基于相交邻域粗糙集的基因微阵列数据分类

孟 军 李 锐 郝 涵
(大连理工大学计算机科学与技术学院 大连 116024)

摘 要 在对基因微阵列数据的特征选择和分类的研究中,粗糙集理论是一个可以消除冗余基因的有效工具。但是传统的粗糙集模型不能很好地处理连续型数值数据,而离散化方法可能会导致信息的丢失。为此,提出了一种基于相交邻域粗糙集模型的属性约简算法,即将传统粗糙集中的距离邻域扩展为相交邻域,采用基于集合的方式来定义近似,以此构建粗糙集模型。在癌症数据集上进行实验,结果表明基于集合近似和相交邻域的粗糙集模型可以取得较好的分类效果,并且通过对选择出的基因进行GO术语分析,进一步证明了该模型的有效性。

关键词 粗糙集,相交邻域,基因微阵列数据
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.008

Gene Microarray Data Classification Based on Intersecting Neighborhood Rough Set

MENG Jun LI Rui HAO Han
(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract In the research of gene microarray data classification and feature selection, rough set theory is an effective tool, as it can eliminate redundant genes. However a drawback in traditional rough set is that it cannot handle with continuous numeric data well, and discretization method may lead to the loss of information. We proposed an attribute reduction algorithm based on intersecting neighborhood rough set, extended the distance neighborhood to intersecting neighborhood and employed the definition of approximation based on set, to build the rough set model. Experimental results on three cancer data sets show that the rough set model based on the set approximate and intersecting neighborhood is effective and efficient. Meanwhile, the analysis of GO terms on selected genes further proves the validity of the model.

Keywords Rough set, Intersecting neighborhood, Gene microarray data

1 引言

随着微阵列和高通量技术的发展,研究者们得到了海量的基因表达数据。如何对这些数据进行分析,从中挖掘出有价值的知识,是生物信息学的一个研究热点[1,2]。

在对基因表达数据的分类研究中发现,简单地使用统计学和机器学习的方法,不能达到很好的分类效果,并且由于微阵列数据的高维度和小样本的特点,采用传统的方法对其进行分类会出现训练时间长和分类准确率偏低等不足。为了解决这些问题,需要对传统的方法进行适当的加工与处理,以适应这种特点。

人们发现,在微阵列数据中有很大一部分基因,对其分类是没有意义的,因此选择合适的基因用于分类是一个关键的问题。基因选择也即特征选择的方法可以分为两大类:Filter方法和Wrapper方法。其中Filter方法独立于所选择的分类器,它是通过某种标准或者统计检验的方法,将基因进行排序,进而选择相应的基因[3]。这种方法的计算速度很快,使得它适用于处理高维度的数据,但是Filter方法与分类器无关,因此筛选出的基因并不真正适用于该分类器。而Wrapper方法与具体的学习算法有关,它通过搜索优化算法来为分类方法找到一个合适的基因子集[4],因而在大多数情况下,Wrapper方法的性能要优于Filter[5,6]。但是其在计算方面花费的代价更高;另外,该方法由于不是直接基于基因表达值的差异,因此所选出的基因可能不具有明确的生物学意义。

然而,经过Filter和Wrapper方法的处理,仍然会存在冗余的信息,因为在相似生物通路下的基因有相近的得分[7]。粗糙集理论可以用来消除这些冗余信息[8],它是一种处理不精确、不一致、不完备等数据信息的数学方法,已经被应用于模糊模型的结构辨识、微阵列数据特征选择和分类等领域[9-11]。给定一个具有离散属性值的数据集,使用粗糙集理论可以从原始的属性中选取一个子集,并且使信息损失降至最低。

属性离散化是粗糙集的一个重要研究方向。现实生活中的数据一般都是连续型的,但传统的粗糙集模型只能处理离散形式的数据,有很大的局限性。所以在属性约简之前,需要对基因进行离散化,但这可能会导致信息丢失。为了弥补传

