



面向单调特征选择的改进模糊排序互信息

皮 洪¹, 罗 川¹⁺, 李天瑞², 陈红梅²

1. 四川大学 计算机学院, 成都 610065

2. 西南交通大学 计算机与人工智能学院, 成都 611756

+ 通信作者 E-mail: cluo@scu.edu.cn

摘 要: 缺失值、异常值等低质信息的大量存在使得实际应用中单调分类任务通常不满足一致性单调约束。然而, 现有的面向单调分类的特征选择算法中用于评估特征相关性的互信息度量准则在处理不一致单调分类任务时不满足特征集合与度量准则之间的单调性限制关系。针对此问题, 首先对经典的模糊排序信息熵进行了改进, 使之满足单调性限制关系。进一步定义了模糊排序互信息(nFRMI)用于评估特征之间的单调一致性, 并证明了该度量准则在一致性和不一致性单调分类任务中均满足单调性约束, 此外结合最大相关最小冗余(mRMR)准则提出了两阶段的特征选择算法。算法第一阶段利用nFRMI计算每个特征的特征重要度, 然后根据特征重要度得到特征排序; 第二阶段采用Wrapper方法选出具有最优分类性能的特征子集。仿真实验从UCI挑选了四个单调分类数据集, 并从分类精度指标对所提出的特征选择算法的有效性进行了验证。

关键词: 特征选择; 单调分类; 不一致; 模糊信息熵; 排序互信息

文献标志码: A **中图分类号:** TP181

Improved Fuzzy Rank Mutual Information for Monotonic Feature Selection

PI Hong¹, LUO Chuan¹⁺, LI Tianrui², CHEN Hongmei²

1. College of Computer Science, Sichuan University, Chengdu 610065, China

2. School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China

Abstract: The large amount of low-quality information, such as missing values and outliers, makes monotonic classification tasks usually not satisfy the monotonic consistency. However, for the existing ordered feature selection algorithm, the mutual information used to evaluate the feature relevance doesn't satisfy the monotonic restriction between feature sets and the metrics when the monotonic classification tasks are inconsistent. In response to the problem, this paper firstly improves the classical fuzzy rank information entropy to make it satisfy the monotonicity restriction. The novel fuzzy rank mutual information (nFRMI) is further defined to evaluate the monotonic consistency of features, and it is proven that the metric criterion satisfies the monotonicity in both consistent and inconsistent monotonic classification tasks. In addition, a two-stage feature selection algorithm is proposed in combination with the minimal-redundancy-maximal-relevance (mRMR) criterion. In the first stage of the algorithm, it calculates significance of each feature based on nFRMI, and then obtains the feature ranking based on the feature significance; the Wrapper method is used to select the feature subset with the optimal classification performance in the second stage. Simulation experiments are conducted on four UCI datasets to verify the effectiveness of the proposed algorithm from the perspective of classification accuracy.

Key words: feature selection; monotonic classification; inconsistent; fuzzy information entropy; rank mutual information

基金项目: 国家自然科学基金(62076171, 61573292, 61976182)。

This work was supported by the National Natural Science Foundation of China (62076171, 61573292, 61976182).

收稿日期: 2021-05-08 **修回日期:** 2021-06-30

从不同决策值之间是否存在有序结构的角度, 可将分类问题分为有序分类及名义分类^[1]。对于有序分类任务, 不同决策值之间存在着有序关系。例如, 学生的成绩可以分为5个等级, 分别是不合格、合格、中等、良好、优秀, 在这5个等级中存在有序关系, 优秀表示最好的等级, 而不合格则是最差的等级。当两个样本的条件特征值存在优劣关系, 并且其决策特征值也存在对应的优劣关系, 那么这两个样本之间存在单调约束。在有序分类的基础上, 假设所有条件特征与决策特征之间都存在单调性依赖约束, 那么该有序分类问题称为单调分类问题。单调分类任务广泛存在于信用评价、医疗诊断、风险评估等决策领域, 具有重要的研究意义和实际应用价值。

特征选择是从原始特征中选择最具类别辨识性的特征以降低特征维度的数据处理过程^[2-4]。近年来, 面向单调分类任务的特征选择方法逐渐成为了人们关注的热点问题, 并取得了一些研究成果。Hu等将最大决策边界准则引入到具有单调性约束的有序分类问题中, 提出了基于ReliefF和Simba的单调特征选择算法^[5]。Pan等定义了特征的单调依赖度并提出了通过使用梯度下降最大化模糊偏好依赖的特征选择算法^[6]。考虑到同时含有名义分类特征和单调分类特征的异构有序分类问题, Pan和Hu进一步定义了基于异构单调分类一致性的特征相关性度量函数, 并利用基因算法对异构单调数据集中的最优特征子集进行求解^[7]。Villela等基于最大决策边界稀疏分类器, 提出了一种基于容许有序搜索策略的包裹式有序特征选择算法^[8]。

互信息作为信息论中一种常用的度量, 通常用于评估两个随机变量之间的相关性。在名义分类任务中, 已经提出了大量基于互信息的特征选择算法^[2,9-13]。Qian等针对不完备数据集提出了一种基于互信息的特征选择算法, 并证明了互信息的单调性^[2]。Kwak等提出了一种计算连续输入特征及离散输出类别间的互信息的方法, 并将其应用到输入特征贪心选择算法中^[11]。Battiti探讨了互信息准则在评估候选特征集的应用场景, 并利用互信息准则来选择一组具有高度代表性的特征, 将其作为神经网络分类器的输入数据^[12]。Peng等结合mRMR(minimal-redundancy-maximal-relevance)及其他方法(如包装器)提出了一种两阶段的特征选择方法, 能够以极低的代价选出一组具有高度代表性的特征^[13]。为了将基于互信息的特征相关性度量准则引入到基于单调性约束的有

