# 运用信息增益和不一致度进行填补的属性约简算法

# 李虹利 蒙祖强

(广西大学计算机与电子信息学院 南宁 530004)

摘 要 针对不完备、不一致性数据的属性约简是数据挖掘研究的一个重要内容。将信息增益,不一致度相结合,提出一种面向不完备不一致性数据的属性约简算法。首先,介绍了信息增益,定义了不一致度的概念与算法公式,并给出了基于二者对数据进行填补的方法;然后,基于该填补方法,以最大不一致度条件下的信息增益为权值,以不一致度为属性约简的启发信息,给出属性约简算法;最后,通过实验证明了所提算法的有效性。

关键词 信息增益,填补,属性约简,不一致性,不完备

中图法分类号 TP181

文献标识码 A

**DOI** 10. 11896/j. issn. 1002-137X, 2018, 10, 040

# Attribute Reduction Algorithm Using Information Gain and Inconsistency to Fill

LI Hong-li MENG Zu-qiang

(College of Computer and Electronic Information, Guangxi University, Nanning 530004, China)

Abstract The attribute reduction of incomplete and inconsistent data is a major content of data mining. Combining information gain and inconsistent degree of data, this paper proposed an attribute reduction algorithm for incomplete and inconsistent data. First, the information gain is introduced, and the concept and algorithm formula of inconsistent degree are defined. Besides, the method of data filling based on information gain and inconsistent degree is given. Then, based on this data filling method, the attribute reduction algorithm is provided with the information gain under the condition of taking the maximum inconsistent degree as the weight and inconsistent degree as heuristic information. Finally, the experimental results demonstrate the effectiveness of the proposed algorithm.

Keywords Information gain, Filling, Attribute reduction, Inconsistent, Incomplete

# 1 引言

Pawlak于 1982 年提出的粗糙集理论[1],能有效地解决不确定、不完整知识的表达和推理,是当前人工智能领域内持续研究的热点之一,广泛应用于机器学习和决策分析等领域。

随着信息技术的快速发展,数据信息量呈指数增长,这就很容易产生许多不完备和不一致性数据,不完备数据包含缺失值或者丢失值[2],缺失值给数据分析和处理带来了极大的挑战,其使得原始数据不能提供完整的知识获取和知识表示,因此无法准确地得到先验知识和分类信息,数据分类和处理的准确性也由此受到较大的影响。造成数据丢失或缺失的主要原因有:数据获取时的人为疏忽、人工遗漏或人工丢失、数据文本的损坏和存储介质被破坏等。因此,在数据预处理中大多数都是利用粗糙集理论和统计学理论对不完备数据进行填补或者删除。相关科研成果主要有:文献[3]提出了新型关系矩阵的填补算法;文献[5]提出了基于属性重要度的填补算法;文献[6]提出了一种建立相

似关系并用最大相似度的方法得到最佳相似个体的填补算 法;文献[8]提出了一种通过计算不完备数据整体对象的相异 程度并结合其聚类的结果进行填补的算法;文献「97基于马氏 距离和信息熵的概念来计算最近邻基因的权值,从而得到缺 失值的填补算法;文献「10]提出了新的最近邻填补算法;文献 [11]提利用属性约简并结合改进的相似度与概率填充方法对 不完备数据进行填充的算法;文献[12]在数据集中用各属性 之间的关系,并结合信息增益的数据填充方法,考虑了各个属 性之间的关系但是没有考虑到对象关于属性集的不一致度的 关系;文献[13]计算包含缺失值的属性的信息增益,并将其作 为权值来预测缺失值的可能值,再以最大概率选择填补的缺 失值。由于在计算包含缺失值的属性的信息增益时,可能得 到不完全准确的信息增益,因此结合上述研究进行优化。综 合考虑信息增益和不一致度对数据本身的影响后,提出结合 信息增益和不一致度的填补算法(Filling Algorithm Combining Information Gain and Inconsistent Degree, IGIDFA).

对于属性约简,众多学者在粗糙集的基础上进行研究和拓展:Kryszkiewicz<sup>[14]</sup>提出了一种根据容差关系进行属性约减的算法;王国胤<sup>[15]</sup>认为 Kryszkiewicz 提出的容差关系是建

到稿日期: 2017-08-10 返修日期: 2017-11-16 本文受国家自然科学基金项目(61762009,61363027),广西自然科学基金项目(2015GXNSFAA139292)资助。

**李虹利**(1990-),男,硕士,主要研究方向为数据挖掘、机器学习;**蒙祖强**(1974-),男,博士,教授,主要研究方向为人工智能、数据挖掘与知识发现、智能决策、智能信息处理,E-mail;mengzuqiang@163.com(通信作者)。

立在未知值等于其他任意属性值之上的,这会让两个个体在 未明确同一属性的条件下被误判在同一个类别中,因此他提 出了一种改进的限制容差关系,并由此提出属性约减算法;付 昂等[16] 用相容类划分条件属性,用模糊近似集来决策分类, 先得到相容类划分,再计算条件信息熵,根据删除属性前后的 信息熵值大小是否相等来进行属性约简,适用于小规模和大 规模数据集的属性约简;陶志等[17]提出了决策属性支持度的 定义,并求出相对核属性,然后将核属性放入到遗传算法的初 始种群中,加入惩罚函数以保证最佳搜索效果,再通过选择运 算和变异运算,结合最优个体的保护,最后得到相对属性约 简。然而这些算法都没有对数据关于属性的不一致性程度进 行研究。数据的矛盾、不相容性反映了数据的不一致性,这对 属性约简而言是需要重点关注的。因此,文中提出了一种结 合信息增益和不一致度的约简算法 IGIDRA (Reduction Algorithm Based on Information Gain and Inconsistency Degree)。

## 2 相关概念

本节主要介绍了信息增益和粗糙集正域等概念<sup>[11,14,18-20]</sup>,并定义了不一致度和相应的算法公式。

定义 1 一个决策系统 DS = (U,A,V,f),其中,U为非空对象集,也称为论域;A是属性集合, $A = C \cup D$ ,C是非空条件属性集,D是非空决策属性集, $C \cap D = \emptyset$ ; $V = \bigcup_{a \in A} V_a$ , $V_a$ 是属性a的值域; $f:U \times A \rightarrow V$ ,是一个信息函数,用于为每个对象映射属性值。对于  $\forall a \in A$ , $x \in U$ ,有 f(x,a) = \*,则称决策系统为不完备决策系统,否则称为完备决策系统。