到稿日期:2014-04-20 返修日期:2014-05-27 本文受辽宁省自然科学基金项目(20130200029)资助。
孟 军(1964-),女,博士,副教授,主要研究方向为机器学习与数据挖掘,E-mail:mengjun@dlut.edu.cn;李 锐(1991-),男,硕士生,主要研究方向为数据挖掘与粗糙集理论;郝 涵(1991-),男,硕士生,主要研究方向为粗糙集与粒计算理论。

统粗糙集模型不能直接操作数据值的不足，胡清华等人[12]利用粒计算中的邻域，将粗糙集中的等价关系扩展为邻域关系，建立邻域粗糙集模型。在此模型的基础上，人们相继提出了一系列的属性约简方法，以直接用于处理数值型数据，如2012年闵帆等人提出通过误差范围来构建邻域粗糙集[19]。

本文依据基因微阵列数据的特点，提出基于相交邻域的粗糙集模型，给出一种基于正域的、满足向前删除策略的基因选择方法，用于处理基因微阵列数据的基因选择与分类问题。通过在癌症基因数据集上的实验，分析在不同近似方法、邻域构建方式下特征提取的不同情况，比较不同方法的分类效果及各自适应的数据集。

2 问题定义及描述

2.1 粗糙集理论

粗糙集理论于1982年由波兰数学家Pawlak提出，它的主要思想是在保持数据分类能力不变的前提下，对数据进行属性约简、特征提取等操作。在该理论中，用于分类或属性约简的数据集可以用决策表的形式表示，具体的相关定义如下[13]。

定义1 给定一个决策表 $DT=(U,A=C\cup D,\{V_a\},f_a)_{a\in A}$，其中 U 为非空有限对象集合，称为论域；C 和 D 为两个非空有限集，分别代表条件属性集和决策属性集，并且满足 $C\cup D=A$；V 是所有属性的值域集合，$V=\bigcup_{a\in A}V_a$，V_a 表示属性 a 的值域集合；f 是所有属性的信息函数，$f=\bigcup_{a\in A}f_a$，其中，f_a 表示从 U 到 V_a 的映射。

定义2 如果对象 x,y 在属性 a 上的取值相等，即 $f_a(x)=f_a(y)$，则称 x 和 y 在属性 a 上是不可分辨的，属性子集 $B\subseteq A$ 在 U 上的不可分辨关系定义为：

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

这种关系是一种等价关系，且满足 $IND(B)=\bigcap_{b\in B}IND(b)$。式(1)表示等价关系 B 将 U 划分的等价类集合。

定义3 对于任一对象子集 $X\subseteq U$ 和属性子集 $R\subseteq C$，X 的下近似和上近似分别定义如下：

$$\underline{R}(X)=\{x\in U\mid [x]_R\subseteq X\} \quad (2)$$

$$\overline{R}(X)=\{x\in U\mid [x]_R\cap X\neq\varnothing\} \quad (3)$$

也可定义为：

$$\underline{R}(X)=\bigcup\{[x]_R\in U/R\mid [x]_R\subseteq X\} \quad (4)$$

$$\overline{R}(X)=\bigcup\{[x]_R\in U/R\mid [x]_R\cap X\neq\varnothing\} \quad (5)$$

在等价关系下，上述两种定义方式是等价的，第一种是基于点的近似；第二种是基于集合的近似。集合 X 的正域、负域和边界域分别定义如下：

$$POS_R(X)=\underline{R}(X) \quad (6)$$

$$NEG_R(X)=U-\overline{R}(X) \quad (7)$$

$$BND_R(X)=\overline{R}(X)-\underline{R}(X) \quad (8)$$

$POS_R(X)$ 是根据 R 判定一定属于 X 的 U 中对象的集合；$NEG_R(X)$ 是根据 R 判定一定不属于 X 的 U 中对象的集合；$BND_R(X)$ 是不能确定是属于 X 还是属于 X 的补的对象集。

