



· 不确定性与知识获取 ·

面向模糊邻域动态特征交互的特征选择方法

徐久成^{1,2}, 牛武林^{1,2}, 段江豪^{1,2}, 张 杉^{1,2}, 白 晴^{1,2}

(1. 河南师范大学 计算机与信息工程学院, 河南 新乡 453007;

2. 智慧商务与物联网技术河南省工程实验室, 河南 新乡 453007)

摘要 针对大部分基于模糊邻域信息系统的特征选择方法,在选入特征时只是单调的按照特征顺序进行特征交互的问题,以及大多数特征度量函数仅从代数观或信息观的视角构造度量函数的缺陷,提出了一种基于模糊邻域动态特征交互的特征选择方法。首先,引入模糊邻域互信息计算特征相关度,并根据特征相关度将特征顺序重组。其次,分析特征动态交互的过程,通过模糊邻域互信息以及模糊邻域条件互信息,按照特征排序依次计算特征之间的冗余及动态交互程度。最后,为改善大多数特征度量函数构造视角单一的缺陷,提出一种多视角下的依赖模糊邻域混合互信息度量函数。在8个公共数据集上与现有的7种属性约简算法进行对比实验,实验结果表明,该算法消除了冗余特征,同时提高了数据分类的准确度。

关键词 特征选择;模糊邻域粗糙集;动态特征交互;互信息;条件互信息

中图分类号:TP181 DOI:10.16152/j.cnki.xdxbzr.2025-02-009

Feature selection method for dynamic feature interaction in fuzzy neighborhood system

XU Jiucheng^{1,2}, NIU Wulin^{1,2}, DUAN Jianghao^{1,2}, ZHANG Shan^{1,2}, BAI Qing^{1,2}

(1. College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China;

2. Engineering Lab of Intelligence Business & Internet of Things, Henan Province, Xinxiang 453007, China)

Abstract In order to solve the problem that most feature selection methods based on fuzzy neighborhood information system only monotonically interact with features according to the order of features when selecting features, and that most feature measurement functions only construct metric functions from the perspective of algebraic view or information view, a feature selection method based on dynamic feature interaction in fuzzy neighborhood is proposed. Firstly, fuzzy neighborhood mutual information is introduced to calculate the feature correlation degree and reorganize the feature order according to the feature correlation degree. Secondly, the process of dynamic interaction between features is analyzed, and the degree of redundancy and dynamic interaction between features is calculated according to the order of features through fuzzy neighborhood mutual information and fuzzy neighborhood conditional mutual information. Finally, in order to improve the defect of single perspective in the construction of most feature metric functions, a hybrid mutual information metric function relying on fuzzy neighborhoods from multiple perspectives is proposed. Experimental results show that the pro-

收稿日期:2024-09-28

基金项目:国家自然科学基金(61976082,62076089,62002103)

第一作者:徐久成,男,博士,教授,从事粒计算、数据挖掘和生物信息学研究,xjc@htu.edu.cn。

posed algorithm eliminates redundant features and improves the accuracy of data classification while comparing with seven existing feature reduction algorithms on eight public datasets.

Keywords feature selection; fuzzy neighborhood rough set; dynamic feature interaction; mutual information; conditional mutual information

在大数据时代,数据不断产生并随着通讯技术的发展快速传播,涵盖了决策分析、认知建模、医疗诊断、军事建设和过程控制等各个领域。然而,在实际应用中,真正有用的数据仅占到一小部分,大量冗余信息对数据的分类造成了显著的影响。因此,如何从所有特征属性中挖掘出最优的特征约简子集被认为是各种学习任务中值得研究的课题^[1-2]。

粗糙集自提出以来,便在数据挖掘中备受关注。作为一种经典的知识发现方法,粗糙集理论可直接对数据所包含的信息进行推理^[3]。然而,传统粗糙集难以直接处理数值型数据,经离散化处理的数据又可能会丢失一部分信息。模糊邻域粗糙集能够有效解决这一问题,并在特征选择方面凭借其理论优势显示出较强的特征度量能力^[4-5]。近年来,学者们致力于提高模糊及邻域粗糙集模型在特征选择中的分类性能与泛化能力^[6-7]。Shi等利用广义模糊邻域信息系统,提出了3种多粒度可变精度模糊邻域粗糙集模型^[8]。Hu等通过研究模糊信息粒的关系和结构,提出了基于包含度的模糊粗糙集模型,定义了适用于混合型数据的特征重要度度量^[9]。Wang从数学的角度提出了一种基于三角范数和模糊隐含的模糊粗糙集模型^[10]。杨洁等从多粒度视角,构建出一种代价敏感的模糊邻域粗糙集多粒度近似模型^[11]。

然而,大多数特征选择方法忽略了特征间因相互依赖而产生的相互作用,从而导致一些重要信息的丢失^[12-13]。鉴于此,特征选择过程中特征的交互性逐渐引起了研究者的关注^[14]。Hagar等对组合测试的研究表明,大约90%的软件故障是由参数之间的相互作用触发的^[15]。Wang等考虑了特征之间的相互作用,有效地选择了具有相互作用的特征^[16]。Zhou等将特征交互与流结合,提出了一种基于特征交互的流特征选择方法^[17]。在生物医学研究中,生理和病理过程的变化往往受到分子间相互作用的影响,因此,在许多疾病的诊断、治疗和预防中,辨识交互特征具有关键性的意义^[18]。

现有的大多数特征交互算法在进行特征交互的过程中仅仅是根据特征的先后顺序进行交互,影响最终约简结果的准确性,且大多数基于模糊邻域信息系统下的特征度量函数构造视角单一,得出的度量函数大多仅从信息观或代数观的一个方面出发,不能完全度量特征的不确定性。针对以上问题,本文提出了一种基于模糊邻域动态特征交互的特征选择算法(a feature selection algorithm based on fuzzy neighborhood dynamic feature interaction, FNDFI)。该算法首先根据特征相关度对特征顺序进行重组。其次,通过整合代数观和信息观,在特征相关性、冗余性和动态交互性的基础上提出多视角依赖模糊邻域混合互信息度量,为不确定性度量提供了一个新的思路和方法。最后,在8个公共数据集上与现有的7种属性约简算法进行对比实验,实验分析表明所提算法的可行性和有效性。本文的主要贡献包括3个方面:

1) 在模糊邻域信息系统中考虑动态特征交互问题,可更好地处理现实生活中大量存在的模糊数据;

2) 优化了特征交互的过程,在选入条件特征前,先根据条件特征对决策类的相关性从大到小依次排序,之后再依次进入冗余和动态特征交互分析,避免了传统特征交互条件特征单调性选入的问题;

3) 在模糊邻域信息系统中引入信息观和代数观并综合特征相关性、冗余性及动态交互性度量,提出基于模糊邻域信息系统下的多视角依赖模糊邻域混合互信息度量。

1 相关工作

1.1 模糊邻域粗糙集

给定一个模糊邻域信息系统 $FNDS = \langle U, C, D, V, \delta \rangle$, $U = \{x_1, x_2, \dots, x_n\}$ 是一个非空有限的样本空间。 $C = \{c_1, c_2, \dots, c_m\}$ 是样本集的条件特征集, D 是样本的决策特征集。 V 为全体特征的值域, δ 为模糊邻域半径,用于设置阈值筛选样本间

的模糊相似关系, $\delta \in (0, 1)$ 。 $U/D = \{D_1, D_2, \dots, D_l\}$ 表示 D 将 U 划分为 l 个等价类。

