

DOI:10.13232/j.cnki.jnju.2024.01.011

不完备数据集的邻域容差互信息选择集成分类算法

李丽红^{1,2,3*}, 董红瑶^{1,2,3,6}, 刘文杰⁴, 李宝霖^{1,2,3}, 代 琪⁵

(1. 华北理工大学理学院, 唐山, 063210; 2. 河北省数据科学与应用重点实验室, 华北理工大学, 唐山, 063210;
3. 唐山市工程计算重点实验室, 华北理工大学, 唐山, 063210; 4. 华北理工大学人工智能学院, 唐山, 063210;
5. 中国石油大学(北京)自动化系, 北京, 102249; 6. 首钢矿业公司职工子弟学校, 唐山, 064404)

摘 要:针对不完备混合信息系统的分类问题,结合粒计算中的邻域容差关系和互信息理论,定义邻域容差互信息的概念,并利用集成学习的思想,提出不完备数据集的邻域容差互信息选择集成分类算法.该算法首先根据缺失属性得到信息粒,划分粒层构建粒空间,在不同的粒层上使用以 BP 神经网络作为基分类器的集成算法,构建新的基分类器;然后,根据每个信息粒的缺失属性计算出关于类属性的邻域容差互信息,来衡量各个信息粒的重要度,并根据基分类器预测准确率以及邻域容差互信息重新定义基分类器权重;最后,根据预测样本对基分类器加权集成预测分类结果,并与传统的集成分类算法进行对比分析.对于部分不完备混合型数据集,新提出的集成分类算法能有效提升分类准确率.

关键词:不完备混合信息系统,邻域容差互信息,集成学习,分类

中图分类号:TP181

文献标志码:A

Neighborhood-tolerance mutual information selection ensemble classification algorithm for incomplete data sets

Li Lihong^{1,2,3*}, Dong Hongyao^{1,2,3,6}, Liu Wenjie⁴, Li Baolin^{1,2,3}, Dai Qi⁵

(1. College of Science, North China University of Science and Technology, Tangshan, 063210, China; 2. Hebei Province Key Laboratory of Data Science and Application, North China University of Science and Technology, Tangshan, 063210, China;
3. Tangshan Key Laboratory of Engineering Computing, North China University of Science and Technology, Tangshan, 063210, China; 4. College of Artificial Intelligence, North China University of Science and Technology, Tangshan, 063210, China; 5. Department of Automation, China University of Petroleum, Beijing, 102249, China; 6. Shougang Minning workers' Children School, Tangshan, 064404, China)

Abstract: In order to solve the classification problem of incomplete mixed information systems, the concept of neighborhood-tolerance mutual information is defined by combining neighborhood-tolerance and mutual information theory in granular computing, and a selective ensemble classification algorithm based on neighborhood-tolerance mutual information is proposed by using ensemble learning. In this algorithm, information particles are obtained according to the missing attributes, and the space is constructed by dividing the particles into different layers. A new base classifier is constructed by integrating the BP neural network as the base classifier on different layers. Then, the neighborhood-tolerance mutual information about class attributes is calculated according to the missing attributes of each information particle to measure the importance of each information particle, and the weight of the base classifier is redefined according to the prediction accuracy of the base classifier and the neighborhood-tolerance mutual information. Finally, based on the predicted samples, the weighted ensemble prediction results of base classifier are analyzed and compared with the traditional ensemble classification algorithm. For

基金项目:河北省数据科学与应用重点实验室项目(10120201),唐山市数据科学重点实验室项目(10120301)

收稿日期:2023-09-27

* 通讯联系人, E-mail: 22687426@qq.com

partial incomplete mixed data sets, the proposed ensemble classification algorithm can effectively improve the classification accuracy.

Key words: incomplete hybrid information system, neighborhood - tolerance mutual information, ensemble learning, classification

在大数据时代,数据具有不确定性、动态更新、不完备性等特点.其中,数据挖掘领域常用的UCI数据库中有40%左右的数据集是不完备的.针对不完备数据集的分类问题,可以通过简单删除法或者填充法将不完备数据集进行处理,再用完备的数据集做进一步的分类,但这种方法不能保证数据是否为随机缺失,从而对分类精度产生影响^[1].尽管近年来对不完备数据集的研究逐渐增多,但目前不完备数据集中的大部分分类算法是针对只含有离散属性值的数据集设计的^[2].然而,有一种常见的情况是数据集为既含有离散型属性又含有连续型属性的混合形式,如从商业、医疗、银行、人口普查和生物科学中收集的数据,都是混合型的.对于医学数据,除了血压和血糖等连续型属性外,还包括性别和是否对某种药物过敏等离散型属性,人口普查收入数据包括职业、教育水平和婚姻状况等离散型属性,以及年龄、工资和每周工作时间等连续型属性.针对连续型属性值通常采用离散化的计算方式将连续型数值直接转化为离散型数值,这样会带来信息损失,从而影响分类准确率^[3].所以学者提出了直接处理不完备混合型数据集的方式,如利用相容关系^[4]、限制邻域关系^[5]和邻域容差关系^[6]等,不同的数据结构采用不同的逼近机制和粒化方式构建粒空间再进行下一步数据操作^[7].对于不完备混合型信息系统的分类问题,为进一步提升其分类性能,将集成分类方法引入研究.

目前针对不完备混合型信息系统的集成分类算法研究较少.Krause and Polikar^[8]首次提出Learn+MF集成算法处理不完备数据集的分类问题,子分类器在随机特征子集上进行训练,这种方法相对复杂,效率较低.因为集成分类算法针对不完备数据集的分类问题具有较好的冗余性而且适用性广,它不会因为对数据集假设不当使最终构建的模型产生偏差,而且可以充分利用数据

集的信息,所以,用集成算法处理不完备数据集的问题相继被提出^[9].Chen et al^[10]与吕靖和舒礼莲^[11]提出一种基于不完备数据集的不完备特征组合的集成框架,该方法不需要任何关于缺失数据的假设,但没有考虑不同特征子集重要程度的差异.在一般集成框架的基础上,通过考虑特征重要度,提出了多粒度集成方法(Multi-Granularity Integration Method, MGNE)^[12],然而,对于含有大量不完整样本的数据集,该方法性能有待提高,同时,随着缺失值数量的增加,这些算法非常耗时.为克服传统集成学习技术的高计算成本的不足,集成剪枝是一种常见提升性能的方法^[13-15].Yan et al^[16]针对不完备数据集提出一种选择性神经网络集成分类算法,与传统神经网络集成算法在保证精度的前提下相比,提高了算法效率.并且针对不完备混合数据集的分类问题,传统的集成分类算法在赋予各个子分类器权重时,仅考虑数据集中所含样本的多少和属性的维数,并没有考虑不同属性或属性组合对最终分类结果的贡献度.因此,如何有效地衡量不完备混合系统中属性对分类结果的贡献度,从而更加合理地计算基分类的权重提高分类的准确率有待进一步完善和解决.