定义 2 一个不完备决策系统 DS = (U, C, D, V, f),令  $B \subseteq C$ ,在 U 上的容差关系 T(B)为:  $T(B) = \{(x,y) \in U \times U \mid \forall a \in B, f(x,a) = f(y,a) \lor f(x,a) = * \lor f(x,a) = * \rbrace$ ,  $H_B(x) = \{y \in U \mid (x,y) \in T(B)\}$ ,  $T_B(x)$ 为对象 x 在属性集 B下的相容类。

定义 3 一个决策系统 DS = (U, C, D, V, f),令  $X \subseteq U$ ,  $B \subseteq C \cup D, X$  的上近似集为  $: B^{-}(X) = \{x \in U \mid T_{B}(x) \cap X \neq \emptyset\}$ , X 的下近似集为  $: B_{-}(X) = \{x \in U \mid T_{B}(x) \subseteq X\}$ 。

定义 4 一个决策系统 DS = (U, C, D, V, f),D 的 C 正域记为: $POS_C(D) = \bigcup_{X \in U(D)} C_-(X)$ 。

定义 5 一个决策系统 DS = (U, C, D, V, f),对于属性  $a \in C$ ,如果有  $POS_B(D) = POS_{B-\{a\}}(D)$ ,则称属性 a 关于子 集 B 是冗余的;对于属性子集,令  $B \subseteq C$ ,如果有  $POS_B(D) = POS_C(D)$ ,且  $\forall B' \subseteq C$ ,  $POS_B'(D) \neq POS_C(D)$ ,则称 B 是 C 的一个属性约简。

定义 6 一个决策系统 DS = (U, C, D, V, f), 若决策属性集 D 仅有一个属性 d,令  $D = \{d_1, d_2, d_3, \cdots, d_k\}$ ,  $d_i$  为属性 d 的第 i 个取值,  $p_i$  为  $d_i$  在对象集 U 中出现的概率,即  $p_i = \frac{|d_i|}{|U|}$ ,则对象集 U 所包含的信息熵定义为:

$$Entropy(U) = -\sum_{i=1}^{k} p_{i} \log_{2} p_{i}$$
 (1)

定义 7 一个决策系统 DS = (U, C, D, V, f), 若  $a \subseteq C$ , 属性 a 有 m 个非重复的取值, 属性 a 将对象集 U 划分为 m 个对象子集 $\{U_1, U_2, U_3, \cdots, U_m\}$ , 对象集 U 按属性 a 划分的信息

增益为:

$$Gain(U,a) = Entropy(U) - Entropy_a(U)$$

$$= Entropy(U) - \sum_{i=1}^{m} \frac{|U_i|}{|U|} Entropy(U_i)$$
 (2)

定义 8 一个完备决策系统 DS = (U, C, D, V, f),令  $\forall x \in U, B \subseteq C$ ,对象 x 关于属性集 B 的不一致度  $\Gamma(x)$ 和所有对象关于属性集 B 的不一致度之和  $S\Gamma$  为:

$$\Gamma(x) = \frac{|POS_B(x) \cap T_D(x)|}{|POS_B(x)|} * W_B(x)$$
(3)

$$S\Gamma = \sum_{i=1}^{n} \frac{|POS_B(x_i) \cap T_D(x_i)|}{|POS_B(x_i)|} * W_B(x_i)$$
(4)

$$WB(x) = \sum_{i=1}^{k} W_{B_i}(x), W_B(x_i) = \sum_{i=1}^{m} W_{B_i}(x_i)$$
 (5)

当  $W_B(x)$ <1 时:

$$W_B(x) = \frac{W_B(x)}{|B|} \tag{6}$$

同理,当 $W_B(x_i)$ <1时:

$$W_B(x_i) = \frac{W_B(x_i)}{|B|} \tag{7}$$

本文将对象关于属性集的不一致度统一简称为不一致度。其中, $\Gamma(x)$ 的取值在[0,1]之间,若对象 x 在属性集 B 下是完备一致的,则  $\Gamma(x)=1$ ; 若对象 x 在属性集 B 下是完备不一致的,则  $0 < \Gamma(x) < 1$ ; 若对象 x 在属性集 B 下是不完备不一致的,则  $0 < \Gamma(x) < 1$ 。  $S\Gamma$  的取值在[1,|U|]之间,若对象集 U 在属性集 B 下是完备一致的,则  $S\Gamma=|U|$ ;若对象集 U 在属性集 B 下是完备一致的,则  $S\Gamma<|U|$ ;若对象集 U 在属性集 B 下是完备不一致的,则  $S\Gamma<|U|$ ;若对象集 U 在属性集 B 下是不完备不一致的,则  $1 < S\Gamma<|U|$ 。 |U|为对象集 U 的大小。

 $|POS_B(x) \cap T_D(x)|$  为对象 x 关于属性集 B 的正域和对象 x 关于决策属性集 D 的正域的交集的个数, $W_B(x)$  为对象 x 关于属性集 B 的权值乘积,当  $W_B(x)$  <1 时意味着该对象 x 有缺失值,所对应的权值应该为  $W_B(x)$ /属性集 B 的大小,因为其每个属性要共同分担缺失值的权值大小,当  $W_B(x)$ =1 时表明此时对象 x 不存在缺失值,|B| 为属性集 B 的大小。 $|POS_B(x_i) \cap T_D(x_i)|$  为对象  $x_i$  关于条件属性集 B 的正域和对象  $x_i$  关于决策属性集 D 的正域的交集的个数, $|POS_B(x_i)|$  为对象  $x_i$  关于条件属性集 B 的正域的个数, $W_B(x_i)$  为对象  $x_i$  关于条件属性集 B 的权值乘积。

定义 9 一个不完备决策系统 DS = (U, C, D, V, f),令  $\forall b \in C, x \in U, f(x,b) = *$ ,整个决策系统的缺失比率  $Miss-Ratio = \frac{\Sigma |f(x,b)|}{|U||C|}$ ,决策系统的不一致率  $T = \frac{|\Gamma(x)|}{|U|}$ ,其中  $|\Gamma(x)|$ 为  $\Gamma(x)$  < 1 的个数,|U| 和 |C| 分别为对象集的大小和条件属性集的大小, $\Sigma |f(x,b)|$  为对象 x 所有缺失属性值的个数之和。

