智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

不完备数据中面向特征值更新的增量特征选择方法

唐荣，罗川，曹潜，王思朝

引用本文:
唐荣, 罗川, 曹潜, 等. 不完备数据中面向特征值更新的增量特征选择方法[J]. 智能系统学报, 2021, 16(3): 493–501.
TANG Rong, LUO Chuan, CAO Qian, et al. Incremental approach for feature selection in incomplete data while updating feature values[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(3): 493–501.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202006045

DOI: 10.11992/tis.202006045

不完备数据中面向特征值更新的增量特征选择方法

唐荣, 罗川, 曹潜, 王思朝
(四川大学 计算机学院, 四川 成都 610065)

摘 要: 实际应用中, 数据常常表现出不完备性和动态性的特点。针对动态不完备数据中的特征选择问题, 提出了一种基于相容粗糙集模型和信息熵理论的增量式特征选择方法。首先, 建立了不完备信息系统中特征值动态更新时论域上条件划分与决策分类的动态更新模式, 分析了作为特征重要度评价准则的不完备相容信息熵的增量计算机制, 并将该机制引入到启发式最优特征子集搜索过程中特征重要度的迭代计算, 进一步设计了不完备数据中面向特征值动态更新的增量式特征选择算法。最后, 在标准 UCI 数据集上从分类精度、决策性能和计算效率 3 个方面对文中所提出的增量算法的有效性和高效性进行了实验验证。

关键词: 特征选择; 维度约简; 粗糙集; 信息熵; 不完备数据; 缺失值; 启发式搜索; 增量学习
中图分类号: TP18 **文献标志码:** A **文章编号:** 1673−4785(2021)03−0493−09

中文引用格式: 唐荣, 罗川, 曹潜, 等. 不完备数据中面向特征值更新的增量特征选择方法 [J]. 智能系统学报, 2021, 16(3): 493–501.
英文引用格式: TANG Rong, LUO Chuan, CAO Qian, et al. Incremental approach for feature selection in incomplete data while updating feature values[J]. CAAI transactions on intelligent systems, 2021, 16(3): 493–501.

Incremental approach for feature selection in incomplete data while updating feature values

TANG Rong, LUO Chuan, CAO Qian, WANG Sizhao
(College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: In practical application, data often exhibits incomplete and dynamic characteristics. For the feature selection problem in dynamic incomplete data, an incremental feature selection method based on the tolerance rough set model and information entropy theory is proposed. First, the update patterns of conditional partition and decision classification are established based on the variation of feature values in incomplete information systems. The incremental computing mechanism of incomplete tolerance information entropy as the evaluation criterion of feature importance is built subsequently. Such an incremental mechanism is integrated into the iterative calculation of feature importance during the heuristic search of optimal feature subset, and an incremental feature selection algorithm for dynamic variation of feature values is developed. Finally, the effectiveness and efficiency of the proposed incremental algorithm are verified on several standard UCI datasets in terms of classification accuracy, decision performance, and computing efficiency.

Keywords: feature selection; dimensional reduction; rough set; information entropy; incomplete data; missing values; heuristic search; incremental learning

特征选择的目标是在给定评价标准下选择非冗余的特征子集, 其作为一项重要的数据预处理步骤, 能够有效地提高数据分析模型的准确性和高效性, 在数据挖掘与知识发现中起着重要的作用[1]。

数据中存在一些缺失值是一种非常普遍的现象, 缺失值给数据中的分类知识带来了不一致性问题[2]。粗糙集理论是一种能够有效应对不精确、不一致信息的数据建模与知识获取工具。近年来, 人们基于粗糙集理论针对不完备数据的特征选择问题进行了深入的分析和讨论。Kryszkiewicz[3]

收稿日期: 2020−06−27.
基金项目: 国家自然科学基金项目 (62076171); 四川省科技厅应用基础研究计划项目 (2019YJ0084).
通信作者: 罗川. E-mall: cluo@scu.edu.cn.

认为不完备数据中缺失值应是已有值域中的某一特征值，进而提出了一种基于广义差别矩阵的特征选择方法。Parthalin 等[4]为了保留分类所产生的不一致决策区域，研究了基于容差粗糙集的特征选择方法。Meng 等[5]讨论了不一致不完备决策系统中基于区分矩阵的特征选择方法。Grzymala-Busse 等[6]将缺失值考虑为丢失值和不在乎值，提出了基于广义特征关系粗糙集模型的特征选择方法。Qian 等[7]提出了一种高效的正向近似加速器，用于加速不完备数据特征选择的启发式特征搜索过程。Dai[8]为了处理不完备数值型数据，建立了一种新的容差模糊粗糙集模型，并提出了基于差别矩阵的特征选择方法。Yang 等[9]定义了多准则决策系统中相似优势关系的概念，提出了 4 种基于差别矩阵的近似分布约简方法。Liang 等[10]提出了一种不完备信息系统中基于粗糙熵的启发式特征选择算法。Qian 等[11]基于不完备信息系统中的最大一致块概念，提出了一种新的组合信息熵用于度量信息系统的不可分辨能力。Dai 等[12]在不完备决策系统中提出一种新的满足单调性约束的条件信息熵。Zhao 等[13]提出了一种新的邻域容差条件熵，并将其应用于混合不完备数据中的特征选择问题。

