

基于相容粗糙集的改进的基因特征选择方法

焦娜

(华东政法大学信息科学与技术系 上海 201620)

摘要 在基因表达数据中,有效的基因选择方法是癌症基因数据研究的重要内容。粗糙集是一个去掉冗余特征的有效工具。由于基因表达数据的连续性,为了避免运用粗糙集方法所必须的离散化过程带来的信息丢失,将相容粗糙集应用于基因的特征选择,提出基于相容粗糙集的基因特征选择方法,并在此方法基础上进一步对粗糙集的边界域进行研究,提出了基于相容粗糙集的改进的基因特征选择方法。在两个标准的基因表达数据上进行实验,结果表明与传统的基因特征选择方法相比,所提方法能够有效提高分类精度。

关键词 粗糙集,相容关系,基因特征选择,基因表达数据,癌症分类

Evolutionary Gene Selection Based on Tolerance Rough Set Theory

JIAO Na

(Department of Information Science and Technology, East China University of Political Science and Law, Shanghai 201620, China)

Abstract Gene selection is to select the most informative genes from the whole gene set, which is a key step of the discriminant analysis of microarray data. Rough set theory is an efficient mathematical tool for further reducing redundancy. The main limitation of traditional rough set theory is the lack of effective methods for dealing with real-valued data. However, gene expression data sets are always continuous. This has been addressed by employing discretization methods, which may result in information loss. This paper investigates one approach combining feature ranking together with features selection based on tolerance rough set theory. Moreover, this paper explores the other method which can utilize the information contained within the boundary region to improve classification accuracy in gene expression data. Compared with gene selection algorithm based on rough set theory, the proposed methods are more effective for selecting high discriminative genes in cancer classification task.

Keywords Rough set theory, Tolerance relation, Gene selection, Gene expression data, Cancer classification

1 引言

DNA 微阵列技术的发展为疾病的诊断、新药物的研制和新疗法的探索带来一场革命。DNA 微阵列的广泛应用提供了海量的基因表达数据,癌症基因表达数据分析成为基因表达数据分析的重要研究内容之一。癌症基因表达数据分析使得人们能够在正常和疾病两种状态下进行基因表达状况的比较研究,从而帮助人们更好地进行癌症诊断与分类、病理分析、对“因”下药和癌症防治等相关课题的研究。尤其是在恶性肿瘤的癌前诊断方面,有助于发现新的癌症类别,或者把已知的恶性肿瘤归类到某个已知的分类中。基因表达数据主要有两个特点:1)基因数量大(从几千个到几万个);2)样本数量少(从几个到几十个,最多几百个)。

在 DNA 微阵列实验中,虽然基因的数目成千上万,但实际上影响样本分类的,往往只有很少一部分的关键基因,其他的基因往往是不相关的或是冗余的。过多的基因会导致噪声的增加,增加分类算法的复杂性,并影响到分类的效果,而且与分类无关的基因将会干扰相关基因的信息。因此通过微阵列数据找到在不同类别中有显著不同表达的基因子集。最常

用的差异表达基因识别方法主要有两类:基于 filter 的选择方法和基于 wrapper 的选择方法。基于 filter 的选择方法^[1]常常使用数据本身的内在特性作为评价基因的准则对样本所包含的全体基因进行筛选,可有效降低冗余。t-检验和 F-检验是两种有效的选择方法。而基于 wrapper 的选择方法^[2]则根据分类器的某种性能来评价基因或基因子集的重要性,但 wrapper 方法的主要缺点是在基因的选择过程中需要反复地调用分类算法以评估基因的分类性能,结果往往造成极高的计算量。

通过过滤算法(filter)选择出来的若干个基因可能具有较强的相关性,粗糙集的属性约简能删除这种冗余,能用更少的属性集合来有效区分决策表中的对象。自 1982 年 Pawlak 提出粗糙集^[3-15]以来,它已经被广泛应用在特征选择和分类领域中^[3,4]。很多研究工作者致力于使用粗糙集理论进行属性约简的研究^[5-15]。然而,粗糙集处理的是离散化的数据,基因表达数据却往往都是连续的。有一种方法是将基因表达数据先进行离散化^[16,17],但离散化过程必定会造成某种程度的信息损失。而相容粗糙集则放宽了等价类的严格划分条件,从而避免了一定程度的信息丢失。利用相容粗糙集方法对属性