#### 3 不完备数据的填补方法

数据的不完备性导致了不能准确、完整地反映数据本身的特性,无论使用监督学习还是无监督学习的方法进行各种统计学的知识分析,都只能得到片面的、相对有效的信息;而人为进行的数据填补只能相对无限地接近原数据本身,并不能正确地代表原数据本身;数据的不一致性也给属性约简的研究带来了极大的困扰和阻碍。因此,文中提出一种对不完

备不一致数据的填补算法,使填补后的数据能保持原数据本身的不一致性特征,更有利于属性约简。

本文提出的数据填补算法和传统的数据填补算法的出发点不同,传统的数据填补算法是为了使不完备数据填补后变成完备数据,让不一致性数据,填补后尽可能地变成一致性数据,以提高数据的分类准确率。这样使用传统的数据填补算法可能会导致对数据集的属性进行过多或过少的约简,进而影响不完备不一致性数据属性约简的整体性能。而本文对数据进行填补是为了使不完备不一致性数据填补之后更接近原数据本身,即让不一致性数据填补后也能继续保持数据不一致性的变化和趋势,因此对不完备不一致性数据的属性约简并不会造成影响。

IGIDFA 算法主要用于计算填补每列属性缺失值后的信息增益和不一致度,在快速排序后通过比较选取最佳的属性填补值,再更新缺失值所对应的权值。采用快速排序算法,是因为其平均时间复杂度较好;信息熵是度量属性的信息量,而信息增益是度量信息熵的减少量,信息熵越小,对象集对决策属性的分布越纯;信息增益越大,则属性确定性越大,划分的对象子集更纯,更能提高其分类能力。不一致度反映了数据集对象的不一致性程度,而不一致度之和反映了整体对象集不一致性程度的高低,因此在填补数据时考虑每列属性的信息增益和不一致度之和非常重要。IGIDFA 算法的步骤如算法1所示。

# 算法 1 IGIDFA 算法

输入:决策系统 DS=(U,C,D,V,f)

输出:新的决策系统 DS

Step1 初始化新决策系统 DS=Ø,信息增益数组 InfoGain[j][k]=0,k 为第 j 列不同属性值的个数,不一致度数组 AiiArray[j] [k]=0 和权值数组 Weigh[i][j]=1,将对象 x 属性缺失值的下标添加到缺失值数组 MissArray[j][m]中,其中 m 为第 j 列属性缺失值的个数;

Step2 遍历条件属性集  $C_j$ ,令对象 x 第 j 列的属性缺失值 AttValue 为  $C_{jk}$ ,此时计算第 j 列的信息增益 InfoGain[j][k] 和整个对象集 X 的不一致度之和 AiiArray[j][k];

Step3 对 InfoGain 和 AiiArray 数组进行快速排序,得到填补第 j 列 的 k 个不同属性值时的最大信息增益和最大不一致度之和的 数组下标  $j_1$  和  $j_2$ ;

Step4 如果  $j_1=j_2$ ,则第 j 列的属性缺失值的填补值为  $C_{j1}$ ,否则为  $C_{j2}$ ;

 Step5
 只更新第 i 行第 j 列的属性缺失值的 Weigh[i][j]=InfoGain [j][ji];

Step6 输出新的决策系统 DS',得到填补后的数据集,算法结束。

算法的性能分析: Step1 中的初始化操作的时间复杂度为 O(|U||C|); Step2 中  $AttValue = C_{j,k}$ , Step4 中比较  $j_1 = j_2$  的时间复杂度均为 O(h|C|),  $h = \max\{C_j\}$ ; Step2 中计算 InfoGain[j][k]和 AiiArray[j][k]需要遍历整个对象集U、条件属性集C和决策属性集D,因此所用时间复杂度为 O(h|U|(|C|+|D|)); Step3 中对 InfoGain 和 AiiArray 数组快速排序的时间复杂度为  $O(|C|hlog_2|C|h)$ ; Step5 中更新对象有属性缺失值的权值所用的时间复杂度为 O(h|U|)。因此,IGIDFA 算法的最大时间复杂度为 O(h|U|(|C|+|D|)), $h = \max\{C_j\}$ 。

# 4 面向不完备不一致数据的约简算法

文献[21]对不完备不一致性数据的属性约简是将不一致性数据和一致性数据分开处理,最后合并并提取最优选择作为规则约简,但是有一个重要的前提条件是不能改变原有属性对应的等价类。如果前提条件发生变化,那么对应的最优选择和规则约简也将发生变化,这是需要进一步考虑和优化的。对于不完备数据的属性约简,文献[22]首先考虑没有缺失值或者缺失值较少的属性,再考虑缺失值较多的属性来计算对应的属性的重要性的做法存在不足,因为包含缺失值对应的属性列也可能是关键属性列,且对决策系统具有非常重要的作用,所以需要首要考虑有包含缺失值的属性列。因此,本文在文献[23-24]和上述的研究的基础上,首先使用 IGID-FA 算法填补不完备不一致性数据,将缺失值填补后对应的信息增益值作为权值,没有缺失值的权值设为 1;再对属性列的信息增益大小排序后作为属性待约简序列,用  $\Gamma(x)$ 作为启发函数的属性约简算法(IGIDRA)。

IGIDRA 算法主要用于计算 IGIDFA 算法对数据集填补后每列的信息增益,升序排序后得到一个属性约简序列集合,属性的信息增益越小,其所反映的分类能力越弱,该属性被约简的可能性越大。计算每个对象 x 在约简一个属性前、后的不一致度  $\Gamma(x)$  和  $\Gamma'(x)$ ,通过比较决定该属性能否被约简,再继续计算  $\Gamma(x)$  和  $\Gamma'(x)$ ,比较并进行属性约简,最后得到一个最优约简集。

IGIDRA 算法的步骤如算法 2 所示。

### 算法 2 IGIDRA 算法

输入:决策系统 DS=(U,C,D,V,f)

输出:约简集 Redu(C)