2.2 基于粗糙集的属性约简

基因选择是一个属性约简的过程，即在保证数据分类能力不变的前提下，尽可能地除去冗余的信息。具体可分为相对约简和绝对约简[13]。

定义4 给定一个决策表 $DT=(U,A=C\cup D,\{V_a\},f_a)_{a\in A}$，对于 $\forall a\in R\subseteq C$，若 $IND(R)=IND(R-\{a\})$，则称 a 为 R 中不必要的，否则就为必要的。若每个 a 都是 R 中必要的，则称 R 为独立的，否则称为依赖的。

定义5 对于一个决策表 $DT=(U,A=C\cup D,\{V_a\},f_a)_{a\in A}$，决策属性 D 将 U 划分成 S 个等价类，令属性子集 $R\subseteq C$，决策属性 D 的 R 正域定义为：

$$POS_R(D)=\bigcup_{i=1}^{S}\underline{R}(D_i) \quad (9)$$

定义6 对于 $\forall a\in R\subseteq C$，若 $POS_{R-\{a\}}(D)=POS_R(D)$，则称 a 为 R 中相对 D 不必要的，否则称 a 为 R 中相对 D 必要的。若对于所有的 $a\in R$ 都为 R 中相对 D 必要的，则称 R 为相对 D 独立的。

定义7 若 R 是相对 D 独立的，并且 $POS_R(D)=POS_C(D)$，则称 R 为 C 的一个相对于 D 的约简，简称相对约简。对于属性子集 R，若 R 是独立的，且 $IND(R)=IND(C)$，则称 R 为 C 的一个绝对约简。

2.3 基于相容关系的粗糙集模型

等价关系约束下的经典粗糙集模型只能处理离散化形式的数据，具有明显的局限性。基于相容关系的粗糙集模型，可以方便地处理数值型数据。对粗糙集模型进行扩展，提出了基于相交邻域的粗糙集模型。

定义8 给定论域 U 和其上的相容关系 T，对于 $\forall x\in U$，关联一个在 T 上与 x 有关的对象 y 的集合 $N_T(x)$：

$$N_T(x)=\{y\mid (x,y)\in T \text{ 且 } y\in U\} \quad (10)$$

称 $N_T(x)$ 为对象 x 在 U 上关系 T 的邻域，映射 $T:x\to N_T(x)$ 叫做相容邻域系统[14]。

定义9 给定二元组 $K=(U,\alpha)$，其中 α 由一个或多个相容关系构成。若 α 是一个相容关系，那么 K 为相容近似空间；若 α 由多个相容关系构成，那么 K 为相容知识库[14]。在关系子集 $\beta\subseteq\alpha$ 上的二元关系用 T_β 表示：

$$T_\beta=\{(x,y)\mid (x,y)\in U\times U,\forall T\in\beta \text{ 且 } x\in N_T(y)\} \quad (11)$$

定义10 对象 x 在关系集 β 上的最近邻元素集合定义为：

$$N_\beta(x)=\{y\mid (x,y)\in T_\beta \text{ 且 } y\in U\} \quad (12)$$

$K=(U,\alpha)$ 是一个相容知识库且 $\beta\subseteq\alpha$，对于 $\forall X\subseteq U$，其下、上近似分别用 $\underline{\beta}(X)$ 和 $\overline{\beta}(X)$ 表示，基于点的近似定义为：

$$\underline{\beta}(X)=\{x\mid N_\beta(x)\subseteq X \text{ 且 } x\in U\} \quad (13)$$

$$\overline{\beta}(X)=\{x\mid N_\beta(x)\cap X\neq\varnothing \text{ 且 } x\in U\} \quad (14)$$

基于集合的近似定义为：

$$\underline{\beta}(X)=\bigcup\{N_\beta(x)\mid N_\beta(x)\subseteq X \text{ 且 } x\in U\} \quad (15)$$

$$\overline{\beta}(X)=\bigcup\{N_\beta(x)\mid N_\beta(x)\cap X\neq\varnothing \text{ 且 } x\in U\} \quad (16)$$