序分类问题中, Hu等定义了排序互信息(rank mutual information, RMI)及模糊排序互信息(fuzzy rank mutual information, FRMI), 以评估单调分类任务中条件特征与决策特征之间的单调一致性, 并基于这两种度量提出了面向单调分类的特征选择算法^[1]。许行等在Hu的算法的基础上, 首先利用双向有序互信息生成不同的决策树, 然后集成其分类规则得到最终的决策结果^[14]。然而, 实际应用中由于低质信息的存在, 使得单调分类任务中往往存在决策不一致性。Hu等所定义的基于RMI及FRMI的特征一致性度量准则在处理不一致单调分类任务时不满足特征集合与度量准则之间的单调性限制关系, 进而影响特征重要性排序的启发式搜索过程。鉴于此, 本文在Hu等提出的RMI和FRMI的基础上提出了一种新颖的不一致单调分类中满足单调性限制关系的模糊排序条件熵以及模糊排序互信息, 并将这种模糊排序互信息与最大相关最小冗余准则结合起来, 实现了不一致单调分类任务的特征选择问题。在UCI公开数据集上的仿真实验进一步验证了本文提出的特征选择算法具有更好的单调分类性能。

1 基本概念

本章介绍了单调有序分类、有序关系以及模糊排序互信息等相关基本概念。

1.1 单调有序分类

在粗糙集理论中, 三元组 $DS = \langle U, A, D \rangle$ 是一个决策系统, 其中 U 被称为论域, 表示样本的非空有限集合; A 表示条件特征的非空有限集; D 表示决策特征集。令决策特征集为 $D = \{d_1, d_2, \dots, d_k\}$, 若任意 $d_i \in D$ 是名义型特征, 则 $\langle U, A, D \rangle$ 是一个名义分类决策系统。若决策特征满足关系 $d_1 < d_2 < \dots < d_k$, 则称 $\langle U, A, D \rangle$ 是一个有序分类决策系统。进一步, 若样本的条件特征与决策特征之间存在单调约束关系, 即对 $\forall x \in U$, $v(x, A) \leq v(x', A)$ 有 $f(x) \leq f(x')$, 其中 $v(x, A)$ 表示样本 x 在特征集 A 上的取值, f 为决策函数。那么, 就称 $\langle U, A, D \rangle$ 是一个单调分类决策系统。为了便于讨论, 本文仅讨论有序关系增加的相关问题。

给定一个单调分类决策系统 $\langle U, A, D \rangle$, 特征子集 $B \subseteq A$, 则条件特征与决策特征的有序关系定义为:

$$R_B^{\leq} = \{(x_i, x_j) \in U \times U | v(x_i, a_l) \leq v(x_j, a_l), \forall a_l \in B\} \quad (1)$$

$$R_D^{\leq} = \{(x_i, x_j) \in U \times U | v(x_i, D) \leq v(x_j, D)\} \quad (2)$$

$\forall x_i \in U$, 样本 x_i 关于特征子集 B 的有序信息粒

可表示为:

$$[x_i]_B^{\leq} = \{x_j \in U | (x_i, x_j) \in R_B^{\leq}\} \quad (3)$$

$\forall x_i \in U$, 样本 x_i 关于决策特征 D 的有序信息粒可表示为:

$$[x_i]_D^{\leq} = \{x_j \in U | (x_i, x_j) \in R_D^{\leq}\} \quad (4)$$

给定单调分类决策系统 $\langle U, A, D \rangle$, 论域 $U = \{x_1, x_2, \dots, x_n\}$ 。定义 $R = \{r_{ij}\}_{n \times n}$ 是 U 上的一个有序关系, 其中 r_{ij} 的值将根据样本 x_i 与 x_j 在条件特征 a 上的取值分为三种情况:

- (1) 当 $v(x_i, a) > v(x_j, a)$ 时, $r_{ij} \in [0, 0.5)$;
- (2) 当 $v(x_i, a) = v(x_j, a)$ 时, $r_{ij} = 0.5$;
- (3) 当 $v(x_i, a) < v(x_j, a)$ 时, $r_{ij} \in (0.5, 1.0]$ 。

使用 Logsig Sigmoid 函数来计算样本 x_i 和 x_j 之间的有序隶属度 r_{ij} :

$$r_{ij} = \frac{1}{1 + e^{\frac{k(v(x_i, a) - v(x_j, a))}{2}}} \quad (5)$$

样本 x_i 的模糊有序集可表示为:

$$[x_i]_R^{\leq} = r_{i1}/x_1 + r_{i2}/x_2 + \dots + r_{in}/x_n \quad (6)$$

其中, r_{ij} 能够反映样本 x_i 劣于 x_j 的程度^[15]。

1.2 已存在的模糊排序互信息

为了处理单调分类决策系统中的特征选择问题, Hu 等^[1]从信息论的角度定义了一种模糊排序条件熵和模糊排序互信息。

给定单调分类决策系统 $DS = \langle U, A, D \rangle$, 特征子集 $B \subseteq A$, $R = \{r_{ij}\}_{n \times n}$ 是论域 U 上关于特征子集 B 的有序关系。 $[x_i]_B^{\leq}$ 是样本 x_i 的关于特征子集 B 的有序信息粒。决策系统 DS 的模糊排序熵定义如下:

$$RH_R^{\leq}(U) = - \sum_{i=1}^n \frac{1}{n} \ln \frac{|[x_i]_R^{\leq}|}{n} \quad (7)$$

