Computer Science and Application 计算机科学与应用, 2020, 10(11), 1952-1961
Published Online November 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1011206

Hans汉斯

基于联合粒度属性约简信息损失的研究

薛欢欢[1]，郭　步[1]，隋龙飞[2]，王群笑[1]

[1]嘉兴学院南湖学院，浙江　嘉兴
[2]嘉兴职业技术学院，浙江　嘉兴
Email: 1530043379@qq.com

收稿日期：2020年10月27日；录用日期：2020年11月11日；发布日期：2020年11月18日

摘　要

随着互联网技术的迅速发展，社会进入了大数据时代。数据不仅类型多种多样，结构错综复杂还具有动态变化的特点。如何从海量数据中快速获取有价值的信息是当前亟待解决的问题。粗糙集是一种处理数据不确定性的数据评价方法。属性约简是粗糙集理论的一个重要核心应用。本文将围绕属性约简后信息损失量进行研究，从而找寻一种属性约简算法，在约简后既能保持数据分类准确率较高且信息损失较少。本文借助知识粒度的概念和约简算法，引入联合粒度，并将其运用到属性约简过程，进一步得出基于联合粒度属性约简算法。然后运用其算法对决策表系统进行约简，得出该算法在保持分类准确率不变的情况下，其信息损失量降至较低。最后通过UCI数据集进行仿真实验探究，从而验证了该方法的准确性和有效性。

关键词

属性约简，知识粒度，联合粒度，信息损失

Research on Information Loss of Attribute Reduction Based on Joint Granularity

Huanhuan Xue[1], Bu Guo[1], Longfei Sui[2], Qunxiao Wang[1]

[1]Nanhu College of Jiaxing University, Jiaxing Zhejiang
[2]Jiaxing Vocational and Technical College, Jiaxing Zhejiang
Email: 1530043379@qq.com

Received: Oct. 27th, 2020; accepted: Nov. 11th, 2020; published: Nov. 18th, 2020

Abstract

With the rapid development of Internet technology, the society has entered the era of big data. The

文章引用: 薛欢欢, 郭步, 隋龙飞, 王群笑. 基于联合粒度属性约简信息损失的研究[J]. 计算机科学与应用, 2020, 10(11): 1952-1961. DOI: 10.12677/csa.2020.1011206

data is not only of various types and structures, but also of dynamic change. How to quickly obtain valuable information from massive data is an urgent problem to be solved. Rough set is a data evaluation method to deal with data uncertainty. Attribute reduction is an important core application of rough set theory. This paper will focus on the amount of information loss after attribute reduction, so as to find an attribute reduction algorithm, which can keep the data classification accuracy higher and information loss less after reduction. In this paper, the concept of knowledge granularity and reduction algorithm, the introduction of joint granularity, and its application to the process of attribute reduction, further get the attribute reduction algorithm based on joint granularity. Then the algorithm is used to reduce the decision table system. It is concluded that the information loss of the algorithm is reduced to a low level while the classification accuracy remains unchanged. Finally, the accuracy and effectiveness of this method are verified by the simulation experiment of UCI data set.

Keywords

Attribute Reduction, Knowledge Granularity, Union Granularity, Information Loss

1. 引言

粗糙集理论[1]作为一种数据分析方法，在定性和定量分析的基础上用来描述决策系统中的不确定性。属性约简[2] [3] [4] [5]，也称特征选择，是粗糙集理论的重要核心应用。它是从信息系统的所有属性中选择部分重要度较高的属性，同时不减少数据分类的准确性。属性约简已被广泛应用于许多研究领域，如模式识别、数据挖掘与机器学习等。在数据挖掘[6]领域中，属性约简是一种常用的数据处理工具，主要对数据中的冗余属性进行删除操作，使其达到更高的分类性能且不损失数据表的任何信息。