另一方面，实际应用中数据随时间的推移呈现出动态更新的变化趋势，数据的采集与分析是一个不断优化升级的动态过程。面向动态数据的高效特征选择方法成为了当前人们普遍关注的一个研究热点。增量技术可以利用已有计算结果进行特征选择增量计算，以发现新的特征子集，从而避免重新计算整个特征空间以获取新的特征子集[14-15]。近年来，许多学者通过将增量学习技术引入到特征选择问题中，对动态数据环境下的高效特征选择方法进行了广泛深入的研究。Xu 等[16]将特征选择问题转化为 0-1 整数规划问题，提出了一种对象更新条件下的动态特征选择方法。Qian 等[17]设计了一种新的基于相对不可辨识对象对的属性重要度度量方式，并提出了动态粒度空间下的基于序贯三支决策模型的增量特征选择方法。Yang 等[18]分析了对象动态变化时相对可辨识关系的增量更新机制，提出了基于模糊粗糙集的动态特征选择算法。Lang 等[19]提出了覆盖信息系统中基于相关族的动态特征选择方法。Wei 等[20]设计了基于辨识矩阵和压缩辨识矩阵的增量特征选择算法，以获得数据动态变化时最优的特征子集。Zeng 等[21]基于高斯核模糊粗糙集模型，研究了混合信息系统的动态特征选择方法。Liang 等[22]提出了信息熵的批增量递推计算机制，可用于多个数据集之间信息熵的高效融合。Shu 等[23]针对含有缺失值的不完备数据，提出了基于正域的增量特征选择算法。Xie 等[24]提出了 3 种不完备数据中相容类的更新策略，并设计了相应的增量特征选择算法。考虑到动态数据中特征值存在频繁的修改和更新操作，Wang 等[25]针对完备数据集研究了特征值动态更新时信息熵的增量更新机制，进一步设计了相应的动态特征选择算法。刘吉超等[26]针对不完备数据中数据集维数动态增加的情形，分析了互补信息熵的更新机制，进而提出了一种增量特征约简算法。钱进等[27]提出一种基于正域处理面向成组对象集的增量式特征选择算法。综合上文所述，大部分研究者针对完备决策系统的动态更新特征选择问题进行深入的研究，鲜有对不完备决策系统动态更新特征问题研究。基于正域处理不完备决策系统的特征选择存在无法处理边界域中的样本分类的不确定性问题。信息熵作为度量信息不确定性的度量标准，有助于不完备数据特征选择问题研究，而引入增量计算机制可以加速特征选择过程，有效减少计算时间。本文针对不完备决策系统，设计了一种面向特征值动态更新的特征选择算法。文中首先分析了特征值更新时不完备决策系统中相容类和决策类的动态变化模式，并以此给出了条件信息熵的增量计算机制，进而设计了基于增量条件信息熵的动态特征算法，最后通过实验验证进一步说明了算法的有效性和高效性。

1　基本概念

粗糙集理论中，信息系统表示为一个四元组 $S=(U,A,V,f)$，其中，U 表示对象的非空有限集合，称为论域；A 表示特征的非空有限集合，即特征集；V_a 表示特征 $a\in A$ 的值域，并且有 $V=\cup_{a\in A}V_a$；对任意 $a\in A$ 和 $x\in U$，$f:U\times A\to V$ 是一个信息函数，通过信息函数给每一个对象 $x\in U$ 一个特定的特征值 $f(x,a)\in V_a$，$a\in A$。决策系统表示为 $\mathrm{DS}=(U,C\cup\{d\},V,f)$，其中，$C$ 代表条件特征的非空有限集合；d 表示决策特征。在实际应用中，信息系统中某些对象的特征值容易丢失，如果一个信息系统中 V 包含缺失的特征值，记作“*”，那么该信息系统被称为不完备信息系统 (incomplete information system, IIS)；对于决策系统来说，如果 $*\in V_C,*\notin V_d$，称这样的决策系统为不完备决策系统 (incomplete decision system, IDS)；对于 $*\notin V_C,*\notin V_d$ 这样的决策系统，称为完备决策系统。

完备信息系统中条件特征的任何子集 $P\subseteq C$ 可诱导一种不可辨识关系 $\mathrm{IND}(P)$，定义为

$$\mathrm{IND}(P)=\{(x,y)\in U\times U|\forall a\in P, f(x,a)=f(y,a)\}$$

$\mathrm{IND}(P)$ 是具有自反性、对称性与传递性的等价关系。等价关系 $\mathrm{IND}(P)$ 将论域 U 划分为等价类的集合, 表示为 $U/\mathrm{IND}(P)=\{[x]_P|x\in U\}$, 其中 $[x]_P=\{y|(x,y)\in \mathrm{IND}(P)\}$。为了处理含有缺失值的不完备决策系统, Kryszkiewicz 提出一种新的二元关系 $T(P), P\subseteq C$, 定义为

$$T(P)=\{(x,y)\in U\times U|\forall a\in P, f(x,a)=f(y,a)\vee f(x,a)=*\vee f(y,a)=*\}$$

$T(P)$ 是具有自反性和对称性, 但不具有传递性的相容关系。在 P 下任意一个对象 $x\in U$ 的相容类定义为 $T_P(x)=\{y\in U|(x,y)\in T(P)\}$。$U/T(P)$ 表示相容类集合 $\{T(P)|x\in U\}$。$U/T(P)$ 中构成论域 U 上的一个覆盖, 对于论域中任意一个对象 $x\in U, T_P(x)\neq\varnothing$, 并且 $\cup_{x\in U}T_P(x)=U$。给定一个不完备决策系统 $\mathrm{IDS}=(U,C\cup\{d\},V,f)$, 决策属性 d 将对象分类为 m 个确定互斥的子集 $U/d=\{D_1,D_2,\cdots,D_m\}$。目标决策概念 $D_i\in U/D$ 的上、下近似集定义为

$$\underline{\mathrm{apr}}(D_i)=\{x\in U|T_P(x)\subseteq D_i\}$$
$$\overline{\mathrm{apr}}(D_i)=\{x\in U|T_P(x)\cap D_i\neq\varnothing\}$$

基于粗糙集理论的特征选择方法根据特征重要度的不同度量标准, 可笼统地归纳为依赖性度量、一致性度量、距离度量和信息度量。前面 3 种度量方法都局限于数据的实际值, 对含有噪声或缺失值的数据处理十分敏感。而基于信息论的度量方法仅关注随机变量的概率分布, 不关注其实际值, 成为了高维数据中常用的特征重要度度量方式。借鉴香农熵的传统定义形式, Dai 等[8,12]定义了一种新的满足单调性的条件熵来度量不完备决策系统协调程度的不确定性。给定一个不完备决策系统 $\mathrm{IDS}=(U,C\cup\{d\},V,f)$, 其中, $U=\{x_1,x_2,\cdots,x_n\}$; $U/T(P)=\{T_P(x_i)|i=1,2,\cdots,n\}$; $U/d=\{D_1,D_2,\cdots,D_m\}$。决策特征 d 关于条件特征子集 P 的条件熵定义为

$$H(d|P)=-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|T_P(x_i)\cap D_j|}{|U|}\log_2\frac{|T_P(x_i)\cap D_j|}{|T_P(x_i)|}\tag{1}$$

