

DOI:10.13232/j.cnki.jnju.2019.04.013

## 基于模糊区分矩阵的结直肠癌基因选择

李 藤<sup>1</sup>, 杨 田<sup>2,4</sup>, 代建华<sup>2</sup>, 陈 鸽<sup>3\*</sup>

(1. 中南林业科技大学物流与交通学院, 长沙, 410004;

2. 湖南师范大学智能计算与语言信息处理湖南省重点实验室, 长沙, 410081;

3. 中南大学湘雅医院, 长沙, 410008; 4. 国防科技大学系统工程学院, 长沙, 410073)

**摘 要:** 由于低分化肿瘤很难通过常规组织病理学诊断发现, 而结合基因检测的手段可以准确筛选出针对特定肿瘤的致病基因, 因此基因选择是进行肿瘤分类和临床治疗的关键问题. 肿瘤基因表达数据具有样本小、维度高的特征, 现有的基因选择算法在分类精度和计算效率上还有待提高. 在模糊粗糙集理论的基础上进行区分矩阵模糊化, 并依此设计了模糊区分矩阵属性约简算法. 相比于经典的区分矩阵, 模糊化的区分矩阵能够体现不同属性对于两个对象区分程度的差异, 从而选择区分程度更高的属性而获得更好的分类效果. 数值实验表明该方法提高了肿瘤基因数据的分类精度, 且降低了计算耗时. 实验采用 kNN 分类器进行结直肠癌 (Colon Microarray) 分类特征基因选择实验, 从 2000 个特征基因中筛选出了五个结直肠癌发病相关的关键基因, 且分类精度高达 88.06%.

**关键词:** 模糊粗糙集, 粗糙集, 模糊区分矩阵, 基因选择

中图分类号: TP311

文献标识码: A

## Colon characteristic gene selection based on fuzzy discernibility matrix

Li Teng<sup>1</sup>, Yang Tian<sup>2,4</sup>, Dai Jianhua<sup>2</sup>, Chen Ling<sup>3\*</sup>

(1. College of Logistics and Transportation, Central South University of Forestry and Technology, Changsha, 410004, China; 2. Hunan Provincial Science and Technology Project Foundation, Hunan Normal University,

Changsha, 410081, China; 3. Xiangya Hospital of Central South University, Changsha, 410008, China;

4. College of Systems Engineering, University of Defense Science and Technology, Changsha, 410073, China)

**Abstract:** Since poorly differentiated tumors are difficult to be diagnosed by conventional histopathology, through gene selection can accurate screen disease-causing genes for specific tumors, therefore gene selection has become a key issue in tumor classification and clinical treatment. Tumor gene expression data usually contains thousands of genes but a small number of samples. On the basis of fuzzy rough set theory, the concept of discernibility matrix fuzzification is proposed in this paper. Compared with the classical discernibility matrix, the fuzzy discernibility matrix can reflect the difference in the degree of the two objects distinguished by different attributes, so that the attributes with higher degree of distinction can be selected for better classification effect. Numerical experiments show that this method improves the classification accuracy of tumor gene data and reduces the computation time. In this study, kNN classifier was used for

基金项目: 中国博士后科学基金 (2017T100795), 湖南省自然科学基金 (2017JJ2408), 湖南省重点研发计划 (2018SK2129)

收稿日期: 2019-05-28

\* 通讯联系人, E-mail: 50766131@qq.com

the gene selection of Colon cancer (Colon Microarray), five key genes related to Colon cancer were screened from 2000 feature genes and the classification accuracy was as high as 88.06%.

**Key words:** fuzzy rough sets, rough sets, fuzzy discernibility matrix, gene selection

肿瘤是一种系统生物学疾病,严重威胁人类的生命健康,其发病机制尚不完全清楚. 临床诊疗晚期肿瘤疗效欠佳,早期诊断和治疗是当前提高预后最直接、最有效的方法. 目前结肠癌的诊断主要依靠光镜下形态学和免疫组织化学,但光镜下结肠高分化腺癌与正常组织或低级别瘤变区别不大,可能造成漏诊,而低分化腺癌有时与间叶组织来源肿瘤难以区分,需增加免疫组化检查. 免疫组化依靠特定来源的蛋白表达水平确定来源,但因技术原因,时间较长. 此外,单次组织学切片无法获知“交界性肿瘤”恶变过程的潜在遗传异常,无法预估病变发展的生物学过程,因此基因检测成为癌症治疗中的关键诊疗方式. 研究表明,基因表达谱中与特定肿瘤疾病密切相关的特征基因数量非常有限,筛选出与特定肿瘤相关的基因是进行肿瘤分类研究和实现药物靶向治疗的关键所在.

近些年发展的生物芯片技术可以同时测定不同样本中成千上万的基因表达水平,为研究基因表达谱与肿瘤疾病分类之间的关系提供了数据基础. 然而基因数据的获取难度大且成本高,导致基因数据存在样本小、维度高、噪声大且冗余基因多等显著特征,给基于基因表达谱的肿瘤分类问题带来了巨大的挑战<sup>[1]</sup>,同时极大地激发了许多学者对基于基因表达的肿瘤分类研究的兴趣,成为目前的研究热点之一.<sup>[2-8]</sup>

现有的基因选择方法有很多,这些方法大致上可以分为三类:过滤法、封装法和嵌入法<sup>[9]</sup>. 过滤法一般作为一种独立于分类器的预处理方法,其中基于粗糙集基因选择方法就是一种典型的过滤法,即根据某些标准分析相关基因的特征从而对这些基因进行排序,进而计算每个基因的信息增益. 通常这些评价标准包括:相关系数、距离度量、信息增益和一致性<sup>[4]</sup>. Golub et al<sup>[10]</sup>最早提出信噪比函数来评价