Step1 计算填补后数据集的各列信息增益,升序排序后得到一个约 简集 Redunew(B);

Step2 生成新约简集  $Redu_{temp}(B) = Redu_{new}(B) - \{a\}$ ,原约简集  $Redu_{new}(B) = Redu_{new}(B)$ ,count=0;

Step3 计算对象 x 在约简集  $Redu_{new}$  (B) 和  $Redu_{temp}$  (B) 时的  $POS_B(x), POS_B'(x)$  和  $T_D(x),$  并求  $POS_B(x) \cap T_D(x)$  和  $POS_R'(x) \cap T_D(x)$ ;

Step4 计算  $W_B(x)$ 和  $W_B'(x)$ ;

Step5 计算约简集  $Redu_{new}(B)$ 和  $Redu_{temp}(B)$ 上的  $\Gamma(x)$ 和  $\Gamma'(x)$ ;

Step6 比较  $\Gamma(x)$ 和  $\Gamma'(x)$ ,若  $\Gamma'(x)$ 》 $\Gamma(x)$ ,count++,转 Step3,计 算下一个对象 x 对应的值,否则转 Step2,在 Step2 中生成新 约简集  $Redu_{temp}(B) = Redu_{new}(B) - \{b\}$ ;

Step8 当 Redu<sub>temp</sub>(B)=Redu<sub>new</sub>(B)时,算法结束,此时得到 Redu(C)= Redu<sub>temp</sub>(B)。

分析算法的性能: Step1 计算各列信息增益的时间复杂 度为 O(|U||C|),升序排序所用的时间复杂度为  $O(|C|\log_2|C|)$ ;Step2 和 Step7 的时间复杂度均为 O(|C|);Step3 计算  $POS_B(x)$ , $POS_B'(x)$  所需的时间复杂度为 O(h|U||C|), $h=\max\{|Redu_{new}(B)|,|Redu_{temp}(B)|\}$ ,最坏的情况是数据集的每个属性均未被约简,并且每次都执行了整个循环,此时时间复杂度为  $O(|U||C|^2)$ 。在每次进行属性约简时判

断  $\Gamma(x)$ 和  $\Gamma'(x)$ ,不符合条件即退出当前内循环,故即使属性集没有被约简,也很少每次都执行了整个内外循环。因此,大多数情况下时间复杂度都小于  $O(|U||C|^2)$ 。  $POS_B(x) \cap T_D(x)$  和  $POS'_B(x) \cap T_D(x)$  的时间复杂度为 O(|U||C|); Step4,Step5 和 Step6 所需的时间复杂度均为 O(|U||C|); Step8 的时间复杂度为 O(h|U||C|),由 max{ $|Redu_{new}(B)|,|Redu_{temp}(B)|}$ ,最坏情况下的时间复杂度为  $O(|U||C|^2)$ 。

## 5 实验分析

序号

2

voting-records

本次实验环境为: Intel Pentium CPU G3240 @ 3.10 GHz, RAM 4GB, Windows7 系统, 实验采用 Java 语言实现。

#### 5.1 IGIDFA 算法实验

为了验证 IGIDFA 算法的填充效果,本文采用 4 个 UCI 数据集<sup>1)</sup>来进行数据填补, chess 数据集是完备一致的数据集,有 36 个条件属性; audiology 数据集是不完备的数据集,有 69 个条件属性; soybean 数据集和 voting-records 数据集是不完备不一致性数据集,前者有 35 个条件属性,后者有 16 个条件属性,4 个数据集的具体情况如表 1 所列。

表 1 4个 UCI 数据集 Table 1 Four UCI data sets

数据集名称	N	NC	ND	S	T
chess	3196	36	2	0	0
soybean	683	35	19	0.10	0.12
audiology	226	69	24	0.02	0

16

0.06

本文统一令数据集的样本个数为N,条件属性集个数为NC,决策属性集个数为ND,缺失比率为S,不一致率为T,约简后条件属性个数为RC。

435

首先,对表 1 中的数据集多次随机抽取样本;然后,设置 缺失值使数据集的缺失比率达到  $5\% \sim 50\%$ ,将每次填补后的缺失值与填补前的对应样本值进行比较,若对缺失值进行填补后的值与设置缺失值前的值一样,即视为填补正确。假设每次填补正确的值的总个数为 TN,数据集的总样本数为 N,缺失比率为 S,填补正确率 V=TN/(N\*S)。

基于各列出现频率最高的值对缺失值进行填补的算法简称为 MPF 算法,文献[25]用模糊加权相似性的填补算法,称为 FWSDC 算法,本文用 MPF 算法、IGIDFA 算法和 FWSDC 算法分别对 4 个数据集进行 100 次不同缺失比率下的数据填补,得到的结果如表 2 所列。

由于 soybean 数据集的缺失比率约为 10%,因此表 2 中 缺失比率为 5%时的填补正确率用符号"一"表示。因为voting-records 数据集本身并不完备, 缺失率约为 6%, 大于 5%,不能获得缺失率为5%时的准确率,所以在表2中的对 应表格里的填补正确率也用符号"一"表示。图 1-图 4 分别 为使用上述 3 种算法对 chess, soybean, audiology 和 votingrecords 数据集在不同缺失比率下进行填补后的填补正确率 的结果图。由表 2、图 1-图 4 可以看出,随着缺失比率的增 加,填补正确率大部分会逐渐下降。这是因为缺失比率的增 加使得对应的数据集缺失值增大,增加了数据缺失变化的多 元性和可能性,因此对数据的填补增加了难度和不确定性。 整体来看,使用 IGIDFA 算法比使用 MPF 算法的填补正确率 更高,但是低于 FWSDC 算法的填补正确率,这是由于 FWS-DC 算法用最大相似性的对象的属性值来填补缺失值,而 IGIDFA 算法是用最大不一致度的对象的属性值来填补缺失 值。对于数据的填补,不能只考虑其填补正确率,还需要用分 类算法检验使用3个填补算法填补后的数据的填补效果,接 着进行如下实验。

