文章编号: 1001-0920(2018)08-1407-08 DOI: 10.13195/j.kzyjc.2017.0512

基于聚类离散化和变精度邻域熵的属性约简

陈迎春†, 李 鸥, 孙 昱

(信息工程大学 信息系统工程学院，郑州 450000)

摘 要: 针对传感网采集数据的不完备性,利用数据本身特点,通过定义类簇指标,提出基于改进K-means聚类算法的数据离散化方法,以减小噪声、孤立点和不完备数据集对决策识别结果产生的影响;然后,通过引入互信息熵的属性重要度度量和变精度修正系数,提出基于互信息熵的变精度邻域粗糙集属性约简启发式算法,整合变精度和邻域粗糙集的优势,在减小约简算法计算复杂度的同时提高决策系统识别精度. 仿真结果表明了算法在提高决策系统识别精度和降低其计算复杂度方面的有效性,模拟环境测试进一步验证了其工程适用性.

关键词: 变精度粗糙集；邻域粗糙集；K-means聚类；互信息熵；属性约简

中图分类号: TP18 **文献标志码:** A

Attribute reduction based on clustering discretization and variable precision neighborhood entropy

CHEN Ying-chun†, LI Ou, SUN Yu

(College of Information System Engineering，Information Engineering University，Zhengzhou 450000，China)

Abstract: To deal with the incompleteness of sensor network data, an attribute discretization method is proposed, based on the improved K-means clustering algorithm. In the method, a cluster index is defined and the data characteristics of each attribute are utilized in order to reduce the influences of noise, outliers and incomplete data sets on recognition results. Then, through the introduction of the mutual information entropy and variable precision correction coefficient, an attribute reduction heuristic algorithm is proposed, which integrates the advantages of the variable precision rough set and neighborhood rough set. The computational complexity is reduced and the recognition accuracy is improved by using the algorithm. Simulation results show the effectiveness of the proposed algorithms in dealing with the recognition accuracy and the computational complexity of the decision recognition system. The simulated environment test further verifies the applicability of the proposed algorithms.

Keywords: variable precision rough set； neighborhood rough set； K-means clustering； mutual information entropy； attribute recuction

0 引 言

随着通信智能的快速发展和通信环境的日趋复杂,将传感网感知技术应用于决策识别领域愈加广泛[1-3]. 数据样本越来越多,数据价值密度越来越低,数据格式各异,如何在保证数据处理精度条件下进行数据约简,从海量数据中有效去除噪声、冗余或不重要的信息,从而提高决策识别精度和时效性,已成为大数据背景下传感网数据应用的关键问题.

粗糙集理论,主要有代数和信息论两种观点[4],以其能够定量分析处理不精确、不一致、不完整信息与知识的能力,被广泛应用于决策支持与分析、数据挖掘、知识发现、机器学习及传感网、物联网等领域,而决策信息系统的约简已成为粗糙集理论和应用研究的焦点问题之一[5]. 基于粗糙集理论的数据属性约简,仅依赖所需处理的数据集合本身,就能够在保持数据决策识别能力不变的情况下删除冗余和不重要的数据. 与目前常用的高维数据降维算法PCA[6-7]相比,可以不改变数据属性的物理意义而降低决策识别系统数据处理量,提高系统处理效率[8-9]. 但是,经典粗糙集理论只能应用于离散值属性,其上、下近似概念定义在严格等价关系上,实际数据往往难以满足[10-15]. 为了解决这一问题,粗糙集理论出现了大量

收稿日期: 2017-04-26；**修回日期:** 2017-09-03.

基金项目: 国家自然科学基金项目(61601516)；国家科技重大专项项目(2014ZX03006003).

责任编委: 阳春华.

作者简介: 陈迎春(1979−), 女, 副教授, 博士生, 从事传感器网络、数据融合分析等研究； 李鸥(1961−), 男, 教授, 博士生导师, 从事认知网络、无线自组织网络等研究.

†通讯作者. E-mail: springer_2002@163.com

扩展模型[10-11,16-17],不仅指明了其与形式概念分析等其他不确定信息处理理论的关系,还指出了知识的粗糙性实质上是对其所含信息更深层次的刻画,为粗糙集理论实际应用提供了更广的理论、方法和工具. 文献[14]的邻域粗糙集模型将严格等价关系扩展为非等价关系;文献[15]又进一步扩展了邻域关系和相容关系,将粗糙集理论应用于不完备混合型数据集. 为了进一步解决数据离散化的误差问题和属性间无函数关系的数据分类问题,文献[11,18]指出,变精度粗糙集模型从集合包含度的视角给出了决策粗糙集模型的一个特例,其通过引入一个误差精度,在一定程度上放宽了粗糙集理论对不可分辨关系的要求[19],可对样本属性进行一定的噪声容忍,兼顾不一致性和随机性,使不确定性量度性质更佳[20-21]. 文献[22-23]进一步将变精度粗糙集与邻域粗糙集相结合,用邻域粒化的思想推广了变精度粗糙集模型;文献[24]对邻域熵扩展为变精度邻域熵的原理进行了证明;文献[23]使用基于正域的属性重要度作为属性核的约简标准;而文献[4-5,15]则证明了粗糙集理论中约简的信息论观点包含了代数观点. 因此,采用基于信息熵理论的属性约简可更全面地处理系统的不确定性,获得高效的属性约简算法[15,25-26].