根据式(1), 通过从特征子集 P 删除某个特征 a 引起的条件熵的变化大小, 可定义特征的重要度度量函数:

$$\mathrm{sig}(a,P,d)=H(d|P)-H(d|P-\{a\})$$

2　不完备数据集中特征值更新的增量特征选择

当不完备决策系统中特征值发生动态更新时, 由特征子集所诱导的相容关系和由决策特征所诱导的等价关系会随之变化, 进而使得特征度量准则条件熵发生变化。下面, 首先分析一组对象的特征值发生更新时相容类和决策类的变化情况。由于条件熵的计算与相容类和决策类中的对象顺序无关, 为了方便阐述, 下文中假设决策系统中发生特征值修改的对象集合为 $\{x_i|i=p+1,p+2,\cdots,q\}$, 则更新后不完备决策系统中相容类的更新为

$$U/T'(P)=\{T'_P(x_i)|i=1,2,\cdots,p,p+1,\cdots,q,q+1,\cdots,k,k+1,\cdots,n\}$$

式中: $x_i(i=1,2,\cdots,p)$ 表示相容类保持不变的对象, 即 $T'_P(x_i)=T_P(x_i)$; $x_i(i=p+1,p+2,\cdots,q)$ 表示特征值发生变化的对象, 其相容类需根据定义计算, 即 $T'_P(x_i)=\{x_l\in U|(x_i,x_l)\in T'(P)\}$; $x_i(i=q+1,q+2,\cdots,k)$和 $x_i(i=k+1,k+2,\cdots,n)$ 分别表示相容类可能出现的两种更新模式的对象集合。对于对象集合 $x_i(i=q+1,q+2,\cdots,k)$, 相容类更新为 $T'_P(x_i)=(T_P(x_i)-T_P^-(x_i))\cup T_P^+(x_i)$, 对于对象集合 $x_i(i=k+1,k+2,\cdots,n)$, 相容类更新方式为 $T'_P(x_i)=T_P(x_i)-T_P^-(x_i)$, 其中, $T_P^-(x_i)=\{x_l|\{x_i,x_l\}\in T(P), p+1\leqslant l\leqslant q\}$; $T_P^+(x_i)=\{x_l|\{x_i,x_l\}\in T'(P), p+1\leqslant l\leqslant q\}$。

不完备决策系统中决策特征值发生变化后决策类的更新为

$$U/d=\{D'_j|j=1,2,\cdots,q_1,q_1+1,\cdots,q_2,q_2+1,\cdots,m\}$$

式中: $D'_j(j=1,2,\cdots,q_1)$ 表示决策类保持不变的对象, 即 $D'_j=D_j$; $D'_j(j=q_1+1,q_1+2,\cdots,q_2)$ 表示需要从当前决策类中删除特征值更新的可辨识对象, 即 $D'_j=D_j-D_j^-$; $D'_j(j=q_2+1,q_2+2,\cdots,m)$ 表示不仅需要删除特征值发生更新的可辨识对象集合, 同时要增加新特征值下不可辨识的对象集合, 即 $D'_j=(D_j-D_j^-)\cup D_j^+$, 其中 $D_j^-=\{x_l|\exists x_r\in D_j, f(x_l,d)=f(x_r,d), p+1\leqslant l\leqslant q\}$, $D_j^+=\{x_l|\exists x_r\in D_j, f(x_l,d)=f(x_r,d), p+1\leqslant l\leqslant q\}$。

根据上文中不完备决策系统中特征值动态修改后相容类和决策类的更新模式, 下面进一步可得相容类与决策类交集的动态更新方式。

对于任意对象 $x_i(i=1,2,\cdots,p)$ 的相容类和决策类 $D'_j(j=1,2,\cdots,m)$, 有 $|T'_P(x_i)\cap D'_j|=|T_P(x_i)\cap D_j|$ 成立, 对于任意对象 $x_i(i=p+1,p+2,\cdots,q)$, 由于其相容类需根据定义计算, 无法利用已有结果, 因此相容类和决策类 $D'_j(j=1,2,\cdots,m)$ 的交集表示为 $|T'_P(x_i)\cap D'_j|$。

对于任意对象 $x_i(i=q+1,q+2,\cdots,k)$ 的相容类和决策类 $D'_j(j=1,2,\cdots,m)$, 其交集的更新模式为

$$|T'_P(x_i)\cap D'|=\begin{cases}|T_P(x_i)\cap D_j|, & 1\leqslant j\leqslant q_1\\ |T_P(x_i)\cap D_j|-|T_P^-(x_i)\cap D_j|-|T_P(x_i)\cap D_j^-|+|T_P^-(x_i)\cap D_j^-|, & q_1+1\leqslant j\leqslant q_2\\ |T_P(x_i)\cap D_j|-|T_P^-(x_i)\cap D_j|-|T_P(x_i)\cap T_P^-(x_i)|+|T_P^-(x_i)\cap D_j^-|+|T_P^+(x_i)\cap D_j^+|, & q_2+1\leqslant j\leqslant m\end{cases}$$

对于任意对象 $x_i(i=k+1,k+2,\cdots,n)$ 的相容类和决策类 $D'_j(j=1,2,\cdots,m)$, 其交集的更新模式为

$$|T'_P(x_i)\cap D'_j|=\begin{cases}|T_P(x_i)\cap D_j|, & 1\leqslant j\leqslant q_1\\ |T_P(x_i)\cap D_j|-|T_P^-(x_i)\cap D_j|-\\ |T_P(x_i)\cap D_j^-|+|T_P^-(x_i)\cap D_j^-|, & q_1+1\leqslant j\leqslant m\end{cases}$$

