2]
(上海电力学院能源与环境工程学院 上海 200090)[3]

摘 要 依据基因表达谱建立有效肿瘤分类模型的关键在于准确找出决定样本类别的一组特征基因。粗糙集理论已成功应用于肿瘤分类特征基因选取中。然而,粗糙集方法处理连续值的基因表达谱数据集所必需的离散化过程会使得部分信息丢失,对所选取的特征基因的分类精度造成一定影响。因此,曾提出基于互信息的模糊粗糙集基因表达谱数据集特征基因的选取算法。然而,该算法计算代价较高,当所选取的基因数较多时难以实现。为此,对该算法进行了改进,从最大相关性和最重要性(最小冗余)两方面对互信息进行了近似替代计算,大大降低了算法的复杂度,提高了算法的效率。以急性白血病亚型(leukemia)、直肠癌(colon)和乳腺癌(Breast)分类特征基因选取为例进行实验,然后分别采用1NN和SVM分类器进行特征基因分类精度检验,结果证实了新方法的可行性和有效性。

关键词 特征选取,模糊粗糙集,互信息,基因表达谱数据集
中图法分类号 TP18 **文献标识码** A

Fast Approach to Mutual Information Based Gene Selection with Fuzzy Rough Sets

XU Fei-fei[1] WEI Lai[2] DU Hai-zhou[1] WANG Wen-huan[3]
(College of Computer and Information Engineering, Shanghai University of Electric Power, Shanghai 200090, China)[1]
(College of Information Engineering, Shanghai Maritime University, Shanghai 201303, China)[2]
(Institute of Energy and Environment Engineering, Shanghai University of Electric Power, Shanghai 200090, China)[3]

Abstract Feature selection is an essential step to perform cancer classification with DNA microarrays. Rough set theory has already been successfully applied to gene selection. To avoid losing information by discretization of continuous gene expression data in rough set theory, the theory of fuzzy rough sets is applied to gene selection. A fuzzy rough attribute reduction algorithm based on mutual information was proposed and applied to gene selection. The cost of computation of the algorithm is too high to be carried out if the number of the selected genes is large. This paper raised an approximate replacement of computation of the mutual information, from both maximum relevance and maximum significance. The novel method improves the efficiency and decreases the complexity. Extensive experiments were conducted on three public gene expression datasets. The experimental results confirm the efficiency and effectiveness of the algorithm.

Keywords Feature selection, Fuzzy rough sets, Mutual information, Gene expression data

1 引言

近年来,随着肿瘤基因表达谱技术的出现,肿瘤学的研究进入了一个全新的阶段。通过DNA芯片可以在一次实验中同时获得组织样本中成千上万个基因的表达水平,即肿瘤基因表达谱[1]。如何对肿瘤基因表达谱进行有效的分析,挖掘和发现其中蕴含的信息和知识,是当前生物信息学研究的重点课题[2,3]。肿瘤基因表达谱数据的一个显著特点是样本维数过高,每个样本都记录了组织细胞中所有可测基因的表达水平,而实际上只有少数基因才真正同样本分类有关,这些基因被称为分类特征基因。分类特征基因选取问题是肿瘤基因表达谱分析研究的核心内容,是建立有效分类模型的关键所在,同时也是发现肿瘤分类与分型的基因标记物及药物治疗潜在靶点的重要手段。

目前已有很多学者针对分类特征基因选取问题进行了研究[4-12],然而如何在表达谱成千上万个基因中有效地选出样本的分类特征基因,一直是肿瘤基因表达谱分析中的难点所在,仍有待深入研究。在模式识别中,基因通常被认为是特征,分类特征基因选取问题即为特征选择问题。特征选择方法一般分为3大类:过滤式(Filter)、封装式(Wrapper)和嵌入式(Embedded)[13]。过滤法与分类算法相互独立,其时间复杂度低,但没有考虑特征集与分类算法之间的关联和相互影

到稿日期:2012-09-09 返修日期:2012-12-22 本文受国家重点基础研究发展计划(973计划)子课题(2009CB219801),上海市教育委员会科研创新项目(12YZ140),上海高校青年教师培养资助计划(sdl11003),上海电力学院人才引进基金(K-2011-002)资助。
徐菲菲(1983-),女,博士,讲师,主要研究方向为模糊粗糙集理论、数据挖掘等,E-mail:Xufeifei1983@hotmail.com;魏 莱(1980-),男,博士,讲师,主要研究方向为高维仿生信息几何学、流形学习等;杜海洲(1980-),男,硕士,讲师,主要研究方向为数据挖掘等;王文欢(1980-),女,博士生,讲师,主要研究方向为节能诊断、数据挖掘。

响，因此精度不高。经典的过滤法包括 Fisher Ratio[14]、互信息[14]、Mahalanobis[15]和 t 值过滤[16]等。封装法则直接利用分类算法的训练准确率评估特征子集，因此可对不同的分类器选出最适应的近似最优的特征子集，但其计算量很大，在高维、高噪的数据中容易产生过拟合现象。嵌入式方法将特征选择算法本身作为组成部分嵌入到学习算法中，特征选择过程通过向逻辑公式表达式中加减特征来实现，其时间复杂度较低，但准确性不高。最典型的有决策树算法，如 Quinlan 的 ID3 和 C4.5[17,18]以及 Breiman 的 CART 算法。