文献[27]指出,利用互信息增益度量属性重要度,具有更有效的计算理论. 因此,本文以传感网数据决策识别系统为应用背景,针对不完备混合决策系统,提出一种基于互信息熵的变精度邻域粗糙集属性约简启发式算法,从信息论的角度出发,结合变精度粗糙集和邻域粗糙集的优势,综合考虑属性全集对决策结果的影响,实现决策识别系统处理精度的提高和整体计算复杂度的降低. 为了满足粗糙集理论离散值属性要求,本文又针对不同识别目标传感器数据具有一定区分度、不同传感器数据格式各异且有空值、采集易受噪声影响等特点,提出一种基于改进K-means聚类算法的数据离散化方法. 该方法在降低传统K-means算法对初始聚类中心敏感性的同时,以一种新的类簇指标为度量,针对各单一属性数据进行离散化,减小噪声和孤立点对聚类结果产生的影响,克服统一阈值设定无法兼顾不同属性数值分散程度而造成决策识别精度下降的弊端.

1 基本概念

针对不完备混合决策系统$\text{IMT}=\langle U, C\bigcup D, V, f\rangle$[15],论域$U=(x_1,x_2,\cdots,x_n)$表示非空有限样本集,$C$和$D$分别为条件属性集和决策属性集且$C\bigcap D=\varnothing$,$V=\bigcup\limits_{a\in C\bigcup D}V_a$为属性$a$的值域,包含连续数值数据、离散符号数据和空值,信息函数$f: U\times(C\bigcup D)\to V$,表示样本与其属性取值的映射关系.

定义1 给定IMT,$B\subseteq C\bigcup D$,$\forall x_i\in U$,样本x_i在属性子集B上的广义邻域为[15]

$$\delta_B(x_i)=\{x_j\in U|(x_i,x_j)\in \text{TN}_B\}. \tag{1}$$

其中:$\delta_B(x_i)$为广义邻域粒子,TN_B为广义邻域关系.

定义2 广义邻域模型可变精度阈值$\beta\in(0.5,1]$的上近似和下近似分别为[22]

$$\overline{N^\beta}X=\{x_i|I(\delta_B(x_i),X)>1-\beta,x_i\in U\}, \tag{2}$$

$$\underline{N^\beta}X=\{x_i|I(\delta_B(x_i),X)\geqslant\beta,x_i\in U\}. \tag{3}$$

其中:针对集合A和集合B,$I(A,B)=\text{Card}(A\bigcap B)/\text{Card}(A)$,$A,B\neq\varnothing$,Card表示集合的势. 当$A=\varnothing$或者$B=\varnothing$时,$I(A,B)=0$,$I(A,B)$表示$A$包含于$B$的程度且$0\leqslant I(A,B)\leqslant 1$. $\underline{N^\beta}X$表示U中以不小于β的分类精度划分到决策类上的邻域信息粒子的集合[23].

定义3 对于IMT,条件属性子集$B\subseteq C$,$\forall c\in C-B$,则c对于D的重要度为[23]

$$\text{SIG}(c,B,D)=\gamma^\beta_{B\bigcup\{c\}}(D)-\gamma^\beta_B(D). \tag{4}$$

其中:$\gamma^\beta_B(D)$表示决策属性D对条件属性子集B的β近似依赖度,若采用基于正域的属性重要度标准,则$\gamma^\beta_B(D)=\text{Card}(\text{POS}(B,D,\beta))/\text{Card}(U)$,正域$\text{POS}(B,D,\beta)=\bigcup\underline{N^\beta}X$. 根据条件属性重要度,通过启发式算法去除对决策识别结果影响不大的属性,达到数据约简的目的.

2 基于K-means聚类的数据离散化

将连续属性离散化,是数据处理中的重要一环,其效果好坏直接影响数据分析结果[28]. 数据离散化算法按照是否考虑类别信息分为有监督和无监督方法[28],按照考虑单个属性还是全局属性分为局部和全局方法[28-29],按照是否考虑结果生成过程分为静态和动态方法[28,30]等. 为了满足粗糙集理论的离散化属性值要求,根据多传感器识别目标所采集数据通常具有异类目标数据之间差异较大的特点,本文提出一种基于K-means聚类的数据离散化方法. 该方法首先将大量高维属性数据点按其类簇指标划分为多个簇,提取每个簇的类簇标签;然后使用该标签代替簇中所有数据实现离散化,通过充分利用数据本身的特点减小离散化带来的误差. 本离散化方法处于属性约简前的数据预处理阶段,在离散化过程中并没有考虑数据类别信息,而是通过对每一维属性数据特点的具体考量实现全体数据的离散化,属于无监督、局部的静态离散化方法.