式中, $|[x_i]_R^{\leq}| = \sum_j r_{ij}$ 是模糊有序集 $[x_i]_R^{\leq}$ 的模糊基数,

$RH_R^{\leq}(U)$ 也可记为 $RH_B^{\leq}(U)$ 。

给定单调分类决策系统 $DS = \langle U, A, D \rangle$, 特征子集 $B \subseteq A$, B 和 D 的模糊联合排序熵定义如下:

$$RH_{B \cup D}^{\leq}(U) = - \sum_{i=1}^n \frac{1}{n} \ln \frac{|[x_i]_B^{\leq} \cap [x_i]_D^{\leq}|}{n} \quad (8)$$

式中, $|[x_i]_B^{\leq} \cap [x_i]_D^{\leq}| = \min(r_{i1}, s_{i1}) + \min(r_{i2}, s_{i2}) + \dots + \min(r_{in}, s_{in})$ 。

给定单调分类决策系统 $DS = \langle U, A, D \rangle$, 特征子集 $B \subseteq A$, 给定 B 后, D 的模糊条件排序熵定义如下:

$$RH_{D|B}^{\leq}(U) = - \sum_{i=1}^n \frac{1}{n} \ln \frac{|[x_i]_B^{\leq} \cap [x_i]_D^{\leq}|}{|[x_i]_B^{\leq}|} \quad (9)$$

值得注意的是, 当单调分类任务满足一致性时, 上述定义的条件排序熵满足单调性。然而, 在实际应用中, 单调分类任务往往是不一致的, 即 $[x_i]_B^{\leq} \not\subseteq [x_i]_D^{\leq}$ 。换句话说, 对于 $C \subseteq B \subseteq A$, $RH_{D|C}^{\leq}(U) \geq RH_{D|B}^{\leq}(U)$ 不成立。

给定单调分类决策系统 $DS = \langle U, A, D \rangle$, 对于特征子集 $B \subseteq A$, 模糊排序互信息定义如下:

$$FRMI_{B,D}(U) = - \sum_{i=1}^n \frac{1}{n} \ln \frac{|[x_i]_B^{\leq}| \times |[x_i]_D^{\leq}|}{n \times |[x_i]_B^{\leq}| \cap [x_i]_D^{\leq}|} \quad (10)$$

从式(9)中可以得出 $RH_{D|B}^{\leq}(U)$ 并不总是满足单调性, 由此可以推断出上述模糊排序互信息并不满足单调性。因此, 它不适用于度量不一致单调分类决策系统中条件特征与决策特征间的单调一致性。

2 单调分类决策系统中的模糊排序互信息

本章将给出一种模糊排序互信息的定义, 用于度量单调分类任务中的条件特征与决策特征间的单调一致性。这种新的定义能够克服单调分类任务不一致时模糊排序互信息不单调的缺点^[16-17]。

为了定义新的模糊排序互信息, 本文首先给出新的模糊排序熵及模糊条件排序熵的定义以及相关性质。

定义 1 (模糊排序熵) 给定单调分类决策系统 $DS = \langle U, A, D \rangle$, $U = \{x_1, x_2, \dots, x_n\}$, 特征子集 $B \subseteq A$, U 关于 B 的模糊排序熵定义为:

$$RH_B^{\leq}(U) = - \sum_{i=1}^n \frac{|[x_i]_B^{\leq}|}{n} \ln \frac{|[x_i]_B^{\leq}|}{n} \quad (11)$$

定义 2 (模糊条件排序熵) 给定单调分类决策系统 $DS = \langle U, A, D \rangle$, $U = \{x_1, x_2, \dots, x_n\}$, 特征子集 $B \subseteq A$, 给定 B 后, D 的模糊条件排序熵定义为:

$$RH_{D|B}^{\leq}(U) = - \sum_{i=1}^n \frac{|[x_i]_B^{\leq} \cap [x_i]_D^{\leq}|}{n} \ln \frac{|[x_i]_B^{\leq} \cap [x_i]_D^{\leq}|}{|[x_i]_B^{\leq}|} \quad (12)$$

下面将讨论模糊条件排序熵的一些重要的性质。

定理 1 (单调性) 给定一个单调分类系统 $\langle U, A, D \rangle$, 特征子集 $C \subseteq B \subseteq A$, 则:

$$RH_{D|B}^{\leq}(U) \leq RH_{D|C}^{\leq}(U)$$

证明 首先, 将 B 的有序信息粒改写为如下形式:

$$\begin{aligned} [x_i]_B^{\leq} &= \sum_{x_j \in U} [x_i]_B^{\leq}(x_j) = \sum_{x_j \in [x_i]_B^{\leq}} [x_i]_B^{\leq}(x_j) + \\ &\sum_{x_j \in U - [x_i]_B^{\leq}} [x_i]_B^{\leq}(x_j) = [x_i]_B^{\leq} \cap [x_i]_D^{\leq} + [x_i]_B^{\leq} \cap (U - [x_i]_D^{\leq}) \end{aligned}$$