前期学者们提出了许多属性约简准则和约简算法，几乎所有的约简准则都宣称保持分类不变、信息无损失。然而，本文从信息熵角度[6]分析，定量分析删除的冗余属性可能含有一定的信息量，可以得出属性约简可能会导致信息的损失。为了探究不同约简准则信息损失量的差异，提出联合粒度属性约简算法，同时对比了正区域约简、知识粒度约简和该算法的约简准则对数据分类、信息损失量的影响，从而验证此算法的有效性。

2. 基础知识

给定一个信息系统[1] $IS=(U,A)$ 中，U 是论域，A 是论域 U 上的条件属性集。在条件属性集中任取条件属性 $m\in A$ 都存在一个函数 $m:U\to F_m$，与其相对应，F_m 为属性 m 的值域。U 中每个元素称为个体、对象或行。

定义 2.1 [5]对于任意的属性子集 $B\subseteq A$ 和任何个体 $x\in A$ 都对应着如下的信息函数：

$$Inf_B(x)=\{(a,a(x)):a\in B\}$$

B-不分明关系(或称为不可区分关系)定义为:

$$IND(B)=\{(x,y):Inf_B(x)=Inf_B(y)\}$$

任何满足 $IND(B)$ 的 2 个元素 x, y 都不能由 B 的任何子集区分，$[x]_B$ 表示由 x 引导的 $IND(B)$ 等价类。

2.1. 属性约简

定义 2.2 [6]在决策系统 $DS=(U,C,d)$ 中，$B\subseteq C$ 是 DS 的基于正区域的相对约简当且仅当 $B\subseteq C$ 满足以下两个条件：

(1) $POS_B(d)=POS_C(d)$；

(2) 对于任意的 $a\in B$，都有 $POS_{B-\{a\}}(d)\neq POS_C(d)$.

定义 2.3 [6]给定一个决策系统 $DS=(U,A,d)$，$B\subseteq A$ 称为该决策系统 DS 的一个基于条件熵的相对约简当且仅当 $B\subseteq A$ 满足如下两个条件：

(1) $H(DS,\{d\}\mid B)=H(DS,\{d\}\mid A)$；

(2) 对任意的 $S\subset B$，均都有 $H(DS,\{d\}\mid S)\neq H(DS,\{d\}\mid A)$。

2.2. 知识粒度的基本概念

定义 2.4 [7]给定一个信息系统 $IS=(U,A,V,f)$，对于任意属性子集 $M\subseteq A$ 导出的划分 $U/M=\{R_1,R_2,R_3,\cdots,R_n\}$，则该信息系统对于属性子集 M 的知识粒度的定义为

$$GD(M)=\sum_{i=1}^{n}\frac{|R_i|^2}{|U|^2}$$

定义 2.5 [7]设给定一个信息系统 $IS=(U,A,V,f)$，如果存在属性子集 $R_1,R_2\subseteq A$，则有属性子集 R_1 相对于子集 R_2 的知识粒度的定义为

$$GD(R_1\mid R_2)=GK(R_2)-GK(R_1\cup R_2)$$

定义 2.6 [7]对于给定决策系统 $DS=(U,A\cup D,V,f)$，若属性子集 $R\subseteq A$ 是当且仅当满足以下两个条件：

(1) $GD(D\mid A)=GK(D\mid R)$；

(2) 对于任意的 $m\in R$，$GK(D\mid R-\{m\})\neq GK(D\mid R)$。

称 $R\subseteq A$ 为该决策系统的一个知识属性约简。

定义 2.7 [8]在给定的信息系统 $IS=(U,A\cup D,V,f)$ 中，若存在属性子集 $c\in A$，则属性 $c\in A$ 与决策属性 D 的相似度定义为：

$$s(c,D)=\frac{\left|IND(D\cup\{c\})\right|}{\sqrt{\left|IND(c)\right|}\cdot\sqrt{\left|IND(D)\right|}}$$

2.3. 知识粒度的启发式属性约简算法[8]