K-means 聚类具有算法实现快速、简单,能有效处理大数据集等优点[31],广泛应用于大数据分析[32],是数据挖掘领域的经典算法之一. 但经典 K-means 算法对初始聚类中心敏感,易使结果不稳定或陷入局部最优;而基于密度的数据处理方法由于可以有效反映数据的分布特征,被大量应用于有监督和无监督数据处理中[33]. 因此,基于密度的聚类中心选取方法成为优化 K-means 聚类算法的有效手段之一[34]. 本文借鉴文献[34]的方法,选取类簇指标随 k 值变化曲线的最低点(最小值)所对应的 k 值作为最终类簇个数. 同时,为减小计算复杂度,本算法采用类簇平均质心距离的平均值作为类簇指标,综合考虑样本距离和样本分散程度,以减小噪声和孤立点对决策识别结果的影响. 针对数据集中的空值,可根据类簇密度和上下样本进行插值. 而 k 作为决策类别数,在实际传感网应用中,通常具有一定的先验信息.

设样本数据集 $X=\{x_i|i=1,2,\cdots,n\}$,则类簇平均质心距离的平均值 E 可表示为

$$E=\frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{n_i}d(x_j,\mathrm{core}_i). \tag{5}$$

其中: k 为类簇的数量; n_i 为第 i 类 C_i 中样本数量; $d(\cdot)$ 为距离函数,一般为欧氏距离; $\mathrm{core}_i=\dfrac{1}{n_i}\sum\limits_{x\in C_i}x$ 为 C_i 的聚类中心.

数据离散化算法流程如图 1 所示.

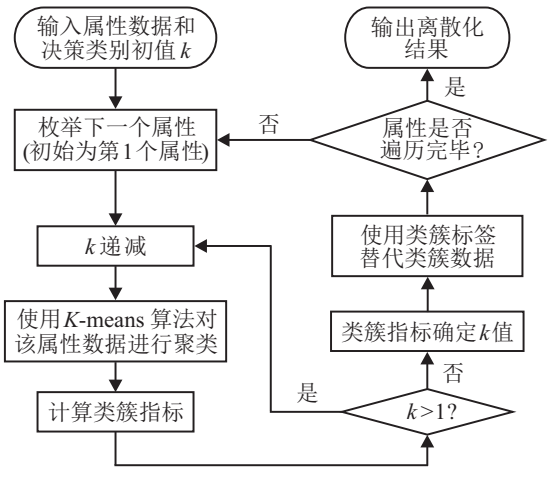

图 1　基于 K-means 聚类的数据离散化算法流程

以某一属性为例,基于 K-means 数据离散化算法步骤如下.

输入: 属性数值和 k 的最大值,这里通过一个递减的计数器枚举 k 的取值;

输出: 属性的离散化数值.

Step 1: 针对该属性和一个固定的 k 值,对所有属性数据进行 K-means 聚类,结果暂存;

Step 2: 计算类簇平均质心距离的平均值作为类簇指标;

Step 3: 判断 k 是否大于 1,若是,则 $k-1$,返回 Step 1,否则前往 Step 4;

Step 4: 计算该属性类簇指标随 k 值变化的最小值,确定 k 最终取值,并从暂存的数据中选择对应 k 值的聚类结果;

Step 5: 使用类簇标签代替类簇中数据的值,将连续属性离散化.

在上述算法中,针对每一个样本属性,根据文献[34]的思想,其计算时间主要消耗在样点距离的计算上,时间复杂度为 $O(|U|k\ln|U|)$. 而总体算法需要遍历每一个样本属性,因此,该离散化算法的时间复杂度为 $O(|C||U|k\ln|U|)$. 在空间复杂度方面,可以通过在硬盘暂存的方式节省内存空间.

3　基于互信息熵的属性约简启发式算法

本文以求取 IMT 的决策表核属性集为目标,以属性重要性为启发信息,提出基于互信息熵的前向属性约简启发式算法.

定义 4　给定 IMT, $B\subseteq C\bigcup D$,则属性子集 B 和决策属性 D 的条件熵定义为[15]

$$H_\delta(D|B)=-\sum_{i=1}^{|U|}\frac{1}{|U|}\log_2\frac{|\delta_D(x_i)\bigcap\delta_B(x_i)|}{|U|}. \tag{6}$$

文献[24]证明了可变精度粗糙集下 β 近似等价于 U 中以不小于 β 的分类精度进行决策类划分的邻域下近似. 条件熵 $H_\delta(D|B)$ 表示在已知参数集 B 的情况下,因不确定性规则而使决策状态 D 存在平均不确定度,它根据所有条件属性等价类和决策属性等价类的交集来计算熵值,没有区分确定性规则和不确定性规则[25]. 本文借鉴文献[25]的思想,引入变精度粗糙集模型,通过设定可变精度阈值 β,将由噪声数据引起的弱确定性规则过滤出来.

定义 5　给定 IMT, $B\subseteq C\bigcup D$,精度阈值 $\beta\in(0.5,1]$,则决策状态 D 相对于参数集 B 的变精度条件熵定义为

$$H_\delta^\beta(D|B)=-\sum_{i=1}^{|U|}\frac{1}{|U|}\log_2\frac{|\delta_D^\beta(x_i)\bigcap\delta_B^\beta(x_i)|}{|U|}+\frac{|U-V_\delta^\beta|}{|U|}\log|U|, \tag{7}$$

其中集合 V_δ^β 表示参数集合 B 在邻域范围内到决策状态 D 的对应关系中,相应变精度下近似 $\underline{N^\beta}X$ 的并.