基因的优缺点和肿瘤分子分型的差异;Zhang et al<sup>[11]</sup>基于 ReliefF (Relief Family of algorithms)<sup>[12]</sup>和 MRMR (Minimal-Redundancy-Maximal-Relevance)<sup>[13]</sup>算法设计了新的基因选择算法;Chen et al<sup>[4]</sup>通过调整邻域参数对基因数据进行粒度划分,并在邻域粗糙集的理论基础上提出了并熵的概念,用以评价基因数据的不确定性,这一方法在基因选择上取得了很好的分类效果. 而封装器本质上是一个分类器,它将分类的准确性作为选择最佳基因子集的标准<sup>[4]</sup>. Guyon et al<sup>[14]</sup>代表性地提出了基因选择的递归特征消除算法 (A Recursive Feature Elimination algorithm for gene selection, SVM-RFE),该算法通过递归地消除支持向量机的参数,而成功地应用于基因选择. 但是封装法对分类器很敏感,性能不稳定且时间复杂度通常比较高<sup>[4]</sup>. 除这两种方法之外,嵌入法也得到了不少学者的关注,惩罚支持向量机 (Penalized Support Vector Machine, PSVM)是最有效的嵌入法之一. PSVM通过将 SVM 与惩罚函数相结合很好地应用到基因选择和分类上<sup>[15]</sup>. 通过构造不同的惩罚函数可以构建不同的 PSVM 模型,代表性的有最小绝对收缩和选择算子 (the Least Absolute Shrinkage and Selection Operator, LASSO)<sup>[16]</sup>和平滑剪切绝对偏差惩罚 (the Smoothly Clipped Absolute Deviation penalty, SCAD)<sup>[17]</sup>. 而采用 SCAD 惩罚的 PSVM 模型的效果取决于恰当的调节参数<sup>[5]</sup>.

基因选择是从成千上万的基因数据中找到肿瘤发病的关键基因,本质上可以看作一个数据预处理过程. 1982 年 Pawlak 提出的粗糙集理论是处理模糊和不确定信息的有效工具,无需先验知识即可有效进行数据预处理,因而在特征选择中扮演重要角色<sup>[18-21]</sup>. 但由于 Pawlak 粗糙集是建立在等价关系的基础上,需要对数据

进行离散化,会导致信息丢失. 模糊集是 Zadeh 在 1965 年提出的,它在处理连续型和混合型数据时不需要进行数据离散化处理,可以获得更好的分类结果. 为提高模型的学习能力、避免离散化, Dubois and Prade<sup>[22]</sup> 提出模糊粗糙集的概念,有效克服了连续型或混合型数据离散化处理问题,更加完整地保存了连续型属性的分类信息. 模糊粗糙集理论中的特征提取问题成为近年的研究热点,大量基于模糊粗糙集理论的特征选择算法被提出. Jensen and Shen<sup>[23]</sup> 最先将经典粗糙集模型中的依赖函数引入模糊案例中,提出一种基于模糊粗糙集的属性约简算法. Hu et al<sup>[24]</sup> 将信息熵扩展到模糊粗糙集以评估特征和标签之间的相关性,并利用新的信息熵计算模糊粗糙近似空间的不确定性<sup>[25]</sup>. Chen et al<sup>[26]</sup> 和 Tsang et al<sup>[27]</sup> 将传统区分矩阵的概念引入模糊粗糙集并设计了相应的属性约简算法,该方法是将粗糙集进行了模糊化,但不是对区分矩阵进行模糊化,建立的仍然是经典区分矩阵,即区分矩阵的元素仍然是经典集合. Dai et al<sup>[28]</sup> 利用模糊相似关系从样本对的角度进行特征提取. Wang et al<sup>[29]</sup> 通过引入两个参数来调控模糊依赖度函数,改善了模糊粗糙依赖度仅能获取最大依赖度的不足,解决了传统模糊粗糙集模型中错误分类的问题. 这些方法都能有效的进行特征提取,存在各自的优缺点.

模糊依赖度方法空间复杂度低,鲁棒性强,缺点为:(1) Wang et al<sup>[29]</sup> 指出经典模糊依赖度法只保留最大依赖函数而不能保持样本在它自身的决策类中隶属度最大,可能出现错误分类. 其模糊决策类的定义是基于所有特征的模糊邻域产生的,这意味着需要通过计算所有模糊邻域来生成模糊决策类,增加了计算成本. (2) Wang et al<sup>[29]</sup> 对依赖度模型进行了修改,计算效率有所提高,但模糊决策类的生成没有得到改进,且设置的参数多,计算成本依然很高.

区分矩阵在分类表现上存在一定优势,缺点为:(1) 针对样本规模的计算复杂度,尤其是空间复杂度比较高,对计算机内存要求高.

(2) 经典区分矩阵从区分的角度考虑属性的重要度,只要对象对在该属性下的差异超过给定的阈值,则该属性相对于这个对象对的评分为 1,否则评分为 0. 但是不同属性对于对象对存在区分程度的差异,经典区分矩阵忽略了这种差异,只要满足区分条件,即使是两个区分程度差异非常大的属性也会在特征选择过程中不加区分地给予同样的评估值,这样会导致那些区分效果更好的属性无法被优先选择. 对区分矩阵进行模糊化处理则可以体现属性区分程度差异的量化,因此选择出的属性是区分程度最大的,大大提高了分类精度. 尤其对于肿瘤基因数据,如果区分效果更好的基因被漏选,势必会影响基因研究的方向和肿瘤治疗方案的选择.