根据知识属性约简的定义可得出一个属性约简算法，算法具体步骤如表 1。

目前很多约简算法均要先求核属性，但该算法不需要。此算法在一定程度上减少了求核属性的工作量，同时时间复杂度相对缩减。

3. 联合粒度属性约简算法

受表 1 算法的启发，以知识粒度[7]作为启发式函数进行遍历搜索，选取知识粒度值的大小进行比较，得出一个新算法—基于联合粒度属性约简算法。算法的具体描述如表 2 所示。

Table 1. Heuristic attribute reduction algorithm based on knowledge granularity [8]
表 1. 基于知识粒度的启发式属性约简算法[8]

输入：一个简单决策表 $DT=(U,C\cup D,V,f)$。
输出：条件属性 C 相对于决策属性 D 的一个相对约简 $B\in RED_D(C)$。

1. 计算每个条件属性与决策属性的相似度 $s(c_i,D)$。
2. 然后根据每个条件属性与决策属性的相似度的值的大小将条件属性集进行降序排列 $C=\{c_1,c_2,\cdots,c_n\}$,其中n是条件属性的个数。
3. 对属性约简 $RED_D(C)$ 赋初始值 $RED_D(C)=C$，冗余属性 $C_p=\varnothing$，对此，检验 $C_1=\varnothing$，然后对集合C中的每个条件属性属性 c_i，令 $C_m=RED_D(C)-c_i$，执行下述操作：
3.1. 计算每个条件属性属性 c_i 与集合 C_m 中条件属性 c_j 的相似度 $s(c_i,c_j)$，然后与前面的 $s(c_i,D)$ 进行比较，若 $s(c_i,D)\le s(c_i,c_j)$，则将条件属性 c_j 赋值给 C_p，将集合 $RED_D(C)-c_j$ 赋值给 C_1；
3.2. 接下来计算 $s(C_1,D)$ 是否与 $s(C,D)$ 相等，如果相等，则删除条件属性 c_j，将集合 C_1 赋给 $RED_D(C)$，反之将保留 c_j，$RED_D(C)$ 保持不变；
3.3. 如果 C_m 中的条件属性没有完成遍历，则直接转至步骤3.1，若完成遍历，则令 $i=i+1$，$C_m=RED_D(C)-c_i$，如果 $C_m=\varnothing$，转步骤4，否则，转步骤3.1；
4. 算法结束，输出相对属性约简集合 $RED_D(C)$。

Table 2. Attribute Reduction Algorithm Based on Joint Granularity
表 2. 基于联合粒度属性约简算法

输入：一个完备的决策信息系统 $DT=(U,A\cup\{d\},V,f)$
输出：属性约简结果 $P=RED_A$

1. 初始化 $RED_A=\varnothing$；
2. 根据知识粒度的定义，计算所有剩余属性的知识粒度值 $GK(A_i)$；
3. If $GK(A_i)<1$ then
4. 寻找第 2 步中的知识粒度最小的属性，记为 $\min(A_i)$；
5. $RED_A=RED_A\cup\min(A_i)$；
6. 跳入第 2 步；
7. Else 跳转步骤 10；
8. End if
9. 返回属性约简 RED_A。算法结束。

基于联合粒度属性约简算法不同于上述表 1，它主要从知识属性、粒计算[9]角度进行约简，可以直接对不同知识属性进行粒度和重要度的比较，同时使得约简过程更加精细，约简结果更加精确。

属性约简信息损失[9]

定义 3.1 [9] [10] [11]在一个给定的决策系统 $DS=(U,A,d)$ 中，设 A 和 $\{d\}$ 在论域 U 上导出的划分分别为 X 和 Y，其中 $X=U/A=\{X_1,X_2,\cdots,X_N\}$，$Y=U/\{d\}=\{Y_1,Y_2,\cdots,Y_M\}$，$p(X_i)=\dfrac{|X_i|}{|U|}$，$p(Y_j)=\dfrac{|Y_j|}{|U|}$，$p(X_i,Y_j)=\dfrac{|X_i\cap Y_j|}{|U|}$，$p(Y_j\mid X_i)=\dfrac{|X_i\cap Y_j|}{|X_i|}$，$i=1,2,\cdots,N$，$j=1,2,\cdots,M$．$A$ 和 $\{d\}$ 的信息熵[12]分别可定义为：

$$H(DS,A)=-\sum_{i=1}^{N}p(X_i)lbp(X_i)$$

$\{d\}$ 与 A 的联合熵[13]即可定义为：

$$H\left(DS,A,\{d\}\right)=-\sum_{i=1}^{N}\sum_{i=1}^{M}p\left(X_i,Y_j\right)lbp\left(X_i,Y_j\right)$$