定义 6　给定 IMT,条件属性子集 $B\subseteq C$,精度阈值 $\beta\in(0.5,1]$, $\forall c\in C-B$,则 c 对于 D 的重要度为

$$\mathrm{SIG}(c,B,D)=I_\delta^\beta(D;B\bigcup\{c\})-I_\delta^\beta(D;B)=$$
$$H_\delta^\beta(D)-H_\delta^\beta(D|B\bigcup\{c\})-(H_\delta^\beta(D)-H_\delta^\beta(D|B))=$$
$$H_\delta^\beta(D|B)-H_\delta^\beta(D|B\bigcup\{c\}),\tag{8}$$

其中$I_\delta^\beta(D;B)$表示决策状态D与参数集B的互信息熵. 属性重要度其实为互信息的增量,增量越大,说明在已知属性B的条件下, 属性c对决策D越重要.

在前向启发式约简算法中,逐个选择最重要的属性并添加到核属性集中,直到$I_\delta^\beta(D;B)=I_\delta^\beta(D;C)$. 本文借鉴文献[26]和PCA的思想,进一步引入变精度修正系数ε用于降低算法复杂度. 若满足$I_\delta^\beta(D;B)/I_\delta^\beta(D;C)\geqslant\varepsilon$,则认为启发式算法结束. 一般可取$\varepsilon\in[0.9,1]$.

具体算法步骤如下.

输入: 不完备的混合决策系统IMT, 邻域半径δ, 变精度修正系数ε,可变精度阈值β;

输出: 条件属性C的一个相对约简Red.

Step 1: 依据第2节基于K-means聚类的数据离散化方法,对条件属性数据进行归一离散化处理.

Step 2: 计算$I_\delta^\beta(D;C)$.

Step 3: 令$j=1,Y=C,C$的D核$\mathrm{Core}_D(C)=\varnothing$,重复Step 3.1 ~ Step 3.5:

Step 3.1: 计算$G=H_\delta^\beta(D|C)$;

Step 3.2: 任取$g\in Y$,计算$G'=H_\delta^\beta(D|C-\{g\})$;

Step 3.3: 若$G/G'\geqslant\varepsilon$,则$\mathrm{Core}_D(C)=\mathrm{Core}_D(C)\bigcup\{g\}$;

Step 3.4: $Y=Y-\{g\}$;

Step 3.5: $j=j+1$,若$j>|C|$,则终止,转Step 4, 否则转Step 3.1.

Step 4: 令Red=$\mathrm{Core}_D(C)$,重复Step 4.1~Step 4.3:

Step 4.1: 对$g_i^*\in C-\mathrm{Red},1\leqslant i\leqslant|C-\mathrm{Red}|$,计算$\mathrm{SIG}(g_i^*,\mathrm{Red},D)$;

Step 4.2: 取$\mathrm{SIG}(g_r^*,\mathrm{Red},D)=\max\limits_{i=1}^{|C-\mathrm{Red}|}\mathrm{SIG}(g_i^*,\mathrm{Red},D)$,令Red= $\mathrm{Red}\bigcup\{g_r^*\}$;

Step 4.3: 若$\dfrac{I_\delta^\beta(D;\mathrm{Red})}{I_\delta^\beta(D;C)}\geqslant\varepsilon$,则终止,转Step 5, 否则转Step 4.1.

Step 5: 输出所求的约简属性Red.

在上述算法中,Step 1的计算复杂度是第2节中分析的$O(|C||U|k\ln|U|)$. Step 2互信息熵的计算消耗主要集中在等价类计算上,时间复杂度为$O(|C||U|\ln|U|)$. Step 3和Step 4采用在论域U中逐步增量获取约简属性的方法,是以熵的计算为基础的,最坏的情况就是遍历样本所有条件属性,其计算复杂度为$O(|C|^2|U|\ln|U|)$. 在K-means聚类离散化时,类簇个数k一般远小于$|C|$和$|U|$,因此,本文算法的时间复杂度主要由Step 3和Step 4决定,为$O(|C|^2|U|\ln|U|)$. 在空间复杂度方面,相比于文献[23]基于正域的属性重要度算法,虽然引入了互信息熵的计算存储空间,但却减少了正域的计算存储空间,其空间复杂度的增加主要集中在Step 1数据离散化方面. 正如第2节所述,该算法的空间复杂度在可控的有限范围内.

4 仿真结果与分析

为了测试本文算法性能,首先选择来自UCI数据库的8个数据集(见表1)进行实验测试,然后再通过搭建模拟环境进行实际测试. 表1中的8个数据集涉及日常生活和医学等领域,均是通过采集样本属性数据实现对样本类别的决策识别,与本文传感器网络数据的表现形式及处理目标一致. 实验过程中将约简后的数据集作为NBC(Naive bayes classifier)分类器的输入,采用10折交叉验证的方法输出目标识别结果,以检测属性约简算法在决策识别系统中的有效性.

表1 数据集描述

数据集	实例数	条件属性	决策类别
Wine	178	13	3
Dermatology	366	34	6
Breast cancer	699	10	2
Credit approval	690	15	2
Vehicle	846	18	4
Iono	351	34	2
Mushroom	8 124	22	2
Hepatitis	155	19	2

4.1 基于K-means聚类的数据离散化方法验证

对于数据集中的数据,首先要进行基于K-means聚类的数据离散化. 为了验证该离散化方法的有效性,本文以Wine数据集为例,选择初始$k=5$,经过改进的K-means方法归一离散化后的数据如表2所示.