表 2 4 个数据集填补后的正确率

Table 2 Accuracy rate after filling four data sets

41. /L (1. →s /						填补正	.确率/%					
缺失比率/	chess				soybean			audiology	7	voting-records		
/0	MPF	IGIDFA	FWSDC	MPF	IGIDFA	FWSDC	MPF	IGIDFA	FWSDC	MPF	IGIDFA	FWSDC
5	51.69	77.15	81.28	_	_	_	64.94	88.35	92.48	_	_	_
10	50.28	77.16	81.27	41.51	55.18	73.82	64.45	88.40	92.28	53.18	51.84	55.48
15	47.89	77.18	81.24	41.01	59.22	71.24	49.61	88.25	92.26	53.27	51.83	55.66
20	43.69	77.13	81.24	38.92	59.12	71.07	49.68	88.25	92.18	53.22	52.06	55.78
25	41.72	77.16	81.26	38.48	59.22	71.08	51.34	88.30	92.17	53.55	51.97	55.90
30	38.96	77.13	81.26	36.22	59.02	71.13	53.48	88.27	92.14	53.52	51.91	56.10
35	36.68	77.15	81.26	35.71	59.21	71.02	54.07	88.29	92.14	53.29	51.97	56.27
40	34.60	77.15	81.26	35.71	59.34	70.93	56.99	88.36	92.15	50.81	52.07	56.33
45	30.62	77.13	81.27	34.26	59.10	70.95	60.04	88.31	92.11	47.64	51.86	56.44
50	29.19	77.15	81.26	33.98	59.28	70.92	63.53	88.29	92.08	43.89	52.04	56.46

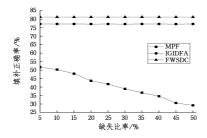


图 1 chess 数据集填补后的正确率

Fig. 1 Accuracy rate after filling chess data set

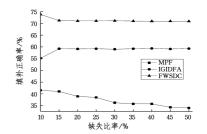


图 2 soybean 数据集填补后的正确率

Fig. 2 Accuracy rate after filling soybean data set

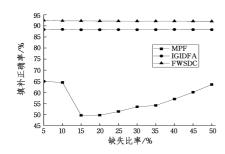


图 3 audiology 数据集填补后的正确率

Fig. 3 Accuracy rate after filling audiology data set

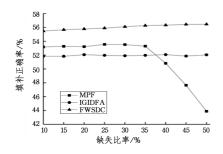


图 4 voting-records 数据集填补后的正确率

Fig. 4 Accuracy rate after filling voting-records data set

本文对 chess, soybean, audiology 和 voting-records 4 个数据集分别在不同的缺失比率下使用 MPF 算法、IGIDFA 算法和 FWSDC 算法填补后,再采用 C4.5 算法和 SVM 算法分

别进行十折交叉验证,所获得的分类准确率如表 3 和表 4 所列。其中,soybean 数据集和 voting-records 数据集自身的缺失比率都超过了 5%,因此在缺失比率为 5%时不能获得算法的分类准确率,用符号"一"表示。图 5 一图 8 分别为使用C4.5 算法和 SVM 算法对 chess,soybean,audiology 和 voting-records 4 个数据集在不同缺失比率下填补后的数据进行分类的结果图。

由表 3一表 4 和图 5一图 8 可知,4 个数据集使用 IGID-FA 算法填补后 C4.5 和 SVM 的分类准确率大多数比使用 MPF 算法填补后 C4.5 和 SVM 的分类准确率高。除了 soybean 数据集使用 IGIDFA 算法填补后的分类准确率比 FWS-DC 算法填补后的分类准确率低外,其他 3 个数据集使用 IGIDFA 算法的分类准确率整体要高于使用 FWSDC 算法填 补后的分类准确率。对完备的 chess 数据集使用 IGIDFA 算 法填补后有 70%的 C4.5 算法的分类准确率高于使用 FWS-DC 算法填补后的分类准确率;对不完备的 audiology 数据集 使用 IGIDFA 算法填补后,有 80%的 C4.5 算法和 30%的 SVM 算法的分类准确率高于使用 FWSDC 算法填补后的分 类准确率:不完备不一致的 voting-records 数据集的不一致率 达到 0.67,在填补后仍然具有较高的分类准确率,使用 IGID-FA 算法填补后有 66.7%的 C4.5 算法和 77.8%的 SVM 算 法的分类准确率高于使用 FWSDC 算法填补后的分类准确 率,这说明使用 IGIDFA 算法进行数据填补是有效的。

表 3 分类准确率(1)

Table 3 Classification accuracy rate (1)

	分类准确率/%											
缺失比率/		chess							soy	bean		
/0	C4.5	SVM	C4.5	SVM	C4.5	SVM	C4.5	SVM	C4.5	SVM	C4.5	SVM
5	96.96	91.75	96.96	91.84	96.99	93.50	_	_	_	_	_	_
10	94.53	89.64	94.54	89.75	94.52	91.22	92.56	88.69	92.56	89.40	92.31	92.93
15	91.89	87.38	91.89	87.48	91.88	88.83	88.06	83.73	88.09	83.83	88.56	88.96
20	89.73	85.47	89.78	85.56	89.70	86.80	83.68	79.30	83.76	79.20	84.67	85.13
25	87.35	83.36	87.40	83.35	87.33	84.59	78.48	74.15	78.78	74.05	80.23	80.68
30	84.92	81.13	84.97	81.17	84.99	82.44	74.12	69.18	74.18	69.20	76.09	76.71
35	82.66	78.90	82.71	78.96	82.65	80.26	69.96	64.42	70.02	64.03	72.06	72.65
40	80.26	76.58	80.34	76.72	80.31	78.10	65.56	58.96	65.73	58.68	67.76	68.58
45	77.86	74.35	77.92	74.45	77.92	75.91	60.91	53.34	61.15	53.30	63,66	64.81
50	75.50	72.26	75.53	72.29	75.56	73.74	56.88	48.71	56.92	78.04	59.45	60.83

表 4 分类准确率(2)

Table 4 Classification accuracy rate (2)