本文受上海高校青年教师培养资助计划(hdzf10008),华东政法大学科研项目(11H2K034)资助。

焦娜(1977—),女,博士,讲师,主要研究方向为人工智能、模式识别、粗糙集、数据挖掘等, E-mail: zdx_jn@163.com。

值为连续值的基因进行选取,能最大限度地保持数据集的分类能力。

本文根据相容度,综合相容度、相容关系和新的属性依赖度提出两种基于相容粗糙集基因特征选择方法。这两种方法主要由两个部分组成:第一步用 t -检验的过滤算法计算每个基因的分值,选择评分靠前的若干个基因组成条件属性集合;第二步用相容粗糙集对选择出的基因进行属性约简。提出的第二种基于相容粗糙集的改进的基因特征选择方法在第一种基于相容粗糙集的基因特征选择方法基础上,引入了边界域的距离函数,该函数对边界域中的不确定信息进行进一步计算。然后,在两个标准的基因表达数据上进行实验。最后,采用 KNN 和 C5.0 作为分类器分别进行分类测试。结果表明,本文提出的方法提取的基因组比粗糙集提取的基因组更能准确分类。

2 基于粗糙集的基因特征选择方法

2.1 基本概念

基因表达数据中有一个类别属性,因此可以把基因表达数据看成粗糙集里面的决策表。

定义 1 决策表 $T=(U, C \cup D, V, f)$, 其中 U 为非空有限论域; C 为条件属性集, D 为决策属性集; $V = \bigcup_{a \in C \cup D} V_a$ 为属性值域, 其中 V_a 为属性 a 的值域; $f: U \times (C \cup D) \rightarrow V$ 为信息函数。

定义 2 给定决策表 $T=(U, C \cup D, V, f)$, $B \subseteq C \cup D$, 论域 U 关于不可分辨关系 B 定义为:

$$IND(B) = \{(x, y) \in U \times U \mid \forall b \in B, b(x) = b(y)\} \quad (1)$$

显然, $IND(B)$ 是一个等价关系。

定义 3 给定决策表 $T=(U, C \cup D, V, f)$, $B \subseteq C \cup D$, 论域 U 关于不可分辨关系 B 的划分定义为:

$$U/IND(B) = \{[x]_B \mid x \in U\} \quad (2)$$

式中, $[x]_B$ 表示 U 中所有与 x 在关系 $IND(B)$ 下是等价的元素构成的集合。

定义 4 给定决策表 $T=(U, C \cup D, V, f)$, $X \subseteq U$, $B \subseteq C$, X 关于 B 的上、下近似集分别记为: $\bar{B}X, \underline{B}X$:

$$\underline{B}X = \bigcup \{[x]_B \mid [x]_B \subseteq X\} \quad (3)$$

$$\bar{B}X = \bigcup \{[x]_B \mid [x]_B \cap X \neq \emptyset\} \quad (4)$$

定义 5 给定决策表 $T=(U, C \cup D, V, f)$, $X \subseteq U$, $B \subseteq C$, 决策属性集 D 相对条件属性集 B 的正区域、负区域和边界域分别定义为:

$$POS_B(D) = \bigcup_{X \in U/IND(D)} \underline{B}X \quad (5)$$

$$NEG_B(D) = U - \bigcup_{X \in U/IND(D)} \bar{B}X \quad (6)$$

$$BND_B(D) = \bigcup_{X \in U/IND(D)} \bar{B}X - \bigcup_{X \in U/IND(D)} \underline{B}X \quad (7)$$

定义 6 给定决策表 $T=(U, C \cup D, V, f)$, 决策属性集 D 相对条件属性集 B 的分类质量, 即依赖度定义为:

$$\gamma_B(D) = |POS_B(D)| / |U| \quad (8)$$

2.2 t -检验

基因表达数据的特点是样本少、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, 与分类无关的基因, 称为“无关基因”或“噪声基因”。

因此, 一般可以先对基因表达数据进行过滤。在两个类别中, 分布差异较大的样本数据能够提供样本的分类信息。

所以本文采用 t -检验^[18,19], 先选取分布差异较大的若干个基因。对于二元分类问题, 有如下的 t -统计。

定义 7 给定基因 a , t -检验定义为:

$$t(a) = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \quad (9)$$

式中, n_i, μ_i 和 σ_i 分别为第 i ($i=1, 2$) 类的样本数、基因 a 的均值和标准方差。

2.3 基于粗糙集的基因特征选择方法

基于粗糙集的基因特征选择方法是由 t -检验及粗糙集属性约简^[15]两部分组成的。首先, 通过 t -检验方法选出评分靠前的 n 个基因组成条件属性集合; 然后对所有选出的属性进行离散化; 最后, 对每一个属性进行判断, 删除掉冗余的属性, 直到满足终止条件, 即得到一个约简。