定义 3.2 [9] [10] [11]给定 $DS=(U,A,d)$ 是一个决策系统，$B\subseteq A$ 是该决策系统的一个约简，则 $B\subseteq A$ 属性约简的信息损失定义如下：

$$\Delta(B)=H\left(DS,A\right)-H\left(DS,B\right)$$

$B\subseteq A$ 属性约简的信息损失率可定义如下：

$$s(B)=\frac{\Delta(B)}{H\left(DS,A\right)}\times100\%=1-\frac{H\left(DS,B\right)}{H\left(DS,A\right)}\times100\%$$

命题 1 给定一个决策系统 $DS=(U,A,d)$，若 $B_1\subseteq A$ 是基于正区域的相对约简，$B_2\subseteq A$ 是基于知识粒度的属性约简，$B_3\subseteq A$ 是基于联合粒度的属性约简，则有 $\Delta(B_1)\geq\Delta(B_2)\geq\Delta(B_3)$.

证明：根据知识粒度约简、联合粒度约简及正区域相对约简可证。

例 1 下表所示 $DS_1=(U,C\bigcup D,V,f)$ 是一个决策表，如表 3 所示。U 为论域，其中 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$，$C=\{e_1,e_2,e_3,e_4\}$ 是条件属性，$D=\{d\}$ 是决策属性集。

Table 3. Decision System $DS_1=(U,A,d)$

表 3. 决策系统 $DS_1=(U,A,d)$

U	e_1	e_2	e_3	e_4	d
x_1	0	1	2	1	0
x_2	0	2	1	1	1
x_3	1	0	1	1	0
x_4	1	0	1	1	1
x_5	1	1	1	1	0
x_6	1	1	1	1	1
x_7	1	1	0	1	0
x_8	1	1	0	0	1

$$U/C=\left\{\{x_1\},\{x_2\},\{x_3,x_4\},\{x_5,x_6\},\{x_7\},\{x_8\}\right\}$$

$$U/\{d\}=\left\{\{x_1,x_3,x_5,x_7\},\{x_2,x_4,x_6,x_8\}\right\}$$

由正区域的定义可得：$POS_C(d)=\{x_1,x_2,x_7,x_8\}$。

根据条件属性依赖度的计算公式可得：$\gamma_C(D)\neq1$，所以该决策表是不相容决策表。

(1) 根据知识粒度的属性约简算法可得约简后为：$R_1=RED_C=\{e_2,e_3,e_4\}$.

根据属性约简的信息损失的计算方法可得：

$$H\left(DS_1,C\right)=-\sum p\left(X\right)lbp\left(X\right)=2.5$$

$$H\left(DS_1,R_1\right)=-\sum p\left(X\right)lbp\left(X\right)=2.5$$

故 R_1 的信息损失量为：$\Delta(R_1)=0$。

(2) 由正区域约简算法可得：$R_2=\{e_2,e_3,e_4\}$，$R_3=\{e_1,e_3,e_4\}$

$$H\left(DS_1,R_3\right)=-\sum p\left(X\right)lbp\left(X\right)=2$$

故 R_2 的信息损失量为 0。

R_3 的信息损失量为 $\Delta\left(R_3\right)=0.5$。

(3) 根据联合粒度属性约简算法可得：$R_4=\{e_2,e_3,e_4\}$

故 R_4 的信息损失量为 0。

例 2 下表所示为一个决策表，$DS_2=(U,A,d)$ 是一个决策系统，如表 4 所示。U 为论域，$A=\{\alpha_1,\alpha_2,\alpha_3,\alpha_4\}$ 是条件属性，$\{d\}$ 是决策属性集。

