基于模糊粗糙集的肿瘤分类特征基因选取

徐菲菲 苗夺谦 魏 莱
(同济大学计算机科学与技术系 上海 201804)
(同济大学嵌入式系统与服务计算教育部重点实验室 上海 201804)

摘 要 依据基因表达谱有效建立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。粗糙集理论作为一种新的软计算方法能够保持在原数据集的分类能力不变的基础上,对属性极大约简,从大量基因中找到对分类有效的基因。由于基因表达谱数据集的连续性,为了避免运用粗糙集方法所必需的离散化过程带来的信息丢失,尝试将模糊粗糙集应用于特征基因的选取,提出了基于互信息的模糊粗糙集属性约简算法,运用于基因表达谱数据集的基因选取。然后分别采用KNN和C5.0分类器进行特征基因分类性能进行检验。以急性白血病亚型(leukemia Microarray)和直肠癌(colon Microarray)分类特征基因选取为例进行实验,结果表明了上述方法的可行性和有效性。

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

Feature Selection for Cancer Classification Based on Fuzzy Rough Sets

XU Fei-fei MIAO Duo-qian Wei Lai
(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)
(Key Laboratory of Embedded System & Service Computing, Ministry of Education of China, Tongji University, Shanghai 201804, China)

Abstract Feature selection is an essential step to perform cancer classification with DNA microarrays, for there are a large number of genes from which to predict classes and a relatively small number of samples. Rough set theory is a tool for reducing redundancy in information systems, thus successful application of rough set to gene selection is of great significance. Fuzzy rough set was introduced to avoid losing information caused by discretization of continuous gene expression data which is needed in rough set theory. A novel gene selection method called IMIBAFRAR was improved to reduce the computation of mutual infor-mation. Then KNN and C5.0 were applied to validate the classification performance of the genes selected for distinguishing different tissue type. The work was applied to two public gene expression datasets: leukemia and colon. Experimental results show the selected genes don't reflect the classification ability of the original genes. Compared with the unreduced genes and the genes selected by classical rough set method, our method leads to significantly improved recognition accuracy. Meanwhile, computational complexity is reduced.

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

1 引言

随着大规模基因表达谱技术的推广,人们利用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[1]。依据DNA芯片测定的基因表达谱建立有效的分类模型,在分子水平上实现对肿瘤类型及亚型的准确识别,对肿瘤的诊断和治疗具有重要意义[2,3]。然而,数据集中的每个样本都记录了组织样本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与本分类相关。如何发现对样本分类而言至关重要的一组基因作为样本的分类特征基因,是建立有效分类模型的关键所在,同时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。

鉴于肿瘤分类特征基因选取的重要性,目前已经出现了针对该问题的大量研究文献[4-7]。粗糙集理论[8]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从数据中发现隐含的知识,揭示数据潜在的规律。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[9],在肿瘤分类特征基因选取方面取得了一些较好的结果[10]。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[11-13],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[14]结合了模糊集和粗糙集[8]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,

到稿日期:2008-04-17 本文受国家自然科学基金项目(60475019),国家自然科学基金重点项目(60534060),国家重点基础研究发展计划(973计划)(2003CB316902),2006年博士学科点专项科研基金(20060247039)资助。
徐菲菲(1983—),女,博士研究生,研究方向为粗糙集理论、数据挖掘,E-mail:xufeifei1983@hotmail.com;**苗夺谦**(1964—),男,博士,教授,博导,研究方向为粗糙集理论、数据挖掘等;**魏 莱**(1980—),男,博士研究生,研究方向为高维仿生信息几何学、流形学习等。