针对上述问题,根据当前利用集成分类算法和粗糙粒化思想处理不完备混合数据的不足及优势,本文提出基于邻域容差互信息的选择集成分类算法(Neighborhood Tolerance Mutual Information Selection Ensemble Classification Algorithm, NTMISECA).首先定义邻域容差互信息,并详细描述基于邻域容差互信息选择集成分类算法的思想和步骤,然后介绍验证该算法采用的实验数据的详细信息与仿真环境,最后对实验结果进行讨论和总结以及阐述未来研究的工作重点.

1 基本原理

1.1 不完备混合型信息系统

定义 1^[17] 设一个混合型信息系统为 $S = (U, A)$. 其中, U 为信息系统的论域, $A = C \cup D$ 称为信息系统的属性集合, $C = C^d \cup C^c$ 称为信息系统的条件属性集合, D 称为信息系统的决策属性集合. 这里的 C^d 为条件属性值是离散型数值, C^c 为条件属性值是连续型数值.

若 $\exists x \in U, x$ 在属性 $a (a \in A)$ 上的取值未知, 通常用“*”表示, 即 $a(x) = *$, 那么此时 S 称为不完备混合型信息系统.

1.2 粒计算 粒计算主要目的在于通过对问题的粒化分解, 使复杂问题得以简单处理, 体现问题处理的多维度、多视角、多层次的思想^[18]. 通过研究粒的产生、粒的性质以及粒化方式, 提出数据处理的数学方法, 支撑模型的建立, 实现计算机程序化处理.

一个本质性问题是基于粒计算理论对信息进行粒化, 不同的粒化方法可以获得不同的粒层信息, 粒空间的结构直接决定目标的求解效率. 常用的粒化方法依据二元关系, 如等价关系、相似关系、领域关系、优势关系等, 同一类样本分配到同一个信息粒中^[19-23]. 一般地, 存在两类造粒过程, 如功能造粒和关系造粒. 如果该构造过程完全基于样本属性, 称为功能粒化; 如果粒化过程基于样本之间的关系, 称为关系粒化. 在粒化过程中, 若给定多个不同的粒化规则, 从多角度、多层次可以得到各不相同的粒层. 本文针对不完备混合数据集, 利用邻域容差关系对数据集粒化处理.

定义 2^[24] (邻域容差关系) 设 $S = (U, A)$ 为不完备混合型信息系统, $A = C \cup D$, 设 $B \subseteq C$ 为属性子集, 且 $B = B^d \cup B^c$. 其中, B^d 表示属性子集中属性值为离散值, B^c 表示属性子集中属性值为连续值. 设邻域为 δ , 则在不完备混合信息系统 S 下, 根据属性子集 B 确定的邻域容差关系为:

$$NT_B^\delta = \left\{ (x, y) \in U^2 \mid \begin{aligned} & (a(x) = * \vee a(y) = * \vee (\Delta_a(x, y) = 0)) \wedge \\ & (\Delta_a(x, y) \leq \delta), \forall a \in B^d, \forall b \in B^c \end{aligned} \right\} \quad (1)$$

其中, $\Delta_a(x, y)$ 和 $\Delta_b(x, y)$ 分别表示对于离散属性和连续属性对象 x 与对象 y 之间的距离度量. 那么对于 $\forall x \in U$, 关于 NT_B^δ 的邻域类定义如下.

$$\delta_B(x) = \{ y \in U \mid (x, y) \in NT_B^\delta \} \quad (2)$$

1.3 信息论 Shannon^[25] 首次在通信领域提出信息熵的概念, 来衡量一个给定的随机事件所包含信息量的大小. 信息熵常被用来作为一个系统信息量的量化指标和属性的分辨力, 也是信息系统中相关性度量的一种常见手段.

互信息是信息论中一种有用的信息度量, 可以用来直接衡量两个变量之间拥有多少相同的信息量. 它可以看成一个随机变量包含另一个随机变量的信息量, 也可以看成一个随机变量确定的情况下另一个随机变量减少的不确定性.

定义 3 (互信息) 若有随机变量 X 和随机变量 Y , 随机变量 X 和随机变量 Y 之间的互信息 $I(X, Y)$ 表示如下:

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \quad (3)$$

其中, $H(X)$, $H(Y)$, $H(X, Y)$ 分别表示变量 X 的信息熵、变量 Y 的信息熵、变量 X 和 Y 的联合信息熵.

若两个随机变量之间的互信息越大, 则说明拥有的共同信息越多. 反之, 这两个随机变量之间拥有的共同信息越少. 如果有多个随机变量 X_1, X_2, \dots, X_n 和随机变量 Y , 那么这组随机变量集合与随机变量 Y 之间的互信息定义为:

$$I(X_1, X_2, \dots, X_n, Y) = H(X_1, X_2, \dots, X_n) + H(Y) - H(X_1, X_2, \dots, X_n, Y) \quad (4)$$

其中, $H(X_1, X_2, \dots, X_n)$, $H(Y)$, $H(X_1, X_2, \dots, X_n, Y)$ 分别为变量 X_1, X_2, \dots, X_n 的联合熵, 变量 Y 的信息熵, 变量 X_1, X_2, \dots, X_n, Y 的联合熵.

定义 4^[24] (邻域容差信息熵) 给定不完备混合型信息系统 $S = (U, A)$, $B \subseteq A$, 邻域半径为 δ , 且 $U/NT_B^\delta = \{NT_B^\delta(x_1), NT_B^\delta(x_2), \dots, NT_B^\delta(x_{|U|})\}$, 定义 B 的邻域容差信息熵为:

$$NTE_\delta(B) = \frac{1}{|U|} \left(1 - \frac{|NT_B^\delta(x_i)|}{|U|} \right) \quad (5)$$

如果 $NT_B^\delta(x_i) = U$, 则 $NTE_\delta(B) = 0$; 如果 $\forall x_i \in U, NT_B^\delta(x_i) = \{x_i\}$, 那么 $NTE_\delta(B) = 1$ —

$\frac{1}{|U|}$, 则 $0 \leq NTE_{\delta}(B) \leq 1 - \frac{1}{|U|}$.

定义 5^[24] (邻域容差联合熵) 给定 $S = (U, A)$ 为不完备混合型信息系统, $A = C \cup D$, $B_1, B_2 \subseteq C$, 设邻域半径为 δ , 并且 $U/NT_B^{\delta} = \{NT_B^{\delta}(x_1), NT_B^{\delta}(x_2), \dots, NT_B^{\delta}(x_{|U|})\}$. 则 B_1 和 B_2 的邻域容差联合熵记为:

$$NTE(B_1, B_2) = \frac{1}{|U|} \left(1 - \frac{|NT_{B_1}^{\delta}(x_i) \cap NT_{B_2}^{\delta}(x_i)|}{|U|} \right) \quad (6)$$

如果设 $U/NT_D = \{NT_D(x_1), NT_D(x_2), \dots, NT_D(x_{|U|})\}$, 对于任意 $B \subseteq C$, D 和 B 的邻域容差联合熵记为:

$$NTE(D, B) = \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_D(x_i) \cap NT_B(x_i)|}{|U|} \right) \quad (7)$$

1.4 集成学习 目前, 一般认为集成学习的研究始于 1990 年, Hansen 和 Salamon 首次提出用神经网络作为基分类器进行集成, 使用该方法可以简单地通过训练多个神经网络将其结果进行结合, 从而对比单个神经网络算法能显著提高学习系统的泛化能力^[23]. 在 Hansen 和 Salamon 之后集成学习得到了广泛的研究.