Table 4. Decision system $DS_2=(U,A,d)$

表 4. 决策系统 $DS_2=(U,A,d)$

U	a	b	c	e	d
e_1	0	1	0	1	0
e_2	1	1	0	0	1
e_3	1	1	0	1	1
e_4	0	1	1	1	0
e_5	0	2	1	0	1
e_6	1	2	0	1	0

$$U/A=\{\{e_1\},\{e_2\},\{e_3\},\{e_4\},\{e_5\},\{e_6\}\},\ U/\{d\}=\{\{e_1,e_4,e_6\},\{e_2,e_3,e_5\}\},$$

由正区域的定义可得：$POS_A\left(d\right)=\{e_1,e_2,e_3,e_4,e_5,e_6\}$。

根据条件属性依赖度计算公式得：$\gamma_A\left(D\right)=1$，所以该决策表是相容决策表。

$$H\left(DS_2,A\right)=-\sum p\left(X\right)lbp\left(X\right)=2.585$$

(1) 基于正区域的相对约简：$S_1=\{a,b\}$

$$H\left(DS_2,S_1\right)=-\sum p\left(X\right)lbp\left(X\right)=1.918$$

故 S_1 的信息损失量为：$\Delta\left(S_1\right)=0.667$。

(2) 基于知识粒度的属性约简：$S_2=RED_A=\{a,b,e\}$

$$H\left(DS_2,S_2\right)=-\sum p\left(X\right)lbp\left(X\right)=2.252$$

故 S_2 的信息损失量为：$\Delta\left(S_2\right)=0.333$。

(3)基于联合粒度的属性约简：$S_3=\{a,b,e\},S_4=\{a,b,c\}$

$$H\left(DS_2,S_4\right)=-\sum p\left(X\right)lbp\left(X\right)=2.252$$

故 S_3,S_4 的信息损失量为：$\Delta\left(S_3\right)=\Delta\left(S_4\right)=0.333$。

4. 仿真实验分析

本节将采用一系列数据集进行验证本文所提出的算法的高效性和准确性。首先将本文提出的联合粒度属性约简算法与知识粒度属性约简算法、正区域相对约简算法[14]对同一组 UCI 数据集进行离散化处理，然后比较三种算法属性约简后的信息损失量从而验证本文算法的有效性。其实验平台的硬件环境为

CPU Intel core i5 和 4 GB 内存，操作系统为 Windows 10 专业版，实验所运用的编程工具为 VC++6.0。

本次实验数据取自 UCI 数据集中的四组典型数据如表 5 所示。首先将每组数据集分成 10 等份，选择 1 份作为测试集，其他 9 份作为训练集，对这 9 份训练集进行联合粒度、知识粒度和正区域约简，然后比较它们约简后的信息损失量。

Table 5. Data set description
表 5. 数据集描述

序号	数据集名称	属性	对象
1	Dermatology	34	366
2	Contraceptive Method Choice	9	1473
3	Mushroom	22	8124
4	Letter Recognition	16	20000

文中采用重庆邮电大学开发的 RIDAS 系统进行属性约简。使用正区域约简算法、知识粒度约简算法和联合粒度约简算法求取不同属性约简集合。根据信息熵的计算公式，根据已有计算信息熵的程序代码[9]，计算给定信息系统的条件属性集合的信息熵及每个约简后的信息熵，可得出不同约简准则的信息损失量的大小以及信息熵与信息损失率之间的关系。

图 1~4 分别表示四组数据集的约简准则与信息损失量仿真实验结果图，其中横坐标表示属性约简结果，纵坐标表示信息损失量大小。图 5~8 分别表示四组数据集约简后集合的信息熵与信息损失率仿真实验图，其中横坐标表示约简后信息熵，纵坐标表示信息损失率大小。