在等价关系下，两种近似定义方式的结果相同；而在相容关系下，基于集合的下近似定义优于基于点的下近似定义。

3 基于相交邻域粗糙集模型的基因选择

通过相交邻域粗糙集模型的构建，提出基因选择方法的具体架构以及基于向前删除策略的基因选择方法。

3.1 秩和检测

微阵列数据的维度一般都在数千甚至上万以上，直接对原始数据进行基因选择代价很高，因此首先需要进行基因初选。主要的思想是通过某种标准来衡量每个基因的分类能力，按照计算得出的值进行排序，选择前 n 个基因。常用的方法有 T-检测、Relief-检测和 Wilcoxon 秩和检测等[2,6]。由于前两者都要求数据满足高斯分布的条件，否则会出现基因的排序结果与基因对样本的真实分类能力的排序不一致等问题，而秩和检测是一个不需要参数的假设检验方法，因此本文采用 Wilcoxon 秩和检测进行基因的初选。

秩和检测的目标是比较两个样本间的差异，它的基本思想是：若两个样本的容量满足 $n_1 \leqslant n_2$，首先将两个样本的数据混在一起，然后按照数据从小到大的顺序，为每一个数据编秩次，最小的数据秩为 1，最大的数据秩为 n_1+n_2。如果两个样本的秩次和相等或接近，那么两个样本无较大差别；如果两个样本的秩次和相差较大，那么两个样本的水平差异较大。设 $\{x_i \mid i=1,2,\cdots,n_1\}$ 和 $\{y_i \mid i=1,2,\cdots,n_2\}$ 是两个独立的随机样本集，令 U 表示为所有 y 观察值大于 x 观察值的个数，当 n_1，n_2 均大于 10 时，检测量定义如下：

$$Z=(U-n_1n_2/2)/\sqrt{n_1n_2(n_1+n_2+1)/12}\sim N(0,1) \quad (17)$$

得出每个基因的 Z 值以后，再计算其相应的 p 值，按照 p 值由大到小排序，选择前 n 个基因。

3.2 相交邻域的定义

在获取对象邻域的过程中，根据何种度量方式计算对象间的相似程度是至关重要的因素，处理实数类型数据时最常用的是欧氏距离。给定一个决策表 $DT=(U,C\cup D,\{V_a\},f_a)_{a\in C}$，任意两点 $x,y\in U$ 在特征子集 $R\subseteq C$ 上的欧氏距离定义为：

$$\Delta(x,y,R)=\sqrt{\sum_{a\in R}(f_a(x)-f_a(y))^2} \quad (18)$$

在特征子集 R 中，根据欧氏距离构建对象 x 的邻域有两种方法：

(1)利用欧氏距离计算出 x 与 U 中其它对象之间的几何距离，将距离小于所给阈值的对象放入 x 的邻域中；

(2)对每一维分别利用欧氏距离计算 x 的邻域，将 x 在所有维上的邻域相交构成 x 的邻域。

其中，前者称为距离邻域，只需要设定一个适当的阈值，是目前常用的邻域构造方法；后者即是本文提出的相交邻域，需要针对每一维给出适当的阈值，用于构建实数型数据对象邻域。对于任意 $x\in U$ 和 $R\subseteq C$，x 在 U 上由特征子集 R 限定的距离邻域定义为：

$$N_R^{\delta}(x)=\{y \mid y\in U,\Delta(x,y,R)\leqslant\delta\} \quad (19)$$

其中，$\delta\geqslant 0$，是给定的阈值。在特征子集 R 中，所有对象及其邻域可以构建一个关系矩阵 $T=(r_{xy})_{n\times n}$，该矩阵满足自反性和对称性，是一个相容关系。x 在 U 上由特征子集 R 限定的相交邻域定义为：

$$N_R^{\cap}(x)=\bigcap\{y \mid y\in U,\Delta(x,y,\{a\})\leqslant\delta_a \text{ 且 } a\in R,\delta_a\in\delta_R\} \quad (20)$$

其中，$\delta_a\in\delta_R$ 是在属性 a 上的阈值，δ_R 是特征子集 R 中每个属性的阈值集合。同上，利用相交邻域构建对象的邻域，也可形成一个相容关系。