同理可得:

$$[x_i]_C^{\leq} = [x_i]_C^{\leq} \cap [x_i]_D^{\leq} + [x_i]_C^{\leq} \cap (U - [x_i]_D^{\leq})$$

令 $[x_i]_B^{\leq} \cap [x_i]_D^{\leq} = b_{i1}$, $[x_i]_B^{\leq} \cap (U - [x_i]_D^{\leq}) = b_{i2}$

$[x_i]_C^{\leq} \cap [x_i]_D^{\leq} = c_{i1}$, $[x_i]_C^{\leq} \cap (U - [x_i]_D^{\leq}) = c_{i2}$

因此,模糊条件排序熵能够写为如下形式:

$$RH_{D|B}^{\leq}(U) = -\frac{1}{n} \sum_{i=1}^n b_{i1} \ln \frac{b_{i1}}{b_{i1} + b_{i2}}$$

$$RH_{D|C}^{\leq}(U) = -\frac{1}{n} \sum_{i=1}^n c_{i1} \ln \frac{c_{i1}}{c_{i1} + c_{i2}}$$

接下来对二元函数 $g(x, y) = -x \ln \frac{x}{x+y}$ ($x > 0, y \geq 0$)

的单调性进行讨论,记 $z = \frac{x}{x+y}$, 则:

$$\begin{cases} \frac{\partial g}{\partial x} = -\frac{y}{x+y} - \ln \frac{x}{x+y} = -1 + z - \ln z \\ \frac{\partial g}{\partial y} = \frac{x}{x+y} > 0 \end{cases}$$

由 $(-1 + z - \ln z)' = 1 - \frac{1}{z}$ 以及 $z \in (0, 1]$ 可得 $\frac{\partial g}{\partial y} \geq 0$ 。

由上述讨论可以看出函数 $g(x, y)$ 分别关于两个变量单调递增。

对 $C \subseteq B$, 由 $[x_i]_B^{\leq} \subseteq [x_i]_C^{\leq}$ 可推导出以下公式:

$$[x_i]_B^{\leq} \cap [x_i]_D^{\leq} \subseteq [x_i]_C^{\leq} \cap [x_i]_D^{\leq}$$

$$[x_i]_B^{\leq} \cap (U - [x_i]_D^{\leq}) \subseteq [x_i]_C^{\leq} \cap (U - [x_i]_D^{\leq})$$

由此可得 $0 \leq b_{i1} \leq c_{i1}, 0 \leq b_{i2} \leq c_{i2}$, 进而 $g(b_{i1}, b_{i2}) \leq$

$g(c_{i1}, c_{i2})$, 因此 $\frac{1}{n} \sum_{i=1}^n g(b_{i1}, b_{i2}) \leq \frac{1}{n} \sum_{i=1}^n g(c_{i1}, c_{i2})$,

即 $RH_{D|B}^{\leq}(U) \leq RH_{D|C}^{\leq}(U)$ 。

定理 2 (最大值) 给定一个单调分类系统 $\langle U, A, D \rangle$, $U = \{x_1, x_2, \dots, x_n\}$, 特征子集 $B \subseteq A$, 如果对 $\forall x_i \in U$ 都有 $[x_i]_B^{\leq} = U$ 以及 $[x_i]_D^{\leq} = n/e$, 则模糊条件排序熵的最大值为 n/e 。

证明 特征子集 $B \subseteq A$, 则有 $[x_i]_A^{\leq} \subseteq [x_i]_B^{\leq} \subseteq U$ 。

由定理 1 可得, 当对 $\forall x_i \in U$, 都有 $[x_i]_B^{\leq} = U$, 此时 $RH_{D|B}^{\leq}$ 的值最大, 结果如下:

$$RH_{D|B}^{\leq} = -\sum_{i=1}^n \frac{[x_i]_D^{\leq}}{n} \ln \frac{[x_i]_D^{\leq}}{|U|} = -\frac{1}{n} \sum_{i=1}^n [x_i]_D^{\leq} \ln \frac{[x_i]_D^{\leq}}{n}$$

令 $g(x) = x \ln \frac{x}{n}, x \in [1, n]$, 则函数 $g(x)$ 的一阶导和二阶导分别为:

$$g'(x) = \ln \frac{x}{n} + 1, g''(x) = \frac{1}{x} > 0$$

由上述两式可得 $g' \geq 0$ 且 g' 单增。

令 $g'(x_0) = 0$ 可得 $x_0 = n/e$, 从而可以得到 $g(x_0) = \min g(x) = n/e \cdot \ln \frac{n/e}{n} = -n/e$, 此时 $RH_{D|B}^{\leq}$ 的值最大, 为 n/e 。

定理 3 (最小值) 给定一个单调分类决策系统

$\langle U, A, D \rangle$, 特征子集 $B \subseteq A$, $RH_{D|B}^{\leq}(U) = 0$ 当且仅当 D 完全依赖于 B 。

证明 (1) 当决策特征 D 完全依赖于条件特征集合 B 时, 该单调分类决策系统满足一致性, 即 $[x_i]_B^{\leq} \subseteq [x_i]_D^{\leq}$, 于是:

$$RH_{D|B}^{\leq} = -\sum_{i=1}^n \frac{[x_i]_B^{\leq}}{n} \ln \frac{[x_i]_B^{\leq}}{[x_i]_B^{\leq}} = 0$$