由于基因数据的样本个数少,所以针对基因数据基于区分矩阵算法的实际内存占用和计算量都不高,甚至比其他算法更低. 本文构造了模糊化的区分矩阵以解决上述问题,首先提出模糊区分度的概念,并针对基因表达谱自身的特点提出一种新的基于模糊区分矩阵的特征提取算法. 数值实验表明该方法能快速处理基因数据,并明显提高肿瘤分类的精度.

## 1 预备知识

**1.1 模糊粗糙集** 在 Pawlak 粗糙集理论中,等价关系是一个非常重要的概念. 但是在模糊粗糙集中,通常用模糊相似关系来代替等价关系.  $U$  上的模糊二元关系  $\bar{R}$  也称之为模糊  $T$ -相似关系,如果满足自反性、对称性和  $T$ -传递性. 而相似划分  $[x]_{\bar{R}}$  是  $U$  上的一个模糊集合,被定义为:

$$[x]_{\bar{R}}(y) = \bar{R}(x, y), y \in U$$

给定一个非空有限集  $U$ ,  $\bar{R}$  是  $U$  上的一个模糊二元关系,用矩阵表示为:

$$M(\bar{R}) = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix}$$

其中,  $r_{ij} \in [0, 1]$  表示  $x_i$  和  $x_j$  之间的关系值. 模糊关系矩阵具有如下性质<sup>[30]</sup>:

- (1)  $\bar{R}_1 = \bar{R}_2 \Leftrightarrow \bar{R}_1(x, y) = \bar{R}_2(x, y)$ ;  
 (2)  $\bar{R} = \bar{R}_1 \cup \bar{R}_2 \Leftrightarrow \bar{R} = \max \bar{R}_1(x, y), \bar{R}_2(x, y)$ ;  
 (3)  $\bar{R} = \bar{R}_1 \cap \bar{R}_2 \Leftrightarrow \bar{R} = \min \bar{R}_1(x, y), \bar{R}_2(x, y)$ ;  
 (4)  $\bar{R}_1 \subseteq \bar{R}_2 \Leftrightarrow \bar{R}_1(x, y) \leq \bar{R}_2(x, y)$ .

**定义 1**<sup>[31]</sup> 假定  $\bar{X}$  是一个模糊集合, 可以表示为:  $\mu_{\bar{X}}(x_1)/x_1, \mu_{\bar{X}}(x_2)/x_2, \dots, \mu_{\bar{X}}(x_n)/x_n$ . 其中  $\mu_{\bar{X}}$  是论域  $\bar{X}$  到  $[0, 1]$  的一个映射, 即:

$$\mu_{\bar{X}}: \bar{X} \rightarrow [0, 1], x \mapsto \mu_{\bar{X}}$$

其中,  $\mu_{\bar{X}}(x_j)$  表示元素  $x_j$  对模糊集合  $\bar{X}$  的隶属程度.

**性质 1**<sup>[30]</sup> 给定一个信息系统  $S = (U, A, D), B, B_1, B_2 \subseteq A$ , 称  $\bar{R}_B$  为属性子集  $B$  产生的模糊二元关系, 它具有如下性质:

- (1)  $\bar{R}_B = \bigcap_{a \in B} \bar{R}_a$ ;  
 (2)  $\bar{R}_{B_1 \cup B_2} = \bar{R}_{B_1} \cap \bar{R}_{B_2}$ .

**定义 2**<sup>[30]</sup> 由模糊二元关系  $\bar{R}$  生成的模糊空间可以定义为:

$$\langle U, \bar{R} \rangle = ([x_1]_{\bar{R}}, [x_2]_{\bar{R}}, \dots, [x_n]_{\bar{R}})$$

其中:

$$[x_i]_{\bar{R}} = \frac{r_{i1}}{x_1} + \frac{r_{i2}}{x_2} + \dots + \frac{r_{in}}{x_n}$$

是  $x_i$  和  $\bar{R}$  产生的模糊等价关系.  $[x_i]_{\bar{R}}$  称之为  $x_i$  的模糊邻域, 且  $r_{ij}$  表示  $x_i$  等价于  $x_j$  的程度. “+”代表元素的并关系, 而  $[x_i]_{\bar{R}}$  的基数可以用式(1)计算:

$$[x_i]_{\bar{R}} = \sum_{j=1}^n r_{ij} \quad (1)$$

Dubois and Prade<sup>[22]</sup> 最早提出了第一种模糊粗糙集模型, 并用模糊二元关系定义了上、下逼近算子. 而 Hu et al<sup>[24]</sup> 给出了混合数据背景下模糊粗糙集的一个定义, 如下所示.

**定义 3**<sup>[25]</sup> 给定  $\langle U, \bar{R} \rangle$  为一个模糊近似空间,  $\bar{X}$  是论域  $U$  上的模糊子集. 那么下逼近算子  $\underline{\bar{R}}\bar{X}$  和上逼近算子  $\overline{\bar{R}}\bar{X}$  可以分别定义为:

$$\underline{\bar{R}}\bar{X} = \{x_i | [x_i]_{\bar{R}} \subseteq \bar{X}, x_i \in U\}$$

$$\overline{\bar{R}}\bar{X} = \{x_i | [x_i]_{\bar{R}} \cap \bar{X} \neq \emptyset, x_i \in U\}$$

其中,  $[x_i]_{\bar{R}} \subseteq \bar{X}$  意为隶属度,

$$\mu_{[x_i]_{\bar{R}}}(x_i) \leq \mu_{\bar{X}}(x_i)$$

而  $[x_i]_{\bar{R}} \cap \bar{X} \neq \emptyset$  则表示:

$$\min \{\mu_{[x_i]_{\bar{R}}}(x_i) \leq \mu_{\bar{X}}(x_i)\} \neq \emptyset$$

## 1.2 粗糙集

**定义 4**<sup>[32]</sup> 假设  $U$  是一个论域,  $R = \{R_1, R_2, \dots, R_m\}$  是  $U$  上的一族广义模糊二元关系, 那么  $(U, R)$  可以称之为一个关系信息系统,  $R$  称之为条件属性集且存在  $\text{Int}R = \bigcap_{i=1}^n R_i$ .