与采用单个学习器的机器学习方法不同, 集成学习方法通过训练多个基学习器, 并将训练结果结合考虑来解决一个问题. 通常集成学习也称为多分类器系统或基于委员会的学习. 一个集成系统由多个基学习器构成, 而基学习器由基学习算法在训练数据上训练获得, 如神经网络、决策树、朴素贝叶斯或其他学习算法. 虽然传统基分类器的种类繁多, 但其分类精度有待提高且容易出现过拟合等. 故集成学习方法很受关注. 通常, 集成学习具有比基学习器更高的预测准确率及更强的泛化能力^[6]. 图 1 表示一个通用的集成学习框架.

2 基于邻域容差互信息的选择集成分类算法

2.1 问题提出 互信息可以衡量两个离散变量之间拥有相同信息量的差异程度, 同样可以度量离散属性集 X 与离散属性集 Y 之间的相关程度.

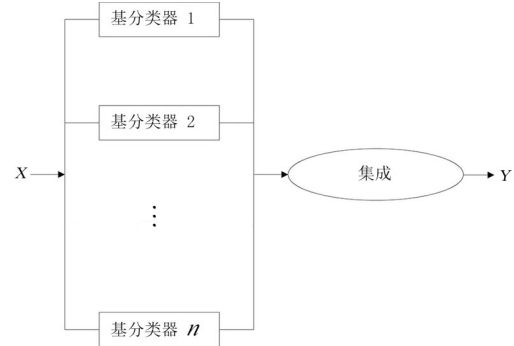


图 1 一个通用的集成学习框架

Fig. 1 A general purpose integrated learning framework

与条件熵不同的是, 属性集 X 与属性集 Y 的互信息越大表明其相关性越大, 反之, 属性集 X 与属性集 Y 的互信息越小表明其相关性越小. 对于既含有离散型属性又含有连续型属性的不完备混合型信息系统, 引入邻域容差关系, 将邻域容差关系和互信息结合, 定义邻域容差互信息的概念来衡量不同属性或属性组合与类别属性之间拥有共同信息量的多少, 为使最终的加权投票结果更加合理, 改进基分类器的预测权重, 提出基于邻域容差互信息的选择集成分类算法.

2.2 算法思想 类比邻域容差条件熵的相关理论, 给出邻域容差互信息的相关定义.

定义 6 (邻域容差互信息) 给定不完备混合型信息系统 $S = (U, A)$, $A = C \cup D$, $B_1, B_2 \subseteq C$, $U/NT_{B_1} = \{NT_{B_1}(x_1), NT_{B_1}(x_2), \dots, NT_{B_1}(x_{|U|})\}$, $U/NT_{B_2} = \{NT_{B_2}(x_1), NT_{B_2}(x_2), \dots, NT_{B_2}(x_{|U|})\}$. B_2 到 B_1 的邻域容差互信息记为:

$$NTI(B_2; B_1) = NTE(B_1) + NTE(B_2) - NTE(B_2, B_1) = \frac{1}{|U|} \left(1 - \frac{|NT_{B_1}(x_i)|}{|U|} - \frac{|NT_{B_2}(x_i)|}{|U|} + \frac{|NT_{B_1}(x_i) \cap NT_{B_2}(x_i)|}{|U|} \right) \quad (8)$$

证明 根据定义 4 得:

$$NTE(B_1) = \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_{B_1}(x_i)|}{|U|} \right)$$

$$NTE(B_2) = \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_{B_2}(x_i)|}{|U|} \right)$$

$$NTE(B_2, B_1) = \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_{B_2}(x_i) \cap NT_{B_1}(x_i)|}{|U|} \right)$$

所以,

$$\begin{aligned} NTI(B_2; B_1) &= NTE(B_1) + NTE(B_2) - NTE(B_2, B_1) = \\ &= \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_{B_1}(x_i)|}{|U|} \right) + \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_{B_2}(x_i)|}{|U|} \right) - \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_{B_2}(x_i) \cap NT_{B_1}(x_i)|}{|U|} \right) = \\ &= \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_{B_1}(x_i)|}{|U|} - \frac{|NT_{B_2}(x_i)|}{|U|} + \frac{|NT_{B_2}(x_i) \cap NT_{B_1}(x_i)|}{|U|} \right) \end{aligned}$$

定义 7 (邻域容差互信息) 设 $U/NT_D = \{NT_D(x_1), NT_D(x_2), \dots, NT_D(x_{|U|})\}$, 对于任意 $B \subseteq C, D$ 和 B 的邻域容差互信息记为:

$$NTI(D; B) = NTE(D) + NTE(B) - NTE(D, B) = \frac{1}{|U|} \left(1 - \frac{|NT_D(x_i)|}{|U|} - \frac{|NT_B(x_i)|}{|U|} + \frac{|NT_D(x_i) \cap NT_B(x_i)|}{|U|} \right) \quad (9)$$

证明 根据定义 4 得: $NTE(D) = \frac{1}{|U|} \left(1 - \frac{|NT_D(x_i)|}{|U|} \right)$

$$NTE(B) = \frac{1}{|U|} \left(1 - \frac{|NT_B(x_i)|}{|U|} \right), NTE(D, B) = \frac{1}{|U|} \left(1 - \frac{|NT_D(x_i) \cap NT_B(x_i)|}{|U|} \right)$$

所以,

$$\begin{aligned} NTI(D; B) &= NTE(D) + NTE(B) - NTE(D, B) = \\ &= \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_D(x_i)|}{|U|} \right) + \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_B(x_i)|}{|U|} \right) - \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_D(x_i) \cap NT_B(x_i)|}{|U|} \right) = \\ &= \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_D(x_i)|}{|U|} - \frac{|NT_B(x_i)|}{|U|} + \frac{|NT_D(x_i) \cap NT_B(x_i)|}{|U|} \right) \end{aligned}$$

如果 $\forall x_i \in U, NT_{B_1}(x_i) \subseteq NT_{B_2}(x_i)$, 则 $NTI(B_2|B_1) = 1$. 如果 $\forall x_i \in U, NT_{B_1}(x_i) = U$ 且 $NT_{B_1}(x_i) = \{x_i\}$, 则 $NTI(B_2|B_1) = 1 - \frac{1}{|U|}$. 因此, $1 - \frac{1}{|U|} \leq NTI(B_2|B_1) \leq 1$.

性质 1 设 $S = (U, A)$ 为不完备混合型信息系统, $A = C \cup D$, 对于任意的 $B_1, B_2 \subseteq C$, 如果 $B_1 \subseteq B_2$, 则 $NTI(D; B_1) \leq NTI(D; B_2)$.