(2) 当 $RH_{D|B}^{\leq} = 0$ 时, 需要证明 D 完全依赖于 B 。

假设 D 不完全依赖于 B , 则 $\exists x_i \in U$ 使得 $[x_i]_B^{\leq} \not\subseteq [x_i]_D^{\leq} \Rightarrow [x_i]_B^{\leq} \cap [x_i]_D^{\leq} \neq [x_i]_B^{\leq} \Rightarrow \frac{[x_i]_B^{\leq} \cap [x_i]_D^{\leq}}{n} \ln \frac{[x_i]_B^{\leq} \cap [x_i]_D^{\leq}}{[x_i]_B^{\leq}} \neq 0 \Rightarrow RH_{D|B}^{\leq} \neq 0$ 。

这与已有条件矛盾, 故假设不成立, 即 D 完全依赖于 B 。

在给出上述定义及性质后, 下面将定义新的模糊排序互信息。

定义 3 (模糊排序互信息) 给定一个单调分类决策系统 $\langle U, A, D \rangle$, 特征子集 $B \subseteq A$ 。 B 和 D 的模糊排序互信息定义如下:

$$nFRMI_{B,D}(U) = RH_D^{\leq}(U) - RH_{D|B}^{\leq}(U) \quad (13)$$

3 特征选择的模糊排序互信息准则

本章首先讨论单调分类决策系统中模糊排序互信息用于特征选择的有效性, 接下来介绍基于模糊排序互信息的特征选择算法。

3.1 模糊排序互信息的有效性

定理 4 (单调性) 给定一个单调分类决策系统 $\langle U, A, D \rangle$, 若特征子集 $C \subseteq B \subseteq A$, 则:

$$nFRMI_{C,D}(U) \leq nFRMI_{B,D}(U)$$

从定理 4 中可以得出新定义的模糊排序互信息在一致单调分类决策系统中及不一致单调分类决策系统中均满足单调性。

定理 5 (有效性) 给定一个单调分类决策系统 $\langle U, A, D \rangle$, 特征子集 $B \subseteq A$ 。 如果决策特征 D 完全依赖条件特征集 B , 那么有:

$$nFRMI_{B,D}(U) = RH_D^{\leq}(U)$$

证明 当 D 完全依赖于 B 时, $[x_i]_B^{\leq} \subseteq [x_i]_D^{\leq}$ 。 由此可得:

$$\begin{aligned} RH_{D|B}^{\leq}(U) &= -\sum_{i=1}^n \frac{[x_i]_B^{\leq} \cap [x_i]_D^{\leq}}{n} \ln \frac{[x_i]_B^{\leq} \cap [x_i]_D^{\leq}}{[x_i]_B^{\leq}} = \\ &= -\sum_{i=1}^n \frac{[x_i]_B^{\leq}}{n} \ln \frac{[x_i]_B^{\leq}}{[x_i]_B^{\leq}} = 0 \end{aligned}$$

于是 $nFRMI_{B,D}(U) = RH_D^{\leq}(U) - 0 = RH_D^{\leq}(U)$ 。

3.2 基于模糊排序互信息的特征选择算法

上述章节提出了用于单调分类决策系统的基于模糊排序互信息的特征度量准则,接下来讨论基于该准则的特征选择算法。

在特征选择过程中,需要考虑两方面:所选特征间的冗余度要尽可能小以及所选特征与决策特征之间的相关性要尽可能大。从相关性及冗余度的角度来说,最优的 m 个特征是与分类最相关的特征。但是最优 m 个特征之间可能存在冗余,因此最相关的 m 个特征不一定比其他 m 个特征得到更好的分类准确率。因此,特征选择应该被分为两个过程:如何度量特征与决策特征之间的相关性以及如何处理特征间的冗余性。特征相关性可以利用本文定义的模糊排序互信息来度量。第二个问题可由下述定义4解决,即采用最大相关最小冗余(mRMR)准则来选择当前第 $|B|+1$ 个特征。

定义4(特征评价函数) 给定单调分类决策系统 $\langle U, A, D \rangle$, 设 $B \subseteq A$ 是所选的条件特征子集, $a_j \in A - B$ 是候选特征, 候选特征 a_j 的评价函数定义如下:

$$e(a_j) = nFRMI_{a_j, D} - \frac{1}{|B|} \sum_{a_i \in B} nFRMI_{a_i, a_j} \quad (14)$$

本文提出了一种两阶段的特征选择算法。第一个阶段根据特征评价函数对所有条件特征进行排序。将从候选特征集中选择最大化 $e(a_j)$ 的特征。被选择的特征与之前所选的特征冗余度小且与决策特征的相关性大。第一阶段的具体过程见算法1。在第二个阶段,利用算法2选择一个最具类别辨识性的特征子集。该特征子集具有最好的分类性能。

算法1 模糊决策系统的特征排序

输入:单调分类决策系统 $DS = \langle U, A, D \rangle$ 。

输出:特征排序 B 。

1. 初始化 $B = \emptyset$, $e(a_j) = 0$;