从而避免了一定程度的信息丢失。利用模糊粗糙集方法对属性值为连续值的基因进行选取，能最大限度地保持原数据集的分类能力。

本文基于模糊粗糙集理论的知识，在分析肿瘤基因表达谱特征的基础上，研究了肿瘤分类特征基因选取问题。首先，本文提出了基于互信息的模糊粗糙集属性约简算法，将其运用于基因表达谱数据，进行基因选取。然后，在常用的两个基因表达谱数据集急性白血病亚型(leukemia)、直肠癌(colon)上分别进行实验，将标准化、模糊化后的数据利用基于互信息的模糊粗糙集属性约简方法筛选出肿瘤分类特征基因。最后，采用 KNN，C5.0 作为分类器分别进行分类测试。结果表明，约简后的基因数据集不会降低分类器的分类能力，并且本方法提取的基因组比粗糙集提取的基因组更能实现对急性白血病两种亚型和是否患有直肠癌的准确分类。

本文其余部分组织如下：在第 2 节介绍信息观点下的粗糙集和模糊粗糙集概念；第 3 节具体描述基于模糊粗糙集的基因选择方法；第 4 节在急性白血病亚型(leukemia)和直肠癌(colon)两个基因表达谱数据集上进行了实验，并分析实验结果；最后得出结论。

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[16]从信息论的角度对粗糙集做出了新的描述。

2.1 粗糙集的概念

文献[16]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[16] 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

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

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

其中 $p(X_i)=\frac{|X_i|}{|U|}$，$i=1,2,\cdots,n$；$p(Y_j)=\frac{|Y_j|}{|U|}$，$j=1,2,\cdots,m$；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[17]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 模糊粗糙集的概念

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。从文献[16]得到启发，对模糊粗糙集在信息观下进行重新表示。

我们首先利用模糊隶属度函数，将粗糙集信息观下的知识定义进行重新定义。

符号假设与定义 1 相同，定义两个隶属度函数：

$$\mu_{X_i}(x_k)=\begin{cases}1, & x_k\in X_i\\ 0, & x_k\notin X_i\end{cases},\mu_{Y_j}(x_k)=\begin{cases}1, & x_k\in Y_j\\ 0, & x_k\notin Y_j\end{cases}$$

那么，$p(X_i)=\frac{|X_i|}{|U|}$ 就可以表示为 $p(X_i)=\frac{\sum_{k=1}^{|U|}\mu_{X_i}(x_k)}{|U|}$，$i=1,2,\cdots,n$；同样有 $p(Y_j)=\frac{\sum_{k=1}^{|U|}\mu_{Y_j}(x_k)}{|U|}$，$j=1,2,\cdots,m$。因此，式(1)就可以写成：

$$H(P)=-\sum_{i=1}^{n}p(X_i)\log p(X_i)=-\sum_{i=1}^{n}\frac{\sum_{k=1}^{|U|}\mu_{X_i}(x_k)}{|U|}\log\frac{\sum_{k=1}^{|U|}\mu_{X_i}(x_k)}{|U|}$$

式(2)同样可以表示为

$$\begin{aligned}H(Q|P)&=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log p(Y_j|X_i)\\&=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}\frac{p(Y_j\cap X_i)}{p(X_i)}\log\frac{p(Y_j\cap X_i)}{p(X_i)}\\&=-\sum_{i=1}^{n}\frac{\sum_{k=1}^{|U|}\mu_{X_i}(x_k)}{|U|}\sum_{i=1}^{m}\frac{\sum_{k=1}^{|U|}\mu_{X_i\cap Y_j}(x_k)}{\sum_{k=1}^{|U|}\mu_{X_i}(x_k)}\log\frac{\sum_{k=1}^{|U|}\mu_{X_i\cap Y_j}(x_k)}{\sum_{k=1}^{|U|}\mu_{X_i}(x_k)}\end{aligned}$$

有了上述定义，我们就可以将粗糙集的信息观推广到模糊粗糙集下。

先给出模糊决策表的定义。

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 $\tilde{A}$ 是由一族模糊属性$\{\tilde{A}^1,\tilde{A}^2,\cdots,\tilde{A}^M,\tilde{A}^{M+1}\}$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，$V\in[0,1]$。我们称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。