基于粗糙集的基因特征选择方法(GSRS)的具体步骤如下。

算法 1 基于粗糙集的基因特征选择算法(GSRS)

1. 计算每个基因的 t -检验值, 选出评分靠前的 n 个基因组成条件属性集合 C 。
2. 离散化属性集 C
3. Set $P=C$
4. do
5. for each $a \in P$
6. if $\gamma_{P-\{a\}}(D) = \gamma_C(D)$
7. $P = P - \{a\}$
8. until $\gamma_{(P-\{a\})}(D) < \gamma_{(C)}(D)$
9. return P

值得提出的是, 基于粗糙集的基因特征选择算法(GSRS)并不是一个完备的属性约简算法, 但是在大多数情况下确实能够有效地得到决策表的约简。

3 基于相容粗糙集的基因特征选择方法

为了避免因粗糙集离散化所带来的信息损失, 本文放松了等价关系的严格定义, 首先, 给出在特定属性上对象间相容度的定义; 然后, 扩展到属性子集上对象间综合相容度的定义; 最后, 在此基础上给出了相容关系定义。同时, 上下近似、正区域、负区域、边界域及依赖度都被重新定义。

3.1 相容度

定义 8^[7,9] 给定决策表 $T=(U, C \cup D, V, f)$, $a \in C \cup D$, $x, y \in U$, 任意两个对象 x, y 关于属性 a 的相容度定义为:

$$S_a(x, y) = 1 - \frac{|a(x) - a(y)|}{|a_{\max} - a_{\min}|} \quad (10)$$

式中, $a_{\max}$ 和 $a_{\min}$ 分别是属性 a 的最大值和最小值。对于属性子集, 综合属性相容度有如下定义。

定义 9 给定决策表 $T=(U, C \cup D, V, f)$, $B \subseteq C \cup D$, $x, y \in U$, 任意两个对象 x, y 关于属性子集 B 的综合相容度定义为:

$$S_{B,\tau} = \left\{ (x, y) \mid \frac{\sum_{a \in B} S_a(x, y)}{|B|} \geq \tau \right\} \quad (11)$$

式中, τ ($0 < \tau \leq 1$) 是一个阈值。

定义 10 给定决策表 $T=(U, C \cup D, V, f)$, $B \subseteq C \cup D$, $0 < \tau \leq 1$, 论域 U 关于 B 的 τ 相容关系定义为:

$$S_{B,\tau}(x) = \{y \mid x, y \in U, (x, y) \in S_{B,\tau}\} \quad (12)$$

定义 11 给定决策表 $T=(U, C \cup D, V, f)$, $X \subseteq U$, $B \subseteq$

$C, 0 < \tau \leq 1$, X 关于 B 的 τ 上、下近似集分别记为: $\overline{B}_\tau X, \underline{B}_\tau X$:

$$\underline{B}_\tau X = \{x | S_{B,\tau}(x) \subseteq X\} \quad (13)$$

$$\overline{B}_\tau X = \{x | S_{B,\tau}(x) \cap X \neq \emptyset\} \quad (14)$$

$\langle \underline{B}_\tau X, \overline{B}_\tau X \rangle$ 为相容粗糙集, 相应的正区域、负区域、边界域及依赖度有如下定义。

定义 12 给定决策表 $T = (U, C \cup D, V, f)$, $X \subseteq U, B \subseteq C, 0 < \tau \leq 1$, 决策属性集 D 相对条件属性集 B 的 τ 正区域、负区域和边界域分别定义为:

$$POS_{B,\tau}(D) = \bigcup_{X \in U/S_{D,\tau}} \underline{B}_\tau X \quad (15)$$

$$NEG_{B,\tau}(D) = U - \bigcup_{X \in U/S_{D,\tau}} \overline{B}_\tau X \quad (16)$$

$$BND_{B,\tau}(D) = \bigcup_{X \in U/S_{D,\tau}} \overline{B}_\tau X - \bigcup_{X \in U/S_{D,\tau}} \underline{B}_\tau X \quad (17)$$