**定义 5**<sup>[32]</sup> 假设  $(U, R)$  是一个模糊关系信息系统, 且有  $R_i \in R$ , 如果  $\text{Int}R = \text{Int}(R - R_i)$ , 则称  $R_i$  在  $R$  中是不必要的, 否则称  $R_i$  是必要的. 对于任意子集  $P \in R$ , 如果  $P$  中任意元素都是必要的且满足  $\text{Int}R = \text{Int}P$ , 则  $P$  中所有必要元素的并称之为  $R$  的核, 用  $\text{Core}(R)$  表示.

**定义 6**<sup>[32]</sup> 令  $S = (U, R)$  是一个关系信息系统,  $U = \{x_1, x_2, \dots, x_n\}$ , 用  $M(U, R)$  来表示一个  $n \times n$  的矩阵  $(c_{ij})$ , 称之为  $(U, R)$  上的区分矩阵, 定义为:

$$c_{ij} = \{R \in R : x_j \notin R(x_i), x_i, x_j \in U\}$$

## 2 基于区分矩阵模糊化的特征选择方法

**定义 7**<sup>[30]</sup> 由模糊二元关系  $\bar{R}$  生成的模糊空间可以定义为:

$$\langle U, \bar{R} \rangle = ([x_1]_{\bar{R}}, [x_2]_{\bar{R}}, \dots, [x_n]_{\bar{R}})$$

其中,

$$[x_i]_{\bar{R}} = \frac{r_{i1}}{x_1} + \frac{r_{i2}}{x_2} + \dots + \frac{r_{in}}{x_n}$$

是  $x_i$  和  $\bar{R}$  产生的模糊等价关系.  $[x_i]_{\bar{R}}$  称之为  $x_i$  的模糊邻域, 且  $r_{ij}$  表示  $x_i$  等价于  $x_j$  的程度.

本文采用 Hu et al<sup>[25]</sup> 提出的上、下逼近算子. 给定  $\langle U, \bar{R} \rangle$  为一个模糊近似空间,  $\bar{X}$  是论域  $U$  上的模糊子集. 那么下逼近算子  $\underline{\bar{R}}\bar{X}$  和上逼近算子  $\overline{\bar{R}}\bar{X}$  可以分别定义为:



$$\bar{R}\bar{X}=\{x_i|[x_i]_{\bar{R}}\subseteq\bar{X}, x_i\in U\}$$

$$\bar{R}\bar{X}=\{x_i|[x_i]_{\bar{R}}\cap\bar{X}\neq\emptyset, x_i\in U\}$$

不同于基于等价关系的区分矩阵法,本文的模糊区分矩阵法是基于模糊二元关系,需要计算论域 $U$ 中所有对象的邻域 $[x_i]_{\bar{R}}$ ,即对于任意 $\bar{R}_i\in\bar{R}$ 需要检查 $x_j$ 属于 $[x_i]_{\bar{R}}$ 的程度. Hu et al<sup>[33]</sup>指出采用后继邻域方法可以获得时间的线性变化,因此依据邻域计算模糊区分度得到模糊区分矩阵设计的算法,时间复杂度比较低.

接下来介绍一个新的概念用来评估特征子集的区分能力. 首先需要决策表的数据进行模糊化处理,生成一个模糊区分矩阵.

给定决策表 $S=(U, A\cup D)$ , 其中 $U=\{x_1, x_2, \dots, x_n\}$ 为非空有限样本集合,  $A=\{a_1, a_2, \dots, a_m\}(\forall a_j\subset\bar{R}, 1\leq j\leq m)$ 为条件属性集,  $D=\{d\}$ 为决策属性集. 首先需要对所有连续属性进行标准化处理,将各属性值标准到 $[0, 1]$ 区间. 本文采用最大-最小归一化方法:

$$a(x)' = \frac{a(x) - a(x)_{\min}}{a(x)_{\max} - a(x)_{\min}} \quad (2)$$

其中,  $a(x)_{\max}$ 表示所有对象在属性 $a$ 下的最大值,  $a(x)_{\min}$ 表示所有对象在属性 $a$ 下的最小值.

判定各属性的区分能力常用的思路是:不在同一个决策类的对象对,当对象对 $x_i$ 和 $x_j(1\leq i, j\leq m)$ 之间距离大于一定的范围,才说明两者之间具有区分能力.

**定义8 模糊区分度** 给定决策表 $\langle U, A\cup D \rangle$ , 其中 $U=\{x_1, x_2, \dots, x_n\}$ 为非空有限样本

集合,  $A=\{a_1, a_2, \dots, a_m\}(\forall a_j\subset\bar{R}, 1\leq j\leq m)$ 为条件属性集,  $D=\{d\}$ 为决策属性集. 不在同一个决策类的对象对 $(x_i, x_j)$ 之间的模糊区分度,其计算公式可以定义为:

$$\mu_a(x_i, x_j) = \begin{cases} \frac{|a(x_i) - a(x_j)| - \delta}{\delta}, & |a(x_i) - a(x_j)| > \delta \\ 0, & |a(x_i) - a(x_j)| \leq \delta \end{cases} \quad (3)$$

其中,  $a(x_i)$ 表属性值,  $\delta$ 表邻域阈值. 而式(3)表示不在同一个决策类的对象 $(x_i, x_j)$ 则需要计算其在属性 $a$ 下的模糊区分度, 否则 $\mu_a(x_i, x_j)$ 直接等于零. 模糊区分度表示属性 $a$ 对于对象之间的区分程度,若对象对 $(x_i, x_j)$ 之间的距离大于 $\delta$ ,则模糊区分度大于零,表示两者之间存在区分差异,否则模糊区分度为零.