基于封装法的分类特征基因选择方法可选择对于分类来说重要的基因组合，而不仅是对单个基因重要性进行排序选择。微阵列数据的特征高维性，使得用穷尽法从所有特征组合中找到最优组合成为 NP 难题。因此，所有的封装法都是指定分类器，以启发式算法逼近分类误差最小解来寻找优化的特征集合。从理论上讲，封装法能够提供比过滤法更好的分类精度[19]。2002 年 Guyon 等人结合支持向量机和递归特征消除方法提出的 SVM-RFE 方法在白血病和结肠癌数据的应用中取得了良好的基因选择效果[21]。粗糙集理论[22]作为一种新的软计算方法，能有效地分析和处理各种不精确、不一致、不完整的数据，通过属性约简方法能提取出与分类相关的特征子集。近年来，粗糙集理论凭借自己的独特优势，开始逐渐应用到生物信息学领域中[23]，在肿瘤分类特征基因选取方面取得了一些较好的结果[24-28]。然而，粗糙集处理的是离散化的数据，基因表达谱数据集却往往都是连续的。一种方法是将基因表达谱数据集先进行离散化[29-31]，但离散化过程必定会造成某种程度的信息损失。而模糊粗糙集[32]结合了模糊集[33]和粗糙集[22]两种理论的优点，将对等价类的精确划分转变为模糊划分，确定对象对每个模糊等价类的隶属度，从而避免了一定程度的信息丢失。利用模糊粗糙集方法对特征进行选取，能最大限度地保持原数据集的分类能力[34-38]。

特征选择的目的是在保持尽可能多的类别信息条件下选取最小的特征子集。互信息可以很好地表示两个随机变量之间的相关性，并且对噪声数据具有很好的鲁棒性[39]。同时，互信息对分类器也具有很好的鲁棒性，理论上说，无论分类器如何，互信息方法均可以选择出最优的特征子集[40]。目前，计算连续属性的互信息大多采用 Parzen 窗密度估计方法。文献[34]从信息论观点下的粗糙集引出信息论下的模糊粗糙集表示，为计算连续属性的互信息提供了一条新的思路。该文献还提出一种基于互信息的模糊粗糙集特征选择方法，并将其成功应用于分类特征基因提取中，然而当选择的基因较多的时候，每次计算互信息的代价很大。这些问题在实际应用中较为显著。假设每个基因平均模糊化产生 c 个类，选取 d 个基因，则计算一次互信息的复杂度为 c^d。当模糊等价类的个数较多时，条件互信息将很难被正确计算。因此，尽管基于互信息的特征选择方法在选取较少的特征时是有效的，但当模糊等价类和所选取的特征个数较多时，该方法是不合适的。

本文提出一种新的特征选择评判标准，并改进了文献[34]的算法，在 3 个常用的基因表达谱数据集急性白血病亚型(leukemia)、直肠癌(colon)和乳腺癌(Breast)上分别进行了实验，将标准化、模糊化后的数据利用文献[34]和本文提出的改进算法筛选出分类特征基因，然后采用 1NN、SVM 作为分类器分别进行分类测试，给出了详细的实验结果。结果表明，改进后的算法与原方法相比，大大提高了时间效率，并且两种方法所选择出的特征子集在分类准确率上相差不大。

本文第 2 节简要介绍了信息观下的粗糙集和模糊粗糙集理论中一些必要的知识；第 3 节介绍了文献[34]提出的基于互信息的模糊粗糙集属性约简算法，在此基础上，分析了新的互信息特征选择评判标准，给出了改进后的算法；第 4 节给出了我们在常用基因数据集急性白血病亚型(leukemia)、直肠癌(colon)和乳腺癌(Breast)上的实验结果，并作出了详细的分析。

2 信息观下的粗糙集和模糊粗糙集

1982 年波兰数学家 Pawlak 提出的粗糙集理论是一种有效的、新的数据处理方法，粗糙集理论认为知识是一种分类能力，从而可以在保持知识分类能力的基础上对其进行约简。但 Pawlak 对粗糙集的描述是建立在代数集合论上的，对于一些运算缺乏直观性，为此文献[41]从信息论的角度对粗糙集做出了新的描述。

2.1 粗糙集

文献[41]阐述了信息观点下的粗糙集理论，并且证明在一致决策表的情况下与 Pawlak 的代数观点下的粗糙集是等价的。

定义 1[41] 设 U 为一个论域，P,Q 为 U 上的两个等价关系(即知识)。P,Q 在 U 上导出的划分分别为 X,Y：$X=\{X_1,X_2,\cdots,X_n\}$，$Y=\{Y_1,Y_2,\cdots,Y_m\}$，则 P,Q 在 U 的子集组成的 σ-代数上定义的概率分布为：

$$[X;p]=\begin{bmatrix} X_1 & X_2 & \cdots & X_n \\ p(X_1) & p(X_2) & \cdots & p(X_n) \end{bmatrix}$$

$$[Y;p]=\begin{bmatrix} Y_1 & Y_2 & \cdots & Y_m \\ p(Y_1) & p(Y_2) & \cdots & p(Y_m) \end{bmatrix}$$

式中，$p(X_i)=\frac{|X_i|}{|U|}$，$i=1,2,\cdots,n$；$p(Y_j)=\frac{|Y_j|}{|U|}$，$j=1,2,\cdots,m$；符号 $|E|$ 表示集合 E 的基数。则知识 P 的熵 $H(P)$ 定义为：

$$H(P)=-\sum_{i=1}^{n}p(X_i)\log_2 p(X_i) \tag{1}$$

知识 Q 相对于知识 P 的条件熵 $H(Q|P)$ 定义为：

$$H(Q|P)=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log_2 p(Y_j|X_i) \tag{2}$$

利用上述知识表示方法，文献[42]提出了一种属性约简算法，该算法能在保持原数据集分类能力的基础上约简冗余的属性。但如上文所述，粗糙集处理的是离散化的数据，要处理属性值为连续值的基因表达谱数据集，必须先将数据离散化，这样就存在信息丢失。为此，我们提出用模糊粗糙集来处理连续属性值的基因表达谱数据集。

2.2 模糊粗糙集

模糊粗糙集理论是对粗糙集理论的推广，它将粗糙集中讨论的对象集合拓展为模糊集，并且将等价关系 R 转换为模糊等价关系 $\mathscr{R}$，扩大了粗糙集理论的应用范围，有着广泛的理论和应用价值。文献[34]对模糊粗糙集在信息观下进行表示。