定义 13 给定决策表 $T = (U, C \cup D, V, f)$, $B \subseteq C, 0 < \tau \leq 1$, 决策属性集 D 相对条件属性集 B 的 τ 分类质量, 即 τ 依赖度定义为:

$$\gamma_{B,\tau}(D) = |POS_{B,\tau}(D)| / |U| \quad (18)$$

相容粗糙集属性约简方法是通过 $\gamma_{B,\tau}(D)$ 来度量属性子集的重要性。

3.2 基于相容粗糙集的基因特征选择方法

基于相容粗糙集的基因特征选择方法同样是由 t -检验及相容粗糙集属性约简两部分组成的。从 t -检验过滤出评分靠前的 n 个基因组成条件属性集合。然后, 从所有条件属性集合开始逐一删除冗余的属性, 直到满足终止条件, 即得到一个约简。

基于相容粗糙集的基因特征选择方法(GSTRS)的具体步骤如下。

算法 2 基于相容粗糙集的基因特征选择算法(GSTRS)

1. 计算每个基因的 t -检验值, 选出评分靠前的 n 个基因组成条件属性集合 C
2. Set $P = C$
3. do
4. for each $a \in P$
5. if $\gamma_{P-\{a\},\tau}(D) = \gamma_{C,\tau}(D)$
6. $P = P - \{a\}$
7. until $\gamma_{P-\{a\},\tau}(D) < \gamma_{C,\tau}(D)$
8. return P

4 基于相容粗糙集的改进的基因特征选择方法

在粗糙集中, 边界域中的信息是不确定的, 但这些不确定的信息往往有助于得到属性约简。在属性约简过程中, 如果只考虑正区域而忽视边界域, 包含在不确定区域中的有用信息有可能会丢失。文献[9]用距离度量函数来决定边界域中的对象与下近似集均值的接近程度。

4.1 距离度量函数

定义 14 给定决策表 $T = (U, C \cup D, V, f)$, $a \in C \cup D, x \in U, X \subseteq U, B \subseteq C, 0 < \tau \leq 1$, X 关于 B 的 τ 下近似集的均值定义为:

$$\underline{B}_\tau X_M = \left\{ \frac{\sum_{x \in \underline{B}_\tau X} a(x)}{|\underline{B}_\tau X|} \mid \forall a \in B \right\} \quad (19)$$

定义 15 给定决策表 $T = (U, C \cup D, V, f)$, $X \subseteq U, B \subseteq C, 0 < \tau \leq 1$, 边界域中的对象相对于 τ 下近似集均值的距离度量函数定义为:

$$DIS_{B,\tau} = \sum_{\mu \in BND_{B,\tau}(D)} \theta_B(\underline{B}_\tau X_M, \mu) \quad (20)$$

式中, $\theta_B(\underline{B}_\tau X_M, \mu) = \sqrt{\sum_{a \in B} f_a(\underline{B}_\tau X_M, \mu)^2}$ 是一个欧氏距离函数($f_a(x, y) = a(x) - a(y)$)。DIS 值的变化越小, 边界域中的对象越不重要。

4.2 基于相容粗糙集的改进的基因特征选择方法

基于相容粗糙集的改进的基因特征选择方法(GSTRSB)与 GSTRS 方法相似, 从 t -检验过滤出评分靠前的 n 个基因组成条件属性集合。然后, 从所有条件属性集合开始逐一删除冗余的属性。每次循环过程中, 可能有许多个冗余的属性, 比较删除每一个冗余属性后 DIS 值的变化, DIS 值的变化越小, 边界域中的对象越不重要, 删除掉使 DIS 变化最小的属性。

基于相容粗糙集的改进的基因特征选择方法(GSTRSB)的具体步骤如下。

算法 3 基于相容粗糙集的改进的基因特征选择算法(GSTRSB)

1. 计算每个基因的 t -检验值, 选出评分靠前的 n 个基因组成条件属性集合 C
2. Set $P = C, T = P, best = \phi$
3. do
4. $P = T$
5. $T = P - \{k\}, \exists k \in P$
6. $best = k$
7. for each $a \in P$
8. if $|DIS_{P-\{a\},\tau}(D) - DIS_{P,\tau}(D)| < |DIS_{T,\tau}(D) - DIS_{P,\tau}(D)|$
9. $T = P - \{a\}$
10. $best = a$
11. until $\gamma_{P-\{best\},\tau}(D) < \gamma_{C,\tau}(D)$
12. return P