定义 1^[19] 给定一个 $FNDS$, $c_1 \in C$, 由 c_1 导出模糊二元关系 R_{c_1} 。当 R_{c_1} 同时满足自反性和对称性时, 称 R_{c_1} 为模糊相似关系。 $\forall x, y \in U$, 由高斯核函数计算的模糊相似关系 R_{c_1} 表示为

$$R_{c_1}(x, y) = \exp\left(-\frac{\|x - y\|^2}{\lambda}\right) \quad (1)$$

i) 自反性: $R_{c_1}(x, x) = 1, \forall x \subseteq U$ 。

ii) 对称性: $R_{c_1}(x, y) = R_{c_1}(y, x), \forall x, y \in U$ 。

定义 2^[19] 给定一个 $FNDS$, 设模糊邻域半径参数为 $\delta, \delta \in (0, 1)$ 。 $\forall x, y \in U$, 样本 x 和样本 y 关于特征 c_1 的模糊邻域相似矩阵为 $[x]_{c_1}^\delta(y)$, 则 x 关于 c_1 的参数化模糊邻域粒表示为

$$\begin{aligned} \mathcal{E}_{c_1}^\delta(x) &= [x]_{c_1}^\delta(y) = \\ &\begin{cases} R_{c_1}(x, y), R_{c_1}(x, y) \geq \delta; \\ 0, R_{c_1}(x, y) < \delta. \end{cases} \end{aligned} \quad (2)$$

定义 3^[19] 给定一个 $FNDS$, $c_1 \in C, U/D = \{D_1, D_2, \dots, D_l\}, i = 1, 2, \dots, l$, 那么 D_i 关于 c_1 的模糊邻域上下近似分别表示为

$$\overline{R}_{c_1}^\delta(D_i) = \{x_i \mid \mathcal{E}_{c_1}^\delta(x_i) \cap D_i \neq \emptyset, x \in U\} \quad (3)$$

$$\underline{R}_{c_1}^\delta(D_i) = \{x_i \mid \mathcal{E}_{c_1}^\delta(x_i) \subseteq D_i, x \in U\} \quad (4)$$

定义 4^[20] 给定一个 $FNDS$, $c_1 \in C, U/D = \{D_1, D_2, \dots, D_l\}, i = 1, 2, \dots, l$, 则 D 关于 c_1 的正域和依赖度分别表示为

$$POS_{c_1}^\delta(D) = \sum_{i=1}^l \underline{R}_{c_1}^\delta(D_i) \quad (5)$$

$$\gamma_{c_1}(D) = \frac{|\overline{POS}_{c_1}^\delta(D)|}{|U|} \quad (6)$$

1.2 模糊邻域互信息

互信息作为信息熵的拓展是一种有效的度量方法, 可用来衡量变量之间的相互依赖程度, 将其用于特征选择中可以从信息观的角度有效度量特征对于决策类的不确定性^[20]。本节基于模糊邻域粒给出模糊邻域互信息的定义和定理。

定义 5^[20] 给定一个 $FNDS$, $c_1 \in C, \mathcal{E}_{c_1}(x_i)$ 为样本 x_i 由 c_1 诱导的模糊邻域粒, 关于 c_1 的模糊邻域熵表示为

$$FNE_R(c_1) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|\mathcal{E}_{c_1}(x_i)|}{|U|} \quad (7)$$

定义 6^[20] 给定一个 $FNDS$, $c_1, c_2 \in C$, $\mathcal{E}_{c_1}(x_i)$ 和 $\mathcal{E}_{c_2}(x_i)$ 分别为样本 x_i 由 c_1 和 c_2 诱导的

模糊邻域粒, c_1 和 c_2 之间的模糊邻域联合熵、模糊邻域条件熵和模糊邻域互信息分别表示为

$$FNE_R(c_1, c_2) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|\mathcal{E}_{c_1}(x_i) \cap \mathcal{E}_{c_2}(x_i)|}{|U|} \quad (8)$$

$$FNE_R(c_1 | c_2) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|\mathcal{E}_{c_1}(x_i) \cap \mathcal{E}_{c_2}(x_i)|}{|\mathcal{E}_{c_2}(x_i)|} \quad (9)$$

$$FNE_R(c_1; c_2) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|\mathcal{E}_{c_1}(x_i)| \cdot |\mathcal{E}_{c_2}(x_i)|}{|U| \cdot |\mathcal{E}_{c_1}(x_i) \cap \mathcal{E}_{c_2}(x_i)|} \quad (10)$$

定理 1^[20] 在模糊邻域信息系统 $FNDS$ 中, $c_1, c_2 \in C$, 存在以下关系。

- 1) $FNE_R(c_1; c_2) = FNE_R(c_2; c_1)$;
- 2) $FNE_R(c_1; c_2) = FNE_R(c_1) - FNE_R(c_1 | c_2)$;
- 3) $FNE_R(c_1; c_2) = FNE_R(c_1) + FNE_R(c_2) - FNE_R(c_1, c_2)$ 。

2 基于动态特征交互下的特征选择

2.1 相关性度量

定义 7^[20] 给定一个 $FNDS$, 设模糊邻域半径为 $\delta, Red \subseteq C$ 为已选特征集, $c_j \in C - Red$ 为候选特征, c_j 与决策类 D 间的相关性表示为

$$FNE(c_j; D) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|\mathcal{E}_{c_j}(x_i)| \cdot |\mathcal{E}_D(x_i)|}{|U| \cdot |\mathcal{E}_{c_j}(x_i) \cap \mathcal{E}_D(x_i)|} \quad (11)$$

$FNE(c_j; D)$ 表示 c_j 对 D 相关性的大小, 因此, 应找到与 D 相关性最大的 c_j , 该策略称为最大相关性准则。

$$\max FNE(c_j; D), c_j \in C - Red。$$

2.2 冗余性度量

虽然上述方法已选择与类别相关的特征, 但仍要考虑候选特征与已选特征之间的冗余, 本节利用模糊邻域互信息衡量候选特征和已选特征之间的冗余度。

定义 8^[20] 给定一个 $FNDS$, 设模糊邻域半径为 $\delta, Red \subseteq C$ 为已选特征集, $c_s \in Red, c_j \in C - Red$ 为候选特征, c_j 与 c_s 之间的冗余性表示为

$$FNE(c_j; c_s) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|\mathcal{E}_{c_j}(x_i)| \cdot |\mathcal{E}_{c_s}(x_i)|}{|U| \cdot |\mathcal{E}_{c_j}(x_i) \cap \mathcal{E}_{c_s}(x_i)|} \quad (12)$$

在 c_j 与 c_s 的特征交互中,为除去 $C - Red$ 中的冗余特征,应遵循最小冗余准则

$$\min FNE(c_j;c_s), c_j \in C - Red, c_s \in Red.$$

2.3 动态特征交互

特征选择方法通常致力于在保留分类信息的前提下,选择与类别相关度最高或冗余度最低的特征子集。然而,特征之间普遍存在的相互作用却被忽略,这使得大部分特征选择方法很难获得最优的结果。鉴于此,本小节重点探讨特征间的动态交互作用,并寻找适当的度量方式。

特征交互在信息论的角度下可表示为在已知某个特征的情况下增加一个新特征,分析原特征对分类所贡献的信息量变化。模糊邻域条件互信息提供了一个有效的方法来衡量这些情况。事实上,在使用模糊邻域条件互信息来衡量特征选择的过程中,候选特征与决策类之间的相关性和已选特征与决策类之间的相关性都是动态变化的,即随着候选特征的不断选入,已选特征和候选特征对于分类所贡献的信息量都会不断变化,这是由于特征之间的相互作用引起的^[21]。

因此,本文采用模糊邻域条件互信息来度量候选特征与决策类之间的动态相关性,并分析了这种相关性随时间变化可能出现的几种情况。

定义9 给定一个 $FNDS$, Red 为已选特征集合, $c_s \in Red, c_j \in C - Red$ 为候选特征。在已知已选特征 c_s 的条件下,当前候选特征 c_j 与决策类 D 的模糊邻域条件互信息表示为

$$FNCMI(c_j;D|c_s) = -\frac{1}{|U|} \cdot \sum_{i=1}^{|U|} \log_2 \frac{|\mathcal{E}_{c_j}(x_i) \cap \mathcal{E}_D(x_i)| \cdot |\mathcal{E}_{c_s}(x_i) \cap \mathcal{E}_D(x_i)|}{|\mathcal{E}_{c_s}(x_i)| \cdot |\mathcal{E}_{c_j}(x_i) \cap \mathcal{E}_{c_s}(x_i) \cap \mathcal{E}_D(x_i)|} \quad (13)$$