通过分析不完备决策系统中相容类和决策类, 以及其交集的动态更新模式, 可得特征值发生修改时决策特征 d 关于任意条件特征子集 P 的条件熵的增量计算机制为

$$H'_U(d|P)=H_U(d|P)+\Delta$$

其中 Δ 的值如下所示:

$$\begin{aligned}\Delta=&-\sum_{i=p+1}^{q}\sum_{j=1}^{m}\frac{|T'_P(x_i)\cap D'_j|}{|U|}\log_2\frac{|T'_P(x_i)\cap D'_j|}{|T'_P(x'_i)|}+\\ &\sum_{i=p+1}^{n}\sum_{j=1}^{m}\frac{|T_P(x_i)\cap D_j|}{|U|}\log_2\frac{|T_P(x_i)\cap D_j|}{|T_P(x_i)|}+\\ &\sum_{i=q+1}^{n}\sum_{j=q_1+1}^{m}\left(\frac{|T_P^-(x_i)\cap D_j|+|T_P(x_i)\cap D_j^-|}{|U|}-\right.\\ &\left.\frac{|T_P^-(x_i)\cap D_j^-|+|T_P(x_i)\cap D_j|+|T_P^+(x_i)\cap D_j^+|}{|U|}\right)\cdot\\ &\log_2\frac{|T'_P(x_i)\cap D'|}{|T'_P(x_i)|}\end{aligned}$$

基于上述分析, 算法 1 给出了不完备决策系统中特征值更新时基于条件熵的增量式特征选择算法来计算新的特征子集。

算法 1　不完备决策系统中基于条件熵的增量式特征选择算法 (IFS-CE-IDS)

输入　不完备决策系统 $\mathrm{IDS}=(U,C\cup\{d\},V,f)$, 原始数据 U 上的特征子集 $\mathrm{RED}\in C$, 以及数据中发生修改对象的集合 ΔU;

输出　特征选择后的特征子集 A。

1) 初始化特征子集 $A=\mathrm{RED}$;

2) 根据特征值更新对象集合 ΔU 更新后 $U/d=\{D'_1,D'_2,\cdots,D'_m\}$, $U/T'(C)=\{T'_C(x_1),T'_C(x_2),\cdots,T'_C(x_n)\}$, $U/T'(A)=\{T'_A(x_1),T'_A(x_2),\cdots,T'_A(x_n)\}$, 计算 $T_P^-(x_i)$、$T_P^+(x_i)$、D_j^-、D_j^+;

3) 计算 $H'_U(d|C)$和$H'_U(d|A)$;

4) 如果 $H'_U(d|C)\neq H'_U(d|A)$ 进入 7), 否则进入 5);

5) 当 $H'_U(d|C)\neq H'_U(d|A)$时, 对任意 $a\in C-A$, 计算 $\mathrm{sig}(a,A\cup\{a\},d)$, 并且选择其中拥有最大 $\mathrm{sig}(a,A\cup\{a\},d)$ 的 a, $A=A\cup\{a\}$;

6) 对任意特征 $a\in A$ 计算 $\mathrm{sig}(a,A,d)$, 如果 $\mathrm{sig}(a,A,d)=0$, 则 $A=A-\{a\}$;

7) 返回 A。

该算法中条件熵的计算时间是 $O(|C||U||\Delta U|)$, 在算法 IFS-CE-IDS 中, 步骤 1)~3) 的计算时间是 $O(|C||U||\Delta U|)$, 步骤 5) 的向特征集 A 中添加特征的计算时间为 $O(|C|^2|U||\Delta U|)$, 步骤 6) 中删除掉冗余特征的时间复杂度为 $O(|A||C||U||\Delta U|)$。因此, 算法 IFS-CE-IDS 总的时间复杂度为 $O(|C||U||\Delta U|+|C|^2|U||\Delta U|+|A||C||U||\Delta U|)=O(|C|^2|U||\Delta U|)$。

3　实验及分析

本文选取了 9 组 UCI 数据集进行性能测试, 数据集详细信息如表 1 所示。对于完备数据集 Car 和 kr-vs-kp, 随机删除原始数据集中 5% 的已知特征值变为缺失值, 使原始完备数据集变为不完备数据集。对含有数值型数据的数据集 Hepatitis、Wisconsin、Dermatology 和 Ozone, 将数值型特征进行了离散化处理。如数据集 Hepatitis 包含 19 个特征, 其中 6 个为数值型特征; 数据集 Wisconsin 含有 1 个数值型特征; 数据集 Dermatology 包含 1 个数值型特征; 数据集 Ozone 都是数值型特征。实验环境配置为: Intel(R)Core(TM)i5-4210M CPU 2.60 GHz, 8 GB 内存, 操作系统为 Windows 10, 程序开发平台为 IntelliJ IDEA, 编程语言为 Java。

表 1　数据集描述

Table 1　Description of the datasets

数据集	样本数	特征数	类别数
Hepatitis	155	20	2
Audiology	226	69	24
Cancer	286	9	2
Soybean	307	35	19
Dermatology	366	34	6
Wisconsin	699	10	2
Car	1 728	6	4
Ozone	2 534	72	2
kr-vs-kp	3 196	36	2

为验证本文所提出算法 IFS-CE-IDS 处理数据集特征值更新问题具有高效性和可行性, 使用传统批量式特征选择算法 HFS-CE-IDS 与算法 IFS-CE-IDS 在 9 组 UCI 数据集上进行测试, 从分类精度、决策性能以及计算效率三方面对传统批量式特征选择算法 HFS-CE-IDS 和 IFS-CE-IDS 进行比较。

3.1　分类精度分析

为比较算法 HFS-CE-IDS 与算法 IFS-CE-IDS 所得特征子集的分类精度, 对表 1 中 9 组数据集选择其中 50% 对象, 并且更新其特征值, 然后分别运行传统批量式算法 HFS-CE-IDS 和增量式算法 IFS-CE-IDS 对特征值更新数据集进行特征选择。使用决策树 J48、Naïve Bayes、SVM

(support vector machines) 分类器验证这两种算法的分类性能。实验结果如表 2~4 所示。

表 2　J48 分类精度比较

Table 2　J48 classification accuracy comparison　%

数据集	HFS-CE-IDS算法	IFS-CE-IDS算法
Hepatitis	70.32±0.457 0	67.10±0.481 8
Audiology	30.53±0.204 1	30.09±0.205 8
Cancer	60.14±0.511 4	60.14±0.511 4
Soybean	44.66±0.190 7	43.78±0.193 1
Dermatology	41.26±0.375 3	41.53±0.374 5
Wisconsin	72.25±0.444 1	72.39±0.4380
Car	49.83±0.366 7	49.83±0.366 7
Ozone	73.28±0.442 5	73.28±0.442 5
kr-vs-kp	70.06±0.419 5	70.06±0.419 5