2. while $|B| \neq |A|$ do

3. for each $a_j \in A - B$ do

4. 对每个候选特征 a_j 计算

$$e(a_j) = nFRMI_{a_j, D} - \frac{1}{|B|} \sum_{a_i \in B} nFRMI_{a_i, a_j}$$

5. 找出满足 $e(a_{j_0}) = \max_{a_j \in A - B} e(a_j)$ 的特征

6. 令 $B = B \cup \{a_{j_0}\}$

7. 返回 B

算法2 选择最优特征子集的 Wrapper 算法

输入:单调分类决策系统 $DS = \langle U, A, D \rangle$, 特征排序 B 。

输出:最优特征子集 B^* 。

1. 将 B 分为如下子集: $B_1 = \{a_{i1}\}$, $B_2 = \{a_{i1}, a_{i2}\}$, $B_n = \{a_{i1}, a_{i2}, \dots, a_{in}\}$

2. for $k = 1$ to n do

3. 从数据集 DS 中移除不包含在 B_k 中的特征,并将这个新得到的数据集记为 DS_k

4. 在给定分类器下计算数据集 DS_k 的分类性能

5. 从所有数据子集中选出分类性能最好的数据集 DS_{k_0}

6. 令 $B^* = B_{k_0}$

7. 返回 B^*

4 实验分析

本章将通过实验验证所提方法的有效性及其高效性。本文将第3章提出的评价准则 nFRMI 与 Hu 等提出的 FRMI 进行比较。

在接下来的实验部分,本文共使用两种单调分类器对所提方法的分类精度进行性能评测,分别是 OLM(ordered learning model)^[18]和 OCC(ordinal class classifier),均在 Weka 中得以实现。这两种分类器用于在 Wrapper 阶段计算特征子集的分类性能,并由此选择出具有最优分类性能的特征子集。

本章的所有实验均采用10折交叉验证法,将数据集分为10份,轮流将其中9份用作训练数据,剩余1份用作测试数据。每次测试将得到对应的分类精度。10次测试的结果平均值作为最终的算法评价指标。本文选取了4组单调分类数据集进行性能实验,数据集详细信息如表1所示。

表1 数据集描述

Table 1 Description of datasets

数据集	样本数	特征数	分类数
Crx	690	15	2
Ionosphere	351	34	2
MachineCPU	209	6	4
Pima	768	8	2

表2和表3的数据描述了本文算法与对比算法在不同单调分类器作用下所得到的最优特征子集在10折交叉验证中的平均分类精度及标准差,每行中的最优分类精度用粗体表示。FRMI表示该列实验结果是基于Hu等所提出的度量函数的特征选择算法所得到的,nFRMI则表示实验结果是基于本文算法得到的。

表2 OLM分类器下的分类精度比较

Table 2 Comparison of classification

数据集	accuracy by OLM			单位 %
	Raw	FRMI	nFRMI	
Crx	56.91 ± 4.07	73.29 ± 19.27	74.83 ± 13.03	
Ionosphere	63.08 ± 3.70	67.21 ± 6.99	70.30 ± 6.70	
MachineCPU	54.79 ± 8.87	66.10 ± 13.72	68.30 ± 13.43	
Pima	31.39 ± 2.90	41.75 ± 12.36	41.75 ± 12.36	

表3 OCC分类器下的分类精度比较

Table 3 Comparison of classification

数据集	accuracy by OCC			单位 %
	Raw	FRMI	nFRMI	
Crx	84.74 ± 3.86	85.99 ± 3.87	85.51 ± 3.96	
Ionosphere	89.74 ± 4.38	89.89 ± 4.73	91.09 ± 4.15	
MachineCPU	94.20 ± 5.71	94.16 ± 5.79	94.40 ± 5.32	
Pima	74.67 ± 4.52	73.85 ± 4.64	74.87 ± 4.94	

从表2、表3可以看出,分类器的选择会影响最终的分类精度。当选择分类器OLM时,本文算法所得到的特征子集的分类精度在所有数据集上均优于原始数据集。与基于FRMI的特征选择算法相比,基于nFRMI的特征选择算法所得特征子集的分类精度在Crx、Ionosphere及MachineCPU三个数据集上是更优

的,而在Pima数据集上分类精度是相等的。当选择OCC分类器时,基于nFRMI的特征子集的分类精度在所有数据集上均优于原始数据集。与FRMI相比,nFRMI在Ionosphere、MachineCPU及Pima数据集上分类精度更优。

图1和图2分别展示了采用不同特征选择算法进行特征选择时,随着特征数量的增大基于OLM和OCC分类器的平均分类精度的变化趋势。从实验结果可以看出,除了OCC分类器下的Crx数据集,由基于nFRMI的特征选择算法所得特征子集的分类精度优于或等于基于FRMI的特征选择算法。另一方面,选用OLM分类器的Ionosphere数据集以及选用OCC分类器的MachineCPU数据集,基于nFRMI特征选择算法得到的特征子集在特征个数及分类精度两方面均优于基于FRMI的特征选择算法。从数据集Ionosphere可以看出,在特征选择的初始阶段,nFRMI的分类精度劣于原始数据集。随着特征选择的进行,nFRMI的分类精度整体呈现上升趋势,直至优于原始数据集。从上述实验结果可得,基于nFRMI的特征选择算法在不同分类器、不同数据集下均能够提升分类精度并降低数据维度,且在大部分数据集下分类精度优于传统的基于FRMI的特征选择算法。这表明