根据上述 8 张图示可以清楚地看出，本文提出的基于联合粒度属性约简的信息损失远小于基于正区域、基于知识粒度属性约简信息损失，从而说明本文提出的基于联合粒度属性约简算法的有效性，可以有效地解决大数据时代下海量数据的约简，减少约简导致的分类准确率较低、信息损失量较大的问题。

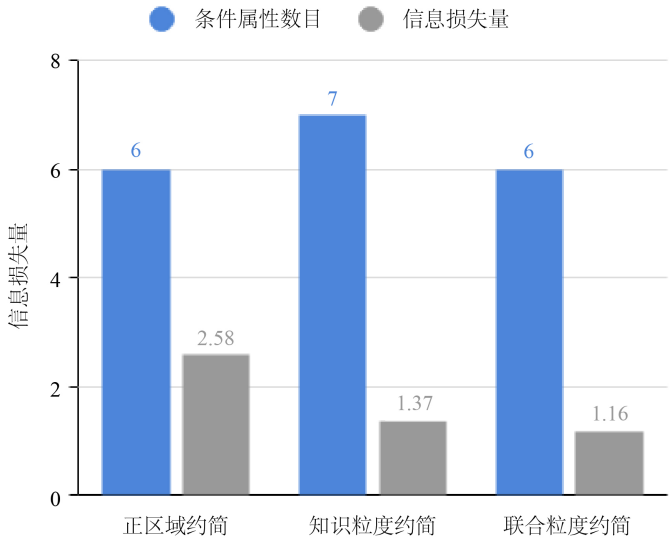

Figure 1. Comparison of dermatology
图 1. Dermatology 比较

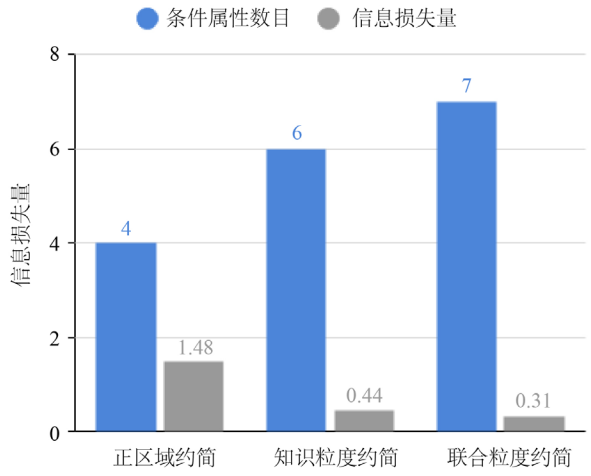

Figure 2. Comparison of conceptual method choices
图 2. Contraceptive Method Choice 比较

