DOI:10.3969/j.issn.1673-4785.201307014
网络出版地址:http://www.cnki.net/kcms/doi/CNKI:23-1538/TP.20131105.1202.006.html

基因表达数据在邻域关系中的特征选择

陈玉明[1],吴克寿[1],李向军[2]
(1. 厦门理工学院 计算机科学与技术系,福建 厦门 361024; 2. 南昌大学 计算机科学与技术系,江西 南昌 330031)

摘　要:基因特征选择是基因表达数据分析中的一种重要方法。粗糙集是一种处理不确定性、不一致性、不精确性数据的有效分类工具,其特点是保持基因表达数据集的分类能力不变,进行基因特征选择。为了避免传统粗糙集特征选择方法所必需的离散化过程带来的信息损失,将邻域粗糙集特征选择方法应用于基因的特征选取,提出了基于邻域粗糙集的基因选择方法。该方法从所有特征出发,根据特征重要度逐步删除冗余的特征,最后得到关键特征组进行分类研究。在 2 个标准的基因表达数据集上进行特征选取,并进行了分类实验,实验结果表明该方法是有效可行的。

关键词:粗糙集;邻域关系;基因表达数据;特征选择;分类
中图分类号: TP391　**文献标志码**:A　**文章编号**:1673-4785(2014)02-0210-04

中文引用格式:陈玉明,吴克寿,李向军. 基因表达数据在邻域关系中的特征选择[J]. 智能系统学报, 2014, 9(2): 210-213.
英文引用格式:CHEN Yuming, WU Keshou, LI Xiangjun. Gene expression data feature selection with neighborhood relation[J]. CAAI Transactions on Intelligent Systems, 2014, 9(2): 209-212.

Gene expression data feature selection with neighborhood relation

CHEN Yuming[1], WU Keshou[1], LI Xiangjun[2]
(1. Department of Computer Science and Technology, Xiamen University of Technology, Xiamen 361024, China; 2. Department of Computer Science and Technology, Nanchang University, Nanchang 330031, China)

Abstract: The selection of an efficient gene feature is a key procedure for analysis of gene expression data. The rough set theory is an efficient classification tool to deal with uncertain, inconsistent and inaccurate gene data. One limitation of the rough set theory is the lack of effective methods for processing real valued data. However, gene expression data sets are always continuous. Discrete methods can result in information loss. This paper investigates an approach to the selection of gene feature on the basis of the neighborhood rough set theory. Starting from all the features, this approach gradually removes the redundant features, and finally gets the key features of the group classification study based on the importance degree of characteristics. To evaluate the performance of the proposed approach, we applied it to two bench mark gene expression data sets which were compared to certain aspects of the feature selections. The experimental results illustrate that our algorithm is more effective for selecting high discriminative genes in cancer classification tasks.

Keywords: rough sets; neighborhood relation; gene expression data; feature selection; classification

美国人类基因组计划(HGP)把基因组信息学定义为:它是一个学科领域,包含着基因组信息的获取、处理、存储、分配、分析和解释的所有方面。基因表达数据分析的对象是在不同条件下,全部或部分基因的表达数据所构成的数据矩阵。通过对该数据矩阵的分析,可以回答一些生物学问题。随着试验技术及仪器的不断改进和基因组数据的急剧增长,现代 DNA 微阵列或芯片技术产生的各种基因表达数据均规模庞大、内容复杂。如何有效地分析利用这些数据成为生物信息学中的挑战性课题。在基因表达数据分析中,基因的数目成千上万,但往往只是

收稿日期:2012-10-26.　网络出版日期:2013-11-05.
基金项目:国家自然科学青年基金资助项目(61103246).
通信作者:陈玉明. E-mail:cym0620@163.com.

很少一部分的关键基因影响样本的分类,其他的基因往往是冗余的或者是不重要的。在设计基因表达数据分类器之前进行特征选择,可以有效降低分类器的时间复杂度,提高分类精度。目前最常用的基因表达数据特征选择方法主要有 2 类:基于过滤算法(filter)的选择方法[1]与基于 wrapper 的选择方法[2]。基于 filter 的基因表达数据特征选择方法使用数据本身的内在特性作为评价基因的准则,但通过 filter 选择出来的若干个基因可能具有较强的相关性。基于 wrapper 的基因表达数据特征选择方法根据分类器的某种性能来评价基因或基因子集的重要性,而基于 wrapper 方法在基因的选择过程中反复调用分类算法,往往造成较高的时间复杂度。