证明 若 $B_1 \subseteq B_2$, 则 $NT_{B_2}(x_i) \subseteq NT_{B_1}(x_i)$. 那么,

$$\begin{aligned} NTI(D; B_1) - NTI(D; B_2) &= \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_D(x_i)|}{|U|} - \frac{|NT_{B_1}(x_i)|}{|U|} + \frac{|NT_D(x_i) \cap NT_{B_1}(x_i)|}{|U|} \right) - \\ &= \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_D(x_i)|}{|U|} - \frac{|NT_{B_2}(x_i)|}{|U|} + \frac{|NT_D(x_i) \cap NT_{B_2}(x_i)|}{|U|} \right) = \\ &= \frac{1}{|U|} \sum_{i=1}^{|U|} \left(1 - \frac{|NT_D(x_i)|}{|U|} - \frac{|NT_{B_1}(x_i)|}{|U|} + \frac{|NT_D(x_i) \cap NT_{B_1}(x_i)|}{|U|} - 1 + \right. \\ &\quad \left. \frac{|NT_D(x_i)|}{|U|} + \frac{|NT_{B_2}(x_i)|}{|U|} - \frac{|NT_D(x_i) \cap NT_{B_2}(x_i)|}{|U|} \right) = \\ &= \frac{1}{|U|} \sum_{i=1}^{|U|} \left(-\frac{|NT_{B_1}(x_i)|}{|U|} + \frac{|NT_D(x_i) \cap NT_{B_1}(x_i)|}{|U|} + \frac{|NT_{B_2}(x_i)|}{|U|} - \frac{|NT_D(x_i) \cap NT_{B_2}(x_i)|}{|U|} \right) \leq 0 \end{aligned}$$

则 $NTI(D; B_1) \leq NTI(D; B_2)$.

对于不完备混合型信息系统,邻域容差互信息可以用来衡量两个变量 X 和 Y 之间共同拥有信息量的多少.若变量 X 和变量 Y 之间邻域容差互信息越小,那么变量 X 和变量 Y 所包含的共同信息越少.极端情况下,当变量 X 和变量 Y 之间的邻域容差互信息为0时,则说明这两个变量是独立的,彼此之间没有任何共同信息.若变量 X 和变量 Y 之间的邻域容差互信息越大,那么变量 X 和变量 Y 所包含的共同信息越多,此时变量 X 和变量 Y 关系密切,其中一个变量变化会对另一个变量产生较大影响.

同样,如果类别属性对于条件属性的邻域容差互信息越大,那么它们的相关程度越大,则它们之间一一映射的程度越高.反之,类别属性和条件属性的邻域容差互信息越小,就有理由认为它们之间近似一一映射的程度很低,则说明条件属性和类别属性之间的相关程度较小.特别地,如果条件属性和类别属性之间的邻域容差互信息为0,表明条件属性即便存在,也无法对最终类别的预测提供任何有效信息.

所以针对不完备混合型信息系统的分类问题,对于既含有缺失的离散型属性又含有缺失的连续型属性的样本可以通过引入邻域容差关系进行处理,结合互信息理论,定义邻域容差互信息来衡量条件属性对于类别属性的重要度,再利用粒化思想和集成分类算法提出基于邻域容差互信息的选择集成分类算法.

利用邻域容差互信息衡量缺失属性对决策分类结果的贡献度,在一个完整的数据集上计算缺失属性与类别属性的邻域容差互信息,属性对类别的贡献程度越大,其作为条件的邻域容差互信息越大;对决策分类结果的贡献度越小,作为条件的邻域容差互信息越小.使用邻域容差互信息、信息粒大小和基分类器的预测准确率来衡量由此信息粒构建的分类器的权重,比仅使用属性维数来衡量基分类器预测的权重更加科学,定义的权重公式如下:

$$w_i = \frac{acc_i |Gra_i| 0.5^{NTI_i}}{\sum acc_i |Gra_i| 0.5^{NTI_i}} \quad (10)$$

其中, w_i 为第 i 个基分类器的预测赋予的权值, acc_i 表示第 i 个基分类器的准确率, $|Gra_i|$ 表示第 i

个信息粒的大小, NTI_i 表示第 i 个信息粒的缺失属性集合对应类别属性的邻域容差互信息.

2.3 算法流程 基于邻域容差互信息的选择集成分类训练流程图如图2所示.基于邻域容差互信息的选择集成分类算法具体步骤如下.

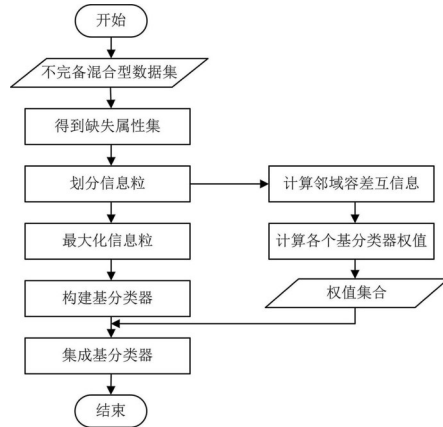


图2 基于邻域容差互信息的选择集成分类训练流程图

Fig. 2 Selective integrated classification training flow chart based on neighborhood tolerance mutual information

步骤1. 根据不完备混合型数据集中的缺失属性对样本进行粒化处理,即把数据集中缺失属性值相同的样本划分到同一信息粒,最终得到若干信息粒.

步骤2. 为了提高预测准确率,充分利用含有缺失属性的数据信息,进行最大化信息粒.首先再次遍历原始数据集,将那些信息粒不含有缺失属性集以及含有缺失属性集的信息粒的属性集合包含在某个信息粒的属性集合中时,把此类信息粒中包含的样本的缺失属性集设置为该信息粒的缺失属性集,形成最大化信息粒.

步骤3. 首先根据定义划分邻域容差类,根据式(5)计算邻域容差信息熵,根据式(7)计算邻域容差联合熵,最后以缺失属性包括连续属性和离散属性作为已知条件,在完整的信息粒上根据式(10)计算基于类别属性的邻域容差互信息.

步骤4. 在各个最大化信息粒上,以非缺失属性作为输入,以BP算法为基分类器的集成分类算法进行集成学习,得到若干个分类预测模型.

步骤5. 使用各个信息粒缺失属性相应的邻域容差互信息、信息粒的大小和子分类器的精度,

根据式(10)计算各个子分类器的权值.

步骤 6. 进行预测. 假设预测数据集中样本的缺失属性集是某个信息粒的缺失属性集的子集, 则可将该样本与这些信息粒相对应的属性集作为对应的子分类器的输入, 经过训练后得出该样本在这些子分类器上的预测类别, 然后再根据这些基分类器的分析结果, 按照步骤 5 的权值公式进行加权集成, 得到最终预测结果.

例 给出不完备混合型数据集, 如表 1 所示, 设置阈值 δ 为 0.5 (阈值过大或过小都会影响粒度的划分. 若阈值过大, 划分的粒会很粗; 若阈值过小, 会导致划分的粒度较细, 失去划分粒层的意义, 加大计算难度. 合理阈值设置为 0.4~0.6, 所以选择 0.5 作为阈值). 计算条件属性集与类别属性的邻域容差互信息.