定理2 $FNCMI(c_j;D|c_s) = FNE(c_j,D) + FNE(D,c_s) - FNE(c_j,D,c_s) - FNE(c_s)$

证明 $FNE(c_j,D) + FNE(D,c_s) - FNE(c_j,D,c_s) - FNE(c_s) =$

$$-\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|\mathcal{E}_{c_j \cup D}(x_i)|}{|U|} - \frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|\mathcal{E}_{D \cup c_s}(x_i)|}{|U|} + \frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|\mathcal{E}_{c_j \cup D \cup c_s}(x_i)|}{|U|} + \frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|\mathcal{E}_{c_s}(x_i)|}{|U|} =$$

$$-\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \left(\frac{|\mathcal{E}_{c_j \cup D}(x_i)|}{|U|} \cdot \frac{|\mathcal{E}_{D \cup c_s}(x_i)|}{|U|} \cdot \frac{|U|}{|\mathcal{E}_{c_j \cup D \cup c_s}(x_i)|} \cdot \frac{|U|}{|\mathcal{E}_{c_s}(x_i)|} \right) = -\frac{1}{|U|} \cdot \sum_{i=1}^{|U|} \log_2 \frac{|\mathcal{E}_{c_j}(x_i) \cap \mathcal{E}_D(x_i)| \cdot |\mathcal{E}_{c_s}(x_i) \cap \mathcal{E}_D(x_i)|}{|\mathcal{E}_{c_s}(x_i)| \cdot |\mathcal{E}_{c_j}(x_i) \cap \mathcal{E}_{c_s}(x_i) \cap \mathcal{E}_D(x_i)|}$$

下面举例说明在模糊邻域信息系统中特征动态交互的过程。

例1 表1 给定一数据经归一化后的模糊邻域信息系统 $FNDS = \langle U, C, D, V, \delta \rangle, U = \{x_1, x_2, x_3, x_4, x_5, x_6\}, C = \{c_1, c_2, c_3, c_4, c_5\}$ 。设 $\delta = 0.8$ 。根据定义7 可得出条件特征与决策类之间的相关性。

表1 一个模糊邻域信息系统

Tab.1 A fuzzy neighborhood information system

U	c_1	c_2	c_3	c_4	c_5	D
x_1	0.842 1	0.191 7	0.572 1	0.257 7	0.619 5	0
x_2	0.571 0	0.205 5	0.417 1	0.030 9	0.326 0	0
x_3	0.313 1	0.108 6	0.310 1	0.432 9	0.239 1	0
x_4	0.423 6	0.122 5	0.352 9	0.319 5	0.326 0	1
x_5	0.623 6	0.762 8	0.802 1	0.742 2	0.456 5	1
x_6	0.578 9	0.505 9	0.491 9	0.407 2	0.304 3	1

在表2中,第1列给出了每个特征相对于决策类 D 的相关性大小。由第1列的数据可知 c_2 对于决策类 D 的相关性最大,所以将 c_2 添加到所选特征集合,并将其他特征按照相关性的从大到小排列。第2列表示 c_2 在已选的基础上同其他候选特征与决策类 D 的模糊邻域条件互信息,其表示当给定 c_2 时,不同候选特征与 D 之间的相关性。可以从第1列和第2列看出,当给定已选特征时,一些候选特征对 D 的相关性会增大或减小。

表2 条件特征与决策类之间的相关性

Tab.2 Correlation between conditional features and decision classes

$FNE(c_i;D)$	$FNCMI(c_i;D c_2)$	$FNCMI(c_2;D c_i)$
0.333 3	0.110 0	0.040 9
0.459 1	-	-
0.203 0	0.207 5	0.456 2
0.138 4	0.207 5	0.069 2
0.244 5	-0.251 6	0.246 7

同样,表2中第3列表示在给定不同候选特征的情况下,所选特征对 D 之间的相关性大小。第3列数据显示,随着候选特征的不断选入,候选特

征和已选特征对于 D 的相关性也时刻在发生变化,因此有必要考虑特征之间的动态交互。

本文将这种动态变化大致分为 3 种情况:

c_j 与 c_s 互补, c_j 与 c_s 相互独立及 c_j 与 c_s 相互抑制。

2.3.1 c_j 与 c_s 互补

c_j 与 c_s 对于 D 的模糊邻域条件互信息大于 c_j 与 c_s 对于 D 的模糊邻域互信息。

i) $FNCMI(c_j; D | c_s) > FNE(c_j; D)$;

ii) $FNCMI(c_s; D | c_j) > FNE(c_s; D)$ 。

i) 表示 c_s 的出现增加了 c_j 对于 D 的相关性。
ii) 也意味着 c_j 的出现增加了 c_s 对于 D 的相关性。
这两种情况说明 c_j 与 c_s 是互补的,也就是说,它们的组合可以为 D 提供更多的分类信息。

2.3.2 c_j 与 c_s 相互独立

c_j 与 c_s 对于 D 的模糊邻域条件互信息等于 c_j 与 c_s 对于 D 的模糊邻域互信息。

i) $FNCMI(c_j; D | c_s) = FNE(c_j; D)$;

ii) $FNCMI(c_s; D | c_j) = FNE(c_s; D)$ 。

i) 的出现表明 c_s 的出现不会影响 c_j 对于 D 的相关性。

ii) 意味着 c_j 的出现不会影响 c_s 对于 D 的相关性。这两种情况意味着 c_j 和 c_s 是彼此独立的,它们的组合不影响彼此相对于 D 的相关性。

2.3.3 c_j 与 c_s 相互抑制

c_j 与 c_s 对于 D 的模糊邻域条件互信息小于 c_j 与 c_s 对于 D 的模糊邻域互信息。

i) $FNCMI(c_j; D | c_s) < FNE(c_j; D)$;

ii) $FNCMI(c_s; D | c_j) < FNE(c_s; D)$ 。

i) 表示 c_s 的出现降低了 c_j 对于 D 的相关性。
ii) 意味着 c_j 的出现同样降低了 c_s 对于 D 的相关性。这两种情况说明 c_j 与 c_s 的组合抑制了它们相对于 D 的相关性。也就是说,它们的组合降低了各自为 D 提供的信息量。

2.3.4 动态交互性度量

基于 c_j 与 c_s 互补、相互独立以及相互抑制 3 种情况,将 $FNCMI(c_j; D | c_s)$ 与 $FNCMI(c_s; D | c_j)$ 相加,构建一种基于模糊邻域条件互信息的动态特征交互性度量(a dynamic feature interaction measure based on fuzzy neighborhood conditional mutual information, DFIM)。

定义 10 设 c_j 和 c_s 分别表示为候选特征和已选特征, Red 为已选特征子集, $c_s \in Red$, D 为决策特征, DFIM 表示为

$$DFIM(c_j) = \sum_{c_s \in Red} |FNCMI(c_j; D | c_s) + FNCMI(c_s; D | c_j)| \quad (14)$$

对于 DFIM 来说,应选择与 c_s 具有最大动态交互的 c_j ,也就是找到与 c_s 最大互补的 c_j 。因此,为找到与 D 交互性最大的候选特征,提出最大交互性准则为

$$\max DFIM(c_j), c_j \in C - Red。$$

2.4 多视角下依赖模糊邻域混合互信息

定义 11 综合上述特征的相关性、冗余性和动态交互性给出多视角下的依赖模糊邻域混合互信息度量函数(fuzzy neighborhoods hybrid mutual information metric function relies on multiple perspectives, FNHMIMP)

$$FNHMIMP(c_j) = \gamma_{c_j}(D) (FNE(c_j; D) - \frac{1}{|Red|} \sum_{c_s \in Red} FNE(c_j; c_s) + \frac{1}{|Red|} DFIM(c_j)) \quad (15)$$

FNHMIMP 度量函数使用模糊邻域互信息来评估候选特征与决策类之间的相关性以及候选特征与已选特征之间的冗余。同时,采用模糊邻域条件互信息来度量候选特征与已选特征之间的动态交互作用。最后,将它们计算后的结果与特征依赖度结合,从代数观和信息观下衡量特征对分类的贡献。其目的是使最终选择的特征子集最具有代表性,并实现相关性、冗余性和动态交互性的权衡。