粗糙集由波兰科学家 Pawlak 于 1982 年提出[3],用于处理不确定、不一致、不精确数据的数学理论工具。现已广泛应用在人工智能、数据挖掘、机器学习等领域[4-7]。然而,Pawlak 粗糙集只能处理离散化的数据,对于现实世界广泛而大量存在的连续数据却缺乏有效的处理能力。基因表达数据也往往都是连续的,目前大多数方法是将基因表达数据先进行离散化[8],离散化过程必定会造成某种程度的信息丢失,并影响分类系统的分类精度。

1　邻域关系

传统粗糙集理论采用等价类形式化地表示知识分类。然而,等价类是基于离散型的数据形成的等价关系划分而得到的,对于连续型的数据并不能构造合适的等价类。因此,下面引入邻域关系处理连续型的基因表达数据,用于基因表达数据的特征选择。

定义 1　设五元组 $I_S=(U,A,V,f,\delta)$ 为邻域信息系统,其中 U 是非空有限集,称为论域,A 是有限特征集,$V=\bigcup_{a\in A}V_a$,V_a 表示特征 a 的值域,$f:U\times A\to V$ 是一个信息函数,即对 $\forall x\in U,a\in A$,有 $f(x,a)\in Va,\delta\in[0,1]$ 为邻域阈值。

定义 2　给定邻域信息系统 $I_S=(U,A,V,f,\delta)$,对于任一 $x,y\in U,B\subseteq A,B=\{a_1,a_2,\dots,a_n\}$,定义 B 上的距离函数 $D_B(x,y)$ 满足:

1) $D_B(x,y)\geqslant 0$,非负;

2) $D_B(x,y)=0$,当且仅当 $x=y$;

3) $D_B(x,y)=D_B(y,x)$,对称;

4) $D_B(x,y)+D_B(y,z)\geqslant D_B(x,z)$。

式中:

$$D_B(x,y)=\left(\sum_{i=1}^{n}(|f(x,a_i)-f(y,a_i)|)^p\right)^{1/p}$$

当 $p=1$ 时,称为曼哈顿距离,当 $p=2$ 时,称为欧氏距离。

定义 3　给定邻域信息系统 $I_S=(U,A,V,f,\delta)$,对于任一 $x\in U$,$B\subseteq A$,定义 x 在 B 上的 δ 邻域 $n_B^\delta(x)$ 为

$$n_B^\delta(x)=\{y\mid x,y\in U,D_B(x,y)\leqslant\delta\}$$

根据距离函数的定义,邻域 $n_B^\delta(x)$ 满足:

1) $n_B^\delta(x)\neq\varnothing$;

2) $x\in n_B^\delta(x)$;

3) $y\in n_B^\delta(x)\Leftrightarrow x\in n_B^\delta(y)$;

4) $\bigcup_{x\in U}n_B^\delta(x)=U$。

定义 4　给定邻域信息系统 $I_S=(U,A,V,f,\delta)$,任一特征子集 $B\subseteq A$ 决定了一个邻域阈值 δ 上的邻域关系 $\mathrm{NR}_\delta(B)$:$\mathrm{NR}_\delta(B)=\{(x,y)\in U\times U\mid D_B(x,y)\leqslant\delta\}$。$U/\mathrm{NR}_\delta(B)$ 构成了 U 的一个邻域划分,称其为 U 上的一簇邻域知识,其中每个邻域划分称为一个邻域类或者邻域知识。上述邻域 $n_B^\delta(x)$ 即为一个邻域类。

2　基于邻域关系的基因选择方法

基于等价关系的信息熵、互信息、粗糙熵等概念度量了知识的粗细程度,也反映了决策系统中的分类能力大小,但主要处理离散型数据的决策系统,对于连续型的数据并不能够直接处理。下面结合邻域关系与邻域类的定义,进一步定义了邻域特征选择概念,用于连续型的基因表达数据的特征选择当中。同时,提出一种基于邻域关系的启发式基因表达数据特征选择算法。

2.1　邻域特征选择

定义 5　定义 $D_T=(U,C\cup D,V,f,\delta)$ 为一个邻域决策表,其中 C 为条件特征,特征值为连续型的数据,邻域阈值为 δ,其邻域划分为 $U/\mathrm{NR}_\delta(C)=\{X_1,X_2,\dots,X_m\}$,$D$ 为决策特征,决策特征是一些决策分类信息,为离散型的数据,以等价关系划分为 $U/D=\{Y_1,Y_2,\dots,Y_n\}$。