表 1 不完备混合型数据集

Table 1 Incomplete mixed data set

样本	属性 a_1	属性 a_2	属性 a_3	属性 a_4	属性 Y
x_1	0.15	1	1	0.2	1
x_2	0.7	0	0	*	1
x_3	0.2	*	*	0.5	1
x_4	0.3	0	0	0.7	2
x_5	0.8	0	0	0.8	0
x_6	0.85	0	*	*	0

根据表中不完备混合数据集的定义以及缺失模式的定义, 样本 x_1, x_4, x_5 无缺失值, 样本 x_2 的缺失属性为 $\{a_4\}$, x_3 的缺失属性为 $\{a_2, a_3\}$, x_6 的缺失属性为 $\{a_3, a_4\}$.

按照缺失属性进行划分, 则 $Granule = \{\{x_1, x_4, x_5\}, \{x_2\}, \{x_3\}, \{x_6\}\}$. 不含缺失属性, 则 $X_1 = \{x_1, x_4, x_5\}$. 缺失属性 $\{a_4\}$, 把不含缺失属性的样本去掉属性 a_4 , 则 $X_2 = \{x_1, x_2, x_4, x_5\}$. 缺失属性 $\{a_2, a_3\}$, 把不含缺失属性的样本去掉属性 a_2, a_3 , 则 $X_3 = \{x_1, x_3, x_4, x_5\}$. 若缺失属性 $\{a_3, a_4\}$, 把不含缺失属性的样本去掉属性 a_3, a_4 , 把缺失属性 $\{a_4\}$ 的样本去掉属性 a_3 , 则 $X_4 = \{x_1, x_2, x_4, x_5, x_6\}$.

首先根据决策属性 D 划分邻域容差类:

$$\begin{cases} NT_D(x_1) = NT_D(x_2) = NT_D(x_3) = \{x_1, x_2, x_3\} \\ NT_D(x_4) = \{x_4\} \\ NT_D(x_5) = NT_D(x_6) = \{x_5, x_6\} \end{cases}$$

$$U/D = \{\{x_1, x_2, x_3\}, \{x_4\}, \{x_5, x_6\}\}$$

根据所有条件属性 $C = \{a_1, a_2, a_3, a_4\}$ 划分邻

域容差类:

$$\begin{cases} NT_C(x_1) = \{x_1\} \\ NT_C(x_2) = \{x_2, x_5, x_6\} \\ NT_C(x_3) = \{x_3\} \\ NT_C(x_4) = \{x_4\} \\ NT_C(x_5) = \{x_2, x_5, x_6\} \\ NT_C(x_6) = \{x_2, x_5, x_6\} \end{cases}$$

$$U/C = \{\{x_1\}, \{x_3\}, \{x_4\}, \{x_2, x_5, x_6\}\}$$

根据条件属性 a_1 划分邻域容差类:

$$\begin{cases} NT_{a_1}(x_1) = NT_{a_1}(x_3) = NT_{a_1}(x_4) = \{x_1, x_3, x_4\} \\ NT_{a_1}(x_2) = NT_{a_1}(x_5) = NT_{a_1}(x_6) = \{x_2, x_5, x_6\} \end{cases}$$

$$U/a_1 = \{\{x_1, x_3, x_4\}, \{x_2, x_5, x_6\}\}$$

根据条件属性 a_2 划分邻域容差类:

$$\begin{cases} NT_{a_2}(x_1) = \{x_1, x_3\} \\ NT_{a_2}(x_2) = \{x_2, x_3, x_4, x_5, x_6\} \\ NT_{a_2}(x_3) = U \\ NT_{a_2}(x_4) = \{x_2, x_3, x_4, x_5, x_6\} \\ NT_{a_2}(x_5) = \{x_2, x_3, x_4, x_5, x_6\} \\ NT_{a_2}(x_6) = \{x_2, x_3, x_4, x_5, x_6\} \end{cases}$$

$$U/a_2 = \{\{x_1, x_3\}, \{x_2, x_3, x_4, x_5, x_6\}, U\}$$

根据条件属性 a_3 划分邻域容差类:

$$\begin{cases} NT_{a_3}(x_1) = \{x_1, x_3, x_6\} \\ NT_{a_3}(x_2) = \{x_2, x_3, x_4, x_5, x_6\} \\ NT_{a_3}(x_3) = U \\ NT_{a_3}(x_4) = \{x_2, x_3, x_4, x_5, x_6\} \\ NT_{a_3}(x_5) = \{x_2, x_3, x_4, x_5, x_6\} \\ NT_{a_3}(x_6) = U \end{cases}$$

$$U/a_3 = \{\{x_1, x_3, x_6\}, \{x_2, x_3, x_4, x_5, x_6\}, U\}$$

根据条件属性 a_4 划分邻域容差类:

$$\begin{cases} NT_{a_4}(x_1) = \{x_1, x_2, x_6\} \\ NT_{a_4}(x_2) = U \\ NT_{a_4}(x_3) = \{x_2, x_3, x_6\} \\ NT_{a_4}(x_4) = \{x_2, x_4, x_5, x_6\} \\ NT_{a_4}(x_5) = \{x_2, x_4, x_5, x_6\} \\ NT_{a_4}(x_6) = U \end{cases}$$

$$U/a_4 = \{\{x_1, x_2, x_6\}, \{x_2, x_3, x_6\}, \{x_2, x_4, x_5, x_6\}, U\}$$

根据条件属性 a_1 和条件属性 a_2 划分邻域容

差类:

$$\begin{cases} NT_{a_1 \cup a_2}(x_1) = \{x_1, x_3\} \\ NT_{a_1 \cup a_2}(x_2) = \{x_2, x_5, x_6\} \\ NT_{a_1 \cup a_2}(x_3) = \{x_1, x_3, x_4\} \\ NT_{a_1 \cup a_2}(x_4) = \{x_3, x_4\} \\ NT_{a_1 \cup a_2}(x_5) = NT_{a_1 \cup a_2}(x_6) = \{x_2, x_5, x_6\} \end{cases}$$

$$U/a_1 \cup a_2 = \{\{x_1, x_3\}, \{x_2, x_5, x_6\}, \{x_1, x_3, x_4\}, \{x_3, x_4\}\}$$

根据条件属性 a_1 和条件属性 a_3 划分邻域容

差类:

$$\begin{cases} NT_{a_1 \cup a_3}(x_1) = \{x_1, x_3\} \\ NT_{a_1 \cup a_3}(x_2) = \{x_2, x_5, x_6\} \\ NT_{a_1 \cup a_3}(x_3) = \{x_1, x_3, x_4\} \\ NT_{a_1 \cup a_3}(x_4) = \{x_3, x_4\} \\ NT_{a_1 \cup a_3}(x_5) = \{x_2, x_5, x_6\} \\ NT_{a_1 \cup a_3}(x_6) = \{x_2, x_5, x_6\} \end{cases}$$

$$U/a_1 \cup a_3 = \{\{x_1, x_3\}, \{x_2, x_5, x_6\}, \{x_1, x_3, x_4\}, \{x_3, x_4\}\}$$