**定义9** 设决策表 $S=(U, A, D, V, f)$ , 其中 $U=\{x_1, x_2, \dots, x_n\}$ 为非空有限样本集合,  $A=\{a_1, a_2, \dots, a_m\}$ 为条件属性集,  $D=\{d\}$ 为决策属性集.  $M=(c_{ij})_{n\times n}$ 为 $S$ 的模糊区分矩阵:

$$c_{ij} = \begin{cases} (\mu_{a_1}(x_i, x_j), \mu_{a_2}(x_i, x_j), \dots, \mu_{a_m}(x_i, x_j)), & d(x_i) \neq d(x_j) \\ (0, 0, \dots, 0), & \text{others} \end{cases} \quad (4)$$

其中,  $c_{ij}$ 为定义在 $A=\{a_1, a_2, \dots, a_m\}$ 上的模糊集,  $c_{ij}(a_t)=\mu_{a_t}(x_i, x_j)$ 表示属性 $a_t$ 在模糊集 $c_{ij}$ 中的隶属度.

利用模糊区分函数进行模糊化处理,生成的模糊区分矩阵 $M$ 可以表示为:

$$M_{n\times n} = \begin{bmatrix} (0, 0, \dots, 0) & (\mu_{a_1}(x_1, x_2), \dots, \mu_{a_m}(x_1, x_2)) & \dots & (\mu_{a_1}(x_1, x_n), \dots, \mu_{a_m}(x_1, x_n)) \\ & (0, 0, \dots, 0) & \dots & (\mu_{a_1}(x_2, x_n), \dots, \mu_{a_m}(x_2, x_n)) \\ & & \dots & (\mu_{a_1}(x_i, x_n), \dots, \mu_{a_m}(x_i, x_n)) \\ & & & (0, 0, \dots, 0) \end{bmatrix}$$

由于模糊邻域生成的二元关系有对称性,因此将模糊区分矩阵 $M$ 定义为上三角矩阵.

**定义10** 给定决策表 $S=(U, A, D, V, f)$ , 其中 $U=\{x_1, x_2, \dots, x_n\}$ 为非空有限样本集合,  $A=\{a_1, a_2, \dots, a_m\}(\forall a_j\subset\bar{R}, 1\leq j\leq m)$ 为条

件属性集,  $D=\{d\}$ 为决策属性集. 属性 $a_t$ 的模糊区分度为:

$$\mu(a_t) = \sum_{i,j=1}^n c_{i,j}(a_t), \forall a_t \in A \quad (5)$$

式(5)表示属性 $a_t$ 对所有对象总的模糊区分度,其值越大表示该属性的区分能力越强.

**定义 11** 给定决策表  $S=(U, A, D, V, f)$ , 其中  $U=\{x_1, x_2, \dots, x_n\}$  为非空有限样本集合,  $A=\{a_1, a_2, \dots, a_m\}$  为条件属性集,  $D=\{d\}$  为决策属性集.

(1) 若  $P \subseteq A$  为满足下列条件的极小子集, 对  $\forall c_{i,j} \neq (0, 0, 0, \dots, 0), \exists a_i \in P, \text{ s. t. } c_{i,j}(a_i) > 0$ , 则  $P$  为  $A$  的约简.

(2)  $\text{Core}(A) = \{a | \mu_a(x, y) > 0, \forall b \in A - \{a\}, \mu_b(x, y) = 0\}$  为  $A$  的核.

下面介绍一种求最大区分度的属性约简启发式算法.

### 3 算法设计

本文基于模糊区分矩阵的属性约简算法 (Reduction algorithm based on Fuzzy Discernibility Matrix, FDM), 其目的是将区分能力更高的属性优先选入属性子集. 该算法分为两个部分, 首先根据第二部分介绍的理论对原始数据进行模糊化处理, 生成模糊区分矩阵 (Step1), 再在模糊区分矩阵的基础上求得属性约简结果 (Step2). 其中,  $a(x_i)$  表属性值,  $\mu_a(x_i, x_j)$  表示属性  $a$  对对象  $(x_i, x_j)$  的模糊区分度,  $\mu(a_i) = \sum_{i,j=1}^n c_{ij}(a_i)$  表示属性  $a_i$  的模糊区分度.

Step1. 生成模糊区分矩阵

输入: 决策表  $\langle U, A, D \rangle$ , 邻域阈值  $\delta$ , 其中  $U = \{x_1, x_2, \dots, x_n\}, A = \{a_1, a_2, \dots, a_m\}, D = \{d\}$ ;

输出: 模糊区分矩阵  $M$ .

(1) 将决策表  $\langle U, A, D \rangle$  的数值标准化到  $[0, 1]$  区

间; 初始化  $M = (c_{ij})_{n \times n}, c_{ij} = (0, 0, \dots, 0)$  是零向量;

(2) if  $f(x_i, d) \neq f(x_j, d)$  and  $|a(x_i) - a(x_j)| > \delta$

(3)  $\mu_a(x_i, x_j) = \frac{|a(x_i) - a(x_j)| - \delta}{\delta}$

(4) end if

Step2. 基于模糊区分矩阵得到属性约简

输入: 模糊区分矩阵  $M$ ;

输出: 约简  $Red$ .