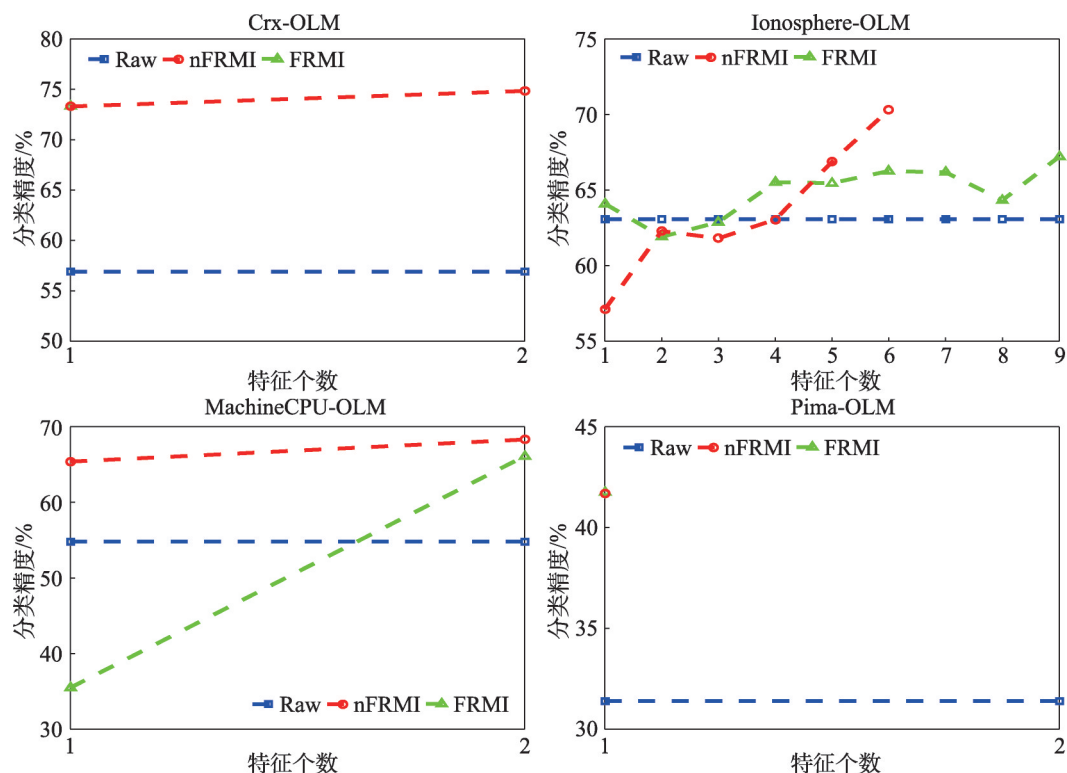


图1 OLM分类器下所选特征的分类精度比较

Fig.1 Comparison of classification accuracies with different number of selected features by OLM

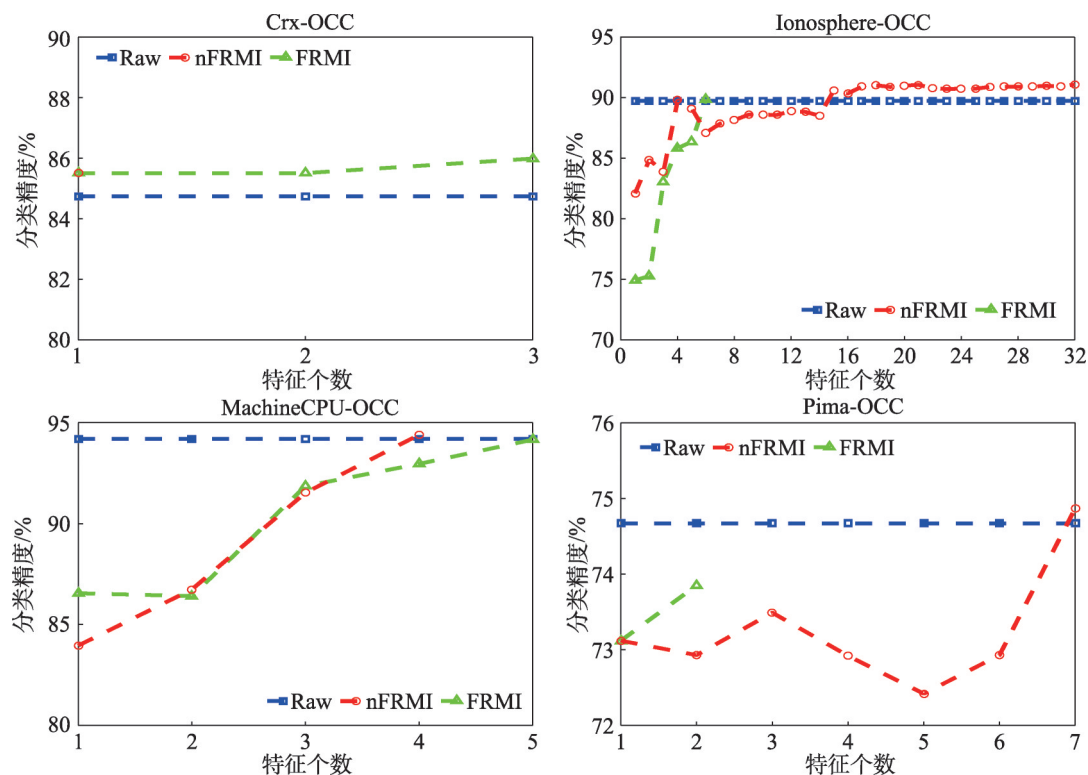


图2 OCC分类器下所选特征的分类精度比较

Fig.2 Comparison of classification accuracies with different number of selected features by OCC