定义 6　设 $D_T=(U,C\cup D,V,f,\delta)$ 为一个邻域决策表,$\forall B\subseteq C$,$X\subseteq U$,记 $U/\mathrm{NR}_\delta(B)=\{B_1,B_2,\dots,B_i\}$,则称 $B_*(X)_\delta=\cup\{B_i\mid B_i\in U/\mathrm{NR}_\delta(B),B_i\subseteq X\}$ 为 X 关于 B 的邻域下近似集,称 $B^*(X)_\delta=\cup\{B_i\mid B_i\in U/\mathrm{NR}_\delta(B),B_i\cap X\neq\varnothing\}$ 为 X 关于 B 的邻域上近似集。

定义 7　设邻域决策表 $D_T=(U,C\cup D,V,f,\delta)$,其中 C 为条件特征,特征值为连续型的数据,邻域阈值为 δ,D 为决策特征,决策特征是一些决策分

类信息,为离散型的数据。定义决策特征 D 对条件特征 C 的邻域依赖度为 $\gamma_C(D)_\delta = |C_*(D)_\delta| / |U|$,其中 $|U|$ 表示集合 U 的基数。

定义 8 设邻域决策表 $D_T = (U, C \cup D, V, f, \delta)$,对 $\forall b \in B \subseteq C$,若 $\gamma_B(D)_\delta = \gamma_{B-\{b\}}(D)_\delta$,则称 b 为 B 中相对于 D 是不必要的;否则称 b 为 B 中相对于 D 是必要的。对 $\forall B \subseteq C$,若 B 中任一元素相对于 D 都是必要的,则称 B 相对于 D 独立。

定义 9 设邻域决策表 $D_T = (U, C \cup D, V, f, \delta)$,若 $\forall B \subseteq C$,$\gamma_B(D)_\delta = \gamma_C(D)_\delta$ 且 B 相对于 D 是独立的,则称 B 是选取的关键特征组,这一特征选取过程称为邻域特征选择。

性质 1 设邻域决策表 $D_T = (U, C \cup D, V, f, \delta)$,若 $B_1 \subseteq B_2 \subseteq \ldots \subseteq C$,则 $0 \leqslant \gamma_{B_1}(D)_\delta \leqslant \gamma_{B_2}(D)_\delta \leqslant \ldots \leqslant \gamma_C(D)_\delta \leqslant 1$。

定义 10 设邻域决策表 $DT = (U, C \cup D, V, f, \delta)$,$\forall a \in C$,$R \subseteq C$,定义 a 相对于 R 的特征重要度为 $\mathrm{Sign}(a, R, D) = \gamma_{R\cup\{a\}}(D)_\delta - \gamma_R(D)_\delta$。

2.2 基于邻域关系的基因选择算法

性质 1 表明邻域依赖度具有单调性,因此可以采用删除法或添加法进行特征选择,基因表达数据可以表示成前面定义的邻域决策表,依据上述邻域特征选择的定义,可设计如下基于邻域关系的基因选择算法。下面以定义 10 的特征重要度为启发式信息设计了一种基于邻域关系的基因选择算法。

算法 GSNRS(基于邻域关系的基因选择算法)

输入:基因表达数据决策表 $D_T = (U, C \cup D, V, f, \delta)$;

输出:D_T 的一个邻域约简 R。

1)计算整个条件特征集 C 相对于决策特征 D 的邻域依赖度为 $\gamma_C(D)_\delta$。

2) $R := C$。

3) 当 $\gamma_R(D)_\delta = \gamma_C(D)_\delta$ 重复:

①对所有的 $a \in R$ 计算特征重要度 $\mathrm{Sign}(a, R, D)$;

②在 R 中选择特征 a 满足特征重要度最小;

③ $R := R - \{a\}$。

4) 输出 R。

在算法中,每次选择特征重要度最小的特征,若去掉它后决策表的邻域依赖度仍然不变,则可以去掉,否则保留下来,依次进行下去,直到得到一个条件特征子集,在其中去掉任何一个特征,决策表的邻域依赖度都会改变,则算法结束,该特征子集即为所选取关键特征组。

3 实验结果与分析

下面选用 2 个标准的基因表达数据集来验证 GSNRS 算法的有效性。2 个标准基因表达数据集分别为 Lymphoma 和 Liver cancer。Lymphoma 数据集包含了 96 个样本,4 026 个特征基因,其中 54 个 Othertype 子类和 42 个 B-celllymphoma 子类。Liver cancer 数据集包含了 156 个样本,1 648 个基因,其中 82 个 HCCs 子类和 74 个 nontumorlivers 子类。实验基因数据集如表 1 所示。