根据条件属性 a_1 和条件属性 a_4 划分邻域容

差类:

$$\begin{cases} NT_{a_1 \cup a_4}(x_1) = \{x_1\} \\ NT_{a_1 \cup a_4}(x_2) = \{x_2, x_5, x_6\} \\ NT_{a_1 \cup a_4}(x_3) = \{x_3\} \\ NT_{a_1 \cup a_4}(x_4) = \{x_4\} \\ NT_{a_1 \cup a_4}(x_5) = NT_{a_1 \cup a_4}(x_6) = \{x_2, x_5, x_6\} \end{cases}$$

$$U/a_1 \cup a_4 = \{\{x_1\}, \{x_3\}, \{x_4\}, \{x_2, x_5, x_6\}\}$$

根据式(9)计算单个属性的邻域容差互信息:

$$\begin{aligned} NTI(D|a_1) &= \frac{1}{6} \left(\frac{3-2}{6} + \frac{3-1}{6} + \frac{3-2}{6} + \right. \\ &\quad \left. \frac{3-1}{6} + \frac{3-2}{6} + \frac{3-2}{6} \right) = \frac{28}{36} \\ NTI(D|a_2) &= \frac{1}{6} \left(\frac{2-2}{6} + \frac{5-2}{6} + \frac{6-3}{6} + \right. \\ &\quad \left. \frac{5-1}{6} + \frac{5-2}{6} + \frac{5-2}{6} \right) = \frac{20}{36} \\ NTI(D|a_3) &= \frac{1}{6} \left(\frac{3-2}{6} + \frac{5-2}{6} + \frac{6-3}{6} + \right. \\ &\quad \left. \frac{5-1}{6} + \frac{5-2}{6} + \frac{6-2}{6} \right) = \frac{18}{36} \\ NTI(D|a_4) &= \frac{1}{6} \left(\frac{3-2}{6} + \frac{6-3}{6} + \frac{3-2}{6} + \right. \\ &\quad \left. \frac{4-1}{6} + \frac{4-2}{6} + \frac{6-2}{6} \right) = \frac{22}{36} \end{aligned}$$

根据式(9)计算属性集合的邻域容差互信息:

$$\begin{aligned} NTI(D|a_1 \cup a_2) &= \frac{1}{6} \left(\frac{2-2}{6} + \frac{3-1}{6} + \frac{3-2}{6} + \right. \\ &\quad \left. \frac{2-1}{6} + \frac{3-2}{6} + \frac{3-2}{6} \right) = \frac{30}{36} \\ NTI(D|a_1 \cup a_3) &= \frac{1}{6} \left(\frac{2-2}{6} + \frac{3-1}{6} + \frac{3-2}{6} + \right. \\ &\quad \left. \frac{2-1}{6} + \frac{3-2}{6} + \frac{3-2}{6} \right) = \frac{30}{36} \\ NTI(D|a_1 \cup a_4) &= \frac{1}{6} \left(\frac{1-1}{6} + \frac{3-1}{6} + \frac{1-1}{6} + \right. \\ &\quad \left. \frac{1-1}{6} + \frac{3-2}{6} + \frac{3-2}{6} \right) = \frac{32}{36} \\ NTI(D|a_1 \cup a_2 \cup a_4) &= \frac{32}{36} \\ NTI(D|a_1 \cup a_3 \cup a_4) &= \frac{32}{36} \end{aligned}$$

3 仿真实验与性能分析

基于Python实现算法仿真,验证所提算法的可行性和有效性.实验设备的基本信息如下.系统环境:CPU Intel i7-10750H;RAM:18 GB;操作系统:Windows 10 专业版;解释器:Python 3.7.10.在3.1中简要介绍实验使用的数据集和对比方法的基本信息.实验结果在3.2中给出并详细分析实验结果.3.3给出非参数统计检验的实验结果,验证与对比算法之间的统计学差异.

3.1 实验设置 从UCI数据库和爱数科公共数据库中选取七个公开获取的数据集进行实验验证,表2为实验数据集的详细信息,其中有一个数据集只包含离散属性,一个数据集只包含连续型数据,三个数据集具有混合型属性.此外,对于数据集E-commerce transportation和Shill Bidding,真实获取的数据集为完备数据集,但为了验证算法的可行性,通过从原始数据集中随机选择5%和10%的已知样本特征值转变为缺失值,形成四个人工不完备数据集.

为了降低由于数据划分带来的随机性,采用十折交叉验证的均值作为模型的最终得分.在所提方法中,采用BP神经网络作为集成分类算法的基分类器,学习率为0.1,隐藏层神经元个数为15,迭代训练次数为15.本实验使用的对比算法包括:极限梯度提升机(XGBoost)、随机森林

(RF)、梯度提升树(GBDT)、自适应 Boosting (AdaBoost)和 Stacking. 需要注意的是,实验使用的分类器均采用 Scikit-learn 学习库的默认参数进行实验.

3.2 实验结果与分析

表 3 至表 6 给出部分数据集通过实验得到的以各个信息粒缺失属性作为条件关于类别属性的邻域容差互信息,其中信息粒缺失属性按照缺失属性从少到多表示. 当所有属性都为离散型属性时,邻域容差关系即为容差关系,例如 Mushroom 数据集,不含有连续型属性,此时阈值设为 0,其缺失属性为一个.

由表 3 至表 6 可以看出,对于同一数据集信息粒中不同的缺失属性集合作为条件的类别属性的邻域容差互信息是不同的,数据集中缺失属性集合包含元素的数量与计算类别属性的邻域容差互信息是无关的. 对于没有缺失属性的数据集,认为其丢失了一个与类属性无关的条件属性,则邻域容差互信息为 0. 若以信息粒缺失属性集合为条件的类别属性的邻域容差互信息较大,说明此缺失属性集合对决策类别的贡献率较大以及携带的信息量较大,对最终的决策类别较为重要. 对于表 3 的 Housing loan 数据集,第二个属性比第七个属性邻域容差互信息大,则认为第七个属性对类属性的影响更大,那么对于缺失第二个属性的预测结果不如缺失第七个属性的预测结果可信度高. 根据实验过程分析对于基分类器的预测准确率与信息粒包含样本的多少是高度相关的,所以预测准确率出现很高或很低的情况. 因此,在定

义基分类器的权重时,充分考虑其邻域容差条件熵,基分类器准确率以及信息粒的大小会更加合情合理,最终加权集成的分类器预测更加准确,构建的集成分类算法也更具有普适性.