3 算法研究

为消除评价函数构造视角单一对特征子集精度的影响,并考虑特征之间的非单调动态相互作用,本节提出了一种基于模糊邻域动态特征交互的特征选择算法,算法描述如下。

输入:模糊邻域信息系统 $FNDS$, 模糊邻域半径 δ , 所选特征个数 k 。

输出:特征约简子集 Red_{best}

- 1 计算 C 下每个特征的相对模糊相似关系 $R_C(x, y)$
- 2 初始化: $Red = \emptyset, Red_{best} = \emptyset, C_1 = \emptyset$
- 3 for each $c_j \in C$ do
- 4 计算 $FNE(c_j; D)$
- 5 end for
- 6 根据每个特征对决策类相关性的的大小依次

```

    将各个特征加入  $C_1$ 
7 将与决策类具有最大相关性的  $c_j$  加入  $Red$ 
8  $C_1 = C_1 - \{c_j\}$ 
9 for each  $c_j \in C_1$  do
10 for each  $c_s \in Red$  do
11 计算  $FNE(c_j; c_s)$ 
12 end for
13 for each  $c_s \in Red$  do
14 计算  $DFIM(c_j)$ 
15 end for
16 计算  $FNHMIMP(c_j)$ 
17 选取  $FNHMIMP(c_j)$  最大的  $c_j$  进入  $Red$ 
18 更新  $C_1$ 
19 end for
20 取  $Red$  中前  $k$  个特征加入  $Red_{best}$ 
```

进一步分析 FNDFI 算法的时间复杂度, 设 m 和 n 分别表示样本数和特征维数。计算所有特征集合下模糊相似关系的时间复杂度为 $O(nm^2)$ 。对于每个特征, 在第一个循环中, 计算特征和决策类之间相关性的时间复杂度为 $O(n)$ 。在下一个双 for 循环中, 时间复杂度为 $O(n^3)$ 。通常样本的数量大于特征的数量, 因此算法 FNDFI 的时间复杂度为 $O(nm^2)$ 。

4 实验结果与分析

为验证所提算法的有效性, 本节选取 8 个公共数据集进行实验, 包括 4 个低维数据集 (UCI) 和 4 个高维数据集 (cancer program dataset)。所有实验数据集的详细信息见表 3。实验在 MATLAB2016b 环境下进行, 使用 KNN (K-Nearest Neighbor) 和 CART (classification and regression tree) 两种分类器, 分类器均采用默认参数值设置。针对这 8 个数据集在 FNDFI 算法上的实验结果, 分析了约简特征在 KNN 和 CART 下的特征约简子集个数与分类精度。实验结果采用了十折交叉验证来进行误差估计, 将数据集分成 10 份, 其中 9 份轮流作为训练数据, 剩余的 1 份作为测试数据。

4.1 Fisher Score 用于高维数据预处理

面对高维数据时, 许多特征选择算法在时间上的开销是巨大的, 这影响了算法在实际问题中的应用。假设一个模糊邻域信息系统 $FNDS = \langle U, C, D, V, \delta \rangle$ 中有 m 个样本和 n 个特征, 并且最

大相关子集中有 t_1 个特征, 约简子集中有 t_2 个已选特征。在高维数据中 $n \gg m$ 和 $n \gg t_1 \gg t_2$ 。因此, 对于高维数据, FNDFI 算法的时间复杂度为 $O(nm^2)$ 。引入 Fisher Score 方法后, FNDS 中的特征减少到 l 个。Fisher Score 方法的时间复杂度为 $O(n)$, 引入 Fisher Score 的 FNDFI 算法的时间复杂度为 $O(lm^2 + n)$ 。由于 $n \gg l$, 故在引入 Fisher Score 方法后, 大大地降低了 FNDFI 对于高维数据的时间开销。

采用 Fisher Score 方法选择重要度最高的前 l 个特征来构建候选特征子集, 然后利用 FNDFI 算法对候选特征子集做进一步处理, 计算 10 个不同维度候选特征子集通过十折交叉验证所获取的在 KNN 和 CART 下的精度来为表 3 中 4 个高维数据集选择合适的候选特征子集。图 1 给出 4 个高维数据集候选特征的维度与平均分类精度之间的变化趋势。根据每个数据集的特点找到最高精度下的特征维数。

表3 数据集信息
Tab. 3 Data set information

数据集	样本数	特征数	类别数
Wine	178	13	3
Wdbc	569	31	2
Sonar	208	60	2
Heart	270	13	2
Colon	62	2 000	2
Leukemia	72	7 129	2
Breast	84	9 216	5
Prostate	136	12 600	2

对于高维数据集 Colon, 分类精度在 100 维时达到最高; 对于高维数据集 Prostate, 分类精度在 50 维时达到最高; 类似地, 数据集 Leukemia 和 Breast 的最佳精度维数分别为 150 维和 200 维。

4.2 模糊邻域参数的选择

特征约简子集的分类精度和大小是作为评价算法优劣的两个重要指标, 本节将探究不同模糊邻域参数值对实验结果的影响, 并在 KNN 和 CART 两个分类器下找到最佳参数。

FNDFI 算法中有两个参数 δ 和 k 。参数 δ 为模糊邻域半径, 用于调整模糊邻域的大小, k 用于控制约简子集的特征数量。为有效涵盖不同数据集的最佳模糊邻域, 同时避免因参数设置过小导致模糊邻域粒的约束过于严格, 在实验中将 δ 的值设置为 0.45 到 0.95, 步长为 0.05。为充分探

究约简子集的特征数量对实验结果的影响,在低维数据集 Wine、Heart、Wdbc 和 Sonar 中 k 分别设置为 13、13、30 和 30。在其余 4 个高维数据集中 k 设置为 50。