表 3　Naïve Bayes 分类精度比较

Table 3　Naïve Bayes classification accuracy comparison　%

数据集	HFS-CE-IDS算法	IFS-CE-IDS算法
Hepatitis	63.87±0.478 9	66.45±0.472 2
Audiology	28.76±0.186 2	30.09±0.185 8
Cancer	65.73±0.491 1	65.73±0.491 1
Soybean	34.55±0.204 8	32.21±0.207 6
Dermatology	46.72±0.336 8	44.54±0.338 2
Wisconsin	73.82±0.450 9	73.10±0.443 6
Car	49.07±0.368 7	49.07±0.368 7
Ozone	69.81±0.488 2	69.65±0.486 9
kr-vs-kp	63.74±0.471 5	63.74±0.471 5

表 4　SVM 分类精度比较

Table 4　SVM classification accuracy comparison　%

数据集	HFS-CE-IDS算法	IFS-CE-IDS算法
Hepatitis	70.32±0.544 8	70.32±0.544 8
Audiology	23.01±0.253 3	24.78±0.250 4
Cancer	64.69±0.594 3	64.69±0.594 3
Soybean	43.19±0.244 5	42.31±0.246 4
Dermatology	46.72±0.421 4	47.54±0.418 2
Wisconsin	73.53±0.514 5	72.82±0.521 4
Car	49.07±0.451 3	49.07±0.451 3
Ozone	73.24±0.517 3	73.28±0.516 9
kr-vs-kp	69.06±0.556 3	69.06±0.556 3

见表 2, 从两种算法在 J48 分类器的分类精度比较可知, 算法 IFS-CE-IDS 在数据集 Hepatitis、Audiology 和 Soybean 上所得的分类精度相较算法 HFS-CE-IDS 所得分类精度差一些, 而在其他 6 个数据集上算法 IFS-CE-IDS 所得分类精度与算法 HFS-CE-IDS 所得分类精度相同甚至更好。从表 3 可知, 在 Naïve Bayes 分类器中, 算法 IFS-CE-IDS 在 9 个数据集上的分类精度结果表明新提出算法在大部分数据集上的分类精度不比算法 HFS-CE-IDS 的分类精度差, 例如在数据集 Cancer、Car 和 kr-vs-kp 上两种算法的分类精度基本相同。

从表 4 可知, 在 SVM 分类器中, 与算法 HFS-CE-IDS 相比, 新提出算法的分类精度在 Hepatitis、Audiology、Cancer、Soybean、Dermatology、Wisconsin、Car、Ozone、kr-vs-kp 等 7 个数据集上相等甚至更好。

实验结果表明, 算法 IFS-CE-IDS 在大部分数据集上能够在特征子集和分类精度上取得和算法 HFS-CE-IDS 相接近的, 甚至更好的结果, 可以证明算法 IFS-CE-IDS 是一种有效的特征选择算法。

3.2　决策性能分析

为检验算法 IFS-CE-IDS 的决策性能, 本文使用文献 [28] 中对不完备数据进行评估所提出的 6 种评估函数评估算法 HFS-CE-IDS 以及算法 IFS-CE-IDS 计算的特征子集的决策性能。

6 种评估函数中, 特征集合 C 下不完备决策系统 $\mathrm{IDS}=(U,C\cup\{d\},V,f)$ 近似准确评估函数定义为

$$a_C(F)=\frac{\sum\limits_{D_j\in U/D}|\underline{\mathrm{apr}}_C(D_j)|}{\sum\limits_{D_j\in U/D}|\overline{\mathrm{apr}}_C(D_j)|}$$

式中: $\underline{\mathrm{apr}}_C=\cup\{x\in U|\mathrm{MC}_C\subseteq D_j,\ D_j\in U/D\},1\leqslant i\leqslant n$, 是下近似值; $\overline{\mathrm{apr}}_C=\cup\{x\in U|\mathrm{MC}_C\cap D_j\neq\varnothing,D_j\in U/D\}$, $1\leqslant i\leqslant n$, 是上近似值; $F=U/D$。其中, MC_C 表示在特征子集 C 下所得最大一致块集合。

不完备决策系统 $\mathrm{IDS}=(U,C\cup\{d\},V,f)$ 特征集合 C 下的一致性度量评估函数的定义为

$$c_C(D)=\frac{\sum_{D_j\in U/D}|\underline{\mathrm{apr}}_C(D_j)|}{|U|}$$

不完备决策系统 $\mathrm{IDS}=(U,C\cup\{d\},V,f)$ 在 $\mathrm{RULE}=\{Z_{ij}|Z_{ij}:\mathrm{des}(X_i)\rightarrow\mathrm{des}(D_j),X_i\in\mathrm{MC}_C,D_j\in\mathrm{MC}_d\}$ 下的确定性度量 α 评估函数定义为

$$\alpha(\mathrm{IDS})=\frac{1}{m}\sum_{N_i}^{m}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{|X_i\cap D_j|}{|X_i|}$$

式中: N_i 是在不完备决策表中由最大一致块 X_i 所诱导得到的决策类数目; $X_i\in U$ 表示在 $P\subseteq C$ 下, $(u,v)\in T(P),\forall u,v\in X_i$, 且不存在一个子集 $Y\in U$, $X_i\subset Y$, $(u,v)\in T(P),\forall u,v\in Y$, 称 X_i 为最大一致块。

不完备决策系统 $\mathrm{IDS}=(U,C\cup\{d\},V,f)$ 在 $\mathrm{RULE}=\{Z_{ij}|Z_{ij}:\mathrm{des}(X_i)\rightarrow\mathrm{des}(D_j),X_i\in\mathrm{MC}_C,D_j\in\mathrm{MC}_d\}$ 下的一致性度量 β 评估函数定义为

$$\beta(\mathrm{IDS})=\frac{1}{m}[1-\frac{4}{|X_i|}\sum_{j=1}^{N_i}|X_i\cap D_j|\mu(Z_{ij})(1-\mu(Z_{ij}))]$$