对于处理不完备混合型数据的集成分类算法,最为典型的是 XGBoost 算法,可以直接处理不完备数据集. 由表 7 的实验结果可以看出,对于不完备混合型数据集的分类问题使用邻域容差互信息选择集成分类算法得到的分类结果的准确率普遍要高于传统的 XGBoost 算法的准确率,其中对于 Housing loan 数据集,用定义的权重公式(10)预测准确率比 XGBoost 算法高 6.1666%,但对于 Adult 数据集,本节提出的算法预测准确率要高于 XGBoost 算法,由于 Adult 数据集缺失属性较少,用插补法处理后使用随机森林、GBDT、AdaBoost、Stacking 算法预测准确率要高一些. 对于其他数据集,本节提出的算法比传统集成分类算法预测准确率也高,例如,对于 Housing loan 数据集,准确率比 GBDT 高 9.9232%,对于 Credit 数据集,准确率比 AdaBoost 高 6.0379% 等. 所以本节提出的基于邻域容差互信息的选择集成分类算法对于解决不完备混合型数据集的分类问题提供了新的思路,在公开的不完备混合数据集上的实验结果证实了本节所提算法的有效性和可行性.

3.3 非参数统计检验 为了进一步验证提出的 NTMISECA 方法与其他对比方法之间的性能差异,使用 Friedman 排名和 Holm's 事后检验方法,在所有实验数据集上,验证模型之间的统计学差异,结果如表 8 所示.

表 2 数据集的详细信息

Table 2 Details of the data set

数据集	样本数	类别数	属性数	连续属性数	离散属性数	属性值缺失率
Housing loan	614	2	13	5	8	1.81%
Adult	32561	2	14	6	8	0.87%
Credit	690	2	15	6	9	0.61%
Mushroom	8124	2	22	0	22	1.33%
Water quality	3276	2	9	9	0	4.38%
E-commerce transportation	10999	2	10	6	4	0
Shill Bidding	6321	2	12	11	1	0

表3 Housing loan数据集缺失属性的邻域容差互信息

Table 3 Neighborhood tolerance mutual information for missing attributes of Housing loan data set

信息粒缺失属性	邻域容差互信息	基分类器预测权重
a_1	0.1253	0.1451
a_3	0.2583	0.1315
a_5	0.1143	0.1527
a_8	0.1088	0.1460
a_9	0.1233	0.1480
a_{10}	0.1699	0.1518
a_5, a_{10}	0.2316	0.1249

表4 Adult数据集缺失属性的邻域容差互信息

Table 4 Neighborhood tolerance mutual information for missing attributes of Adult data set

信息粒缺失属性	邻域容差互信息	基分类器预测权重
a_6	0.3327	0.1945
a_{13}	0.0895	0.2712
a_1, a_6	0.2157	0.3451
a_1, a_6, a_{13}	0.3661	0.1891

表5 Credit数据集缺失属性的邻域容差互信息

Table 5 Neighborhood tolerance mutual information for missing attributes of Credit data set

信息粒缺失属性	邻域容差互信息	基分类器预测权重
a_1	0.2145	0.1639
a_2	0.2859	0.1506
a_{14}	0.2756	0.1502
a_1, a_{14}	0.3940	0.1414
a_6, a_7	0.4671	0.1327
a_1, a_6, a_7	0.4913	0.1296
$a_4, a_5, a_6, a_7, a_{14}$	0.4822	0.1317

表6 Water quality数据集缺失属性的邻域容差互信息

Table 6 Neighborhood tolerance mutual information for missing attributes of Water quality data set

信息粒缺失属性	邻域容差互信息	基分类器预测权重
a_1	0.1330	0.1643
a_5	0.0996	0.0778
a_8	0.0072	0.0262
a_1, a_5	0.1465	0.1534
a_1, a_8	0.1678	0.2756
a_5, a_8	0.1556	0.1762
a_1, a_5, a_8	0.2365	0.1256

根据表8的非参数统计结果,可以发现提出的方法的Friedman排名明显优于其他分类方法。但是,根据常用的显著性差异度量标准($p < 0.05$),提出的方法与使用的对比方法不存在显著的统计学差异,即提出的方法在所有数据集上的性能与对比方法是相近的,差异并不明显。

在实验使用的对比集成学习方法中,均使用决策树作为模型的基分类器,而在提出的NTMI-SECA方法中,使用神经网络作为基分类器。神经网络是一种适用性很强的分类方法,适用于大部分数据集,但是,神经网络对模型的参数很敏感,可以使用经典的参数优化方法或搜索方法选择最优的参数。此外,在基分类器的参数调整方面与使用决策树的模型仍然存在差距。然而,提出的NTMISECA是一种基于粒计算的集成学习框架,其基分类器可以根据实际需要调整。因此,也可以使用单一弱分类器或弱集成学习方法,从而有效地提高提出的方法的分类性能和泛化能力。

4 结论

本章根据粒计算的基本思想,利用集成学习的优势,将邻域容差理论和互信息理论结合,提出一种解决不完备混合型信息系统的分类问题的集成算法,即基于邻域容差互信息的选择集成分类算法。利用传统集成算法对不完备数据集进行分类在权衡各个基分类器的权重时仅考虑数据的维度和属性的多少是不够科学的,不同的属性对类别的贡献程度也是不一样的,所以提出邻域容差互信息的概念来衡量。然后根据粒计算的思想按照缺失属性将数据集划分为不同的信息粒,为充分利用数据信息,将信息粒最大化,并用集成算法训练出基分类器,利用信息粒的大小、邻域容差互信息和基分类器预测准确率来定义基分类器的权重,再次实现加权集成投票。实验表明该算法普遍比传统的集成分类算法预测准确率高。

本文所选用数据集全部为静态数据集,对于动态不完备混合型数据集如何设计集成分类算法,并且对于集成学习算法训练时间会比较长,如

表 7 不同分类器预测不同数据集准确率的对比

Table 7 The accuracy comparison of different classifiers predicting different datasets

数据集	属性值缺失率	NTMISECA	XGBoost	RF	GBDT	AdaBoost	Stacking
Housing loan	1.81%	81.6265%	75.4599%	76.5934%	71.7033%	77.4725%	75.8791%
Adult	0.87%	83.8565%	83.3838%	85.7027%	86.2422%	85.8737%	86.4469%
Credit	0.61%	87.5300%	87.6712%	85.0481%	82.7885%	82.2115%	84.2788%
Mushroom	1.33%	100%	99.7538%	100%	100%	100%	100%
Water quality	4.38%	68.8459%	65.3103%	65.9207%	66.4293%	63.0722%	68.4639%
E-commerce transportation	5%	65.6758%	63.3758%	64.8694%	65.1122%	64.5684%	62.7886%
	10%	65.9581%	63.1212%	64.3565%	64.6657%	63.6746%	59.9897%
Shill Bidding	5%	94.8156%	94.3115%	93.2627%	93.2212%	92.1678%	94.1223%
	10%	93.2296%	90.1354%	90.2319%	92.2376%	91.1324%	93.0034%

表 8 所有分类器的 Friedman 排名和事后检验结果

Table 8 Friedman rankings and postmortem results for all classifiers

分类器	Friedman 排名	未调整 p 值	调整后的 p 值
NTMISECA	2.2222	—	—
Stacking	3.3333	0.207712	0.261440
RF	3.5556	0.130570	0.261440
GBDT	3.7778	0.077760	0.233280
XGBoost	4.0000	0.043820	0.175279
AdaBoost	4.1111	0.032210	0.161048