定义 2 U 是非空有限对象集合，$U=\{x_1,x_2,\cdots,x_N\}$，模糊属性集 $\tilde{A}$ 由一族模糊属性 $\{\tilde{A}^1,\tilde{A}^2,\cdots,\tilde{A}^M,\tilde{A}_{M+1}\}$ 组成，

其中 $D=\{\tilde{A}^{M+1}\}$ 是模糊决策属性，其他为模糊条件属性 $C=\{\tilde{A}^1,\tilde{A}^2,\cdots,\tilde{A}^M\}$。每一个模糊属性可以将论域划分成 p_j 个模糊等价类，即 $F(\tilde{A}^j)=\{\tilde{F}_1^j,\tilde{F}_2^j,\cdots,\tilde{F}_{p_j}^j\}(j=1,2,\cdots,M+1)$，其中 $\tilde{F}_i^j(1\leqslant i\leqslant p_j)$ 为一模糊集。f 是一个 $U\times\tilde{A}$ 到属性值集合 V 上的一个映射，它表示每个对象在每个属性的每个模糊等价类上对应一个值，$V\in[0,1]$。由这样的论域与模糊属性集构成的二维信息表 $S=(U,\tilde{A}=C\cup D,V,f)$ 为模糊决策表。

定义 3 设模糊决策表 $S=(U,\tilde{A}=C\cup D,V,f)$，$P,Q$ 为模糊属性构成的模糊等价关系（也即知识），$U/IND(P)=\{\tilde{X}_1,\tilde{X}_2,\cdots,\tilde{X}_n\}$，$U/IND(Q)=\{\tilde{Y}_1,\tilde{Y}_2,\cdots,\tilde{Y}_m\}$，这里 $\forall\tilde{X}_i\in U/IND(P)$，$\tilde{Y}_j\in U/IND(Q)$ 都是论域 U 上的模糊集，知识 P 的熵定义为：

$$H(P)=-\sum_{i=1}^{n}p(\tilde{X}_i)\log_2 p(\tilde{X}_i)=-\sum_{i=1}^{n}\frac{\sum_{k=1}^{|U|}\mu_{\tilde{X}_i}(x_k)}{|U|}\log_2\frac{\sum_{k=1}^{|U|}\mu_{\tilde{X}_i}(x_k)}{|U|}\tag{3}$$

知识 Q 相对于知识 P 的条件熵 $H(Q|P)$ 定义为：

$$H(Q|P)=-\sum_{i=1}^{n}p(\tilde{X}_i)\sum_{j=1}^{m}p(\tilde{Y}_j|\tilde{X}_i)\log_2 p(\tilde{Y}_j|\tilde{X}_i)\tag{4}$$

其中，$U/IND(P)=\otimes U/IND\{\tilde{A}^j\},\tilde{A}^j\in P$，$U/IND(Q)=\otimes U/IND\{\tilde{A}^j\},\tilde{A}^j\in Q$，$\tilde{T}_1\otimes\tilde{T}_2=\{\tilde{X}\cap\tilde{Y}:\forall\tilde{X}\in\tilde{T}_1,\forall\tilde{Y}\in\tilde{T}_2,\tilde{X}\cap\tilde{Y}\neq\emptyset\}$。此外，$\mu(\cdot)$ 为模糊集的隶属度函数，且 $\mu_{\tilde{T}_1\cap\tilde{T}_2\cap\cdots\cap\tilde{T}_n}(x)=\min\{\mu_{\tilde{T}_1}(x),\mu_{\tilde{T}_2}(x),\cdots,\mu_{\tilde{T}_n}(x)\}$，$\tilde{T}_i$ 是 U 上的模糊集。

将互信息的概念引入到模糊粗糙集中，用于度量模糊决策表中模糊属性的相对重要性。

设模糊决策表 $S=(U,\tilde{A}=C\cup D,V,f)$，$\mathcal{R}$ 是模糊条件属性集合。那么，在 $\mathcal{R}$ 中添加一个模糊属性 $\tilde{A}^j$ 之后互信息的增量为：

$$I(\mathcal{R}\cup\{\tilde{A}^j\};D)-I(\mathcal{R};D)=H(D|\mathcal{R})-H(D|\mathcal{R}\cup\{\tilde{A}^j\})\tag{5}$$

定义 4 设模糊决策表 $S=(U,\tilde{A}=C\cup D,V,f)$，$\mathcal{R}$ 是模糊条件属性集合。则对于任意属性 $\tilde{A}^j\in C-\mathcal{R}$ 的重要性 $SGF(\tilde{A}^j,\mathcal{R},D)$ 定义为：

$$SGF(\tilde{A}^j,\mathcal{R},D)=I(\mathcal{R}\cup\{\tilde{A}^j\};D)-I(\mathcal{R};D)=H(D|\mathcal{R})-H(D|\mathcal{R}\cup\{\tilde{A}^j\})\tag{6}$$

若 $\mathcal{R}=\emptyset$，则 $SGF(\tilde{A}^j,\mathcal{R},D)$ 为 $SGF(\tilde{A}^j,D)=H(D)-H(D|\tilde{A}^j)=I(\tilde{A}^j;D)$，即为模糊属性 $\tilde{A}^j$ 与模糊决策属性 D 的互信息。$SGF(\tilde{A}^j,\mathcal{R},D)$ 的值越大，说明在已知 $\mathcal{R}$ 的条件下，模糊属性 $\tilde{A}^j$ 对于模糊决策属性 D 就越重要。

3 改进的基于互信息的模糊粗糙集特征选择算法

文献[34]提出了一种基于互信息的模糊粗糙集特征选择方法，避免了粗糙集离散化方法带来的信息损失，具体算法如下。

3.1 基于互信息的模糊粗糙集特征选择算法

基于互信息的模糊粗糙集特征选择算法是以 bottom-up 的方式求相对约简的。以空集为起点，依据上述定义的属性重要性，逐次选择最重要的属性添加到集合中，直到终止条件满足。