式中：N_i 是在不完备决策表中由最大一致块 X_i 所诱导得到决策类数目，且 $\mu(Z_{ij})=|X_i\cap D_j|/|X_i|$。

不完备决策系统 $\mathrm{IDS}=(U,C\cup\{d\},V,f)$ 在 $\mathrm{RULE}=\{Z_{ij}|Z_{ij}:\mathrm{des}(X_i)\to\mathrm{des}(D_j),X_i\in\mathrm{MC}_C,D_j\in\mathrm{MC}_d\}$ 下的支持度量 γ 评估函数定义为

$$\gamma(\mathrm{IDS})=\sum_{j=1}^{n}\frac{|D_j|}{N_j|U|}\sum_{k=1}^{N_j}\frac{|X_k\cap D_j|}{|U|}$$

式中 N_j 是条件部分关于 D_j 的最大一致块数。

不完备决策系统 $\mathrm{IDS}=(U,C\cup\{d\},V,f)$ 在 $\mathrm{RULE}=\{Z_{ij}|Z_{ij}:\mathrm{des}(X_i)\to\mathrm{des}(D_j),X_i\in\mathrm{MC}_C,D_j\in\mathrm{MC}_d\}$ 下的覆盖度量 ϑ 评估函数定义为

$$\vartheta(\mathrm{IDS})=\frac{1}{|U|}\sum_{i=1}^{m}\frac{|X_i|}{|U|}$$

选择表 1 中每组数据集中 50% 数据对象，更新其特征值，然后分别运行算法 HFS-CE-IDS 和算法 IFS-CE-IDS 对特征值更新数据集进行特征选择。并且使用上述 6 种评估函数评估每组数据集更新后两种算法特征选择的特征子集的决策性能，实验结果如表 5 所示。

表 5　HFS-CE-IDS 与 IFS-CE-IDS 的度量比较

Table 5　Measurement comparison of HFS-CE-IDS with IFS-CE-IDS

评价指标	算法	Hepatitis	Audiology	Cancer	Soybean	Dermatology	Wisconsin	Car	Ozone	kr-vs-kp
a_C	HFS-CE-IDS	1.0000	1.0000	1.0000	0.8402	1.0000	1.0000	0.0012	0.8471	0.8464
	IFS-CE-IDS	1.0000	1.0000	1.0000	0.8402	1.0000	1.0000	0.0012	0.8471	0.8464
c_C	HFS-CE-IDS	1.0000	1.0000	1.0000	0.9239	1.0000	1.0000	0.0052	0.9183	0.9312
	IFS-CE-IDS	1.0000	1.0000	1.0000	0.9239	1.0000	1.0000	0.0052	0.9183	0.9312
α	HFS-CE-IDS	1.0000	1.0000	1.0000	0.9669	1.0000	1.0000	0.2636	0.9591	0.9678
	IFS-CE-IDS	1.0000	1.0000	1.0000	0.9667	1.0000	1.0000	0.2636	0.9591	0.9678
β	HFS-CE-IDS	1.0000	1.0000	1.0000	0.9372	1.0000	1.0000	0.2295	0.9477	0.9375
	IFS-CE-IDS	1.0000	1.0000	1.0000	0.9369	1.0000	1.0000	0.2295	0.9477	0.9375
γ	HFS-CE-IDS	0.0075	0.0051	0.0019	0.0015	0.0028	0.0019	0.0017	0.0012	0.0004
	IFS-CE-IDS	0.0077	0.0052	0.0019	0.0016	0.0028	0.0019	0.0017	0.0012	0.0004
ϑ	HFS-CE-IDS	0.0067	0.0045	0.0014	0.0016	0.0027	0.0014	0.0034	0.0013	0.0004
	IFS-CE-IDS	0.0067	0.0045	0.0014	0.0016	0.0027	0.0014	0.0034	0.0013	0.0004

从表 5 的实验结果可知，算法 IFS-CE-IDS 与算法 HFS-CE-IDS 相比，在近似准确评估 a_C、一致性度量评估 c_C 以及覆盖度量评估 ϑ 这 3 种评估函数下的评估值是相同的；在数据集 Soybean 上，算法 IFS-CE-IDS 与算法 HFS-CE-IDS 在确定性度量 α 评估函数以及一致性度量 β 评估函数下的评估值不相同，而在其他 8 个数据集上两种算法的确定性度量 α 评估函数以及一致性度量 β 评估函数的评估值是相同的；在支持度量 γ 评估函数所得评估值结果中，两种算法在 Cancer、Dermatology、Wisconsin、Car、Ozone、kr-vs-kp 这 6 个数据集上的评估值相同，虽然在其余 3 个数据集上的评价值不同，但是评估值十分接近。

结合两种算法在 6 种评估函数下的结果，表明算法 IFS-CE-IDS 在大部分数据集下能够取得与算法 HFS-CE-IDS 相同的决策性能。

3.3　效率分析

为验证本文提出的增量式特征选择算法 IFS-CE-IDS 的高效性，采用传统的批量式特征选择算法 HFS-CE-IDS 作比较，该算法是一种与所提出算法基于相同特征评估准则的非增量方法。对表 1 中的每组数据集，依次选择其中的 5%, 10%, 15%, …, 50% 数据对象并更新其对象特征值。同时发生变化的特征值取决于对象特征的值域。当有不同规模的数据对象特征值发生更新，分别使用增量式特征选择算法 IFS-CE-IDS 和传统批量式特征选择算法 HFS-CE-IDS 对数据集进行特征选择，求解特征选择结果。计算时间的比较结果如图 1 所示，图中详细给出两种算法计算时间随数据对象特征值更新规模的变化而发生的变化。