何进一步减少预测时间,提升预测效率也是一个值得研究的问题。

参考文献

- [1] 邓建新,单路宝,贺德强,等. 缺失数据的处理方法及其发展趋势. 统计与决策, 2019, 35(23): 8—34. (Deng J X, Shan L B, He D Q, et al. Processing method of missing data and its developing tendency. Statistics and Decision, 2019, 35(23): 28—34.)
- [2] Tran C T, Zhang M J, Andrae P, et al. An effective and efficient approach to classification with incomplete data. Knowledge-Based Systems, 2018, 154: 1—16.
- [3] 张利亭,冯涛,李欢. 不完备信息系统的直觉模糊决策粗糙集. 郑州大学学报(理学版), 2021, 53(2): 57—65. (Zhang L T, Feng T, Li H. Intuitionistic fuzzy decision rough sets for incomplete information systems. Journal of Zhengzhou University (Natural Science Edition), 2021, 53(2): 57—65.)
- [4] 杨美丽. 基于相容关系的不完整数据集分类方法研究. 硕士学位论文. 合肥: 安徽大学, 2021. (Yang

M L. Incomplete data ensemble classification based on tolerance relationship. Master Dissertation. Hefei: Anhui University, 2021.)

- [5] 刘海峰,续欣莹,申雪芬,等. 基于限制邻域关系的不完备混合决策系统属性约简. 广西师范大学学报(自然科学版), 2013, 31(3): 30—36. (Liu H F, Xu X Y, Shen X F, et al. Attribute reduction of incomplete mixed decision system based on limited neighborhood relation. Journal of Guangxi Normal University (Natural Science Edition), 2013, 31(3): 30—36.)
- [6] Zhao H, Qin K Y. Mixed feature selection in incomplete decision table. Knowledge - Based Systems, 2014, 57: 181—190.
- [7] 梁吉业,钱宇华,李德玉,等. 大数据挖掘的粒计算理论与方法. 中国科学: 信息科学, 2015, 45(11): 1355—1369. (Liang J Y, Qian Y H, Li D Y, et al. Theory and method of granular computing for big data mining. Science in China (Information Sciences), 2015, 45(11): 1355—1369.)
- [8] Krause S, Polikar R. An ensemble of classifiers approach for the missing feature problem//Proceedings of the International Joint

- Conference on Neural Networks, 2003. Portland, OR, USA: IEEE, 2003: 553–558.
- [9] 吕靖, 舒礼莲. 基于 AdaBoost 的不完整数据的信息熵分类算法. 计算机与现代化, 2013(9): 31–34. (Lü J, Shu L L. Incomplete data information entropy classification algorithm based on AdaBoost. Computer and Modernization, 2013, 9: 31–34.)
- [10] Chen H X, Du Y P, Jiang K. Classification of incomplete data using classifier ensembles//2012 International Conference on Systems and Informatics (ICSAI2012). Yantai, China: IEEE, 2012: 2229–2232.
- [11] Yan Y T, Zhang Y P, Zhang Y W. Multi-granulation ensemble classification for incomplete data//9th International Conference on Rough Sets and Knowledge Technology. Springer Berlin Heidelberg, 2014: 343–351.
- [12] Zhang T, Dai Q, Ma Z C. Extreme learning machines' ensemble selection with GRASP. Applied Intelligence, 2015, 43(2): 439–459.
- [13] Ma Z C, Dai Q, Liu N Z. Several novel evaluation measures for rank-based ensemble pruning with applications to time series prediction. Expert Systems with Applications, 2015, 42(1): 280–292.
- [14] Chen T Q, He T, Benesty M, et al. Xgboost: Extreme gradient boosting, 2015, 1(4): 1–4.
- [15] Yan Y T, Zhang Y P, Zhang Y W, et al. A selective neural network ensemble classification for incomplete data. International Journal of Machine Learning and Cybernetics, 2017, 8(5): 1513–1524.
- [16] 彭莉, 张海清, 李代伟, 等. 基于粗糙集理论的不完备数据分析方法的混合信息系统填补算法. 计算机应用, 2021, 41(3): 677–685. (Peng L, Zhang H Q, Li D W, et al. Imputation algorithm for hybrid information system of incomplete data analysis approach based on rough set theory. Journal of Computer Applications, 2021, 41(3): 677–685.)
- [17] 李金海, 王飞, 吴伟志, 等. 基于粒计算的多粒度数据分析方法综述. 数据采集与处理, 2021, 36(3): 418–435. (Li J H, Wang F, Wu W Z, et al. Review of multi-granularity data analysis methods based on granular computing. Journal of Data Acquisition and Processing, 2021, 36(3): 418–435.)
- [18] 李明, 甘秀娜, 王月波. 基于集成学习的决策粗糙集特定类属性约简算法. 计算机应用与软件, 2021, 38(6): 262–270. (Li M, Gan X N, Wang Y B. Class-specific attribute reduction algorithm for decision-theoretic rough sets based on ensemble learning. Computer Applications and Software, 2021, 38(6): 262–270.)
- [19] 杨小平. 粗集中最大相似度的不完备数据补齐. 计算机工程与应用, 2012, 48(36): 164–166. (Yang X P. Completing incomplete data based on maximum similarity in rough sets. Computer Engineering and Applications, 2012, 48(36): 164–166.)
- [20] 姚晟, 陈菊, 吴照玉. 一种基于邻域容差信息熵的组合度量方法. 小型微型计算机系统, 2020, 41(1): 46–50. (Yao S, Chen J, Wu Z Y. Combination measurement method based on neighborhood tolerance information entropy. Journal of Chinese Computer Systems, 2020, 41(1): 46–50.)
- [21] 刘丹, 徐立新, 李敬伟. 不完备邻域多粒度决策理论粗糙集与三支决策. 计算机应用与软件, 2019, 36(5): 145–157. (Liu D, Xu L X, Li J W. Incomplete neighborhood multi-granulation decision-theoretic rough set and three-way decision. Computer Applications and Software, 2019, 36(5): 145–157.)
- [22] 滕书华, 鲁敏, 杨阿锋, 等. 基于一般二元关系的粗糙集加权不确定性度量. 计算机学报, 2014, 37(3): 649–665. (Teng S H, Lu M, Yang A F, et al. A weighted uncertainty measure of rough sets based on general binary relation. Chinese Journal of Computers, 2014, 37(3): 649–665.)
- [23] Hu Q H, Yu D R, Liu J F, et al. Neighborhood rough set based heterogeneous feature subset selection. Information Sciences, 2008, 178(18): 3577–3594.
- [24] He Q, Xie Z X, Hu Q H, et al. Neighborhood based sample and feature selection for SVM classification learning. Neurocomputing, 2011, 74(10): 1585–1594.
- [25] Shannon C E. A mathematical theory of communication. The Bell System Technical Journal, 1948, 27(3): 379–423.

(责任编辑 杨可盛 高善露)