表2 离散化的Wine数据

样本	属性				
	Alcohol	Malic acid	Ash	···	Proline
1	3	1	3	···	1
2	1	1	1	···	1
3	1	2	2	···	1
4	3	1	3	···	1
5	1	2	2	···	3
⋮	⋮	⋮	⋮	⋮	⋮
178	3	3	2	···	2

在表2中,基于K-means聚类的方法对数据集中的每一个属性数据进行离散化,最后选择类簇指标随

k值变化曲线的最小值所对应的$k = 3$作为最终类簇个数,这与Wine 数据集的先验决策类别数3相吻合. 数据经过聚类离散化,不仅降低了数据中噪声和孤立点对决策识别结果的影响,而且还充分显现了传感器网络所采集数据本身的特点——不同目标的感知数据差异较大. 当然,在具有先验决策类别信息时,可以通过预先设定k值的方法减少聚类循环次数,降低算法计算量. 但该离散化方法为无监督的数据属性约简和属性识别提供了一种解决途径.

4.2　基于互信息熵的属性约简启发式算法验证

为了对基于互信息熵的属性约简启发式算法进行验证,本文仍以Wine数据集为例,在对属性数据进行基于K-means聚类离散化的基础上,分别测试邻域半径δ、可变精度阈值β和变精度修正系数ε对约简算法的影响. 测试先从不同邻域半径δ开始,检测约简后的属性个数,以及将约简后的数据作为NBC输入时得到的决策识别精度. 由于$\beta \in (0.5, 1]$,而一般$\varepsilon \in [0.9, 1]$,本文测试的初始值选择$\beta = 0.55$以及$\varepsilon = 0.95$,测试结果如表3所示.

表 3　邻域半径δ不同时决策结果对比

邻域半径δ	决策识别精度/%	约简后的属性个数
0.05	95.1	9
0.1	93.92	6
0.15	94.41	5
0.2	95.52	5
0.25	93.52	6

由表3数据可以看出,当$\delta = 0.2$时,不仅使约简后的属性个数最少,而且还使决策识别精度最高. 因此,在测试可变精度阈值β对决策识别系统性能影响时,设置$\delta = 0.2$且$\varepsilon = 0.95$,测试结果如表4所示.

表 4　可变精度阈值β不同时决策结果对比

可变精度因子β	决策识别精度/%	约简后的属性个数
0.8	95.63	5
0.85	96.93	6
0.9	97.5	6
0.95	97.5	6

在表4中,随着β的不断增加,决策识别精度不断提高. 当$\beta \geqslant 0.9$时,决策识别精度和约简后属性个数基本保持不变. 虽然此时约简后的属性个数比$\beta = 0.8$时增加了1,但决策识别精度却提高了近2个百分点. 因此,在对不同变精度修正因子ε进行测试时,选择初始参数$\delta = 0.2$且$\beta = 0.9$,测试结果如表5所示.

表 5　变精度修正因子ε不同时决策结果对比

变精度修正因子ε	决策识别精度/%	运行时间/s	约简后的属性个数
0.7	94.53	0.06263	4
0.8	96.06	0.07189	5
0.9	97.5	0.08032	6
1	97.5	0.08033	6

观察表5中数据可以发现,随着ε的不断变大,属性约简算法选出的数据属性不断增加,决策识别精度不断提高,但也使运行时间不断变长,这与ε的功能定义相吻合. 当$\varepsilon \geqslant 0.9$时,约简后的属性个数和决策识别精度基本保持不变,运行时间也基本保持稳定,这说明在Wine数据集中,此时所选出的数据属性在互信息熵的定义下已趋于稳定值. 同时,表5也说明,在决策识别精度满足的条件下,可以通过降低ε的值,减少所选数据属性个数,在算法的运行时间与决策识别精度之间进行折中.

在$\delta = 0.2$、$\beta = 0.9$和$\varepsilon = 0.95$时,本文算法与不进行属性约简的决策识别算法、采用文献[6]的PCA属性约简算法、采用文献[15]广义邻域关系下基于条件熵的属性约简算法、采用文献[18]的正域分布保持GA-PRDR约简算法、以及采用文献[23]基于正域的属性重要度约简算法以及采用文献[26]基于互信息的变精度粗糙集属性约简算法的运行结果进行对比,见表6.