在进行基因选择的过程中距离邻域的获取方式会在一定程度上影响不同基因间的相互作用，而有些基因在共同作用时可能会对分类产生负面的效果。当两个基因的阈值取值相同时，距离邻域可能会忽视一些原本与 x 相似的对象。当设定两个不同的阈值时，相交邻域的度量方式更加灵活。因此，传统的基于欧氏距离的距离邻域并不一定是最好的选择，本文为每一个维度指定一个阈值，来计算对象的相交邻域。但是，要对每一个基因取一个合适的阈值，需要一个复杂的参数优化的过程，并结合一定的生物学知识，所以本文对所有维度的阈值取相同的值，后续将深入研究。

3.3 基于相交邻域粗糙集模型的基因选择算法

本文提出基于相交邻域粗糙集模型的基因选择算法，利用集合近似定义，采用向前删除搜索策略。具体的步骤如下：

(1)首先利用秩和检测将所有基因排序，选择前 300 个基因；

(2)采用基于相交邻域粗糙集模型的属性约简算法进行基因选择，获得不同阈值下的基因子集；

(3)利用分类器获得每组基因的最高分类准确率；

(4)根据每组基因的分类准确率排序，得到前 10 组基因作为最终结果。

基于相交邻域粗糙集模型的基因选择具体算法如下：

```
输入:DT=(U,A=C∪D,{Va},fa)a∈A
输出:DT 的一个约简 B
Begin
(1){NC(x)|x∈U};//计算所有对象在 C 上的邻域集合
(2)U/IND(D)={D1,…,Di,…,DS}; //获得决策属性形成的等价类
(3)  POSC(D)=getPRegion(C,U/IND(D)),B=C;
(4)  for each a∈C
(5)     POSB-{a}(D)=getPRegion(B-{a},U/IND(D));
(6)  if POSB-{a}(D)=POSC(D) then
(7)  B=B-{a};
End
```

其中，getPRegion 是计算新属性集下的决策正域的方法，详细描述如下：

```
输入:属性集 B 及决策属性等价类 U/IND(D)
输出:POSB(D)
getPRegion()
Begin
(1){NB(x)|x∈U};  //计算所有对象在属性集 B 上的邻域
(2)  for each Di∈U/IND(D)
(3)    for each x∈U
(4)       if NB(x)⊆Di then
(5)          POSB(Di)=POSB(Di)∪NB(x);
(6)    POSB(D)=POSB(D)∪POSB(Di);//正域合并
End
```

该属性约简算法的时间复杂度为 $O(|C|^2|U|^2)$，与目前大多数基于粗糙集模型的基因选择算法时间复杂度相同。其中，$POS_B(D)$ 的计算方式是影响算法结果的核心因素。

4 实验结果与分析

4.1 数据集及实验方法

本文利用 3 个常用的癌症数据集 Colon、Leukemia 和 Lung 进行实验验证，分别从 http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi，http://www.datatang.com/data/17097 和 http://www.datatang.com/data/17103 处下载。数据集性质如表 1 所列。

表 1　数据集性质

数据集	基因数	样本数	类 1 个数	类 2 个数
Colon	2000	62	40	22
Leukemia	7129	72	47	25
Lung	12533	181	150	31

采用 4 种方法进行了实验，分别是基于点近似和距离邻域的、基于点近似和相交邻域的、基于集合近似和距离邻域的、基于集合近似和相交邻域的 4 种粗糙集模型。为了后续表示方便，在这里分别将 4 种模型标记为：

App1_Neighbor1：基于点近似和距离邻域；

App1_Neighbor2：基于点近似和相交邻域；

App2_Neighbor1：基于集合近似和距离邻域；

App2_Neighbor2：基于集合近似和相交邻域。

由于得到的基因子集与设定的阈值有关，因此令 δ 取值范围为 0.01 到 1，以步长 0.01 递增，可以获得 100 组基因子集。分别利用 C4.5、KNN 和 SVM 分类器进行缺一验证，根据各个分类器的准确率，选择排序靠前且子集中基因个数少于 10 的作为最终结果。