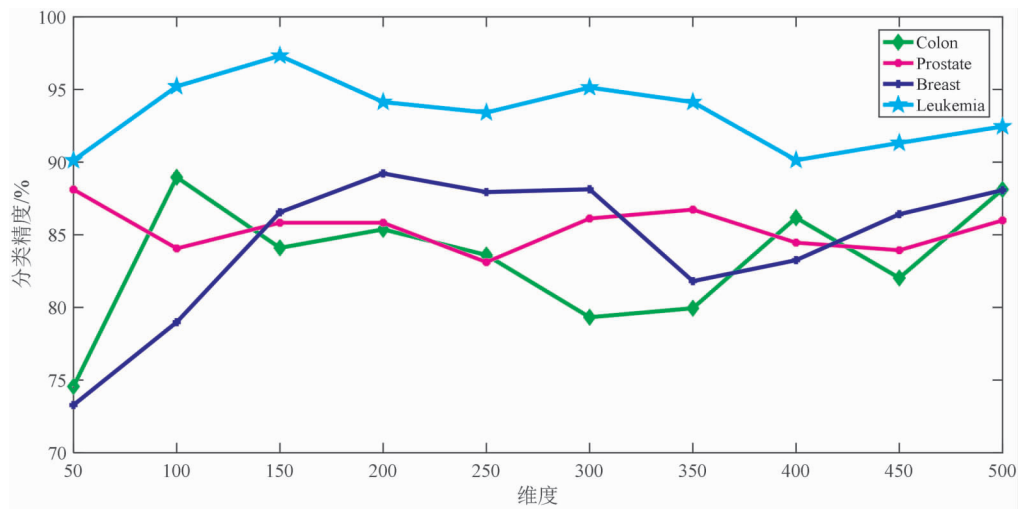
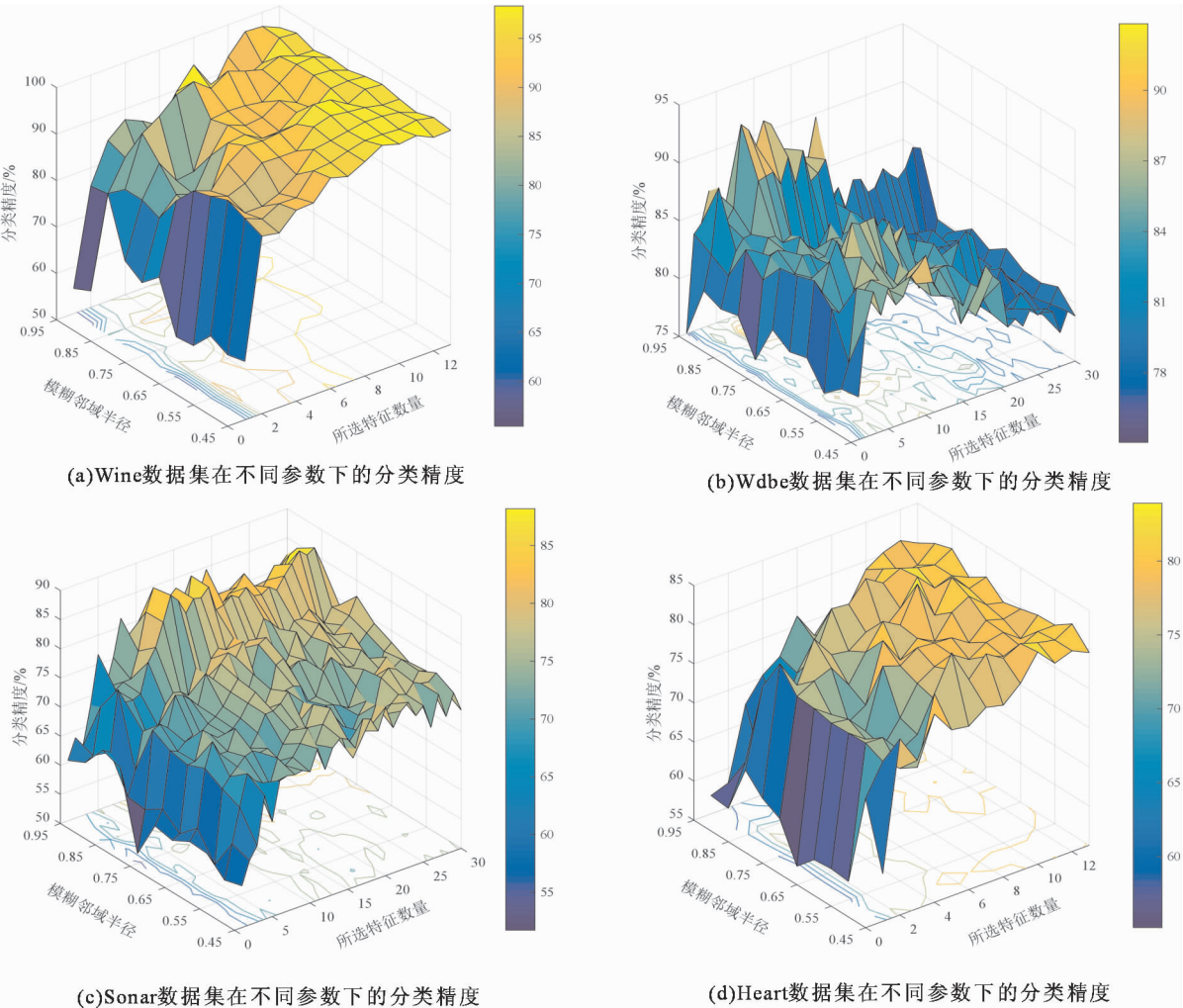


图 1 不同维度数据集在 KNN 和 CART 下的平均分类精度

Fig. 1 Average classification accuracies of data sets with different dimensions under KNN and CART

图 2 呈现了 FNDFI 算法在 KNN 分类器中 4 个低维数据集和 4 个高维数据集的分类精度随参数的变化趋势。CART 分类器实验效果与该分类器实验结果大致相同。图 2 中 x 轴表示所选特征数量 k , y 轴表示模糊邻域半径 δ , z 轴表示分类精度。



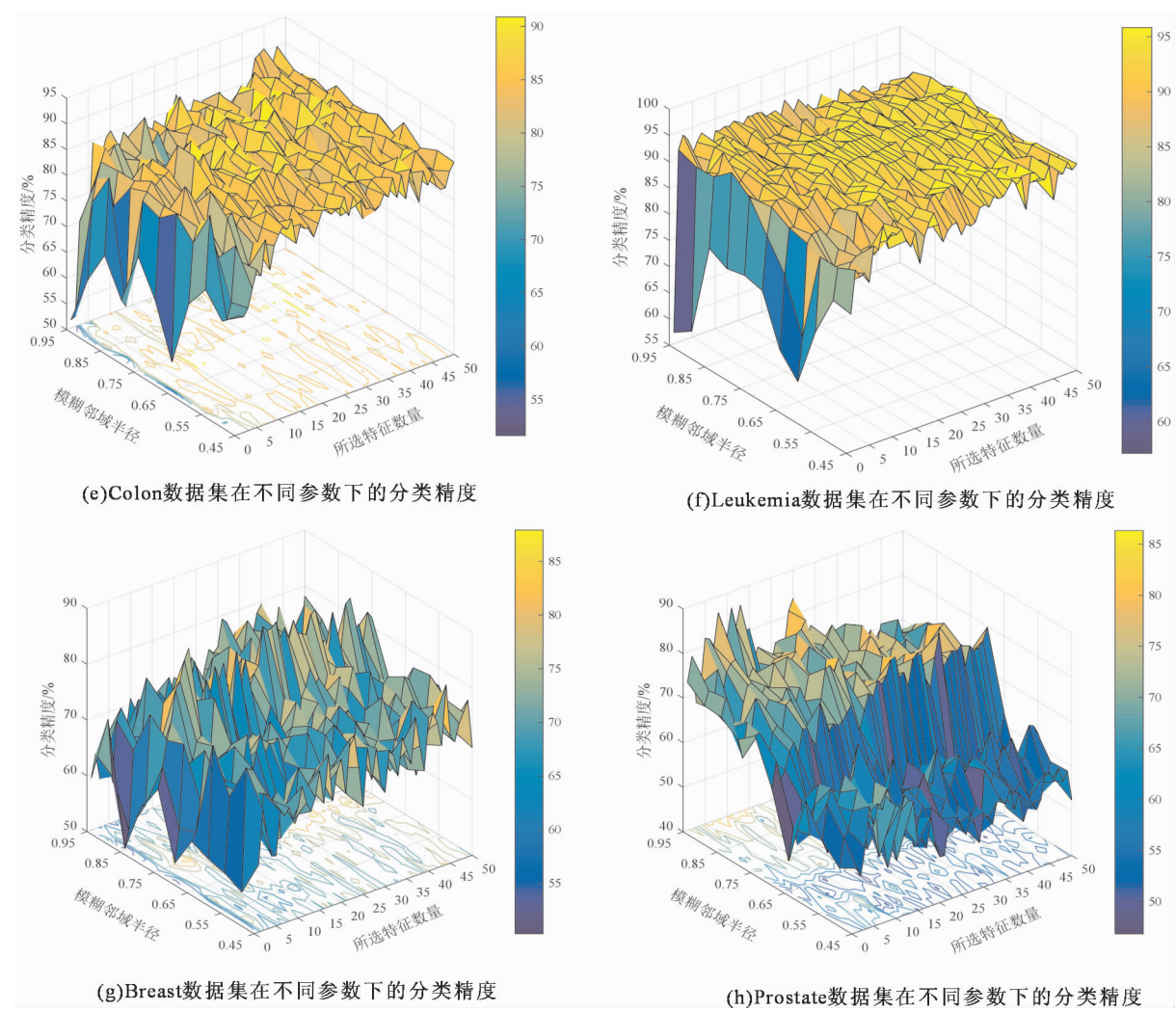


图2 8个数据集在不同参数下 KNN 的分类精度

Fig.2 Classification accuracy of KNN for 8 datasets with different parameters

FNDFI 算法在大多数数据集上的平均分类精度在初始阶段随着 k 的增加而逐渐提高,然后保持平缓或略有下降。在 Wine、Heart 和 Breast 数据集中其平均分类精度变化趋势一直符合先增加然后保持平稳。但对于 Wdbc 数据集,当选到第6个特征时,分类性能已经达到比较高的状态,表明此时的特征子集已经为分类提供了足够多的信息,当再选取其他特征时,这些特征与已选特征存在冗余,因此分类性能呈现出下降趋势。在数据集 Prostate 以及 Sonar 中,明显看出 δ 的变化对分类的效果有着很大的影响。例如在数据集 Prostate 中当 $\delta > 0.8$ 时,比在其他 δ 下的平均分类精度高很多。然而对于一些数据集,例如 Colon 和 Leukemia,参数 δ 对于平均分类精度的影响相对较小。