表 6　7种算法运行结果对比

算法	决策识别精度/%	运行时间/s
不进行属性约简的决策识别算法	97.270	0.1069
文献[26]算法	97.289	0.0805
文献[23]算法	97.290	0.0807
文献[18]算法	97.502	0.0992
文献[15]算法	97.284	0.0805
文献[6]算法	97.110	0.0783
本文算法	97.500	0.0803

由表6中数据可以看出:本文算法通过属性约简选择出重要属性,减少了原始数据中掺杂的噪声影响,与不进行属性约简的决策识别算法相比,决策识别精度有所提高. 同时,由于约简掉一部分属性数据且$\varepsilon = 0.95$,在算法运行时间上明显优于不做属性约简的决策识别算法. 由于文献[23]算法采用变精度粗糙集与邻域粗糙集相结合的方式进行属性约简,使得决策识别精度高于文献[26]的变精度粗糙集属性约简算法和文献[15]的广义邻域条件熵属性约简算法,但文献[26]引入变精度粗熵修正系数,文献[15]对每轮熵值计算都进行了简化,使得文献[23]算法运

行时间长于文献[26]和文献[15]. 本文算法基于互信息熵进行属性约简,采用K-means聚类离散化的方法消除了噪声和孤立点的影响,相比于文献[23]的基于正域属性重要度算法,决策识别精度明显提高. 而且本文算法通过引入变精度修正系数,使得算法运行时间也小于文献[15]和文献[26]. 与文献[18]算法相比,由于文献[18]采用基于遗传的最小约简算法,在决策域分布保持条件下减小了出现约简超集的可能性,使决策识别精度略高于本文算法,但在运行时间上消耗更多. 文献[18]的约简算法其时间复杂度为$O(N|I-\max||C|^2|U|^2)$,N为随机生成的初始种群大小,$|I-\max|$为最大迭代数,远大于本文算法的时间复杂度$O(|C|^2|U|\ln|U|)$;并且,其遗传算法的染色体长度等于条件属性全集所含属性的个数$|C|$. 因此,文献[18]算法的运行时间会随着N的随机性、条件属性个数$|C|$以及样本个数$|U|$的增大而明显增加. 实验发现,在表1的数据集中,除Wine和Hepatitis数据集外,在其他数据集上文献[18]算法的决策识别精度都不高于本文算法,而Wine和Hepatitis数据集的样本个数和条件属性个数都不算大. 因此,在本文进行数据约简以提高决策识别系统时效性的应用背景下,文献[18]算法并不适合,其时间消耗所付出的代价要远大于决策识别精度提高所带来的好处,尤其是在实际工程应用中. 与文献[6]PCA约简方法相比,虽然PCA约简算法运行时间较少,但决策识别精度下降也较多,且PCA的约简方法改变了原始输入属性参数的物理意义,不利于后续决策识别操作的数据分析. 因此,在传感网数据决策识别的应用背景下,本文算法更加适合. 但从表6中不难看出,本文约简算法在算法复杂度方面优势并不十分明显. 虽然本文算法的整体计算量$O(|C|^2|U|\ln|U|)$不由K-means聚类离散化的计算量决定,并且小于基于正域的属性重要度约简算法计算量$O(|C|^2|U|^2)$,但对于数据离散化方法的进一步优化,以及进一步降低算法复杂度,仍是后续研究工作需要解决的问题.

在输入参数不变的情况下,文献[23]的算法和本文算法在其他UCI数据集上的运行结果对比如表7所示.

由表7可以看出,本文算法在决策识别精度方面优于文献[23]的算法. 与表6相比,本文算法在表7中7个数据集上的性能提高更加明显. 这是由于Wine数据集在离散符号数据和空值等方面,其不完备性不如表7的数据集,而本文算法采用K-means聚类离散化方法消除了原始数据集的不完备性,并利用基于熵值的度量方法进行属性约简,使系统决策识别性能明显提高.

表7 不同数据集上决策识别结果对比

数据集	决策识别精度/%	
	文献[23]算法	本文算法
Dermatology	96.08	97.22
Breast cancer	95.86	97.43
Credit approval	82.43	84.69
Vehicle	79.98	83.45
Iono	93.38	95.74
Mushroom	96.42	97.96
Hepatitis	82.52	85.31

4.3 模拟环境测试

为了验证本文算法对实际传感器采集数据处理的有效性,以“车”和“人”两种目标的决策识别为目标,在5 m × 20 m的沿平坦道路范围内布设了2对红外对射传感器、2个震动传感器和2个超声波测距传感器,通过Zigbee通信模块传输到后台数据库进行目标决策识别. 具体测试情况如表8和表9所示.

表8 Zigbee组网测试情况

条目	说明
测试环境	无明显障碍物的平地
测试方案	传感器之间布设距离为20 m
测试结果	组网回传数据成功率为96 %

表9 目标识别测试情况

条目	说明
测试设备	2对红外对射传感器 2个震动传感器 2个超声波测距传感器
测试方案	基础环境设施搭建完毕后, 人或车辆在20 m范围内通过
文献[23]算法测试结果	对人的目标判决进行了100次测试, 决策识别精度为87 %; 对车的目标判决进行了55次测试, 决策识别精度为90.9 %
本文算法测试结果	对人的目标判决进行了100次测试, 决策识别精度为97 %; 对车的目标判决进行了55次测试, 决策识别精度为96.36 %

由表8可以看出,Zigbee组网回传数据成功率为96 %,说明后台数据库中存储的传感网采集数据为不完备数据. 从表9目标识别测试结果可以看出,文献[23]的算法对车的决策识别精度要高于对人的决策

识别精度,其主要原因是由于实际中目标车的多样性和灵活性都不如人,传感器对目标车采集的数据受噪声影响较小,数据的动态变化较人具有更明显的特征. 而本文算法对车和人的决策识别精度不仅都高于文献[23]的算法,并且对车和人的识别精度基本一致,说明本文算法具有较强的抗噪声能力和对不完备数据集的处理能力. 测试结果验证了本文所提出算法的有效性和工程适用性.