4.2　实验结果及 GO 术语分析

本文所有实验是在 Windows 7 系统、2.19GHz 酷睿处理器和 2G 内存环境下进行的。结果如表 2—表 4 所列，Acc. 为其相应的准确率(%)，括号内数字 Num 表示所选基因的个数，这里的准确率和基因个数是所选 10 组基因的平均数。

表 2　Colon 数据集在 3 种分类器下的分类准确率及基因个数

Method	C4.5	KNN	SVM
	Acc. (Num)	Acc. (Num)	Acc. (Num)
App1_Neighbor1	0.8758(6.8)	0.8726(6.0)	0.8629(7.4)
App1_Neighbor2	0.8581(6.6)	0.8774(6.0)	0.8468(7.3)
App2_Neighbor1	0.8758(6.8)	0.8726(6.0)	0.8629(7.4)
App2_Neighbor2	0.8532(6.4)	0.8774(5.9)	0.8468(7.1)

表 3　Leukemia 数据集在 3 种分类器下的分类准确率及基因个数

Method	C4.5	KNN	SVM
	Acc. (Num)	Acc. (Num)	Acc. (Num)
App1_Neighbor1	0.9209(2.9)	0.9819(3.8)	0.9541(6.1)
App1_Neighbor2	0.9347(6.2)	0.9861(5.9)	0.9889(7.0)
App2_Neighbor1	0.9472(3.8)	0.9861(4.3)	0.9653(6.3)
App2_Neighbor2	0.9431(5.1)	0.9903(4.9)	0.9861(6.0)

表 4　Lung 数据集在 3 种分类器下的分类准确率及基因个数

Method	C4.5	KNN	SVM
	Acc. (Num)	Acc. (Num)	Acc. (Num)
App1_Neighbor1	0.9896(5.5)	0.9962(4.9)	1.0(3.2)
App1_Neighbor2	0.9868(3.5)	0.9967(4.1)	1.0(3.1)
App2_Neighbor1	0.989(3.5)	1.0(4.6)	1.0(3.2)
App2_Neighbor2	0.9896(3.2)	0.9994(6.8)	1.0(3.1)

表 2 中，基于距离邻域方法的分类准确率要高于基于相交邻域的方法，而基于集合近似和点近似的方法分类效果差别不大。表 3 中，基于集合近似的方法要优于点近似的方法，基于相交邻域的分类模型在这个数据集上的表现更好。表 4 中，基于集合近似的方法的表现略优于点近似的方法，基于距离邻域和相交邻域的分类模型在这里都获得了较高的分类准确率。另外，可以看到 3 种分类算法下的准确率没有太大的差别，也说明了该方法具有较强的鲁棒性。

以 Leukemia 数据集为例，4 种方法的分类准确率的比较如图 1 所示。

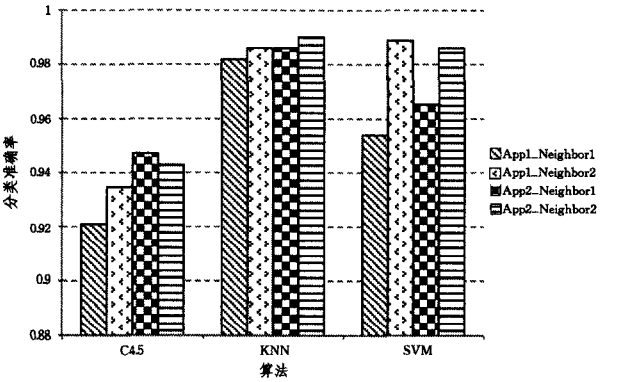

图 1　Leukemia 数据集上的分类准确率比较

总体来看，基于集合近似的粗糙集模型的分类效果要优于基于点近似的粗糙集模型，而基于距离邻域和相交邻域的粗糙集模型有着各自较适应的数据集，这也应证了上文提到的有些基因数据集并不适用于传统的欧氏距离度量方法。