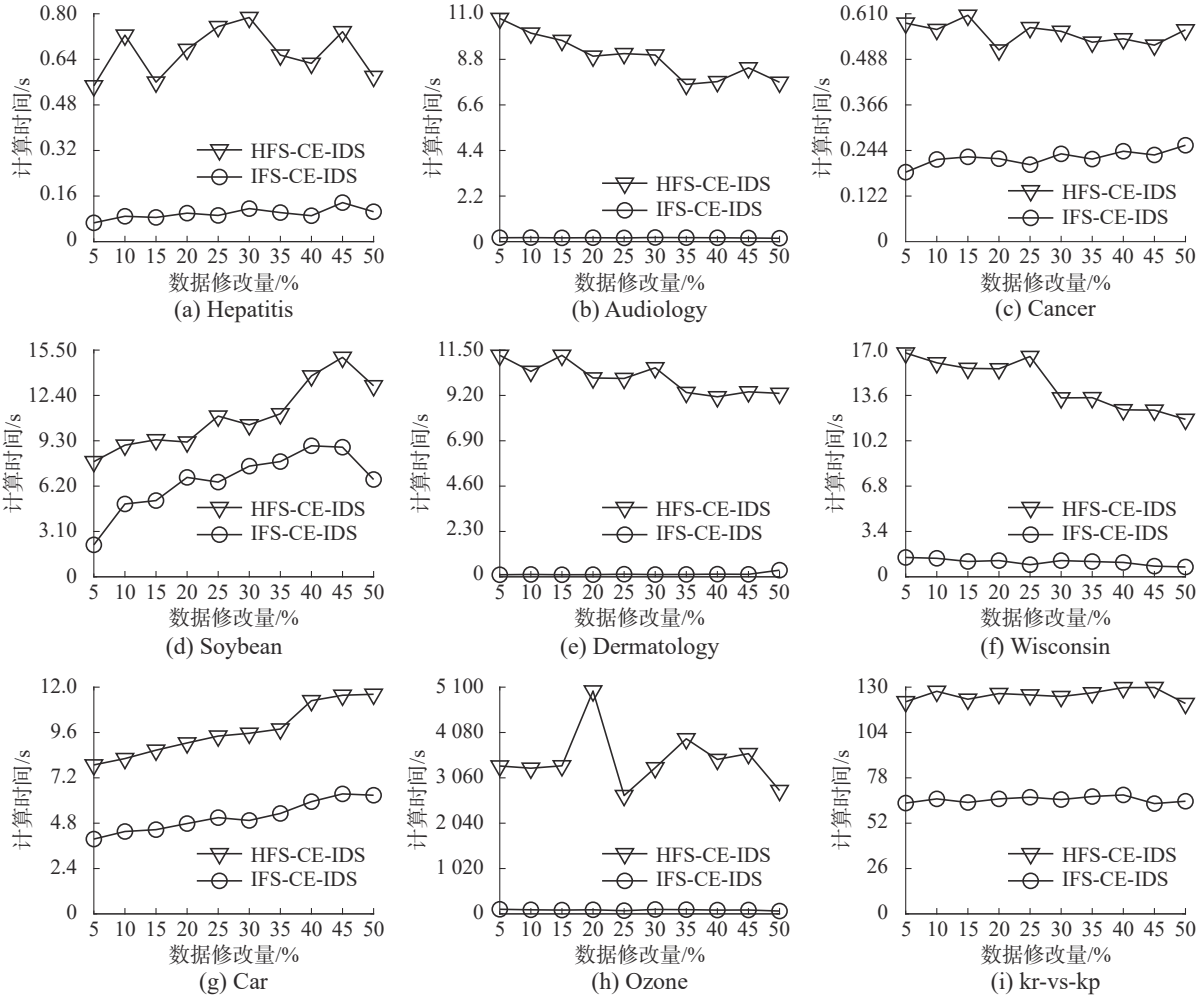


图 1　算法 HFS-CE-IDS 与算法 IFS-CE-IDS 计算时间比较

Fig. 1　Computational time comparison between HFS-CE-IDS and IFS-CE-IDS

从图 1 可知, 当不同规模的数据对象特征值发生更新后, 传统批量式特征选择算法 HFS-CE-IDS 比增量式特征选择算法 IFS-CE-IDS 花费更多时间来选择特征值更新后的特征子集, 主要的原因是增量式算法 IFS-CE-IDS 能够避免重复的计算, 可以利用之前已有的计算结果, 从而使得特征选择的计算效率得以提高。算法 IFS-CE-IDS 在 9 组数据集上的计算效率普遍比算法 HFS-CE-IDS 高, 尤其是在一些数据规模较大的数据集上, 算法 IFS-CE-IDS 的高效性更加明显。比如在 Ozone 数据集上算法 IFS-CE-IDS 的计算效率远优于算法 HFS-CE-IDS 的计算效率。图 1 中两种特征选择算法的计算时间都存在一些波动, 如数据集 Ozone 中对象数据对象特征值更新 20% 时, 算法 HFS-CE-IDS 的计算时间突然变得比其他比例耗时更大, 因为数据集的数值对象特征值发生更新后, 它的相容类与决策类因数据对象特征更新而产生变化, 从而导致计算时间发生波动。

在实验分析中, 通过算法 HFS-CE-IDS 与算法 IFS-CE-IDS 的分类精度、决策性能和计算效率三部分实验结果可知, 算法 IFS-CE-IDS 与算法 HFS-CE-IDS 相比, 在大部分数据集上进行特征选择所得特征子集数量相接近, 两种算法分类精度和决策性能基本相同, 但算法 IFS-CE-IDS 的计算时间小于算法 HFS-CE-IDS, 尤其在数据规模较大的数据集上计算时间的优势更加明显。通过本节分类精度、决策性能和计算效率三部分实验分析, 证明 IFS-CE-IDS 是一种高效的处理数据对象特征值更新问题的增量式特征选择算法。

4　结束语

本文提出了不完备决策系统中面向特征值动态更新的增量式特征选择算法。通过分析不完备决策系统中条件特征值和决策特征值同时更新时相容类和决策类的动态更新模式, 构造了条件信息熵的增量计算机制, 并进一步设计了一种基于动态不完备决策系统的增量式特征选择算法。实验选取了 9 组 UCI 公共数据集, 并通过分类精

度、决策性能以及计算效率 3 个方面与传统批量式特征选择算法进行了性能对比。实验结果表明，本文算法所选择的特征子集与批量式算法在分类精度和决策性能具有基本一致的性能表现。同时，在面对不完备数据中特征值的动态变化环境下，本文算法的计算效率远优于传统批量式算法，可在较短时间内计算出一个可行的特征子集。实验中部分数据集使用算法 IFS-CE-IDS 需要进行特征转化，导致失去部分有效信息，降低算法结果质量，未来将致力于寻求更有效处理混合数据的增量特征算法。

参考文献：

[1] 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.

[2] SLOWINSKI R, VANDERPOOTEN D. A generalized definition of rough approximations based on similarity[J]. IEEE transactions on knowledge and data engineering, 2000, 12(2): 331–336.