然后，给出模糊决策表中知识的信息熵定义。

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，则我们定义知识 P 的熵为

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

$$\log_2 \frac{\sum_{k=1}^{|U|} \mu_{\tilde{X}_i}(x_k)}{|U|} \tag{3}$$

知识 Q 相对于知识 P 的条件熵 $H(Q|P)$ 定义为

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

其中 $U/IND(P)=\otimes U/IND\{\tilde{A}^j\},\tilde{A}^j\in P, U/IND(Q)=\otimes U/IND\{\tilde{A}^j\},\tilde{A}^j\in Q$。我们定义 $\tilde{T}_1\otimes\tilde{T}_2=\{\tilde{X}\cap\tilde{Y}:\forall\tilde{X}\in\tilde{T}_1,\forall\tilde{Y}\in\tilde{T}_2,\tilde{X}\cap\tilde{Y}\neq\emptyset\}$。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。

再将互信息的概念引入到模糊粗糙集中，用其来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 $SGF(\tilde{A}^j,\mathcal{R},D)$ 即 $SGF(\tilde{A}^j,D)=H(D)-H(D|\tilde{A}^j)=I(\tilde{A}^j;D)$ 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。

有了以上一些基本概念，我们就可以给出基于互信息的模糊粗糙集知识相对约简(MIBAFRAR)算法。它是以 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，直到满足终止条件。

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：

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

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 $\mathcal{R}\Leftarrow\mathcal{R}\cup\{\tilde{A}^j\}$；

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

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。计算机处理时，只要 $|I(\mathcal{R};D)-I(\mathcal{R}\cup\tilde{A}^j;D)|$ 的值小于某一阈值，即视为相等。如 $|I(\mathcal{R};D)-I(\mathcal{R}\cup\tilde{A}^j;D)|\leqslant 10^{-3}$，则算法终止，得到 $\mathcal{R}$ 所需要的相对约简。

3 基于模糊粗糙集的基因选择方法

本文尝试用上述基于互信息的模糊粗糙集属性约简算法来进行基因选择，这样可以在一定程度上避免粗糙集离散化所带来的信息损失，具体方法如下。

3.1 模糊化

在模糊粗糙集中，对象属性值可以不必离散化，但需要确定的是对象相对每一属性每一模糊等价类的隶属度，为此我们需要确定每一属性的初始分类。

本文采用代表点算法确定所有对象在各基因的初始分类。具体算法描述如下(ξ 是给定的阈值)：

代表点算法(Leader Algorithm)：

Step1 选择某属性的第一个对象作为初始代表点；

Step2 对该属性的所有对象，

(a)从所有已有的代表点中选择离该对象 CP_i 最近的代表点 L_j；

(b)如果 $D(CP_i,L_j)<\xi$，则将 CP_i 赋给由 L_j 代表的类，否则添加 CP_i 作为新的代表点。

接下来就是确定对象相对每个模糊等价类的隶属度。隶属度函数通常有三角函数、梯形函数、正态分布函数等。

图 1 表示用三角隶属度函数确定对象相对某属性的各模糊等价类的隶属度。在属性模糊化后，我们就可以利用上述的模糊粗糙集属性约简算法来进行基因选择。

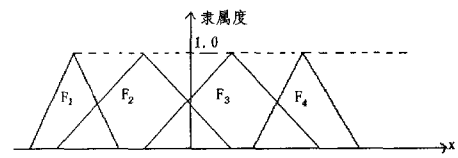

图 1 某基因的等价类的隶属度函数示意图

F_i 表示某一模糊等价类。

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 2 所示。

直肠癌数据集[20](colon Microarray)在 1999 年由 Alon 描述和在网上提供下载,可从下列网站下载 http://microarray. princefon. edn/oncology/affydafa/index. htm/。该数据集是通过在 DNA Microarray 数据中提取所得到的结果。在提取数据之前,需要做前期处理,包括图像扫描、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 x 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。