算法 1 MIBAFRAR（Mutual Information-Based Algorithm for Fuzzy-Rough Attribute Reduction）

Step1 计算模糊决策表中条件属性 C 与决策属性 D 的互信息 I(C;D)；

Step2 令 $\mathcal{R}=\emptyset$，对条件属性集 C−$\mathcal{R}$ 重复：

Step2.1 对每个属性 $\tilde{A}^j\in C-\mathcal{R}$，计算条件互信息 $I(\tilde{A}^j;D|\mathcal{R})$；

Step2.2 选择使条件互信息 $I(\tilde{A}^j;D|\mathcal{R})$ 最大的属性，记作 $\tilde{A}^j$（若同时有多个属性达到最大值，则从中选取一个相似类个数最少的属性作为 $\tilde{A}^j$）；并且 $\mathcal{R}\leftarrow\mathcal{R}\cup\{\tilde{A}^j\}$；

Step2.3 若 $I(C;D)=I(\mathcal{R};D)$，则终止；否则，转 Step2.1；

Step3 最后得到的 $\mathcal{R}$ 就是条件属性 C 相对于 D 的一个相对约简。

属性即为特征，相对约简即为所选取的特征子集。寻找最小知识相对约简是 NP-hard 问题，其复杂性主要是由模糊决策表中的属性组合引起的。对于 MIBAFRAR 算法而言，在最坏情况下，每次所考虑的属性数依次为 $n,n-1,\cdots,1$（n 为模糊决策表的模糊条件属性数），故总次数为 $n+(n-1)+\cdots+1=n(n+1)/2$。

因此，如果忽略对象数对计算时间的影响，那么，在最坏情况下，该算法能够在 $O(n^2)$ 时间复杂性内找到满意的约简。

3.2 一种基于互信息的模糊粗糙集快速特征选取方法

由于肿瘤基因表达谱数据的样本数较少，很难得到比较合理的模糊等价类；并且当已选择的基因较多的时候，每次计算互信息的代价很大。这些问题在实际应用中较为显著。如前文所述，假设每个基因平均模糊化产生 c 个类，选取 d 个基因，则计算一次互信息的复杂度为 c^d。当模糊等价类的个数较多时，条件互信息将很难被正确计算。因此，尽管基于互信息的特征选择方法在选取较少的特征时是有效的，但当模糊等价类和所选取的特征个数较多时，该方法是不合适的。

由于条件互信息的计算比较困难，为了使所选取的特征子集相对于决策具有最大的互信息，一种替代的方法是基于最大相关性的评价标准选取特征[43]。最大相关性是指采用所有选择的模糊属性 $\tilde{A}^j$ 与模糊决策属性 D 的互信息的平均值近似表示所选取的特征子集相对于决策的互信息，即

$$\max R(\mathcal{R},D),R=\frac{1}{|\mathcal{R}|}\sum_{\tilde{A}^j\in\mathcal{R}}I(\tilde{A}^j;D)\tag{7}$$

基于最大相关性所选取的特征很可能具有很大的冗余性，即特征之间的相关性可能很高。当两个特征相互依赖的程度很高时，删除其中一个特征对其区分能力的影响不大。因此，添加以下最大重要性的条件来选取相互不相关的特征：

$$\max S(\mathcal{R},D),\quad S=\frac{1}{|\mathcal{R}|(|\mathcal{R}|-1)}\sum_{\substack{\tilde{A}^i\neq\tilde{A}^j\in\mathcal{R}\\ i<j}}\{I(\tilde{A}^j;D|\tilde{A}^i)+I(\tilde{A}^i;D|\tilde{A}^j)\}\tag{8}$$

结合以上两个约束条件，定义算子 $\Phi(R,S)$ 使得 R 和 S 同时达到最大化：

$$\max\Phi(R,S),\Phi=R+S\tag{9}$$

在实际应用中[43,44]，采用递增的搜索方法寻找根据 $\Phi(\cdot)$ 定义的近似最优的特征子集。给定特征子集 $\mathcal{R}_{d-1}$（已选出 $d-1$ 个特征），我们的目的就是在剩下的特征集 $\{C-\mathcal{R}_{d-1}\}$ 中选取第 d 个特征（使得 $\Phi(\cdot)$ 最大），即满足以下条件：

$$\max_{\tilde{A}^j\in\{C-\mathcal{R}_{d-1}\}}\left[I(\tilde{A}^j;D)+\frac{1}{d-1}\sum_{\tilde{A}^i\in\mathcal{R}_{d-1}}I(\tilde{A}^j;D|\tilde{A}^i)\right]\tag{10}$$

根据上面的讨论，可以得到以下一些结论：

i. 仅最大化式(9)的第一项，即式(7) $\max R(\mathcal{R},D)$，只能

达到最大相关性。式(7)没有考虑到特征之间产生的对目标类 D 的共同作用。

ii. 仅最大化式(9)的第二项，即式(8) $\max S(\mathcal{R},D)$，等价于寻找互相独立的特征，不足以选出具有强区分能力的特征。

iii. 对比式(6)，式(10)可以避免计算多个特征下的条件互信息，仅需要计算两个特征相对于决策类 D 的互信息，可以很容易地得到结果并且使得结果更加精确，这也使得特征选择算法更加有效。

算法 2 基于互信息的模糊粗糙特征选择快速算法

Step1 令 $\mathcal{R}=\emptyset$，对条件属性集 C$-\mathcal{R}$ 重复：

Step1.1 对每个属性 $\tilde{A}^j\in C-\mathcal{R}$，及每个属性 $\tilde{A}^i\in\mathcal{R}$，计算条件互信息之和 $\sum_{\tilde{A}^i\in\mathcal{R}} I(\tilde{A}^j;D|\tilde{A}^i)]$；

Step1.2 选择使 $I(\tilde{A}^j;D)+\frac{1}{d-1}\sum_{\tilde{A}^i\in\mathcal{R}} I(\tilde{A}^j;D|\tilde{A}^i)$ 最大的属性，记作 $\tilde{A}^j$(若同时有多个属性达到最大值，则从中选取一个相似类个数最少的属性作为 $\tilde{A}^j$)；