本文算法对于面向不一致性单调分类任务的特征降维问题是有效的。

5 结束语

本文针对单调有序分类任务,提出了一种新颖的基于改进模糊排序互信息的特征选择算法,该算法能够有效克服已存在的模糊排序互信息在不一致性单调分类任务中不满足单调性约束的缺点。在特征选择过程中,使用最大相关最小冗余准则来确保所选特征之间的低冗余以及与决策特征之间的高相关性。实验部分,将本文算法与基于传统模糊排序互信息的特征选择算法在四个不一致性单调分类数据集上进行了对比实验,分析了两种算法得到的特征子集在两种单调分类器OLM和OCC下的分类精度性能比较。实验结果表明,与基于已存在模糊排序互信息的特征选择算法相比,本文算法在大部分数据集下能得到更优的分类精度,因此本文算法对于不一致性单调分类任务是有效的。

在后续工作中,将对本文所提出的改进模糊排序互信息在噪声数据干扰下的鲁棒性问题进行分析。并针对所提算法的计算效率问题进行优化,将其进一步扩展到面向动态单调分类的高效特征选择。

参考文献:

- [1] HU Q H, PAN W W, ZHANG L, et al. Feature selection for monotonic classification[J]. IEEE Transactions on Fuzzy Systems, 2012, 20(1): 69-81.
- [2] QIAN W B, SHU W H. Mutual information criterion for feature selection from incomplete data[J]. Neurocomputing, 2015, 168: 210-220.
- [3] DASH M, LIU H. Feature selection for classification[J]. Intelligent Data Analysis, 1997, 1: 131-156.
- [4] CHANDRASHKAR G, SAHIN F. A survey on feature selection methods[J]. Computers and Electrical Engineering, 2014, 40: 16-28.
- [5] HU Q H, PAN W W, SONG Y P, et al. Large-margin feature selection for monotonic classification[J]. Knowledge-Based Systems, 2012, 31: 8-18.
- [6] PAN W W, HU Q H, SONG Y P, et al. Feature selection for monotonic classification via maximizing monotonic dependency [J]. International Journal of Computational Intelligence Systems, 2014, 7(3): 543-555.
- [7] PAN W W, HU Q H. An improved feature selection algorithm for ordinal classification[J]. IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences, 2016, 99-A(12): 2266-2274.
- [8] VILLELA S M, LEITE S D C, XAVIER A E, et al. An ordered

- search with a large margin classifier for feature selection[J]. *Applied Soft Computing Journal*, 2021, 100: 106930.
- [9] ESTÉVEZ P A, TESMER M, PEREZ C A, et al. Normalized mutual information feature selection[J]. *IEEE Transactions on Neural Networks*, 2009, 20(2): 189-201.
- [10] VERGARA J R, ESTÉVEZ P A. A review of feature selection methods based on mutual information[J]. *Neural Computing and Applications*, 2014, 24: 175-186.
- [11] KWAK N, CHOI C H. Input feature selection by mutual information based on parzen window[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24(12): 1667-1671.
- [12] BATTITI R. Using mutual information for selecting features in supervised neural net learning[J]. *IEEE Transactions on Neural Networks*, 1994, 5(4): 537-550.
- [13] PENG H C, LONG F H, DING C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(8): 1226-1238.
- [14] 许行, 梁吉业, 王宝丽. 基于双向有序互信息的单调分类决策树算法[J]. *南京大学学报(自然科学)*, 2013, 49(5): 628-636.
- XU H, LIANG J Y, WANG B L. Bi-direction rank mutual information based decision trees for monotonic classification[J]. *Journal of Nanjing University (Natural Science)*, 2013, 49(5): 628-636.
- [15] LIN Y J, HU Q H, LIU J H, et al. Streaming feature selection for multilabel learning based on fuzzy mutual information [J]. *IEEE Transactions on Fuzzy Systems*, 2017, 25(6): 1491-1507.
- [16] DAI J H, WANG W T, XU Q. An uncertainty measure for incomplete decision tables and its applications[J]. *IEEE Transactions on Cybernetics*, 2013, 43(4): 1277-1289.
- [17] ZHANG X, MEI C L, CHEN D G, et al. Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy[J]. *Pattern Recognition*, 2016, 56: 1-15.
- [18] CANO J R, GUTIERREZ P A, KRAWCZYK B, et al. Monotonic classification: an overview on algorithms, performance measures and data sets[J]. *Neurocomputing*, 2019, 341: 168-182.



皮洪(1997—),女,重庆垫江人,硕士研究生,主要研究方向为数据挖掘。

PI Hong, born in 1997, M.S. candidate. Her research interest is data mining.



罗川(1987—),男,河南固始人,博士,副教授,CCF会员,主要研究方向为数据挖掘、粒计算等。

LUO Chuan, born in 1987, Ph.D., associate professor, member of CCF. His research interests include data mining, granular computing, etc.



李天瑞(1969—),男,福建莆田人,博士,教授,博士生导师,主要研究方向为数据挖掘、知识发现、粒计算等。

LI Tianrui, born in 1969, Ph.D., professor, Ph.D. supervisor. His research interests include data mining, knowledge discovery, granular computing, etc.



陈红梅(1971—),女,四川成都人,博士,教授,博士生导师,主要研究方向为数据挖掘、粒计算等。

CHEN Hongmei, born in 1971, Ph.D., professor, Ph.D. supervisor. Her research interests include data mining, granular computing, etc.