表 1 基因表达数据集

Table 1 Gene expression data sets

数据集	基因个数	类别	样本数
Lymphoma	4 026	B-cell	42
Lymphoma	4 026	Other type	54
Liver cancer	1 648	HCCs	82
Liver cancer	1 648	Nontumor livers	74

在 Lymphoma 和 Livercancer 基因表达数据中分别采用文献[9]中粗糙集的特征选择算法 TRS 与本文邻域特征选择算法 GSNRS 进行比较。首先进行预处理,对于有缺失值的数据采用文献[10]的方法进行完备化。基因表达数据集是连续型的数据,对于经典粗糙集特征选择算法,需要对其数据进行离散化,离散化过程采用文献[8]中的方法进行。而本文 GSNRS 特征选择算法,不需要离散化。设邻域参数为 $\delta = 0.1$,特征选择结果如表 2 所示。

表 2 基因数据集特征选择结果

Table 2 Results of gene feature selection

数据集	基因个数	样本数	TRS 算法	GSNRS 算法
Lymphoma	4 026	96	7	6
Liver cancer	1 648	156	6	5

由表 2 可知,TRS 算法在 Lymphoma 数据集中选择出 7 个关键基因,在 Liver cancer 数据集中选择出 6 个关键基因。GSNRS 算法在 Lymphoma 数据集中选择出 6 个关键基因,在 Liver cancer 数据集中选择出 5 个关键基因。下面再比较 2 组基因的分类能力,分别针对选取的关键基因采用 KNN,C5.0 分类器进行分类实验,并用留一交叉法检验分类精确率,实验结果如表 3 所示。

表 3 基因分类精确率

Table 3 Gene classification accuracy rate %

数据集	Lymphoma 本文方法		Liver cancer	
特征选择算法	TRS	GSNRS	TRS	GSNRS
KNN 分类器	93.6	94.9	89.1	91.4
C5.0 分类器	95.1	96.5	91.4	93.2

上述实验结果表明,基于粗糙集的基因选择方

法和基于邻域关系的基因选择方法都能正确提取有效的基因。基于邻域关系的基因选择方法不需要离散化,而且由于避免了离散化过程的造成的信息丢失,提取的特征基因个数较少。在分类精度上,基于邻域关系的基因选择方法提取的基因优于基于粗糙集的基因选择方法提取的基因。

4　结束语

传统粗糙集理论中的特征选择方法往往难以处理连续性的基因表达数据,成为基因表达数据研究中的主要缺陷和障碍。本文针对传统粗糙集理论中难以处理连续数据的缺点,在特征选择中引入邻域关系,定义了邻域依赖度与邻域特征选择等概念,提出了一种基于邻域关系的基因特征选择方法。该特征方法不用对数据进行离散化,避免了信息损失,从而提高了被选择基因的分类准确率。拓展了粗糙集理论的应用范围,为基因表达数据分析技术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'1 Academy of Sciences. [S.l.], USA, 2002: 6567-6572.

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

[3]PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5): 341-356.

[4]BANERJEE M, MITRA S, BANKA H. Evolutinary-rough feature selection in gene expression data[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Application and Reviews, 2007, 37: 622-632.

[5]YANG Ming, YANG Ping. A novel condensing tree structure for rough set feature selection[J]. Neurocomputing, 2008, 71(4/5/6): 1092-1100.

[6]QIAN Yuhua, LIANG Jiye. Positive approximation: an accelerator for attribute reduction in rough set theory[J]. Artificial Intelligence, 2010, 174(9/10): 597-618.

[7]CHEN Yuming, MIAO Duoqian. A rough set approach to feature selection based on power set tree[J]. Knowledge-Based Systems, 2011, 24(2): 275-281.

[8]苗夺谦. Rough set 理论中连续属性的离散化方法[J]. 自动化学报, 2001, 27(3): 296-302.
MIAO Duoqian. A new method of discretization of continuous attributes in rough sets [J]. Acta Automatica Sinica, 2001, 27(3): 296-302.

[9]王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001:24-28.

[10]GRZYMALA-BUSSE J W. Handling missing attribute values[M]. [S.l.]: Springer, 2005: 37-57.

作者简介:

陈玉明,男,1977 年生,副教授,主要研究方向为粒计算、粗糙集、模式识别、数据挖掘等。