Step1.3 若 $I(\mathcal{R};D)=I(\mathcal{R}\cup\tilde{A}^j;D)$，则终止；否则，转 Step1.1，$\mathcal{R}\Leftarrow\mathcal{R}\cup\{\tilde{A}^j\}$。

Step2 最后得到的 $\mathcal{R}$ 就是条件属性 C 相对于 D 的一个相对约简。

需要说明的是，文献[43]提出的 mRMR 算法通过最大化特征子集之间的相关性并且最小化其之间的冗余度来选取特征子集。然而，mRMR 算法对冗余度的衡量并没有考虑到类标签。而本文提出的相关性和重要性两个评价标准均基于类标签。因此，本文提出的算法比已有的 mRMR 方法具有更好的性能。文献[44]提出的 MRMS 算法是建立在代数集合论上的，对一些运算缺乏直观性，为此，本文从信息论的角度，结合模糊粗糙集提出一种基于互信息的模糊粗糙集快速特征提取方法。

4 实验结果与分析

4.1 肿瘤基因表达谱数据描述

基因表达谱是指利用 DNA 芯片所测定的组织样本中基因的表达水平值。本文分析的对象是基因表达谱数据集分析中常用的两个数据集 Leukemia、Colon。Leukemia[4] 是 Golub 等人公布的急性白血病基因表达谱数据集，下载地址 http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi。该数据集共有 72 个急性白血病样本，每个样本均含 7129 个基因的表达数据。其中 47 个样本被诊断为急性淋巴性白血病(acute lymphoblastic leukemia，ALL)，25 个被诊断为急性骨髓性白血病(acute myeloid leukemia，AML)。整个数据集被划分为训练集和测试集，如图 1 所示。

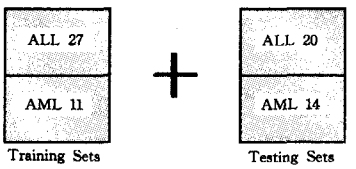

图 1 急性白血病基因表达谱数据集

直肠癌数据集(Colon Microarray)[20] 在 1999 年由 Alon 描述和在网上提供下载，可从网站 http://microarray.princefon.edn/oncology/affydafa/index.htm/下载。该数据集是在 DNA Microarray 数据中提取的。在提取数据之前，需要做前期处理，包括图象扫描、信噪对比、生物学意义上的归一化等等。我们用的数据中有 62 个样本和 2000 个基因表达数据(62 tissues×2000 genes expression values)。在这 62 个样本中有 22 个是正常人，标签为 Positive；40 个是直肠癌病人，标签为 Negative。我们的数据中有 2000 个基因表达数据。

乳腺癌数据集(Breast)[45] 中包含 7129 个基因表达数据和 49 个样本。这些样本中有 25 个是正常的，标签为 positive，其余 24 个样本属于乳腺癌病人，标签为 Negative。

表 1 3 个基因表达数据集的基本信息

数据集	对象个数	属性个数
Leukemia	72	7129
Colon cancer	62	2000
Breast	49	7129

本文以急性白血病的亚型和是否患有直肠癌、乳腺癌的分类为例，对肿瘤基因表达谱数据进行分析。对基于互信息的模糊粗糙集属性约简算法(算法 1)与改进的快速算法(算法 2)进行比较，分析的目标是在提高时间效率的同时，找出决定样本类别的一组分类特征基因，实现对 Leukemia 数据集中 AML 和 ALL 两类样本、Colon 数据集中 Positive 和 Negative 两类、Breast 数据集中 Positive 和 Negative 两类样本的准确分类。

4.2 实验过程

本实验在 windows 环境下采用 matlab 编写，机器配置为 CPU 奔腾 IV 2.4GHz，1MB cache，2GB 内存。

4.2.1 标准化

大量实验表明：基因表达数据在 log 空间里满足正态分布。因此，先将基因表达矩阵中的元素进行对数转换，使其满足正态分布。通过式 $x'_{ij}=\frac{x_{ij}-\overline{x_i}}{\sqrt{\frac{1}{N-1}}\sqrt{\sum_{j=1}^{N}(x_{ij}-\overline{x_i})^2}}$ 分别对每个基因的样本数据进行标准化，使每个基因上的样本满足均值为 0、标准差为 1 的标准正态分布。

4.2.2 t-检验(分类无关基因的过滤)

肿瘤基因表达谱数据集的一个显著特点是样本少、维数高，每个样本都记录了组织细胞中所有可测基因的表达水平。然而只有少数基因才包含了样本具体的类别信息，大部分基因与样本类别并不相关，其作为分类无关基因存在，被称为“无关基因”或“噪声基因”。

尽管大多数过滤法选择出的特征不如封装法或嵌入式方法，但由于后者的极高计算量使得过滤法广泛地应用于特征选择问题的预处理过程。因此一般可以先对基因表达谱数据进行过滤。就模式识别而言，样本数据分布差异较大的属性(参数)提供较多的样本分类信息。所以本文采用 t-检验，先选取分布差异较大的前 200 个基因，以提高实验的整体效率。

4.2.3 模糊化

如前文所述，对基因表达谱数据模糊化首先要对每个属性值聚类。本实验采用等频法，即根据预先设定的所需类别数 k 采用等频法将对象进行聚类。事实上，k 的选取对实验结果影响较大，一般 k 的个数为 2～8 个较好。

k 设定之后，则需要确定每个对象对每个属性的每个区间的隶属度。本实验选取常用的三角隶属度函数确定每个对象对每个属性的每个类的隶属度。三角隶属度函数确定方法如下：选择每个对象对每个属性的每个类的平均值作为三角