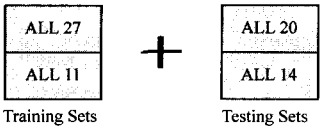

图 2 急性白血病基因表达谱数据集

本文以急性白血病的亚型和是否患有直肠癌的分类为例,对肿瘤基因表达谱数据进行分析。分析的目标是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* 两类样本以及 *colon* 数据集中 *Positive* 和 *Negative* 两类样本的准确分类。

4.2 实验过程

本文的实验总流程图如图 3 所示。

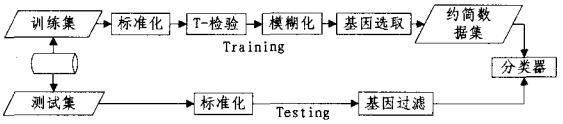

图 3 实验总流程图

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

$$x_{ij}' = \frac{x_{ij} - \overline{x_i}}{\sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(x_{ij} - \overline{x_i})^2}}$$

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,作为分类无关基因存在,称为"无关基因"或"噪声基因"。

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,两个类别中样本数据分布较大差异提供样本的分类信息。所以本文采用 t 检验,先选取分布差异较大的前 50 个基因,提高实验的整体效率。

4.2.3 模糊化

如前所述,对基因表达谱数据模糊化首先要聚类。聚类所采用的代表点算法需要预先设定阈值 ξ。事实上,ξ 的选取对实验结果影响较大,一般使所聚的类个数为 2~8 个较好。本实验选取阈值 ξ 的方法如下:先对每个基因上的样本数据进行排序,两两做差再求和,记为 ω,取 $\xi=5\omega$(当 $\xi=5\omega$ 时,一般可得到 6~8 个类)。

聚类之后,则需要确定每个对象对每个属性的每个类的隶属度。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。再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

经过上述方法得到的模糊决策表则可采用改进的模糊粗糙集基因选择算法,从而选取出肿瘤分类特征基因。由粗糙集属性约简算法和基于模糊粗糙集属性约简算法分别对 leukemia(72 个对象)、colon(62 个对象)数据集提取出的基因组如表 1、表 2 所列。

表 1 leukemia 数据集提取出的特征基因

粗糙集方法		模糊粗糙集方法	
基因名	描述	基因名	描述
M84526_at	DF D component of complement (adipsin)	X17042_at	PRG1 Proteoglycan 1, secretory granule
M89957_at	IGB Immunoglobulin-associated beta (B29)	X69111_at	ID3 Inhibitor of DNA binding 3, dominant negative helix-loop-helix protein
M11722_at	Terminal transferase mRNA	U77948_at	KAI1 Kangai 1
J05243_at	SPTAN1 Spectrin, alpha, non-erythrocytic 1(alpha-fodrin)	M23197_at	CD33 CD33 antigen (differentiation antigen)

表 2 colon 数据集提取出的特征基因

基因名	描述	基因名	描述
X63629	H. sapiens mRNA for p cadherin.	T71025	Human (HUMAN)
J05032	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA	R34698	INTERFERON-INDUCIBLE PROTEIN 9-27 (HUMAN)
H08393	COLLAGEN ALPHA 2 (XI) CHAIN (Homo sapiens)	L11706	Human hormone-sensitive lipase (LIPE) gene, complete cds
U32519	Human GAP SH3 binding protein mRNA, complete cds.	U05040	Human FUSE binding protein mRNA, complete cds
M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6.	R08183	Q04984 10 KD HEAT SHOCK PROTEIN, MITOCHONDRIAL
U09564	Human serine kinase mRNA, complete cds.	T70062	Human nuclear factor NF45 mRNA, complete cds