[3] KRYSZKIEWICZ M. Rough set approach to incomplete information systems[J]. Information sciences, 1998, 112(1/2/3/4): 39–49.

[4] PARTHALÁIN N M, SHEN Qiang. Exploring the boundary region of tolerance rough sets for feature selection[J]. Pattern recognition, 2009, 42(5): 655–667.

[5] MENG Zuqiang, SHI Zhongzhi. Extended rough set-based attribute reduction in inconsistent incomplete decision systems[J]. Information sciences, 2012, 204: 44–69.

[6] GRZYMALA-BUSSE J W, CLARK P G, KUEHNHAUSEN M. Generalized probabilistic approximations of incomplete data[J]. International journal of approximate reasoning, 2014, 55(1): 180–196.

[7] QIAN Yuhua, LIANG Jiye, PEDRYCZ W, et al. An efficient accelerator for attribute reduction from incomplete data in rough set framework[J]. Pattern recognition, 2011, 44(8): 1658–1670.

[8] DAI Jianhua. Rough set approach to incomplete numerical data[J]. Information sciences, 2013, 241: 43–57.

[9] YANG Xibei, YANG Jingyu, WU Chen, et al. Dominance-based rough set approach and knowledge reductions in incomplete ordered information system[J]. Information sciences, 2008, 178(4): 1219–1234.

[10] LIANG Jiye, XU Zongben. The algorithm on knowledge reduction in incomplete information systems[J]. International journal of uncertainty, fuzziness and knowledge-based systems, 2002, 10(1): 95–103.

[11] QIAN Yuhua, LIANG Jiye, WANG Feng. A new method for measuring the uncertainty in incomplete information systems[J]. International journal of uncertainty, fuzziness and knowledge-based systems, 2009, 17(6): 855–880.

[12] DAI Jianhua, WANG Wentao, XU Qing. An uncertainty measure for incomplete decision tables and its applications[J]. IEEE transactions on cybernetics, 2013, 43(4): 1277–1289.

[13] ZHAO Hua, QIN Keyun. Mixed feature selection in incomplete decision table[J]. Knowledge-based systems, 2014, 57: 181–190.

[14] RAGHAVAN V, HAFEZ A. Dynamic data mining[C]// Proceedings of the 13th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. New Orleans, Louisiana, USA, 2000.

[15] LI Tianrui, RUAN Da, GEERT W, et al. A rough sets based characteristic relation approach for dynamic attribute generalization in data mining[J]. Knowledge-based systems, 2007, 20(5): 485–494.

[16] XU Yitian, WANG Laisheng, ZHANG Ruiyan. A dynamic attribute reduction algorithm based on 0-1 integer programming[J]. Knowledge-based systems, 2011, 24(8): 1341–1347.

[17] QIAN Jin, DANG Chuangyin, YUE Xiaodong, et al. Attribute reduction for sequential three-way decisions under dynamic granulation[J]. International journal of approximate reasoning, 2017, 85: 196–216.

[18] YANG Yanyan, CHEN Degang, WANG Hui, et al. Incremental perspective for feature selection based on fuzzy rough sets[J]. IEEE transactions on fuzzy systems, 2018, 26(3): 1257–1273.

[19] LANG Guangming, CAI Mingjie, FUJITA H, et al. Related families-based attribute reduction of dynamic covering decision information systems[J]. Knowledge-based systems, 2018, 162: 161–173.

[20] WEI Wei, WU Xiaoying, LIANG Jiye, et al. Discernibility matrix based incremental attribute reduction for dynamic data[J]. Knowledge-based systems, 2018, 140: 142–157.

[21] ZENG Anping, LI Tianrui, LIU Dun, et al. A fuzzy rough set approach for incremental feature selection on hybrid information systems[J]. Fuzzy sets and systems, 2015, 258: 39–60.

[22] LIANG Jiye, WANG Feng, DANG Chuangyin, et al. A group incremental approach to feature selection applying rough set technique[J]. IEEE transactions on knowledge

and data engineering, 2014, 26(2): 294–308.

[23] SHU Wenhao, SHEN Hong. Incremental feature selection based on rough set in dynamic incomplete data[J]. Pattern recognition, 2014, 47(12): 3890–3906.

[24] XIE Xiaojun, QIN Xiaolin. A novel incremental attribute reduction approach for dynamic incomplete decision systems[J]. International journal of approximate reasoning, 2018, 93: 443–462.

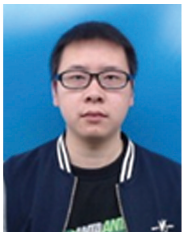
[25] WANG Feng, LIANG Jiye, DANG Chuangyin. Attribute reduction for dynamic data sets[J]. Applied soft computing, 2013, 13(1): 676–689.

[26] 刘吉超, 王锋, 宋鹏. 缺失数据的维数增量式特征选择[J]. 计算机工程与应用, 2019, 55(17): 95–99.
LIU Jichao, WANG Feng, SONG Peng. Dimension incremental feature selection algorithm for missing data[J]. Computer engineering and applications, 2019, 55(17): 95–99.

[27] 钱进, 朱亚炎. 面向成组对象集的增量式属性约简算法[J]. 智能系统学报, 2016, 11(4): 496–502.
QIAN Jin, ZHU Yayan. An incremental attribute reduction algorithm for group objects[J]. CAAI transactions on intelligent systems, 2016, 11(4): 496–502.

[28] QIAN Yuhua, DANG Chuangyin, LIANG Jiye, et al. On the evaluation of the decision performance of an incomplete decision table[J]. Data & knowledge engineering, 2008, 65(3): 373–400.

作者简介:



唐荣, 硕士研究生, 主要研究方向为数据挖掘与知识发现、粒计算与粗糙集。

罗川, 副教授, 博士, 中国人工智能学会粒计算与知识发现专业委员会委员, 中国计算机学会会员、中国人工智能学会会员, 主要研究方向为数据挖掘与知识发现, 粒计算与粗糙集。主持国家自然科学基金项目 2 项, 中国博士后科学基金 2 项。发表学术论文 40 余篇。

曹潜, 硕士研究生, 主要研究方向为数据挖掘、粒计算与粗糙集。