隶属度函数每个等价类的最高点，即纵轴为 1 的点，再选择相邻两个类，即较小类的最大值和较大类的最小值的中点作为纵轴为 0.5 的点，构造三角函数；然后根据基因表达谱数据的取值确定每个样本属于某个属性的某个类的隶属度，最后得到一张模糊决策表。

4.2.4 基于模糊粗糙集的基因选择

经过上述方法得到模糊决策表后，就可采用基于互信息的模糊粗糙集基因选择算法从中选取出肿瘤分类特征基因。终止条件为误差不超过 0.01。对 Leukemia(72 个对象)数据集的训练集(38 个样本)进行实验，模糊化区间数 k 分别取 2，3，4，5，6，7，8，将原模糊粗糙集属性约简方法(算法 1)与改进后的方法(算法 2)分别进行基因选择，记录运行时间及所选取的基因，如图 2、图 3 所示。

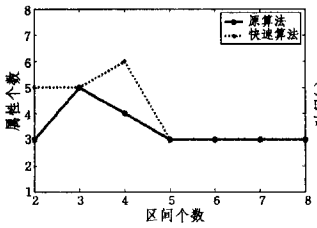

图 2 Leukemia 数据集模糊化区间数所对应选取的基因个数

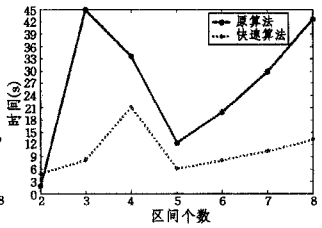

图 3 Leukemia 数据集模糊化区间数选取基因所需时间

由上图可看出，模糊化区间数与选取的基因个数无关，模糊化区间数越多，笛卡尔乘积对应的类也越多，但所需时间并非越多。对同一个算法来说，在选取基因数相同的条件下，一般随着模糊化区间数的增加，所需时间也增加。而选取的基因数越多，一般所需时间越多。由上图可知，改进的算法所需的时间远低于原方法所需时间。

将上述选取的基因分别用 1NN 与 SVM 作分类器对 38 个训练集进行训练，再将测试集的 34 个样本进行测试，得到的平均分类准确率及基因选择所需的平均时间如图 4 所示。

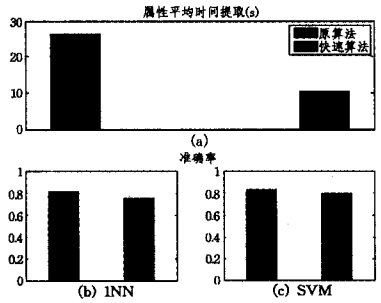

图 4 leukemia 数据集基因选择平均时间与分类准确率

实验过程中发现，当 $k=4$ 即模糊区间数为 4 时，选取 1NN 作分类器的原方法和改进方法的分类准确率均为最高，分别为 88.24%和 100%。当 $k=2$ 时，原方法选取的基因数为 3 个，$k=3$ 时，原方法选取的基因数为 5 个，$k=4$ 时，原方法选取的基因数为 4 个，但分类准确率均为 88.24%。此外，我们发现两种方法在 $k=7$ 与 $k=8$ 时所选取的基因均一样，此时选用 1NN 与 SVM 分类器的分类准确率均不高。实验发现，随着模糊化区间数的变化，改进的方法所选取的前两个基因与原方法选取的前两个基因均相同。

表 2 $k=4$ 时原方法所选取的基因

选取的基因	描述	1NN 分类准确率
M94556_at	SSBP Single-stranded DNA-binding protein	88.24%
M21494_at	CKM Creatine kinase, muscle	
D38548_at	KIAA0076 gene	
M55131_at	CFTR Cystic fibrosis conductance regulator	

表 3 $k=2$ 时原方法选取的基因

选取的基因	描述	1NN 分类准确率
U82671_cds2_at	HSP1-A gene extracted from Homo sapiens	88.24%
X59405_at	MCP Membrane cofactor protein muscle	
M55131_at	CFTR Cystic fibrosis conductance regulator	

表 4 $k=4$ 时改进的方法选取的基因

选取的基因	描述	1NN 分类准确率
M94556_at	SSBP Single-stranded DNA-binding protein	100%
M21494_at	CKM Creatine kinase, muscle	
M91670_at	Ubiquitin carrier protein (E2-EPF) mRNA	
M31169_s_at	GB DEF=Propionyl-CoA carboxylase beta-subunit gene, partial cds	
U12471_cds1_at	Thrombospondin-p50 gene extracted from Human thrombospondin-1 gene, partial cds	
X83705_s_at	GB DEF=C-sis proto-oncogene	

表 5 $k=2$ 时改进方法选取的基因

选取的基因	描述	1NN 分类准确率
U82671_cds2_at	HSP1-A gene extracted from Homo sapiens	85.29%
X59405_at	MCP Membrane cofactor protein muscle	
U12465_at	RPS11 Ribosomal protein S11	
HG2825-HT2949_at	Ret Transforming Gene	
M55131_at	CFTR Cystic fibrosis conductance regulator	

将我们提出的两个算法(算法 1、算法 2)与粗糙集属性约简算法在 leukemia(72 个对象)、colon(62 个对象)和 breast(49 个对象)数据集上进行实验，采用“留一法”(leave-one-out cross validation，LOOCV)进行分类准确率统计，并将结果与未约简的数据集的分类准确率进行对比，结果见表 6。

表 6 分类准确率

数据集	分类器	方法 1	方法 2	方法 3	方法 4
Leukemia	1NN	95.7%	93.1%	**95.8%**	94.8%
	SVM	97.1%	95.3%	**97.6%**	93.4%
Colon	1NN	**80.6%**	79%	79%	77.3%
	SVM	81.3%	83.3%	82.1%	**82.5%**
Breast	1NN	83.8%	84.2%	**85.7%**	85.3%
	SVM	85.3%	85.1%	83.7%	**86.9%**

表 7 对应方法名

方法 1	不进行属性约简
方法 2	基于互信息的粗糙集方法
方法 3	基于互信息的模糊粗糙集方法(算法 1)
方法 4	改进的互信息模糊粗糙集方法(算法 2)