						分类准	确率/%					
缺失比率/ %		audiology							voting-	records		
/0	C4.5	SVM	C4.5	SVM	C4.5	SVM	C4.5	SVM	C4.5	SVM	C4.5	SVM
5	75.98	77.93	76.15	78.29	76.52	78.73	_	_	_	_	_	_
10	73.44	75.00	73.49	75.44	72.58	75.12	94.40	94.47	94.40	94.71	94.53	94.12
15	69.80	71.75	69.80	71.73	68.95	71.59	92.43	92.40	92.64	92.88	92.73	92.40
20	66.87	68.32	66.96	69.16	66.74	69.30	90.63	90.70	91.30	91.25	90.91	90.79
25	64.54	65.34	63.97	65.49	63.36	65.91	89.12	89.26	89.44	89.34	89.17	89.06
30	60.27	61.98	60.37	62.41	60.41	62.97	87.19	87.35	87.63	87.45	87.51	87.51
35	57.67	58.46	57.78	59.54	57.25	59.59	85.49	85.54	86.27	86.18	85.89	85.86
40	54.48	55.67	54.45	56.36	53.79	56.32	83.74	83.74	84.31	84.27	84.09	84.05
45	51.38	52.43	51.41	53, 25	51.00	53,55	82.21	82.20	82.29	82.19	82.36	82.39
50	48.67	49.54	48.40	50.22	48.28	50.46	79.37	79.30	80.96	80.89	80.67	80.76

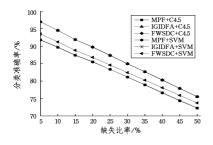


图 5 chess 数据集的分类准确率

Fig. 5 Classification accuracy of chess data sets

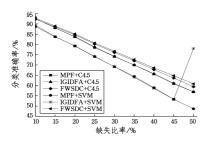


图 6 soybean 数据集的分类准确率

Fig. 6 Classification accuracy of soybean data sets

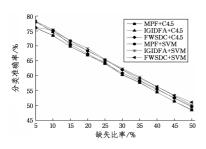


图 7 audiology 数据集的分类准确率

Fig. 7 Classification accuracy of audiology data sets

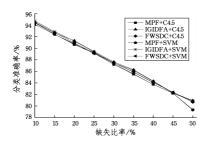


图 8 voting-records 数据集的分类准确率

Fig. 8 Classification accuracy of voting-records data sets

# 5.2 IGIDRA 算法实验

为了验证 IGIDFA 算法与 IGIDRA 算法结合后的约简效果,首先对文献[14]中的经典数据进行填补和约简(见表 5),填补后得到表 6;再对表 6 的数据进行属性约简,得到约简集{Size,Max-speed}。

表 5 不完备不一致决策系统

Table 5 Incomplete and inconsistent decision system

Car	Price	Mileage	Size	Max-speed	d
1	High	High	Full	Low	Good
2	Low	*	Full	Low	Good
3	*	*	Compact	high	Poor
4	Low	*	Full	High	Good
5	*	*	Full	High	Excel
6	Low	High	Full	*	Good

表 6 填补后的不完备不一致决策系统

Table 6 Incomplete and inconsistent decision system after filling

Car	Price	Mileage	Size	Max-speed	d
1	High	High	Full	Low	Good
2	Low	High	Full	Low	Good
3	High	High	Compact	high	Poor
4	Low	High	Full	high	Good
5	High	High	Full	high	Excel
6	Low	High	Full	Low	Good

表 5 到表 6 的填充过程如下:

决策系统属性 Price 有 2 个属性值{High,Low},将属性 Price 列的第 3 行和第 5 行填充为 High 时,属性 Price 列的信息增益 InforGain[0][0]=0.459,整个对象集 X的不一致度之和 AiiArray[0][0]=0.333。

将属性 Price 列的第 3 行和第 5 行填充为 Low 时, InforGain[0][1]=0.109, AiiArray[0][1]=0.6667。

此时属性 Price 列最大信息增益的下标 i=0,其最大的不一致度之和的下标 j=1,根据前文 IGIDFA 算法的 Step 4,对属性 Price 列选取 Low 进行填充。

属性 Mileage 只有 1 个属性值{High},直接填充即可。

属性 Max-speed 有 2 个属性值{Low, High}。当填充值为 Low 时, InforGain[3][0]=0.459, AiiArray[3][0]=0.8333;当填充值为 High 时, InforGain[3][1]=0.2516, Aii-Array[3][1]=0.8333。

此时属性 Price 列最大信息增益的下标 j=0,其最大的不一致度之和的下标 j=0,根据前文 IGIDFA 算法的 Step 4,对属性 Max-speed 列选取 Low 进行填充。

然后再采用 6 个 UCI 数据集进行 IGIDRA 约简算法的验证。使用 IGIDRA 算法属性约简后的结果如表 7 所列。表 7 中有 2 个完备一致的数据集 chess 和 mushroom,1 个不完备数据集 audiology,以及 3 个不完备不一致数据集 primary-tumor, voting-records 和 soybean。

表 7 使用 IGIDRA 算法属性约简后的结果

Table 7 Attribute reduction results by using IGIDRA algorithm

序号	数据集名称	N	NC	ND	S	T	RC
1	chess	3196	36	2	0	0	29
2	mushroom	8124	22	2	0	0	5
3	audiology	226	69	24	0.02	0	14
4	primary-tumor	339	17	21	0.04	0.42	16
5	voting-records	435	16	2	0.06	0.68	11
6	soybean	683	35	19	0.10	0.12	12

由表 7 可知, soybean, audiology 和 mushroom 数据集属性约简的个数最多, audiology 数据集被约简了 55 个属性, soybean 数据集被约简了 23 个属性, mushroom 数据集被约简了 17 个属性, chess 数据集被约简了 7 个属性, voting-records 数据集和 primary-tumor 数据集也分别被有效地约简,从而说明了 IGIDRA 约简算法的合理性。

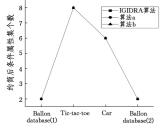
将文献[26]和文献[27]中的约简算法分别称为算法 a 和算法 b,将其与本文的 IGIDRA 算法进行属性约简的比较,结果如表 8 和图 9 所示。由表 8 和图 9 可知,3 种算法对 Balloon

database(1)数据集和 Balloon database(2)数据集都约简了 2 个属性,对 Tic-tac-toe 数据集都约简了 1 个属性,对 Car 数据集约简后的属性个数也均为 6,因此图 9 中 3 条线段最终重合,这说明使用 IGIDRA 约简算法具有与算法 a 和算法 b 相同的约简结果。基于表 8 中的 4 个数据集,比较不同算法的属性约简时间,结果如图 10 和表 9 所示。