表4 为8个数据集在两个分类器下选择的最优

模糊邻域半径值和最优特征约简子集大小的组合。

表4 不同分类器下8个数据集的最佳模糊邻域半径值和特征子集的大小

Tab.4 The optimal fuzzy neighborhood radius values and feature subset sizes for 8 datasets under different classifiers

数据集	CART	KNN
Wine	(0.50,8)	(0.45,9)
Wdbc	(0.85,7)	(0.9,6)
Sonar	(0.85,11)	(0.85,10)
Heart	(0.55,7)	(0.5,7)
Colon	(0.75,9)	(0.75,10)
Leukemia	(0.50,8)	(0.45,8)
Breast	(0.70,8)	(0.75,8)
Prostate	(0.95,9)	(0.95,8)

4.3 FNDFI 算法在低维数据集上的结果

为了验证算法 FNDFI 在处理低维数据时的效果,在实验中选择 4 个低维数据集(Wine、Wdbc、Sonar 和 Heart)作为实验对象,分别分析 FNDFI 算法在 KNN 分类器和 CART 分类器下的实验结果,与 FNRDCI^[20]、DMRA^[22]、FSNTDJE^[23]、IF-

PR^[24]、PDJE-AR^[25]、FRFS^[26]和 MIFS^[27]7 种算法进行比较。表 5 给出了 8 种算法特征约简子集的大小,表 6 和表 7 分别给出了 8 种算法在 KNN 分类器和 CART 分类器下的分类精度,其中 ODFN 和 ODCA 分别为在分类器下原始数据的特征约简子集大小和分类精度。

表 5 8 种算法的特征约简子集大小

Tab.5 Feature reduction subset size of 8 algorithms

数据集	ODFN	FNRDCI	FSNTDJE	IFPR	PDJE-AR	FRFS	MIFS	DMRA	FNDFI
Wine	13	7.9	8.8	6.2	5.7	11.4	10.9	12.1	9.0
Wdbc	31	9.0	16.9	17.1	15.7	24.4	15.8	27.9	6.8
Sonar	60	9.8	19.7	13.0	18.4	19.5	17.1	35.3	10.7
Heart	13	9.3	10.9	8.4	9.6	10.7	11.4	11.8	8.3
Average	29.2	9.0	11.2	11.2	12.4	16.5	13.8	21.8	8.7

表 6 低维数据上 8 种算法的 KNN 分类精度

Tab.6 KNN classification accuracies of 8 algorithms on low dimensional data

单位:%

数据集	ODCA	FNRDCI	FSNTDJE	IFPR	PDJE-AR	FRFS	MIFS	DMRA	FNDFI
Wine	91.95	97.92	90.57	94.29	94.94	93.30	94.74	91.08	98.30
Wdbc	91.24	96.21	94.66	93.11	95.61	94.32	93.37	94.04	92.83
Sonar	86.53	78.74	69.75	70.55	72.40	73.57	74.07	72.14	88.18
Heart	75.18	82.28	77.41	78.89	80.37	75.45	74.44	76.97	83.92
Average	86.22	88.79	83.10	84.21	85.83	84.16	84.16	83.56	90.80

表 7 低维数据上 8 种算法的 CART 分类精度

Tab.7 CART classification accuracies of 8 algorithms on low dimensional data

单位:%

数据集	ODCA	FNRDCI	FSNTDJE	IFPR	PDJE-AR	FRFS	MIFS	DMRA	FNDFI
Wine	87.62	91.52	80.27	87.27	91.01	85.44	85.92	84.92	93.67
Wdbc	92.10	94.25	91.24	92.96	94.02	89.90	94.25	92.08	91.83
Sonar	67.46	76.04	67.07	68.71	69.04	72.60	72.24	73.62	78.18
Heart	77.73	80.66	76.53	74.81	80.00	73.70	74.07	74.44	83.92
Average	84.23	85.62	78.78	80.93	83.51	80.41	81.62	81.27	86.90

从表 5 可以看出,算法 FNDFI 的平均特征约简子集大小为 8.7,较其余算法比为最优。在数据集 Wdbc 和 Heart 上,算法 FNDFI 的特征约简子集最少,分别为 6.8 和 8.3。虽然算法 PDJE-AR 和 FNRDCI 在数据集 Wine 和 Sonar 上的特征约简子集优于算法 FNDFI,但平均分类精度均小于算法 FNDFI。从表 6 和表 7 可以看出,算法

FNDFI 在 KNN 和 CART 分类器下的平均分类精度均最高,分别为 90.80% 和 86.90%。在 KNN 分类器下高于其他算法 2% ~7.7%,在 CART 分类器下的平均分类精度高于其他算法 1.3% ~8.1%。
更为直观地展示 FNDFI 算法的有效性,绘制了 8 种算法在 4 个低维数据集下的盒形图。从图 3 和图 4 可以看出,算法 FNDFI 在两个分类器上的

平均分类精度(中位数)表现最佳,其上下四分位数也显示出较高的分类性能。算法 FNRDCI 虽然在两种分类器下的分类精度也较高,但就箱体而言 FNRDCI 算法对比于 FNDFI 算法箱体较长,表