4.3 实例分析

为说明相容粗糙集两种属性约简的方法, 本文对表 1 所列的决策表进行约简。表 1 中前 3 个属性是连续值的条件属性, 最后一个为离散值的决策属性。设 $\tau = 0.8, C = \{a, b, c\}, D = \{d\}$ 。

表 1 数据集

Object	a	b	c	d
1	0.3	0.4	0.2	R
2	0.3	1	0.6	A
3	0.4	0.3	0.4	R
4	0.9	0.4	0.7	R
5	0.9	0.7	0.7	A
6	1	0.4	0.7	A

相容类分别是:

$$U/S_{D,\tau} = \{\{1, 3, 4\}, \{2, 5, 6\}\}$$

$$U/S_{C,\tau} = \{\{1\}, \{2\}, \{3\}, \{5\}, \{4, 6\}\}$$

$$U/S_{C-\{a\},\tau} = U/S_{\{b,c\},\tau} = \{\{1\}, \{2\}, \{3\}, \{5\}, \{4, 6\}\}$$

$$U/S_{C-\{b\},\tau} = U/S_{\{a,c\},\tau} = \{\{1\}, \{2\}, \{3\}, \{5\}, \{4, 6\}\}$$

$$U/S_{C-\{c\},\tau} = U/S_{\{a,b\},\tau} = \{\{1, 3\}, \{4, 6\}, \{2\}, \{5\}\}$$

$$U/S_{C-\{a,b\},\tau} = U/S_{\{c\},\tau} = \{\{1\}, \{2\}, \{3\}, \{4, 5, 6\}\}$$

$$U/S_{C-\{a,c\},\tau} = U/S_{\{b\},\tau} = \{\{1, 3, 4, 6\}, \{2\}, \{5\}\}$$

$$U/S_{C-\{b,c\},\tau} = U/S_{\{a\},\tau} = \{\{1, 2, 3\}, \{4, 5, 6\}\}$$

GSTRS 方法得到的约简结果:

对于属性集 C 的 τ 下近似为:

$$\underline{C}_\tau\{1,3,4\}=\underline{\{a,b,c\}}_\tau\{1,3,4\}=\{x|S_{\{a,b,c\},\tau}(x)\subseteq\{1,3,4\}\}=\{1,3\}$$

$$\underline{C}_\tau\{2,5,6\}=\underline{\{a,b,c\}}_\tau\{2,5,6\}=\{x|S_{\{a,b,c\},\tau}(x)\subseteq\{2,5,6\}\}=\{2,5\}$$

因此,决策属性集 D 相对条件属性集 C 的 τ 正区域为:

$$POS_{C,\tau}(D)=\bigcup_{X\in U/S_{D,\tau}}\underline{C}_\tau X=\underline{C}_\tau\{1,3,4\}\cup\underline{C}_\tau\{2,5,6\}=\{1,2,3,5\}.$$

决策属性集 D 相对条件属性集 C 的 τ 依赖度为:

$$\gamma_{C,\tau}(D)=\frac{|POS_{C,\tau}(D)|}{|U|}=\frac{|\{1,2,3,5\}|}{|\{1,2,3,4,5,6\}|}=\frac{4}{6}.$$

对于属性集 $C-\{a\}$,决策属性集 D 相对条件属性集 $C-\{a\}$ 的 τ 依赖度为:

$$\gamma_{C-\{a\},\tau}(D)=\frac{|POS_{C-\{a\},\tau}(D)|}{|U|}=\frac{|\{1,2,3,5\}|}{|\{1,2,3,4,5,6\}|}=\frac{4}{6}$$

$$\gamma_{C-\{a\},\tau}(D)=\gamma_{\{b,c\},\tau}(D)=\gamma_{C,\tau}(D)=\frac{4}{6}$$

因此,可以在属性集 C 中删除属性 a 。类似地,决策属性集 D 相对条件属性集 $\{b,c\}-\{b\}$ 的 τ 依赖度为:

$$\begin{aligned}\gamma_{\{b,c\}-\{b\},\tau}(D) &= \frac{|POS_{\{b,c\}-\{b\},\tau}(D)|}{|U|} = \frac{|\{1,2,3\}|}{|\{1,2,3,4,5,6\}|} \\ &= \frac{3}{6}\end{aligned}$$

$$\gamma_{\{b,c\}-\{b\},\tau}(D)=\frac{3}{6}<\gamma_{C,\tau}(D)=\frac{4}{6}$$

最后,算法终止,GSTRS 方法得到的约简结果为 $\{b,c\}$ 。

GSTRSB 方法得到的约简结果:

决策属性集 D 相对条件属性集 C 的 τ 正区域、边界域为:

$$POS_{C,\tau}(D)=\bigcup_{X\in U/S_{D,\tau}}\underline{C}_\tau X=\underline{C}_\tau\{1,3,4\}\cup\underline{C}_\tau\{2,5,6\}=\{1,2,3,5\}$$

$$BND_{C,\tau}(D)=\bigcup_{X\in U/S_{D,\tau}}\overline{C}_\tau X-\bigcup_{X\in U/S_{D,\tau}}\underline{C}_\tau X=\overline{C}_\tau\{1,3,4\}\cup\overline{C}_\tau\{2,5,6\}-\{1,2,3,5\}=\{1,2,3,4,5,6\}-\{1,2,3,5\}=\{4,6\}$$

X 关于 C 的 τ 下近似集的均值为:

$$\begin{aligned}\underline{C}_\tau X_M &= \left\{ \frac{\sum_{a\in C_\tau X} a(x)}{|\underline{C}_\tau X|} \mid \forall a\in C \right\} \\ &= \left\{ \frac{\sum a(1),a(2),a(3),a(5)}{|4|}; \frac{\sum b(1),b(2),b(3),b(5)}{|4|}; \frac{\sum c(1),c(2),c(3),c(5)}{|4|} \right\} \\ &= \left\{ \frac{0.3+0.3+0.4+0.9}{|4|}; \frac{0.4+1+0.3+0.7}{|4|}; \frac{0.2+0.6+0.4+0.7}{|4|} \right\} \\ &= \{0.475; 0.6; 0.475\}\end{aligned}$$

边界域中的对象相对于 τ 下近似集均值的距离度量函数为:

$$\begin{aligned}obj4 &= \sqrt{f_a(\underline{C}_\tau X_M, 4)^2 + f_b(\underline{C}_\tau X_M, 4)^2 + f_c(\underline{C}_\tau X_M, 4)^2} \\ &= \sqrt{(0.475-0.9)^2 + (0.6-0.4)^2 + (0.475-0.7)^2} \\ &= 0.52\end{aligned}$$

$$\begin{aligned}obj6 &= \sqrt{f_a(\underline{C}_\tau X_M, 6)^2 + f_b(\underline{C}_\tau X_M, 6)^2 + f_c(\underline{C}_\tau X_M, 6)^2} \\ &= \sqrt{(0.475-1)^2 + (0.6-0.4)^2 + (0.475-0.7)^2} \\ &= 0.605\end{aligned}$$

$$\begin{aligned}DIS_{C,\tau}(D) &= \sum_{\mu\in BND_{C,\tau}(D)} \theta_C(\underline{C}_\tau X_M, \mu) = 0.52 + 0.605 \\ &= 1.125\end{aligned}$$

类似地, $DIS_{C-\{a\},\tau}(D)=1.66$, $DIS_{C-\{b\},\tau}(D)=1.05$, $DIS_{C-\{c\},\tau}(D)=0.95$ 。

$$|DIS_{C-\{a\},\tau}(D)-DIS_{C,\tau}(D)|=|1.66-1.125|=0.535,$$

$$|DIS_{C-\{b\},\tau}(D)-DIS_{C,\tau}(D)|=|1.05-1.125|=0.075,$$

$$|DIS_{C-\{c\},\tau}(D)-DIS_{C,\tau}(D)|=|0.95-1.125|=0.175.$$

$best=b$ 。

决策属性集 D 相对条件属性集 C 的 τ 依赖度为:

$$\gamma_{C,\tau}(D)=\frac{|POS_{C,\tau}(D)|}{|U|}=\frac{|\{1,2,3,5\}|}{|\{1,2,3,4,5,6\}|}=\frac{4}{6}$$

对于属性集 $C-\{b\}$,决策属性集 D 相对条件属性集 $C-\{b\}$ 的 τ 依赖度为:

$$\gamma_{C-\{b\},\tau}(D)=\frac{|POS_{C-\{b\},\tau}(D)|}{|U|}=\frac{|\{1,2,3,5\}|}{|\{1,2,3,4,5,6\}|}=\frac{4}{6}$$

$$\gamma_{C-\{b\},\tau}(D)=\gamma_{\{a,c\},\tau}(D)=\gamma_{C,\tau}(D)=\frac{4}{6}$$