表 8 属性约简的比较结果

Table 8 Comparison results of attribute reduction

					约简后条件属性数			
序号	数据集名称	N	NC	ND	IGIDRA 算法	算法 a	算法 b	
1	Balloon database(1)	20	4	2	2	2	2	
2	Balloon database(2)	20	4	2	2	2	2	
4	Tic-tac-toe	958	9	2	8	8	8	
3	Car	1728	6	4	6	6	6	



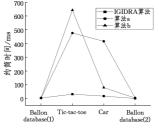


图 9 属性约简的比较结果 (折线图)

图 10 属性约简时间的比较 结果(折线图)

Fig. 9 Comparison results of Fig. 10 Comparison results of attribute reduction(line chart) attribute reduction time(line chart)

表 9 属性约简时间的比较结果

Table 9 Comparison result of attribute reduction time

序号	数据集名称	约简时间/ms					
オサ	数据来名价	IGIDRA 算法	算法 a	算法 b			
1	Balloon database(1)	4	4	3			
2	Balloon database(2)	1	4	4			
4	Tic-tac-toe	32	476	643			
3	Car	18	416	80			

由表 9 和图 10 可知,使用 IGIDRA 约简算法的属性约简时间远远小于算法 a 和算法 b 所需的时间。

综上所述,使用 3 种算法对数据集进行填补后,再用 C4.5 算法和 SVM 算法进行分类预测,使用 IGIDFA 算法对 不完备不一致数据集填补后的分类准确率要高于使用 MPF 算法和 FWSDC 算法填补后的分类准确率。再通过属性约简和约简时间的对比,证明了 IGIDRA 约简算法是实用、高效的。本文提出的 IGIDRA 约简算法对完备一致决策系统和不完备不一致决策系统均适用,约简效果也较为理想。因此,IGIDFA 算法和 IGIDRA 算法具有一定的实用价值。

结束语 目前,对完备一致数据的研究较多,对不完备不一致性数据研究还相对较少,而本文提出了一种新的思想来对不完备不一致性数据进行研究和思考,尽可能在保持原数据原本特征的条件下进行数据填补,再在最大不一致度条件下以信息增益为权值,对有缺失值的对象的权值进行特殊处理,最后进行属性约简,以获得与其他不填补数据进行属性约简时相同的效果,这说明 IGIDFA 填补算法和 IGIDRA 约简算法都是有效的。现在是云计算和大数据时代,下一步需要

对不完备不一致性混合型大数据进行深入研究和拓展,并继续对属性约简算法进行优化,使其达到更优的性能。

# 参考文献

- [1] PAWLAK Z. Rough Sets: Theoretical Aspects of Reasoning about Data[M]. Kluwer Academic Publishers, 1991, 9:24-26.
- [2] STEFANOWSKI J. TSOUKIÀS A. Incomplete Information Tables and Rough Classification [J]. Computational Intelligence, 2001,17(3);545-566.
- [3] LIU P, QIU T R, XIONG X X, et al. An Incomplete Data Filling Approach Based on a New Valued Tolerance Relation[J]. Open Automation & Control Systems Journal, 2014, 6(1):1456-1462.
- [4] JIN C M, E X, MU H J, et al. Data Filling Method Based on New Relationship Matrix [J]. Computer Engineering, 2011, 37(19):28-31. (in Chinese) 金成美, 鄂旭.穆海军, 等. 一种基于新型关系矩阵的数据填补方法[J]. 计算机工程, 2011, 37(19):28-31.
- [5] WU K K,PAN W. Attribute significance based imputation method[J]. Computer Engineering and Design, 2016, 37 (3): 725-730. (in Chinese)
  - 吴康康,潘巍. 基于属性重要度的数据补齐方法[J]. 计算机工程与设计,2016,37(3):725-730.
- [6] KIRAN P M, RAO A P, RATNAMALA B. An Efficient Approach for Filling Incomplete Data[C]//National Conference on Advances in Computer Science and Applications with International Journal of Computer Applications (NCACSA 2012). 2012;23-27.
- [7] YANG X P. Completing incomplete data based on maximum similarity in Rough sets[J]. Computer Engineering and Applications, 2012, 48(36):164-166. (in Chinese) 杨小平. 粗集中最大相似度的不完备数据补齐[J]. 计算机工程

与应用,2012,48(36):164-166.

- [8] WU S, FENG X D, SHAN Z G. Missing Data Imputation Approach Based on Incomplete Data Clustering[J]. Chinese Journal of Computers, 2012, 35(8):1726-1738. (in Chinese) 武森, 冯小东, 单志广. 基于不完备数据聚类的缺失数据填补方法[J]. 计算机学报, 2012, 35(8):1726-1738.
- [9] YANG T,LUO J W, WANG Y, et al. Missing value estimation for gene expression data based on Mahalanobis distance [J]. Computer Applications, 2005, 25(12): 2868-2871. (in Chinese) 杨涛,骆嘉伟,王艳,等. 基于马氏距离的缺失值填充算法[J]. 计算机应用, 2005, 25(12): 2868-2871.
- [10] KIM K Y,KIM B J,YI G S. Reuse of imputed data in microarray analysis increases imputation efficiency[J]. Bmc Bioinformatics,2004,5(1):160.
- [11] CHEN Z K, YANG Y D, ZHANG Q C, et al. Novel algorithm for filling incomplete data of internet of things based on attribute reduction [J]. Computer Engineering and Design, 2013, 34(2):418-422. (in Chinese)
  - 陈志奎,杨英达,张清辰,等.基于属性约简的物联网不完全数据填充算法[J].计算机工程与设计,2013,34(2):418-422.
- [12] ZHANG H X. Missing data imputation: Information gain based on approach[J]. Computer Engineering and Design, 2006, 27(24):4810-4812. (in Chinese)