4.2.5 分类结果与分析

由图 3 和图 4(a)可看出，改进的算法大大提高了原基于互信息的模糊粗糙集属性约简算法的效率，并且在 Leukemia

数据集上所提取的基因在 1NN 与 SVM 作分类器时平均准确率与原方法相差不大。由于模糊化区间数选取对分类结果影响较大，实验表明当区间数为 4 时，两种方法的分类准确率均为最高，尤其是改进算法的分类准确率达到 100%。当区间数较多时，分类准确率下降。改进方法与原方法所选取的前两个基因均相同。

理论上，未经过约简的基因分类能力强，准确率应该高。从实验结果看，若采用 1NN 作分类器，未约简的 Colon 基因分类准确率比其它方法均高一些。但在其它数据集上，未约简的基因分类准确率比其它方法略低，但相差不大，这可能是因为基因表达谱数据高于 40%的数据不是反映真实的值，是噪声数据。而无论是粗糙集方法还是模糊粗糙集方法，都对基因表达谱数据进行了基因选择，去除了噪声数据，所以比未约简数据的准确率高。由于模糊粗糙集方法对基因表达谱数据进行了模糊化处理，避免了粗糙集离散化过程中的信息丢失，因此算法具有很强的鲁棒性，从而分类能力比粗糙集选取的基因组略强。

由表 6 可以看出，无论粗糙集方法还是模糊粗糙集方法提取的基因，都能够保持整个基因数据集的分类能力，并且在采用 SVM 作分类器时，在 Colon 和 Breast 数据集上粗糙集方法提取的基因分类精度略高于模糊粗糙集提取的基因组。

改进后的方法所提取的基因组与原基于互信息的模糊粗糙集属性约简算法提取出的基因组相比，分类精度相差不大，采用 SVM 作分类器时，在 Colon 和 Breast 数据集上经过改进的模糊粗糙集方法提取的基因均比原方法提取的基因分类准确率有所提高。

对 Leukemia 数据集来说，采用留一法的平均分类准确率比训练集(38 个样本)训练后再用剩余 34 个测试样本测试的分类准确率高。

上述实验表明，无论粗糙集还是模糊粗糙集提取的基因，都能够保持整个基因数据集的分类能力，并且模糊粗糙集由于避免了粗糙集离散化过程的信息丢失，提取的特征基因分类精度优于粗糙集方法提取的基因。改进后的算法大大提高了基因选取的效率，并且所提取的基因组分类准确率与原方法提取的基因组分类准确率相差不大。尤其选用 SVM 作分类器时，新方法提取的特征基因能得到比原方法提取的基因组更高的分类准确率。

结束语 基于互信息的模糊粗糙集属性约简方法进行基因选择时避开了离散化过程，因此减少了信息损失，从而相对于基于粗糙集理论属性约简方法选择的基因有较好的准确率。然而，基于互信息的模糊粗糙集属性约简方法计算代价较高，当所选取的基因组个数较多时，该方法无法实现。因此，本文对互信息的计算从最大相关性与最大独立性两方面考虑，采用了近似替代的方法，大大减少了计算互信息的代价。实验结果表明了这一点，并且还表明新的基于互信息的模糊粗糙集属性约简方法选择出的基因与原方法得到的基因分类准确率大致相近，尤其选用 SVM 作分类器时，分类准确率比未约简的数据集还高。这说明新方法具有较优的时空效率以及很好的抗噪性能。然而，属性模糊化过程中不同的隶属度函数的选择以及聚类方法的选择对约简结果具有一定的影响，实验中选取的隶属度函数不一定是最优的，这也是我们下一步要研究的问题。

参 考 文 献

[1] Lander E S. Array of hope[J]. Nature Genetics, 1999, 21(Suppl): 3-4

[2] Ramaswamy S, Golub T R. DNA microarrays in clinical oncology[J]. Journal of Clinical Oncology, 2002, 20(7): 1932-1941

[3] Derisi J, Penland L, Brown P O, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer[J]. Nature Genetics, 1996, 14(4): 457-460

[4] Gloub 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

[5] Khan J, Wei J S, Ringner 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

[6] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2000, 46(13): 389-422

[7] Tibshirani R, Hastie T, Narasimhan B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression[J]. Proceedings of the National Academy of Science, 2002, 99(10): 6567-6572

[8] Fleuret F. Fast binary feature selection with conditional mutual information[J]. J. Mach. Learning Res, 2004(5): 1531-1555

[9] Hedenfalk I, Duggan D, Chen Y, et al. Gene-expression profiles in hereditary breast cancer[J]. New England Journal oMedicine, 2001, 344(8): 529-548

[10] Li X, Rao S, Zhang T, et al. An ensemble method for gene discovery based on DNA microarray data[J]. Science in China(Series C), 2004, 47(5): 396-405

[11] Tang E K, Suganthan P N, Yao X. Gene selection algorithms for microarray data based on least squares support vector machine [J]. BMC Bioinformatics, 2006(7)

[12] Cai Rui-chu, Hao Zhi-feng, Yang Xiao-wei, et al. An efficient gene selection algorithm based on mutual information[J]. Neurocomputing, 2009, 72: 991-999

[13] Kohavi R, John G H. Wrappers for feature subset selection[J]. Artif. Intell., 1997, 97(1/2): 273-324

[14] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. J. Mach. Learning Res., 2003(3): 1157-1182

[15] deSouza M C R, deCarvalho F A T, Tenorio C P. Twopartitional-methods for interval-valued data using mahalanobis distances [J]. Adv. Artif. Intell. Iberamia, 2004, 3315: 454-463

[16] Chang C F, Wai K M, Patterton H G. Calculating the statistical significance of physical clusters of co-regulated genes in the genome: the role of chromatin in domain-wide gene regulation[J]. Nucl. Acids Res., 2004, 32(5): 1798-1807