示 FNRDCI 算法在实际应用中分类精度不稳定。综上所述,FNDFI 算法在平均特征约简子集最小和平均分类精度最高的情况下还表现出了较强的稳定性。

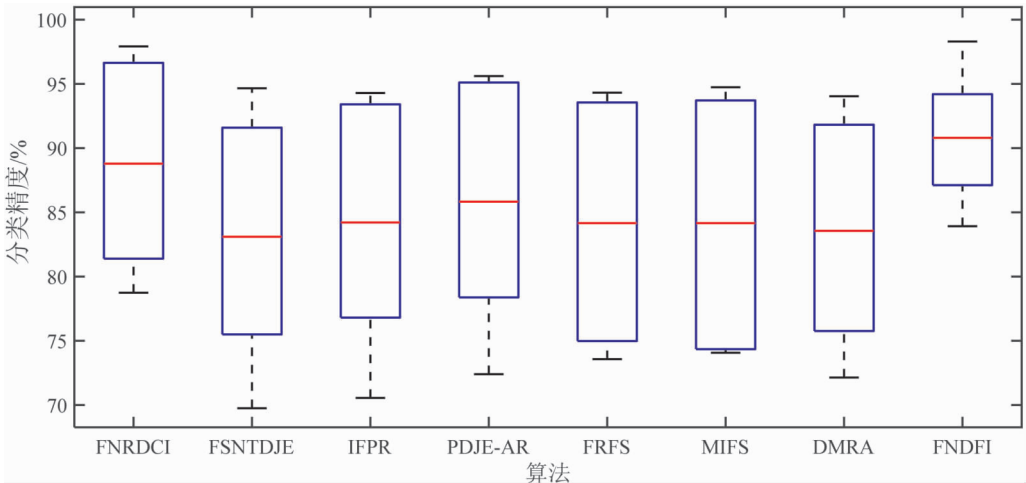


图 3 8 种算法在 KNN 下的盒形图
Fig. 3 Box plot of 8 algorithms under KNN

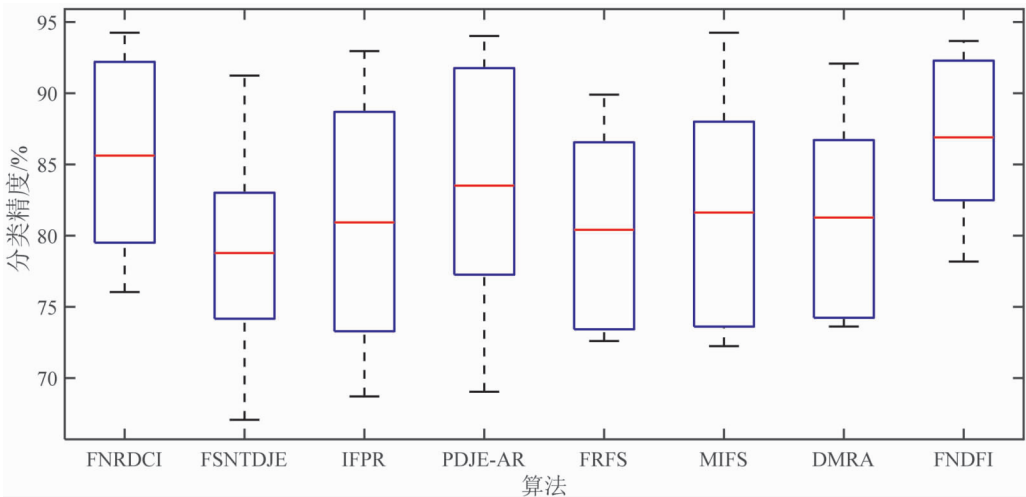


图 4 8 种算法在 CART 下的盒形图
Fig. 4 Box plot of 8 algorithms under CART

4.4 FNDFI 算法在高维数据集上的结果

为了验证算法 FNDFI 在处理高维数据时的有效性,选取了 4 个高维数据集(Colon、Leukemia、Breast 和 Prostate)作为实验对象。分析算法 FNDFI 在 KNN 分类器和 CART 分类器下的实验结果,并与 FNRDCI、FSNTDJE、IFPR、PDJE-AR 和 MIFS 这 5 种算法进行比较。表 8 给出了 6 种算法的特征约简子集的大小,表 9 和表 10 分别给出了 6 种算法在 KNN 分类器和 CART 分类器下的分类精度,其中 ODFN 和 ODCA 分别为在分类器下原始数据的特征约简子集数量和分类精度。

从表 8 中可以看出,算法 FNDFI 的平均特征约简子集大小为 8.5,与其余算法比为最优,6 个算法在平均所选特征数量上相差较小。虽然算法 IFPR 在 Colon 和 Leukemia 两个数据集上平均特征约简子集较小,但从表 9 和表 10 可看出,IFPR 算法在两个分类器下的分类精度明显低于算法 FNDFI。在 KNN 和 CART 两种分类器下算法 FNDFI 的平均分类精度分别为 90.24% 和 88.20%,在 KNN 分类器下高于其他算法 2.3% ~ 16.2%,在 CART 分类器下的平均分类精度高其他算法 0.8% ~ 13.9%。

表 8 6 种算法的特征约简子集大小

Tab. 8 Feature reduction subset size of 6 algorithms

数据集	ODFN	FNRDCI	FSNTDJE	IFPR	PDJE-AR	MIFS	FNDFI
Colon	2 000	9.0	11.5	6.0	6.5	11.0	9.1
Leukemia	7 129	6.6	6.6	6.4	7.6	8.5	7.8
Breast	9 216	11.6	11.6	9.5	10.6	9.7	8.9
Prostate	12 600	8.8	8.5	9.8	9.6	8.5	8.2
Average	7 736	9.0	9.6	7.9	8.6	9.4	8.5

表 9 高维数据上 6 种算法的 KNN 分类精度

Tab. 9 KNN classification accuracies of 6 algorithms on high dimensional data 单位: %

数据集	ODCA	FNRDCI	FSNTDJE	IFPR	PDJE-AR	MIFS	FNDFI
Colon	72.58	86.93	77.42	75.74	85.37	79.58	90.89
Leukemia	73.44	91.30	83.27	80.42	87.52	74.77	95.83
Breast	69.09	86.50	71.43	78.33	75.24	61.96	87.92
Prostate	83.09	87.09	78.71	80.95	79.36	79.69	86.33
Average	74.55	87.96	77.70	78.86	81.87	74.00	90.24

表 10 高维数据上 6 种算法的 CART 分类精度

Tab. 10 CART classification accuracies of 6 algorithms on high dimensional data 单位: %

数据集	ODCA	FNRDCI	FSNTDJE	IFPR	PDJE-AR	MIFS	FNDFI
Colon	59.67	85.38	70.97	75.95	80.65	83.32	90.75
Leukemia	75.56	96.36	72.22	75.56	79.35	90.10	96.66
Breast	70.23	80.74	69.05	73.12	79.52	73.81	77.35
Prostate	69.17	87.17	84.92	76.10	83.12	82.13	88.06
Average	69.66	87.41	74.29	75.18	80.66	82.34	88.20

为更直观地展示 FNDFI 算法的有效性,绘制了 6 种算法在 4 个高维数据集下的盒形图。图 5 和图 6 展示了 6 种算法在 KNN 和 CART 分类器上的分类精度盒形图对比,从中看出算法 FNDFI 在两个分类器的平均分类精度(中位数)表现最佳,其上下四分位数也表现出较高的分类性能,相

比之下,算法 FSNTDJE 和 IFPR 算法整体表现较差,虽然算法 FNRDCI 在 KNN 下的箱体较短但其中位数偏离箱体中间位置且整体低于算法 FNDFI。综上所述,针对 4 个高维数据,算法 FNDFI 在有效减少冗余特征的同时提高了数据的分类精度。