由表 1 和表 2 可知,粗糙集和模糊粗糙集在 leukemia 和 colon 数据集提取出的特征基因个数均相同,但所提取的基因完全不同。

4.2.5 分类结果与分析

实验结果需要比较两方面:一是选择的基因个数;二是分类能力。两种方法选择的基因个数相同,因此比较两组基因的分类能力。分别用 KNN,C5.0 作分类器进行实验,结果如表 3、表 4 所列。

表 3　leukemia 数据集的特征基因分类结果

分类器	未约简各类准确率		未约简总体	粗糙集各类准确率		粗糙集总体	模糊粗糙集各类准确率		模糊粗糙集总
	ALL	AML	分类准确率	ALL	AML	分类准确率	ALL	AML	体分类准确率
KNN	95.7%	100%	97.2%	95.7%	88.2%	93.1%	93.6%	96%	95.8%
C5.0	95.7%	100%	97.2%	97.9%	88.0%	94.4%	100%	96%	98.6%

表 4　colon 数据集的特征基因分类结果

分类器	未约简各类准确率		未约简总体	粗糙集各类准确率		粗糙集总体	模糊粗糙集各类准确率		模糊粗糙集总
	Neg	Pos	分类准确率	Neg	Pos	分类准确率	Neg	Pos	体分类准确率
KNN	90%	63.6%	80.6%	82.9%	72.6%	79.0%	80%	77.3%	79%
C5.0	95%	72.7%	87.1%	95.0%	81.8%	90.3%	92.5%	90.9%	91.9%

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 KNN 作分类器，则模糊粗糙集提取的基因分类准确率均比未约简的基因分类准确率略低，但相差不大，同时高于粗糙集方法提取的基因。这说明模糊粗糙集方法提取的基因能够保持整个基因数据集的分类能力，并且在采用 KNN 作分类器时由于避免了粗糙集离散化过程中的信息丢失，分类精度优于粗糙集方法提取的基因。

而采用 C5.0 作分类器时，经过模糊粗糙集提取后的基因均比未约简的基因分类准确率有所提高。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，故算法具有很强的鲁棒性，从而分类能力比未约简基因组更强。由表 3 可知，采用粗糙集提取 *leukemia* 数据集的特征基因比未约简的基因分类准确率有所降低，而表 4 表明在 *colon* 数据集选出的基因分类准确率比未约简基因略高，但低于用模糊粗糙集提取的基因。因此，模糊粗糙集提取的肿瘤分类特征基因用 C5.0 作分类器时，比原未经过约简的基因组和用粗糙集提取的特征基因具有更高的分类准确率。

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。尤其选用 C5.0 作分类器时，模糊粗糙集提取的特征基因能得到比整个基因数据集更高的分类准确率。

结束语　本文应用基于模糊粗糙集理论的属性约简方法进行基因选择，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（两种方法选择的基因个数相同）有更好的准确率，实验结果表明了这一点。并且实验结果还表明，经过模糊粗糙集选择出的基因与原未约简的基因分类准确率大致相近，在使用同类分类器时，提高分类效率的同时几乎不影响原数据集的分类准确率。尤其选用 C5.0 作分类器时，分类准确率比原数据集还高，这说明模糊粗糙集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. Nature Genetics, 1999, 21(Suppl): 3-4

[2] Ramaswamy S, Gloub T R. DNA microarrays in clinical oncology. 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. 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. 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. 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. 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 // Proceedings of the National Academy of Science. 2002, 99(10): 6567-6572

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

[9] Baxevanis A D, Ouellette B F F. Bioinformatics —— A Practical Guide to the Analysis of Genes and Proteins. Tsinghua University Press, 2000

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

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

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

[13] 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. 福州大学学报: 自然科学版, 2000, 28(1): 7-10

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

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

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

[17] Xu Feifei, Miao Duoqian, Wei Lai. An approach for fuzzy-rough attributes reduction via mutual information // The 4th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'07). Haikou, August 2007 (Accepted)

[18] Alon U, Barkai N, Notterman D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays // Proc. Natl. Acad. Sci. USA. 1999, 96: 6745-6750