为了进一步验证本文所提出方法的有效性，对实验所筛选出的基因通过 GO 术语进行生物学的分析，这里选取的是在阈值 $\delta=0.14$ 条件下(在这个阈值下，选出的基因个数较少，且取得了很高的分类准确率)、Leukemia 数据集得到的基因选择结果，如表 5 所列。

表 5　Leukemia 数据集的基因功能注释

基因	分子功能	生物过程	细胞组成
M23197	protein binding receptor activity	cell adhesion cell-cell signaling negative regulation of cell proliferation	plasma membrane
M31523	protein heterodimerization activity protein homodimerizationactivity	B cell differentiation B cell lineage commitment cell development	Cytoplasm transcription factor complex
J05243	actin binding calcium ion binding calmodulin binding protein binding	actin filament capping apoptotic process axon guidance	extracellular vesicular exosome microtubule cytoskeleton spectrin
M92287	cyclin-dependent protein serine protein binding	cell division T cell proliferation	Cytoplasm nucleus
M11722	DNA binding metal ion binding	DNA modification	Cytoplasm nucleus

从表 5 不难看出，基因 M23197 与细胞的黏附、增殖和细胞间的信号相关[15]，基因 M31523 与 B 细胞的生物过程相关[16]，基因 M92287 与细胞的分类和 T 细胞的增殖相关[17]，而 M11722 则与 DNA 的结合与修正有关系[18]。这说明本文提出的方法所选择的基因同其导致的癌症疾病有密切的关联，进一步证明了本方法的有效性。

结束语　本文提出了基于相交邻域的粗糙集模型，可以很好地对非球型结构分布数据集进行直接处理，相交邻域的每一个属性都可以有不同的阈值，所以它比距离邻域更灵活地适用于各种结构的数据集；并结合基于向前删除策略的属性约简算法，应用于基因选择问题。在 Colon、Leukemia 和 Lung 数据集上进行实验，仅利用筛选出的不到 10 个基因，就获得了较高的分类准确率。除此之外，实验结果表明，基于集合近似的粗糙集比基于点近似的模型可以取得更好的效果；

(下转第 66 页)

2002:79-86

[15] 徐琳宏,林鸿飞,杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报,2007,21(1):96-100
Xu L H, Lin H F, Yang Z H. Text Orientation Identification Based on Semantic Comprehension [J]. Journal of Chinese Information Processing, 2007, 21(1): 96-100

[16] 唐慧丰,谭松波,程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报,2007,21(6):88-94
Tang H F, Tan S B, Cheng X Q. Research on Sentiment Classification of Chinese Reviews Based on Supervised Machine Learning Techniques[J]. Journal of Chinese Information Processing, 2007, 21(6): 88-94

[17] Liu Bing, Hu Min-qing, Cheng Jun-sheng. Opinion Observer: Analyzing and Comparing Opinions on the web[C]// the 14th International Conference on World Wide Web. Chiba, Japan, 2005: 342-351

[18] 周城,葛斌,唐九阳,等. 基于相关性和冗余度的联合特征选择方法[J]. 计算机科学,2012,39(4):181-184
Zhou C, Ge B, Tang J Y, et al. Joint Feature Selection Method Based on Relevance and Redundancy[J]. 2012, 39(4): 181-184

[19] 情感评论语料[EB/OL]. http://www.searchforum.org.cn/tansongbo/senti_corpus.jsp
Semantic Comment Corpus [EB/OL]. http://www.searchforum.org.cn/tansongbo/senti_corpus.jsp

[20] 张启蕊,董守斌,张凌. 文本分类的性能评估指标[J]. 广西师范大学学报:自然科学版,2007,25(2):119-122
Zhang Q R, Dong S B, Zhang L. Performance Evaluation Metric for Text Classifiers[J]. Journal of Guangxi Normal University: Natural Science Edition, 2007, 25(2): 119-122

[21] 王卫玲,刘培玉,初建崇. 一种改进的基于条件互信息的特征选择算法[J]. 计算机应用,2007,27(2):433-435
Wang W L, Liu P Y, Chu J C. Improved feature selection algorithm with conditional mutual information[J]. Journal of Computer Applications, 2007, 27(2): 433-435