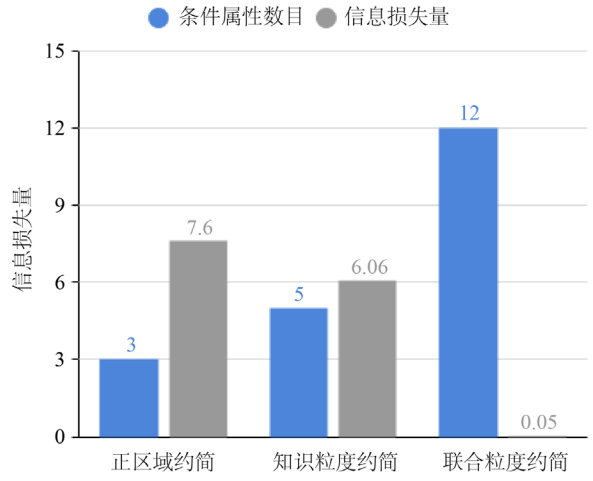

Figure 3. Comparison of mushroom
图 3. Mushroom 比较

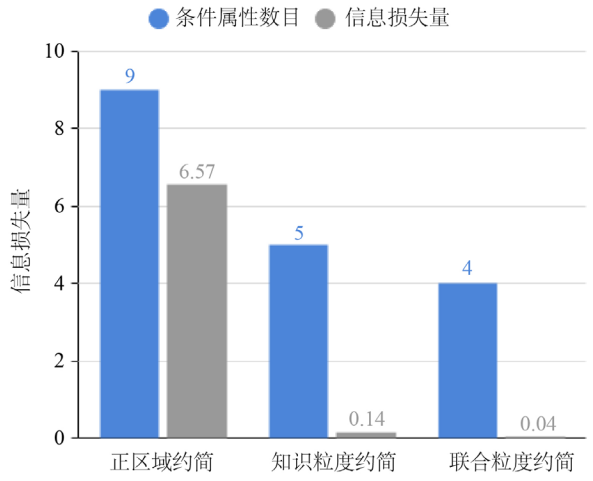

Figure 4. Comparison of letter recognition
图 4. Letter Recognition 比较

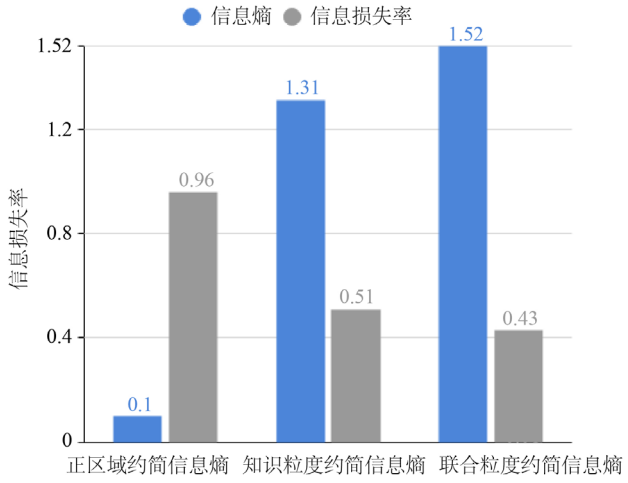

Figure 5. Analysis of dermatology
图 5. Dermatology 分析

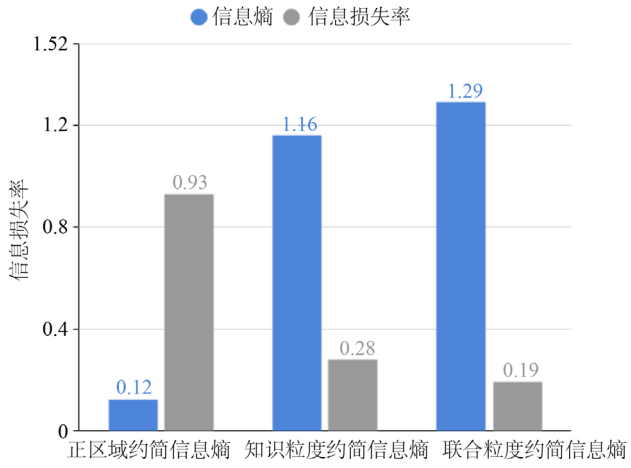

Figure 6. Analysis of conceptual method choice
图 6. Contraceptive Method Choice 分析