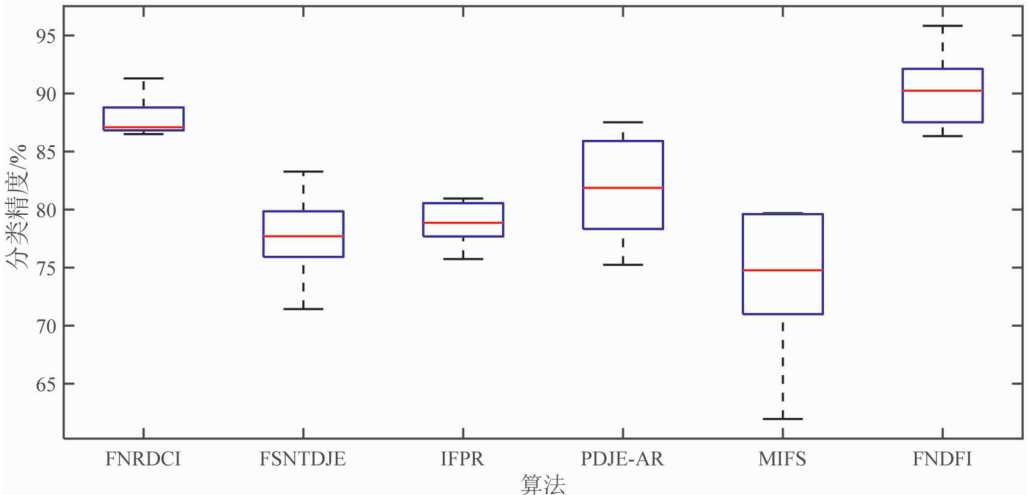


图 5 6 种算法在 KNN 下的盒形图

Fig. 5 Box plot of 6 algorithms under KNN

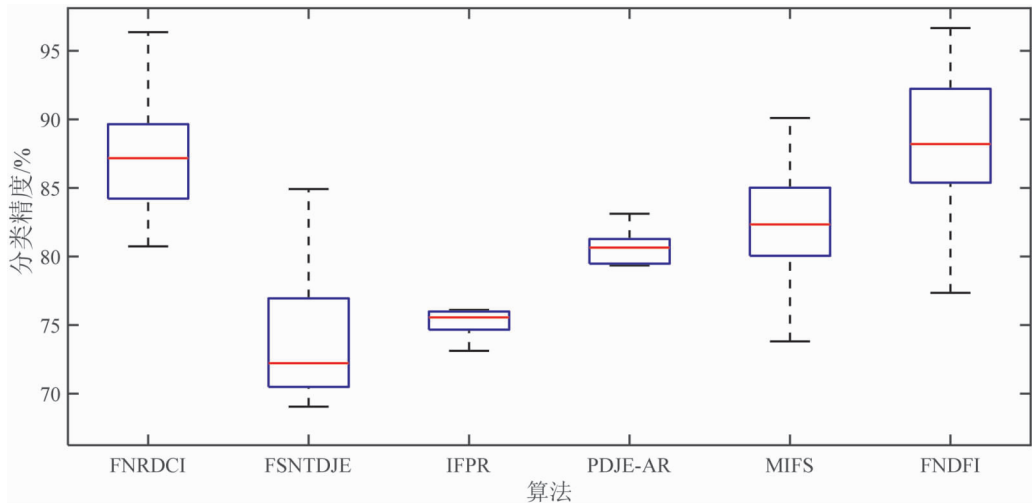


图 6 6 种算法在 CART 下的盒形图

Fig. 6 Box plot of 6 algorithms under CART

5 结语

针对大部分基于模糊邻域信息系统的特征选择方法,在选入特征时只是单调的依次按照特征顺序进行特征交互的问题和构建的度量函数视角单一的缺陷,本文提出一种基于模糊邻域动态特征交互的特征选择算法。首先,基于模糊邻域互信息得到各特征对决策类相关性的 大小,并根据相关性的 大小将各特征从大到小排序;其次,根据特征的排序依次分析各特征之间的冗余以及动态交互程度;然后,将 3 种不同的基于模糊邻域互信息和模糊邻域条件互信息下的度量与依赖度相结合,从代数观和信息观两种视角下提出多视角下的依赖模糊邻域混合互信息度量函数;最后,通过实验来验证所提算法的有效性。尽管算法 FNDFI 在大部分数据集上实现了较高的分类准确率,但其采用穷举法寻找最佳模糊邻域半径的效率不高。因此,未来的研究将聚焦于开发自动确定最优模糊邻域半径的动态特征交互选择算法。

参考文献

[1] AN S, HU Q H, WANG C Z, et al. Data reduction basedon NN-kNN measure for NN classification and regression [J]. International Journal of Machine Learning and Cybernetics, 2022,13(3) : 765-781.

[2] WAN J H, CHEN H M, YUAN Z, et al. A novel hybrid feature selection method considering feature interaction in neighborhood rough set [J]. Knowledge-

Based Systems, 2021, 227(6) :107167.

[3] 冉虹,候婷,贺龙雨,等. 基于模糊邻域系统的模糊粗糙集模型[J]. 计算机科学, 2022,49(S2) :330-334.

RAN H, HOU T, HE L Y, et al. Fuzzy rough set model based on fuzzy neighborhood system[J]. Computer Science, 2022, 49(S2) : 330-334.

[4] 庞婧,姚炳学,李令强. 基于模糊邻域系的 α -粗糙集及其应用[J]. 模糊系统与数学, 2022,36(5) : 158-165.

PANG J, YAO B X, LI L Q. α -rough set based on fuzzy neighborhood system and its application [J]. Fuzzy Systems and Mathematics, 2022, 36(5) : 158-165.

[5] YUAN Z, CHEN H M, XIE P, et al. Attribute reduction methods in fuzzy rough set theory: An overview, comparative experiments, and new directions[J]. Applied Soft Computing, 2021, 107: 107353.

[6] AKRAM M, ALI G, ALCANTUD J C, et al. Parameter reductions in N-soft sets and their applications in decision-making[J]. Expert Systems, 2021, 38(1) : e12601.

[7] ZHANG C, LI D Y, LIANG J Y. Multi-granularity three-way decisions with adjustable hesitant fuzzy linguistic multigranulation decision-theoretic rough sets over two universes[J]. Information Sciences, 2020, 507: 665-683.

[8] SHI Z Q, XIE S R, LI L Q. Generalized fuzzy neighborhood system-based multigranulation variable precision fuzzy rough sets with double TOPSIS method to MADM [J]. Information Sciences: An International

- Journal, 2023, 643: 119251.
- [9] HU Q H, XIE Z X, YU D R. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation[J]. Pattern Recognition, 2007, 40(12): 3509-3521.
- [10] WANG Z H. Fundamental properties of fuzzy rough sets based on triangular norms and fuzzy implications: The properties characterized by fuzzy neighborhood and fuzzy topology[J]. Complex & Intelligent Systems, 2024, 10(1): 1103-1114.
- [11] 杨洁, 匡俊成, 王国胤, 等. 代价敏感的多粒度邻域粗糙模糊集的近似表示[J]. 计算机科学, 2023, 50(5): 137-145.
- YANG J, KUANG J C, WANG G Y, et al. Approximate representation of a cost-sensitive, multi-granularity neighborhood rough fuzzy set[J]. Computer Science, 2023, 50(5): 137-145.
- [12] DUCANGE P, FAZZOLARI M, MARCELLONI F. An overview of recent distributed algorithms for learning fuzzy models in Big Data classification[J]. Journal of Big Data, 2020, 7(1): 1-29.
- [13] WANG D, NIE F P, HUANG H. Feature selection via global redundancy minimization[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(10): 2743-2755.
- [14] TANG X C, DAI Y S, XIANG Y P. Feature selection based on feature interactions with application to text categorization[J]. Expert Systems with Applications, 2019, 120: 207-216.
- [15] HAGAR J D, WISSINK T L, KUHN D R, et al. Introducing combinatorial testing in a large organization[J]. IEEE International Conference on Software Testing, 2015, 48(4): 64-72.
- [16] WANG W J, GUO M, HAN T T, et al. A novel feature selection method considering feature interaction in neighborhood rough set[J]. Intelligent Data Analysis, 2023, 27(2): 345-359.
- [17] ZHOU P, LIP P, ZHAO S, et al. Feature interaction for streaming feature selection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(10): 4691-4702.
- [18] LIN X H, LI C, REN W J, et al. A new feature selection method based on symmetrical uncertainty and interaction gain[J]. Computational Biology and Chemistry, 2019, 83: 107149.
- [19] 孙敏杰. 基于模糊邻域熵的特征基因选择方法研究[D]. 广州: 广州大学, 2023: 7-11.
- [20] XU J C, MENG X R, QU K L, et al. Feature selection using relative dependency complement mutual information in fitting fuzzy rough set model[J]. Applied Intelligence, 2023, 53(15): 18239-18262.
- [21] MA X, XU H, JU C H. Class-specific feature selection via maximal dynamic correlation change and minimal redundancy[J]. Expert Systems with Applications, 2023, 229: 120455.
- [22] CHEN D G, ZHANG L, ZHAO S Y, et al. A novel algorithm for finding reducts with fuzzy rough sets[J]. IEEE Transactions on Fuzzy Systems, 2012, 20(2): 385-389.
- [23] SUN L, WANG L Y, QIAN Y H, et al. Feature selection using Lebesgue and entropy measures for incomplete neighborhood decision systems[J]. Knowledge-Based Systems, 2019, 186: 104942.
- [24] TAN A H, WU W Z, QIAN Y H, et al. Intuitionistic fuzzy rough set-based granular structures and attribute subset selection[J]. IEEE Transactions on Fuzzy Systems, 2019, 27(3): 527-539.
- [25] SUN L, WANG L Y, DING W P, et al. Neighborhood multi-granulation rough sets-based attribute reduction using Lebesgue and entropy measures in incomplete neighborhood decision systems[J]. Knowledge-Based Systems, 2020, 192: 105373.
- [26] JENSEN R, SHEN Q. New approaches to fuzzy-rough feature selection[J]. IEEE Transactions on Fuzzy Systems, 2009, 17(4): 824-838.
- [27] XU J C, QU K L, MENG X R, et al. Feature selection based on multiview entropy measures in multiperspective rough set[J]. International Journal of Intelligent Systems, 2022, 37(10): 7200-7234.

(编辑 张欢)