因此,可以在属性集 C 中删除属性 b 。

$$DIS_{\{a,c\}-\{a\},\tau}(D)=1.2, DIS_{\{a,c\}-\{c\},\tau}(D)=0,$$

$$|DIS_{\{a,c\}-\{a\},\tau}(D)-DIS_{\{a,c\},\tau}(D)|=|1.2-1.05|=0.15$$

$$|DIS_{\{a,c\}-\{c\},\tau}(D)-DIS_{\{a,c\},\tau}(D)|=|0-1.05|=1.05$$

$best=a$ 。

类似地,决策属性集 D 相对条件属性集 $\{a,c\}-\{a\}$ 的 τ 依赖度为:

$$\begin{aligned}\gamma_{\{a,c\}-\{a\},\tau}(D) &= \frac{|POS_{\{a,c\}-\{a\},\tau}(D)|}{|U|} = \frac{|\{1,2,3\}|}{|\{1,2,3,4,5,6\}|} \\ &= \frac{3}{6}\end{aligned}$$

$$\gamma_{\{a,c\}-\{a\},\tau}(D)=\frac{3}{6}<\gamma_{C,\tau}(D)=\frac{4}{6}$$

最后,GSTRSB 算法终止,得到的约简结果为 $\{a,c\}$ 。

5 实验结果与分析

为了验证本文提出算法的有效性,选用 2 个公开的基因表达数据集。Lymphoma 数据集 (<http://llmpp.nih.gov/lymphoma>) 包含了 4026 个基因,96 个样本,其中 54 个 Other type 子类和 42 个 B-cell lymphoma 子类。Colon cancer 数据集 (<http://microarray.princeton.edu/oncology>) 包含有 2000 个基因,62 个样本,其中 40 个 Colon cancer 子类 and 22 个 Normal 子类。实验数据集见表 2。

表 2 两个基因表达数据集

Data sets	Genes	Classes	Samples
Lymphoma	4026	B-cell	42
		Other type	54
Colon cancer	2000	Colon cancer	40
		Normal	22

根据 GRSR、GSTRS 和 GSTRSB 算法,首先,在 Lymphoma 和 Colon cancer 基因表达数据中,通过 t-检验选择出 50 个基因。对于有缺失值的数据用文献[20]方法进行完备化。GSRs 算法需要对连续的属性值用文献[17]的方法进行离散化。对于 GSTRS 和 GSTRSB 算法,设 $\tau=0.9$ 。约简结果见表 3。

(下转第 140 页)

- [4] Kimura M, Saito K, Akano R. Extracting Influential Nodes on a Social Network for Information Diffusion[M]. Data Mining and Knowledge Discovery, 2010; 70-97
- [5] Kempe D, Kleinberg J M, Tardos E. Maximizing the spread of influence through a social network[C]//The Ninth International Conference on Knowledge discovery and Data Mining (KDD). 2003; 137-146
- [6] Ma H, Yang H, Lyu M R. Mining Social Networks Using Heat

- Diffusion Processes for Marketing Candidates Selection[C]//Proceedings of CIKM. 2011; 233-242
- [7] 金迪, 马衍民. PageRank 算法的分析及实现[J]. 计算机应用, 2009, 18(1001): 118-118
- [8] 田家堂, 王铁彤, 冯小军. 一种新型的社会网络影响最大化算法[J]. 计算机学报, 2011, 34(10): 1956-1965
- [9] Leskovec J, Backstrom L, Kleinberg J M. Meme-tracking and the dynamics of the news cycle[C]//KDD. 2009; 497-506

(上接第 128 页)

表 3 Lymphoma 和 Colon cancer 数据集的约简结果

Data sets	Genes	Samples	GSRS	GSTRS	GSTRSB
Lymphoma	4026	96	7	7	7
Colon cancer	2000	62	6	5	5

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。由表 3 可知, GSRS、GSTRS 和 GSTRSB 算法在 Lymphoma 数据集中提取出的特征基因个数均相同。在 Colon cancer 数据集中, GSTRS 和 GSTRSB 算法比 GSRS 算法提取出的特征基因数少。GSTRS 和 GSTRSB 算法在选取基因个数方面优于 GSRS 算法。