(1) 初始化  $Red = \emptyset$ ;

(2) Do while  $\sum_{i=1}^m \mu(a_i) \neq 0$ ;

(3) 找到区分度最大的属性  $a_0$  (即  $\mu(a_0) = \max\{\mu(a_i) | a_i \in A\}$ );

(4)  $\forall j > i$ , if  $c_{ij}(a_0) > 0$  则令  $c_{ij} = (0, 0, \dots, 0)$ ;

(5)  $Red = Red \cup \{a_0\}$

(6) end while

**例 1** 给定表 1 所示的决策表  $\langle U, A, D \rangle$ , 其中  $U = \{x_1, x_2, x_3, x_4\}$  为非空有限样本集合,  $A = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$  为条件属性集,  $D = \{d\}$  为决策属性集. 令  $\delta = 0.36$ , 经过式 (3) 和式 (4) 的模糊化处理, 可以获得表 2 所示的模糊区分矩阵.

表 1 决策表

Table 1 Decision table

$U$	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$D$
$x_1$	0.32	0.77	0.5	0.48	0.1	0.9	0.36	1
$x_2$	0.56	0.98	0.72	0.9	0.4	0.38	0.82	2
$x_3$	0.88	0.42	0.77	0.62	0.54	0.79	0.36	1
$x_4$	0.6	0.67	0.51	0.7	0.3	0.22	0.86	3

表 2 模糊区分矩阵

Table 2 Fuzzy discernibility matrix

$U$	$x_1$	$x_2$	$x_3$	$x_4$
$x_1$	(0, 0, 0, 0, 0, 0, 0)	(0, 0, 0, 0.0938, 0, 0.25, 0.1562)	(0, 0, 0, 0, 0, 0, 0)	(0, 0, 0, 0, 0, 0.5, 0.2188)
$x_2$		(0, 0, 0, 0, 0, 0, 0)	(0, 0.3125, 0, 0, 0, 0.0781, 0.1562)	(0, 0, 0, 0, 0, 0, 0)
$x_3$			(0, 0, 0, 0, 0, 0, 0)	(0, 0, 0, 0, 0, 0.3281, 0.2188)
$x_4$				(0, 0, 0, 0, 0, 0, 0)

用经典的区分矩阵方法得到的区分矩阵如表3所示,矩阵每个位置上的元素是一个精确集.因此,通过析取和合取运算求得的约简为: $Red=\{a_6\}$ 或 $\{a_7\}$ .

根据该算法求出例1的约简和核分别为: $Core(A)=\{a_6\}$ ,  $Red=\{a_6\}$ .可以看出,利用模糊区分矩阵法求得的约简跟经典区分矩阵不一样,这正是因为经典区分矩阵忽略了属性之间区分程度的差异.在例1中属性 $a_6$ 的区分程度比 $a_7$ 的区分程度更高,因此采用模糊区分矩阵法则将 $\{a_6\}$ 作为关键属性;而采用经典的区分矩阵法会同时将 $\{a_6\}$ 和 $\{a_7\}$ 都当作约简,假如在不知道这两个属性的区分程度时, $\{a_7\}$ 被当作关键属性,得到的分类精度将会低一些.

该算法第一步需要计算的矩阵元素是 $\frac{n \times (n-1)}{2}$ ,因此这一步的算法复杂度是 $O(n^2 \times m)$ ;第二步用贪心算法求约简和核,该步骤的时间复杂度为 $O(\min\{n, m\})$ ;因此整个算法的时间复杂度为 $O(n^2 \times m)$ .在实际计算中,如果两个样本属于同一个决策类则不需要针对每一个属性计算它们的区分度,因此实际的计算量远低于 $\frac{n \times (n-1)}{2} \times m$ .

表3 经典区分矩阵

Table 3 Classical discernibility matrix

$U$	$x_1$	$x_2$	$x_3$	$x_4$
$x_1$	$\emptyset$	$\{a_4, a_6, a_7\}$	$\emptyset$	$\{a_6, a_7\}$
$x_2$		$\emptyset$	$\{a_2, a_6, a_7\}$	$\emptyset$
$x_3$			$\emptyset$	$\{a_6, a_7\}$
$x_4$				$\emptyset$

## 4 数值实验

基于基因表达水平的肿瘤分类对肿瘤诊断有重要意义.本文将基于模糊区分矩阵理论的属性约简算法用于肿瘤分类,并具体应用到结直肠癌肿瘤分类上.为评价该模型在连续数据集特征选择上的有效性,将模糊区分矩阵算法

(FDM)与其他几种代表性的模糊粗糙集算法进行比较,它们分别是:基于图论的覆盖决策系统属性约简算法(Reduction algorithm of Covering Decision systems based on Graph theory, CDG)<sup>[34]</sup>、基于邻域判别指数的启发式算法(Heuristic Algorithm based on Neighborhood Discrimination Index, HANDI)<sup>[35]</sup>和基于融合模糊粗糙集的启发式算法(Heuristic algorithm based on Fitting fuzzy Rough Sets, NFRS)<sup>[29]</sup>.由于分类精度对于临床诊疗至关重要,因此本文将基因子集的分类精度作为第一指标,将运行时间作为第二指标.

以上所有方法的数值实验都是在 Matlab R2016a中完成,运行环境:Windows 7 and Intel (R) Core(TM) i5-6200U CPU @ 2.30 GHz,运行内存为4.0 GB.实验中使用的分类器为kNN( $k=3$ ).

**4.1 算法有效性测试** 由于该实验主要是针对基因数据,因此选用类似的数据集用以说明该算法在基因数据上的有效性.该实验采用的数据集均来自UCI(<http://archive.ics.uci.edu/ml/datasets.html>),如表4所示.