- 张红霞. 缺失值填充:基于信息增益的方法[J]. 计算机工程与设计,2006,27(24):4810-4812.
- [13] QIN Z. Information Gain based Algorithm for Filling Missing Data[J]. Microcomputer Information, 2007, 23(12): 180-181. (in Chinese) 覃泽. 基于信息增益的数据库缺失值填充算法[J]. 微计算机信
  - 覃泽. 基于信息增益的数据库缺失值填充算法[J]. 微计算机信息,2007,23(12);180-181.
- [14] KRYSZKIEWICZ M. Rough Set Approach to Incomplete Information System[J]. Information Sciences, 1998, 112(1-4): 39-49.
- [15] WANG G Y. Extension of Rough Set Under Incomplete Information systems[J]. Journal of Computer Research and Development,2002,39(10);1238-1243. (in Chinese) 王国胤. Rough 集理论在不完备信息系统中的扩充[J]. 计算机研究与发展,2002,39(10);1238-1243.
- [16] FU A, WANG G Y, HU J. Information entropy based attribute reduction algorithm in incomplete information systems[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition),2008,20(5):586-592. (in Chinese) 付昂,王国胤,胡军. 基于信息熵的不完备信息系统属性约简算法[J]. 重庆邮电大学学报(自然科学版),2008,20(5):586-592.
- [17] TAO Z, LIU Q Z, LI W M. Attribute reduction based on GA under incomplete information system[J]. Systems Engineering and Electronics, 2007, 29(9):1484-1487. (in Chinese) 陶志,刘庆拯,李卫民. 基于遗传算法的不完备信息系统属性约简方法[J]. 系统工程与电子技术, 2007, 29(9):1484-1487.
- [18] KRYSZKIEWICZ M. Rules in incomplete information systems [J]. Information Sciences, 1999, 113(3-4):271-292.
- [19] XIE H, CHENG H Z, NIU D X. Discretization of Continuous Attributes in Rough Set Theory Based on Information Entropy [J]. Chinese Journal of Computers, 2005, 28(9): 1570-1574. (in Chinese) 谢宏,程浩忠,牛东晓. 基于信息熵的粗糙集连续属性离散化算
- [20] 蒋盛益,李霞,郑琪. 数据挖掘原理与实践[M]. 北京:电子工业出版社,2011,48-58.

法[J]. 计算机学报,2005,28(9):1570-1574.

- [21] FU M L.ZENG H L. Oprimization Selection and Rules Extraction in Inconsistent and Incomplete Information System [J]. Computer Science, 2007, 34(10): 208-211. (in Chinese) 伏明兰,曾黄麟. 一种不一致不完备信息系统的最优选择及规则约简方法研究[J]. 计算机科学, 2007, 34(10): 208-211.
- [22] HE W, LIU C Y, ZHAO J, et al. An Algorithm of Attributes Reduction in Incomplete Information System[J]. Computer Science, 2004, 31(2):117-119. (in Chinese) 何伟,刘春亚,赵军,等. 不完备信息系统下的属性约简算法[J].
  - 何伟,刘春亚,赵军,等. 不完备信息系统下的属性约简算法[J]. 计算机科学,2004,31(2):117-119.
- [23] MENG Z Q, XU K, ZHOU S Q. Maximum distribution reduction and computation methods for incomplete inconsistent decision systems[J]. Journal of Guangxi Normal University(Natural Science Edition),2011,29(3):89-93. (in Chinese) 蒙祖强,许珂,周石泉. 不完备不一致决策系统的最大分布约简及计算方法[J]. 广西师范大学学报(自然科学版),2011,29(3):89-93.
- [24] MENG Z Q, SHI Z Z. A fast approach to attribute reduction in incomplete decision systems with tolerance relation—based rough sets[J]. Information Sciences, 2009, 179(16): 2774-2793.
- [25] MAF M, LIUTT, XUAP. Data completion with rough sets based on fuzzy weighted similarity measure [J]. Computer Engineering and Applications, 2016, 52(9); 62-66. (in Chinese) 马福民,刘涛涛,徐安平. 基于模糊加权相似度量的粗糙集数据补齐方法[J]. 计算机工程与应用, 2016, 52(9); 62-66.
- [26] YANG C Q. The attribute reduction algorithms based on rough sets[J]. Journal of Northwest University(Natural Science Edition),2012,42(2);223-225. (in Chinese) 杨常清. 基于粗糙集的属性约简算法[J]. 西北大学学报(自然科

学版),2012,42(2):223-225.

- [27] YE D Y. An Improvement to Jelonek's Attribute Reduction Algorithm[J]. Acta Electronica Sinca, 2000, 28(12): 81-82. (in Chinese)

  叶东毅 Jelonek 属性约简質注的一个改进[J] 由子学报, 2000.
  - 叶东毅. Jelonek 属性约简算法的一个改进[J]. 电子学报,2000,28(12):81-82.

## (上接第 206 页)

- [10] POLATIDIS N, GEORGIADIS C K. A multi-level collaborative filtering method that improves recommendations [J]. Expert Systems with Applications, 2016, 48:100-110.
- [11] GOMES P, PEREIRA F C, PAIVA P, et al. Using CBR for automation of software design patterns [C] // Proceedings of the 6th European Conference on Advances in Case-Based Reasoning. Heidelberg: Springer, 2002:534-548.
- [12] PAVLIČL, PODGORELEC V, HERIČKO M. A question-based design pattern advisement approach[J]. Computer Science and Information Systems, 2014, 11(2):645-664.
- [13] HASHEMINEJAD S M H.JALILI S. Design patterns selection; an automatic two-phase method[J]. Journal of Systems and Software, 2012, 85(2):408-424.
- [14] NAKATSUJI M, TODA H, SAWADA H, et al. Semantic sensitive tensor factorization [J]. Artificial Intelligence, 2016, 230: 224-245.

- [15] ZHAO Q, ZHOU G, ZHANG L, et al. Bayesian robust tensor factorization for incomplete multiway data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 27(4): 736-748.
- [16] DUANE S, KENNEDY A D, PENDLETON B J, et al. Hybrid Monte Carlo [J]. Physics Letters B, 1987, 195(2): 216-222.
- [17] HUANG W, LEIMKUHLER B. The adaptive verlet method [J]. SIAM Journal on Scientific Computing, 1997, 18(1): 239-256
- [18] CARROLL J D, CHANG J J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition[J]. Psychometrika, 1970, 35(3):283-319.
- [19] NOCK R, NIELSEN F. On weighting clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(8):1223-1235.
- [20] NEIL T. Mobile Design pattern gallery: UI patterns for mobile applications[M]. Sebastopol: O'Reilly Media, 2012.