[17] Quinlan J R. Learning efficient classification procedures and their application to chess end games. Machine Learning: An artificial intelligence approach [M]. San Francisco, CA: Morgan Kaufmann, 1983: 463-482

[18] Quinlan J R. C4. 5: programs for machine learning[M]. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1993, 9(2): 132-136

(下转第 235 页)

邻近区间中采用预测决策方法进行故障的最终确认。实验证明，这种方法能够大大提高机械设备故障预测的准确率，对实际的机械维护有很强的指导意义。

参 考 文 献

[1] Su T, Dy J. A Deterministie Method for Initializing K-Means Clustering[C]// Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence. Boca Raton, Florida, 2009: 784-786

[2] 齐继阳，竺长安. 设备故障智能诊断方法的研究[J]. 仪器仪表学报，2009，27(10)：34-36

[3] Cooley R, Mobasher B, Srivastava J. Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns[R]. TR 97-021. University of Minnesota, Dept. of Computer Science, Minneapolis, 2011: 218-229

[4] 郭岩，白硕，于满泉. Web 使用信息挖掘综述[J]. 计算机科学，2005，32(1)：21-27

[5] Cooley R, Srivastava J. Grouping Web page references into transactions for mining world wide Web browsing patterns[C]// Proceedings of KDEX'97. NewportBeach, CA, USA, 1997: 2-7

[6] 田卫. RS-485 总线分支线短路故障检测技术[J]. 微电子学与计算机，2011，28(4)：116-117

[7] Mannila H, Toivonen H, Verkamo A I. Efficient algorithms for discovering association rules[C]//KDD-94: AAAI Workshop on Knowledge Discovery in Database. Seattle, Washington, 2009: 181-192

[8] Buchner A G, Mulvenna M D. Discovering Internet Marketing Intelligence Through Online Analytical Web Usage Mining[J]. SIGMOD Record, 2008, 27(4): 54-61

(上接第 221 页)

[19] Langleyand P. Selection of relevant features in machine learning [C]// Proceedings of A AAI Fall Symposium on Relevance. 1994

[20] Wang Y, Tetko I V, HallMark A, et al. Gene selection from microarray data for cancer classification-a machine learning approach[J]. Computation Biology and Chemistry, 2005, 29(1): 37-46

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

[22] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Science, 1982, 11: 341-356

[23] 李衍达，孙之荣. 生物信息学基因和蛋白质分析的实用指南[M]. 北京：清华大学出版社，2000

[24] Li Ding-fang, Zhang Wen. Gene selection using rough set theory [C]// Rough Sets and Knowledge Technology 2006 (RSKT 2006). Lecture Notes in Artificial Intelligence, Chongqing, 2006, 4062: 778-785

[25] Skowron A, Komorowski J, Pawlak Z, et al. Rough sets perspective on data and knowledge[M]. Handbook of data mining and knowledge discovery. NewYork: Oxford University Press, 2002

[26] Banerjee M, Mitra S, Banka H. Evolutinary-Rough Feature Selection in Gene Expression Data[J]. IEEE Transaction on Systems, Man, and Cybernetic, Part C: Application and Reviews, 2007, 37: 622-632

[27] Momin B F, Mitra S, Datta G R. Reduct Generation and Classification of Gene Expression Data[C]// Proceeding of First International Conference on Hybrid Information Technology (ICHICT06). 2006: 699-708

[28] Valdes J J, Barton A J. Gene discovery in leukemia revisited: a computational intelligence perspective[C]// Proceedings of the 17th International Conference on Industrial & Engineering Applications of Artificial International Conference & Expert Systems. Springer Verlag, 2004: 118-127

[29] 苗夺谦. 粗糙集理论中连续属性的离散化方法[J]. 自动化学报，2001，27(3)：296-302

[30] 权光日，等. 连续属性空间上的规则学习算法[J]. 软件学报，1999，10(11)：1225-1232

[31] 叶东毅，黄翠微，赵斌. 基于逼近精度的一个粗糙集属性约简算法[J]. 福州大学学报：自然科学版，2000，28(1)：7-10

[32] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets[J]. International Journal of General Systems, 1990, 17: 191-209

[33] Zadeh L A. 模糊集合，语言变量及模糊逻辑[M]. 北京：北京科学出版社，1982

[34] Xu F F, Miao D Q, Wei L. Fuzzy-rough attribute reduction via mutual information with an application to cancer classification [J]. Computers & Mathematics with Applications, 2009, 57(6): 1010-1017

[35] Bhatt R B, Gopal M. On fuzzy-rough sets approach to feature selection[J]. Pattern Recognition Letters, 2005, 26(7): 965-975

[36] Hu Qing-hua, An Shuang, Yu Da-ren. Soft fuzzy rough sets for robust feature evaluation and selection [J]. Information Sciences, 2010, 180(22): 4384-4400

[37] Jensen R, Shen Qiang. Fuzzy-rough data reduction with ant colony optimization[J]. Fuzzy Sets and Systems, 2005, 149(1): 5-20

[38] Chen De-gang, Zhao Su-yun. Local reduction of decision system with fuzzy rough sets. Fuzzy Sets and Systems, 2010, 161(13): 1871-1883

[39] Priness I, Maimon O, Ben-Gal I. Evaluation of gene-expression clustering via mutual information distance measure, BMC Bioinformatics, 2007, 8: 111

[40] Chow T W S, Huang D. Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information [J]. IEEE Trans. Neural Networks, 2005, 16(1): 213-224

[41] 苗夺谦，王珏. 粗集理论中概念与运算的信息表示[J]. 软件学报，1999，2：113-116

[42] 苗夺谦，胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展，1999，36(6)：681-684

[43] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of Max-Dependency, Max-Relevance, and Min-Redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238

[44] Maji P, Paul S. Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data [J]. Int. J. Approx. Reason, 2011, 52(3): 408-426

[45] West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles [C]// Proceedings of the National Academy of Science. USA 98, 2001(20): 11462-11467