表4 数据集的描述

Table 4 Description of the datasets

Datasets	Samples	Features	Classes
Heart	270	13	2
Primary-tumor	339	17	2
Mammographic	961	5	2
Lung cancer	32	56	2
Gastrointestinal	76	698	3
Leukemia	72	7070	2

所有的数值属性首先都需要经过标准化处理到 $[0,1]$ 区间.实验中设置的邻域参数 $\delta$ 表示邻域半径大小, $\delta$ 以步长0.02在0~0.5变化.通过调控邻域参数的大小,可以有效控制各属性对样本的区分能力的差异.由于不同的算法获得最佳分类精度的特征子集不同,本文所展示的分类精度都是最佳分类精度,具体实

验结果如表 5 所示,表中加粗的数据表示该算法测试的对应数据集的分类精度最高,可以看出,基本上所有算法对上述六个医疗数据集进

行属性约简之后,分类精度对比原始精度都有所提高,其中 FDM 算法略优于另外几种算法.

表 5 不同算法的分类精度(%)

Table 5 Classification accuracy of different algorithms(%)

Datasets	Raw	CDG	NFRS	HANDI	FDM
Heart	80.15±7.5	80.86±4.32	81.05±5.74	81.11±9.01	<b>81.30±10.93</b>
Primary-tumor	66.45±4.84	67.77±8.37	<b>69.21±6.83</b>	68.56±4.21	68.12±4.75
Mammographic	75.50±4.42	76.79±3.00	76.69±4.56	<b>77.26±3.05</b>	77.09±2.87
Lung cancer	83.65±21.15	83.75±33.75	80.63±19.37	83.75±33.75	<b>85.63±35.62</b>
Gastrointestinal	54.55±0.00	71.82±17.27	71.36±15.00	71.76±13.24	<b>74.55±15.45</b>
Leukemia	86.63±13.37	97.86±7.38	97.62±7.14	97.86±2.62	<b>98.10±2.86</b>
Average	74.49	79.81	79.43	80.05	<b>80.80</b>

从表 6 的运行时间对比来看,FDM 算法在六个数据集上的平均用时为 5.497 s(黑体字),

时间优势明显,尤其对于高维数据的处理,FDM 算法是非常高效的,耗时较少.

表 6 不同算法求约简的时间(单位:s)

Table 6 Running time of different algorithms (unit: s)

Datasets	CDG	NFRS	HANDI	FDM
Heart	0.811	14.851	6.568	1.108
Primary-tumor	1.810	28.533	14.883	2.012
Mammographic	4.820	49.250	31.247	3.838
Lung cancer	0.156	1.747	0.686	0.094
Gastrointestinal	1.794	281.083	76.485	3.151
Leukemia	27.487	1520.600	168.138	22.776
Average	6.146	316.011	49.668	<b>5.497</b>

**4.2 基因选择实例分析** 基因数据集具有数据维度高、样本数量少的特征,因此从成千上万维的基因中选择出关键基因对肿瘤的诊断至关重要.本文的分析对象是结直肠癌数据(Colon Microarray),下载地址是 <http://featureselection.asu.edu/datasets.php>.关于 Colon 数据集的描述如表 7 所示.

本文从区分的角度,利用模糊区分矩阵方法设计了相应的算法,在 Colon 结直肠癌数据集中,从 2000 个基因中,筛选出了五个与结直肠癌发病相关的基因,它们分别是第 235,341,

表 7 Colon 数据集的描述

Table 7 Description of Colon dataset

Tumor dataset	Gene	Sample	Positive	Normal
Colon	2000	62	40	22

441,1423,1760 个属性,这对肿瘤药物研究和临床诊疗都提供了重要的参考.Chen et al<sup>[4]</sup>通过调整邻域参数对基因数据进行粒度划分,并在邻域粗糙集的理论基础上提出了基于信息熵增益的基因选择方法(The Entropy Gain-based



Gene Selection algorithm for a neighborhood gene dataset, EGGS), 该方法在基因选择实验上取得了较好的分类结果. 在此将本文的实验

结果与 Chen et al<sup>[4]</sup> 的基因选择结果进行了比较, 实验结果如表 8 所示.

表 8 两种基因选择方法的比较

Table 8 Comparison of two gene selection methods

Method	Gene	$\delta$	Feature Selected	kNN
EGGS	2000	0.25	H55933, T58861, H61410, J05032, H06524	86.25
FDM	2000	0.24	T63591, D13315, X83412, J02854, R62438	88.06

上述实验结果表明, 无论是基于信息熵还是区分矩阵提取的基因都能够保持或者提高对于肿瘤患病的分类能力, 显然基于模糊区分矩阵选出的基因子集具有更高的分类精度. 以后将会利用此方法继续研究湖南地区结直肠癌的基因数据.

## 5 结 论

本文基于模糊邻域关系, 提出模糊区分矩阵的概念, 相对于经典的区分矩阵法, 模糊区分矩阵法体现了不同属性区分程度的差异, 并在此基础上筛选具有更强分类能力的属性, 提高分类精度. 实验结果表明基于模糊区分矩阵的特征选择算法具有更高的分类精度, 提高了分类性能, 能精准地筛选出关键基因. 我们将会继续利用此方法研究湖南地区的结直肠癌基因数据.