[22] Platt J C. Fast Training of Support Vector Machines Using Sequential Minimal Optimization[M]// Schoelkopf B, Burges C, Smola A. Advances in Kernel Methods. Cambridge, USA: MIT Press, 1999: 185-208

(上接第 40 页)

并且,不是所有基因微阵列数据都适用于距离邻域的粗糙集模型,有些数据集在相交邻域下的分类效果更加出色。另外,如何对每一个基因设定一个合适的阈值,以及相容关系粗糙集与其它生物知识的结合应用,将是今后的研究重点。

参 考 文 献

[1] Piao Y, Piao M, Park K, et al. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data[J]. Bioinformatics, 2012, 28(24): 3306-3315

[2] Wang Shu-lin, Li Xue-ling, Zhang Shan-wen, et al. Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction[J]. Computers in Biology and Medicine, 2010, 40(2): 179-189

[3] Tong Mu-chen-xuan, Liu Kun-hong, Xu Chun-gui, et al. An ensemble of SVM classifiers based on gene pairs[J]. Computers in Biology and Medicine, 2013, 43(6): 729-737

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

[5] Wang Li, Zhu Ji, Zou Hui. Hybrid huberized support vector machines for microarray classification and gene selection[J]. Bioinformatics, 2008, 24(3): 412-419

[6] Bolon-Canedo V, Sanchez-Marono N, Alonso-Betanzos A. An ensemble of filters and classifiers for microarray data classification[J]. Pattern Recognition, 2012, 45(1): 531-539

[7] Jiao Na, Miao Duo-qian. An efficient gene selection algorithm based on tolerance rough set theory[J]. Data Mining and Granular Computing, 2009, 5908: 176-183

[8] Pawlak Z. Rough sets[J]. Computer and Information Science, 1982, 11(5): 341-356

[9] Jensen R, Shen Q. Fuzzy-rough attribute reduction with application to web categorization[J]. Fuzzy Sets and Systems, 2004, 141(3): 469-485

[10] Paul S, Maji P. Rough set based gene selection algorithm for microarray sample classification[C]// International Conference on Methods and Models in Computer Science. New Delhi, 2010: 7-13

[11] Lu Zheng-cai, Qin Zheng, Zhang Yong-qiang, et al. A fast feature selection approach based on rough set boundary regions [J]. Pattern Recognition Letters, 2014, 36(15): 81-88

[12] 胡清华,于达仁. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报,2008,19(3):640-649
Hu Qing-hua, Yu Da-ren. Numerical Attrrbute Reduction Based on Neighborhood Granulation and Rough Approximation[J]. Journal of Software, 2008, 19(3): 640-649

[13] Pawlak Z. Rough sets: theoretical aspects of reasoning about data[M]. 1991

[14] Meng Jun, Wang Xiu-kui, Wang Peng, et al. Knowledge Dependency and Rule Induction on Tolerance Rough Sets [J]. Journal of Multiple-Valued Logic and Soft Computing, 2013, 20(3/4): 401-421

[15] Orr S J, Morgan N M, Elliott J, et al. CD33 Responses are Blocked by SOCS3 through Accelerated Proteasomal-mediated Turnover[J]. Blood, 2007, 109(3): 1061-1068

[16] Mark P K, Comeils M, Sun X H, et al. A new Homeobox Gene Contributes the DNA Binding Domain of the t(1;19) Translocation Protein in pre-B ALL[J]. Cell, 1990, 60(4): 547-555

[17] Sicinska E, Aifantis I, Laurent L C, et al. Requirement for Cyclin D3 in Lymphocyte Development and T Cell Leukemias [J]. Cancer Cell, 2003, 4(6): 451-461

[18] Mertelsmann R, Steven G, Steinmann G, et al. T-cell Growth Factor (Interleukin 2) and Terminal Transferase Activity in Human Leukemias and Lymphoblastic Cell Lines [J]. Blut, 1981, 43(2): 99-103

[19] Min Fan, William Z. Attribute reduction of data with error ranges and test costs[J]. Information Sciences, 2012, 211: 48-67