5 结 论

为了处理基于传感网数据的决策识别问题,本文针对不完备混合决策系统,首先依据传感器采集的各属性数据本身特点,提出了基于改进K-means聚类算法的数据离散化方法,在消除原始数据集不完备性的同时,减小噪声和孤立点对聚类结果产生的影响. 然后,在归一离散化数据的基础上,提出了基于互信息熵的变精度邻域粗糙集属性约简启发式算法. 该算法引入变精度修正系数,整合变精度粗糙集和邻域粗糙集的优势,在提高决策识别精度的同时降低算法计算复杂度. 仿真结果验证了该算法的有效性,其在决策识别精度和运算时间上都明显优于不进行属性约简的算法.

由于本文算法主要偏重于解决决策识别系统的精确度和时效性问题,所提出的约简算法虽然使决策识别精度得到提高,降低了决策识别系统的整体计算复杂度,但与基于正域的属性重要度约简方法相比,本文约简算法的计算量性能优势不明显. 如何在保证决策识别精度的同时进一步减少计算复杂度,或者实现决策精度与计算复杂度更有效的折中,以及将本文属性约简算法扩展到三支决策,进一步增强算法的实用性,将是后续工作的重点.

参考文献(References)

[1] Ramar C, Rubasoundar K. A survey on data aggregation techniques in wireless sensor networks[J]. Inderscience Publishers, 2011, 6(2): 81-91.

[2] 杨小军,单博炜,梁中华, 等. 具有间歇性观测的无线传感器网络分布式容错目标跟踪[J]. 控制与决策, 2016, 31(6): 1032-1036.
(Yang X J, Shan B W, Liang Z H, et al. Fault tolerant distributed target tracking with intermittent observations in wireless sensor networks[J]. Control and Decision, 2016, 31(6): 1032-1036.)

[3] Ouchi K, Doi M. Indoor-outdoor activity recognition by a smartphone[C]. Proc of the 2012 ACM Conf on Ubiquitous Computing. New York: ACM Press, 2012: 600-601.

[4] 王国胤,于洪,杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 1-8.
(Wang G Y, Yu H, Yang D C. Decision table reduction based on conditional information entropy[J]. Chinese J of Computers, 2002, 25(7): 1-8.)

[5] 王国胤. 决策表核属性的计算方法[J]. 计算机学报, 2003, 26(5): 1-5.
(Wang G Y. Calculation methods for core attributes of decision table[J]. Chinese J of Computers, 2003, 26(5): 1-5.)

[6] 洪斌,邓波,彭甫阳,等. 基于PCA降维的云资源状态监控数据压缩技术[J]. 计算机科学, 2016, 43(8): 19-25.
(Hong B, Deng B, Peng F Y, et al. Data dimension reduction method based on PCA for monitoring data for virtual resources in cloud computing[J]. Computer Science, 2016, 43(8): 19-25.)

[7] 单燕,李玲娟,孙杜静. 基于主成分分析的并行化数据流降维算法研究[J]. 南京邮电大学学报: 自然科学版, 2015, 35(5): 99-104.
(Shan Y, Li L J, Sun D J. Parallel data stream dimensional reduction algorithm based on principal component analysis[J]. J of Nanjing University of Posts and Telecommunications: Natural Science Edition, 2015, 35(5): 99-104.)

[8] Qian Y H, Liang J Y, Pedrycz W, et al. Positive approximation: An accelerator for attribute reduction in rough set theory[J]. Artificial Intelligence, 2010, 174(9): 597-618.

[9] Liang J, Wang F, Dang C, et al. An efficient rough feature selection algorithm with a multi-granulation view[J]. Int J of Approximate Reasoning, 2012, 53(6): 912-926.

[10] 王国胤,姚一豫,于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, 32(7): 1229-1246.
(Wang G Y, Yao Y Y, Yu H. A survey on rough set theory and applications[J]. Chinese J of Computers, 2009, 32(7): 1229-1246.)

[11] 于洪,王国胤,姚一豫. 决策粗糙集理论研究现状与展望[J]. 计算机学报, 2015, 38(8): 1628-1639.
(Yu H, Wang G Y, Yao Y Y. Current research and future perspectives on decision-theoretic rough sets[J]. Chinese J of Computers, 2015, 38(8): 1628-1639.)

[12] Wang C, Shao M, Sun B, et al. An improved attribute reduction scheme with covering based rough sets[J]. Applied Soft Computing, 2015, 26(C): 235-243.

[13] Wang C, He Q, Chen D, et al. A novel method for attribute reduction of covering decision systems[J]. Information Sciences, 2014, 254(5): 181-196.

[14] 唐朝辉,陈玉明. 邻域系统的不确定性度量方法[J]. 控制与决策, 2014, 29(4): 691-695.
(Tang C H, Chen Y M. Neighborhood system uncertainty measurement approaches[J]. Control and Decision, 2014, 29(4): 691-695.)

[15] 徐久成,张灵均,孙林,等. 广义邻域关系下不完备混合决策系统的约简[J]. 计算机科学, 2013, 40(4): 244-

248.
(Xu J C, Zhang L J, Sun L, et al. Reduction in incomplete hybrid decision system based on generalized neighborhood relationship[J]. Computer Science, 2013, 40(4): 244-248.)