### 参考文献

- [1] 叶明全, 高凌云, 伍长荣等. 基于对称不确定性和邻域粗糙集的肿瘤分类信息基因选择. 数据采集与处理, 2018, 33(3): 426—435. (Ye M Q, Gao L Y, Wu C R, et al. Informative gene selection for tumor classification based on symmetric uncertainty and neighborhood rough set. Journal of Data Acquisition and Processing, 2018, 33(3): 426—435.)
- [2] Dai J H, Xu Q. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. Applied Soft Computing, 2013, 13(1): 211—221.
- [3] Wang S L, Li X L, Zhang S W, et al. Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction. Computers in Biology and Medicine, 2010, 40(2): 179—189.
- [4] Chen Y M, Zhang Z J, Zheng J Z, et al. Gene selection for tumor classification using neighborhood rough sets and entropy measures. Journal of Biomedical Informatics, 2017, 67: 59—68.
- [5] Al-Thanoon N A, Qasim O S, Algamal Z Y. Tuning parameter estimation in SCAD - support vector machine using firefly algorithm with application in gene selection and cancer classification. Computers in Biology and Medicine, 2018, 103: 262—268.
- [6] 徐菲菲, 苗夺谦, 魏莱. 基于模糊粗糙集的肿瘤分类特征基因选取. 计算机科学, 2009, 36(3): 196—200. (Xu F F, Miao D Q, Wei L. Feature Selection for Cancer Classification Based on Fuzzy Rough Sets. Computer Science, 2009, 36(3): 196—200.)
- [7] Cao J, Zhang L, Wang B J, et al. A fast gene selection method for multi-cancer classification using multiple support vector data description. Journal of Biomedical Informatics, 2015, 53: 381—389.
- [8] Model F, Adorján P, Olek A, et al. Feature selection for DNA methylation based cancer classification. Bioinformatics, 2001, 17(S1): S157—S164.
- [9] Algamal Z Y, Lee M H. Penalized logistic regression with the adaptive LASSO for gene

- selection in high-dimensional cancer classification. *Expert Systems with Applications*, 2015, 42(23): 9326—9332.
- [10] Golub 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.
- [11] Zhang Y, Ding C, Li T. Gene selection algorithm by combining ReliefF and MRMR. *BMC Genomics*, 2008, 9(S1):S27.
- [12] Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 2003, 53(1—2):23—69.
- [13] Peng H C, Long F H, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(8): 1226—1238.
- [14] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. *Machine Learning*, 2002, 46(1—3):389—422.
- [15] Wang L, Zhu J, Zou H. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, 2008, 24(3): 412—419.
- [16] Ghosh D, Chinnaiyan A M. Classification and selection of biomarkers in genomic data using LASSO. *Journal of Biomedicine and Biotechnology*, 2005, 2005(2):147—154.
- [17] Mylona K, Koukouvinos C, Theodoraki E M, et al. Variable selection via nonconcave penalized likelihood in high dimensional medical problems. *International Journal of Applied Mathematics and Statistics*, 2009, 14:1—11.
- [18] Herawan T, Deris M M, Abawajy J H. A rough set approach for selecting clustering attribute. *Knowledge-Based Systems*, 2010, 23(3):220—231.
- [19] Parthala N M, Shen Q. Exploring the boundary region of tolerance rough sets for feature selection. *Pattern Recognition*, 2009, 42(5):655—667.
- [20] Mi J S, Wu W Z, Zhang W X. Approaches to knowledge reduction based on variable precision rough set model. *Information Sciences*, 2004, 159(3—4):255—272.
- [21] Qian Y H, Liang J Y, Pedrycz W, et al. Positive approximation: an accelerator for attribute reduction in rough set theory. *Artificial Intelligence*, 2010, 174(9—10):597—618.
- [22] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets. *International Journal of General Systems*, 1990, 17(2—3):191—209.
- [23] Jensen R, Shen Q. Fuzzy-rough attribute reduction with application to web categorization. *Fuzzy Sets and systems*, 2004, 141(3):469—485.
- [24] Hu Q H, Yu D, Xie Z X, et al. Fuzzy probabilistic approximation spaces and their information measures. *IEEE Transactions on Fuzzy Systems*, 2006, 14(2):191—201.
- [25] Hu Q H, Yu D R, Xie Z X. Information-preserving hybrid data reduction based on fuzzy-rough techniques. *Pattern Recognition Letters*, 2006, 27(5): 414—423.
- [26] Chen D G, Zhang L, Zhao S Y, et al. A novel algorithm for finding reducts with fuzzy rough sets. *IEEE Transactions on Fuzzy Systems*, 2012, 20(2):385—389.
- [27] Tsang E C C, Chen D G, Daniel S Y, et al. Attributes reduction using fuzzy rough sets. *IEEE Transactions on Fuzzy Systems*, 2008, 16(5): 1130—1141.
- [28] Dai J H, Hu H, Wu W Z, et al. Maximal -discernibility - pair - based approach to attribute reduction in fuzzy rough sets. *IEEE Transactions on Fuzzy Systems*, 2017, 26(4):2174—2187.
- [29] Wang C Z, Qi Y L, Shao M W, et al. A fitting model for feature selection with fuzzy rough sets. *IEEE Transactions on Fuzzy Systems*, 2017, 25(4):741—753.
- [30] Qian Y H, Wang Q, Cheng H H, et al. Fuzzy -rough feature selection accelerator. *Fuzzy Sets and Systems*, 2015, 258:61—78.
- [31] 胡宝清. 模糊理论基础. 第2版. 武汉:武汉大学出版社, 2010, 648.

- [32] Wang C Z, Wu C X, Chen D G. A systematic study on attribute reduction with rough sets based on general binary relations. *Information Sciences*, 2008, 178(9): 2237—2261.
- [33] Hu Q H, Yu D, Xie Z X. Neighborhood classifiers. *Expert Systems with Applications*, 2008, 34(2): 866—876.
- [34] Chen J K, Lin Y J, Lin G P, et al. Attribute reduction of covering decision systems by hypergraph model. *Knowledge - Based Systems*, 2016, 118: 93—104.
- [35] Wang C Z, Hu Q H, Wang X Z, et al. Feature selection based on neighborhood discrimination index. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 29(7): 2986—2999.

(责任编辑 杨可盛)