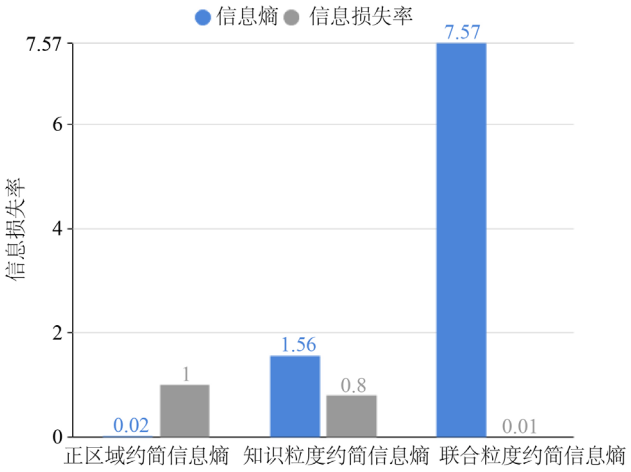

Figure 7. Analysis of mushroom
图 7. Mushroom 分析

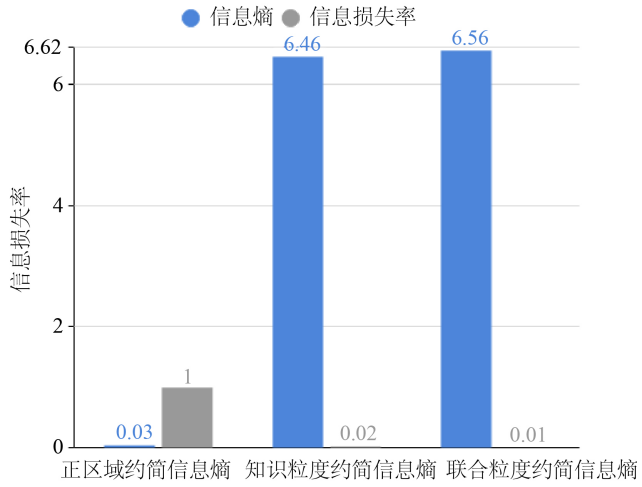

Figure 8. Analysis of letter recognition
图 8. Letter Recognition 分析

5. 结论

本文通过引入联合粒度的概念，并将其运用到粗糙集理论的属性约简[15]中，提出了基于联合粒度属性约简算法，然后分析比较不同的属性约简产生的信息损失，最终得到本文的基于联合粒度的属性约简不仅能保持数据分类的准确性，同时也能维持约简后的信息损失较低的结论。后续我们将继续研究该算法对带权决策表的约简信息损失的影响及进一步的优化。

参考文献

[1] Pawlak, Z. (1982) Rough Sets. *International Journal of Computer and Information Sciences*, **11**, 341-356. https://doi.org/10.1007/BF01001956

[2] Hobbs, J.R. (1985) Granularity. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, Los Angeles, 432-435.

[3] Lin, T.Y. (1997) Granular Computing. Announcement of the BASIC Special Interest Group on Granular Computing.

[4] Zhang, C.C. (2020) Knowledge Granularity Based Incremental Attribute Reduction for Incomplete Decision Systems. *International Journal of Machine Learning and Cybernetics*, **11**, 1141-1157. https://doi.org/10.1007/s13042-020-01089-4

[5] 李旭, 等. 带权决策表的属性约简[J]. 计算机工程与应用, 2020, 56(12): 54-59.

[6] 大数据背景下粗糙集属性约简研究进展[J]. 计算机工程与应用, 2019, 55(6): 31-38.

[7] 基于知识粒化的信息系统增量式属性约简[J]. 模式识别与人工智能, 2019, 38(8): 31-38.

[8] 一种基于知识粒度的启发式属性约简算法[J]. 计算机工程与应用, 2012, 48(36): 31-38.

[9] 邓大勇, 薛欢欢, 苗夺谦, 卢克文. 属性约简准则与约简信息损失的研究[J]. 电子学报, 2017, 45(2): 401-407.

[10] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001.

[11] 腾书华. 基于粗糙集理论的不确定性度量和属性约简方法研究[D]: [博士学位论文]. 长沙: 国防科学技术大学, 2010.

[12] 桑妍丽, 钱宇华. 多粒度决策粗糙集中的粒度约简方法[J]. 计算机科学, 2017, 44(5): 199-205.

[13] 桑妍丽, 钱宇华. 一种悲观多粒度粗糙集中的粒度约简算法[J]. 模式识别与人工智能, 2012, 25(3): 361-366.

[14] 邓大勇, 黄厚宽. 多粒度粗糙集的双层绝对约简[J]. 模式识别与人工智能, 2016, 29(11): 969-975.

[15] 苗夺谦, 李道国. 粗糙集理论、算法与应用[M]. 北京: 清华大学出版社, 2008: 4.