接下来再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, 并用留一交叉检验, 实验结果见表 4。

表 4 Lymphoma 和 Colon cancer 数据集的特征基因分类结果

Data sets	Lymphoma			Colon cancer		
	GSRS	GSTRS	GSTRSB	GSRS	GSTRS	GSTRSB
KNN	93.5%	94.8%	94.8%	79.4%	81.3%	82.9%
C5.0	95.2%	97.4%	97.4%	81.4%	83.6%	85.5%

上述实验结果表明, 无论基于粗糙集的特征选择方法还是基于相容关系的基因选择方法提取的基因都能够保持整个基因数据集的分类能力, 并且基于相容关系的基因选择方法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基于粗糙集的特征选择方法提取的基因。根据距离度量函数进一步确定边界域中对象与下近似集均值的接近程度, 在 Colon cancer 基因表达数据集中, GSTRSB 方法比 GSTRS 方法能够得到更高的分类准确率。在基因分类能力上 GSTRSB 方法优于 GSTRS 和 GSRS 方法。

结束语 粗糙集不能够处理连续性数据的局限性成为它在基因表达数据研究的主要障碍。本文给出了相容度、综合相容度、相容关系和相容关系下依赖度的定义, 提出了基于相容粗糙集的特征选择方法, 进一步给出了距离度量函数, 提出了基于相容粗糙集的改进的基因特征选择方法, 并通过一个实例说明了本文提出的方法。由于本文提出的方法避免了离散化过程, 并根据距离度量函数确定了边界域中的不确定信息, 因此减少了信息损失, 从而相对于基于粗糙集理论的基因特征选择方法选择的基因有更好的准确率。本文提出的方法为基因组的基因选择研究提供了一种新的尝试。

参考文献

- [1] Tibshirani R, Hastie T, Narashiman B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression[C]//Nat'l Academy of Sciences. USA, 2002; 6567-6572
- [2] Kohavi R, John G H. Wrappers for feature subset selection[J]. Artificial Intelligence, 1997; 273-324
- [3] Banerjee M, Mitra S, Banka H. Evolutionary-rough feature selec-

- tion in gene expression Data[J]. IEEE Transaction on Systems, Man, and Cyberneticd, Part C: Application and Reviews, 2007, 37; 622-632
- [4] Momin B F, Mitra S, Datta Gupta R. Reduct generation and classification of gene expression data[C]//Proceeding of First International Conference on Hybrid Information Technology (ICHICT06). New York, 2006; 699-708
- [5] Pawlak Z. Rough sets[J]. International Journal of Information Computer Science, 1982, 11(5): 341-356
- [6] Dubois D, Prade H. Putting rough sets and fuzzy sets together[J]. Intelligent Decision Support, 1992; 203-232
- [7] Jensen R, Shen Q. Tolerance-based and fuzzy-rough feature selection[C]//Proceedings of the 16th International Conference on Fuzzy Systems(FUZZ- IEEE07). 2007; 877-882
- [8] Liang J Y, Li R. Distance: A more comprehensible perspective for measures in rough set theory[J]. Knowledge-Based Systems, 2012, 27; 126-136
- [9] Parthaláin N M, Shen Q. Exploring the boundary region of tolerance rough sets for feature selection[J]. Pattern Recognition, 2009, 42; 655-667
- [10] Yao Y Y, Yao B X. Covering based rough set approximations[J]. Information Sciences, 2012, 200; 91-107
- [11] 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
- [12] Yang X B, Xie J, Song X N, et al. Credible rules in incomplete decision system based on descriptors[J]. Knowledge-Based Systems, 2009, 22; 8-17
- [13] Shen Q, Chouchoulas A. A rough-fuzzy approach for generating classification rules[J]. Pattern Recognition, 2002, 5; 2425-2438
- [14] Ou Yang Y P, Shieh H M, et al. Combined rough sets with flow graph and formal concept analysis for business aviation decision-making[J]. Journal of Intelligent Information Systems, 2011, 36(3); 347-366
- [15] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001
- [16] 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
- [17] Grzymala-Busse J W. Discretization of numerical attributes[M]. Klösgen W, Zytkow J, ed. Handbook of Data Mining and Knowledge Discovery, Oxford University Press, 2002; 218-225
- [18] 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; 531-537
- [19] Wang L P, Feng C, Xie X. Accurate cancer classification using expressions of very few genes[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007, 4; 40-53
- [20] Grzymala-Busse J W, Grzymala-Busse W J. Handling missing attribute values[M]. Maimon O, Rokach L, ed. Handbook of Data Mining and Knowledge Discovery, 2005; 37-57