[16] Li M Z, Wang G Y. Approximate concept construction with three-way decisions and attribute reduction in incomplete contexts[J]. Knowledge-Based Systems, 2016, 91(5): 165-178.

[17] 李磊军,李美争,谢滨, 等. 三支决策视角下概念格的分析和比较[J]. 模式识别与人工智能, 2016, 29(10): 951-960.
(Li L J, Li M Z, Xie B, et al. Analysis and comparison of concept lattices from the perspective of three-way decisions[J]. Pattern Recognition and Artificial Intelligence, 2016, 29(10): 951-960.)

[18] 马希骜,王国胤,于洪. 决策域分布保持的启发式属性约简方法[J]. 软件学报, 2014, 25(8): 1761-1780.
(Ma X A, Wang G Y, Yu H. Heuristic method to attribute reduction for decision region distribution preservation[J]. J of Software, 2014, 25(8): 1761-1780.)

[19] Ziarko W. Variable precision rough set model[J]. J of Computer & System Sciences, 1993, 46(1): 39-59.

[20] Düntsch I, Gediga G. Uncertainty measures of rough set prediction[J]. Artificial Intelligence, 1998, 106(1): 109-137.

[21] Wang J Y, Zhou J. Research of reduct features in the variable precision rough set model[J]. Neurocomputing, 2009, 72(10/11/12): 2643-2648.

[22] 马超,陈西宏,徐宇亮, 等. 广义邻域粗集下的集成特征选择及其选择性集成算法[J]. 西安交通大学学报, 2011, 45(6): 34-39.
(Ma C, Chen X H, Xu Y L, et al. Ensemble feature selection based on generalized neighborhood rough model and its selective integration[J]. J of Xi'an Jiaotong University, 2011, 45(6): 34-39.)

[23] 贾俊芳,张英. 基于属性重要度的变精度邻域粗糙集知识约简[J]. 山西大同大学学报: 自然科学版, 2014, 30(6): 1-3.
(Jia J F, Zhang Y. Knowledge reduction of variable precision neighborhood rough set based on attribute importance degree[J]. J of Shanxi Datong University: Natural Science, 2014, 30(6): 1-3.)

[24] 杨习贝,杨静宇. 邻域系统粗糙集模型[J]. 南京理工大学学报, 2012, 36(2): 291-295.
(Yang X B, Yang J Y. Rough set model based on neighborhood system[J]. J of Nanjing University of Science and Technology, 2012, 36(2): 291-295.)

[25] 文莹,肖明清,王邑,等. 基于信息熵属性约简的航空发动机故障诊断[J]. 仪器仪表学报, 2012, 33(8): 1773-1778.
(Wen Y, Xiao M Q, Wang Y, et al. Aero-engine fault diagnosis based on information entropy attribute reduction[J]. Chinese J of Scientific Instrument, 2012, 33(8): 1773-1778.)

[26] 陈晓,王新民,黄誉,等. 倾转旋翼机飞控系统的变精度粗糙集-OMELM故障诊断方法[J]. 控制与决策, 2015, 30(3): 432-440.
(Chen X, Wang X M, Huang Y, et al. Fault diagnosis for tilt-rotor aircraft flight control system based on variable precision rough set-OMELM[J]. Control and Decision, 2015, 30(3): 432-440.)

[27] Miao D Q, Zhao Y, Yao Y Y, et al. Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model[J]. Information Sciences, 2009, 179(24): 4140-4150.

[28] Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features[C]. Proc of the 12th Int Conf on Machine Learning. San Francisco: Morgan Kaufmann, 1995: 194-202.

[29] Chmielewski M R, Grzymala-Busse J W. Global discretization of continuous attributes as preprocessing for machine learning[J]. Int J of Approximate Reasoning, 1996, 15(4): 319-331.

[30] 刘文军. 基于粗糙集的数据挖掘算法研究[D]. 北京: 北京师范大学数学系, 2004.
(Liu W J. Research on data mining algorithms based on rough sets[D]. Beijing: Department of Mathematics, Beijing Normal University, 2004.)

[31] 翟东海, 鱼江, 高飞, 等. 最大距离法选取初始簇中心的K-means文本聚类算法的研究[J]. 计算机应用研究, 2014, 31(3): 713-719.
(Zhai D H, Yu J, Gao F, et al. K-means text clustering algorithm based on initial cluster centers selection according to maximum distance[J]. Application Research of Computers, 2014, 31(3): 713-719.)

[32] Huang Z X. Extensions to the K-means algorithm for clustering large data sets with categorical values[J]. Data Mining & Knowledge Discovery, 1998, 2(3): 283-304.

[33] Wu D, Shang M S, Luo X, et al. Self-training semi-supervised classification based on density peaks of data[J]. Neurocomputing, 2018, 275(1): 180-191.

[34] 周炜奔,石跃祥. 基于密度的K-means聚类中心选取的优化算法[J]. 计算机应用研究, 2012, 29(5): 1726-1728.
(Zhou W B, Shi Y X. Optimization algorithm of K-means clustering center of selection based on density[J]. Application Research of Computers, 2012, 29(5): 1726-1728.)

(责任编辑: 李君玲)