

基于相似关系的数据库分类不一致程度评价

张伟钢^{1),2)} 潘 泉²⁾ 张洪才²⁾

¹⁾(中国民航大学航空工程学院 天津 300300)

²⁾(西北工业大学自动化学院 西安 710072)

摘 要 作为挖掘算法选择和评价的标准之一,数据集的分类不一致程度一直是分类规则研究中的一项重要内容.然而随着人们对不完备数据集数据挖掘的深入,建立在等价关系上的基于信息熵的评价方法已难以满足实际需要.文中在利用相似关系的基础上,结合证据理论,给出一种基于信任度与似然度的信息粒构建方法,同时构建了类似于不协调度和混淆度的系统分类不一致程度评价方法,并对其相关性质等进行分析与证明.由算例分析可以看出,文中研究结果能够较好地描述缺失环境下的系统分类不一致程度,同时当数据集不存在缺失时,该研究与以往研究具有相同结果.

关键词 相似关系;分类不一致程度;不完备数据库;证据理论;模糊熵

中图法分类号 TP18

Inconsistency Measure of Database with Similarity Relation

ZHANG Wei-Gang^{1),2)} PAN Quan²⁾ ZHANG Hong-Cai²⁾

¹⁾(College of Aeronautical Mechanics and Avionics Engineering, Civil Aviation University of China, Tianjin 300300)

²⁾(School of Automation, Northwestern Polytechnical University, Xi'an 710072)

Abstract As one of the principles to select and appraise data mining algorithm, the inconsistency measure of database received much attention in classification rules discovering. But the classical measure based on information entropy will not meet those needs with the further study of incomplete database since the requirement of equivalence relation may not be satisfied in such condition. This paper gives a method to found information granularity with belief and plausibility measure based on the similarity relation and evidence theory. At the same time, inconsistency measures which are similar to inconsistent and confusion degree of fuzzy entropy are proposed with the proving of their some character. From the proving and simulation, it shows the proposed method will give a well description of inconsistency in incomplete database, and when there is no data missing, it will gives a same result as the previous studies.

Keywords similarity relation; inconsistency; incomplete database; evidence theory; fuzzy entropy

1 引 言

分类规则作为一种基本知识模式,已经成为知

识发现领域的一个焦点.

当前,分类规则研究中的规则发现方法主要包括基于决策树的方法、基于规则的方法、基于统计的方法等.而分类不一致是此类研究中涉及相对较多

的内容. 分类不一致, 又称为分类的不确定性, 是指系统中存在条件属性值相同而决策不同的记录. 由于数据库的分类不一致在一定程度上反映了条件属性不充分或噪声程度^[1], 因此在数据挖掘、机器学习、决策树、粗集理论等诸多领域的研究中受到普遍重视^[2-6]. 一般说来, 系统分类不一致程度往往影响预测精度. 为提高预测精度, 出现了许多针对此类问题的改进方法. 但这些改进方法的计算复杂性往往较高, 因此需要正确判断系统的分类不一致程度并借此选择合适的算法.

对于所有属性为名义尺度或顺序尺度的数据库系统, 粗糙分类熵方法^[7]是当前系统分类不一致评价中使用较多的方法. 该方法是在粗集理论的基础上, 用规则的分类精度代替条件概率以及等价范畴的频率代替相关概率, 结合条件信息熵对数据库分类不一致程度进行评价. 由于该方法与条件信息熵理论联系紧密, 且便于理解, 因此在很多后续的研究中被广泛采用^[8-10].

然而, 由于该理论建立在传统粗集理论的基础上, 因而测量中没有考虑缺失数据的影响. 在现实中, 大多数数据库往往存在不同程度的数据缺失, 因此如何对存在缺失数据的数据库(即不完备数据库), 进行分类规则学习已成为当前研究的一个难点与热点问题, 对此涌现了大量的缺失环境的学习算法^[11-18]. 但是, 从当前研究来看, 关于不完备数据库的分类不一致程度评价研究还显得相对不足.

从当前不完备数据库的相关研究来看, 缺失数据的处理方式是其核心问题之一. 在大部分分类规则学习算法中, 一般采用的缺失数据处理方式可分为以下几种:

- (1) 删除缺失数据所在的纪录;
- (2) 计算缺失数据的可能取值及可能性分布, 即对缺失数据进行估值;
- (3) 将缺失数据作为一个独立的特殊值;
- (4) 利用相似关系代替等价关系.

以上方法中, 除删除纪录的方法以外, 其它方法都是基于缺失数据与其它数据匹配问题的. 其中删除纪录的方法实际是将不完备数据库转化为完备数据库处理, 而将缺失数据作为一个特殊值, 则没有考虑缺失数据的可能取值, 难以反映缺失数据的本质, 因此这两种方法在数据库分类不一致评价的研究中不常见.

利用缺失数据的可能取值与概率分布的方法是当前数据库分类不一致程度评价研究中对缺失数据

进行分析的主要方法. 如, 与分类不一致程度评价具有一定类似性的决策树算法中信息增益的计算问题以及文献^[19]中基于概率论的分类不一致评价方法等. 当先验知识充分时, 此类方法可看作是粗糙熵理论的自然延伸, 因此具有计算精确和便于理解的优点. 然而在大部分应用中, 获取缺失数据的真实分布相对困难, 因此将此类方法用于分类不一致程度评价通常存在一定的偏差. 另外, 此类方法难以用于估计缺失数据对分类不一致程度的影响.

利用相似关系代替等价关系的分析方法在近年的粗集理论研究中获得了普遍重视, 形成了基于相似关系的粗集方法. 从研究结果来看, 该种缺失数据处理方法不涉及对先验信息的需求, 且可根据实际需要构造适当的相似关系以反映对数据相关性的不同假设. 其中与数据库分类不一致程度评价有关的类似研究结果主要有文献^[20-21]给出的在利用类似于 μ^* 测度的相似关系的基础上的数据库缺失程度评价. 然而由于相似关系不构成论域上的划分, 因此与信息熵中要求的概率测度不相洽, 而对于其它许多类型的模糊测度, 缺乏类似于条件信息熵的信息量计算方式, 因此难以构建相关测量.

信息检索领域涉及较多的是缺失数据的匹配问题. 在该类研究中, 当先验知识有限时, 缺失数据与其它数据匹配问题的分析往往是结合证据理论进行的, 将缺失数据取值分别划分为取值的不确定性和取值的不精确性, 并进一步利用似然测度和信任测度等方法对数据的不精确性进行处理^[22-25]. 据此我们认为, 对数据库的分类不一致程度用某种乐观或保守的匹配规则分别进行评价更能反映缺失数据的本质, 而基于这两种匹配方式之间的差异可作为解决第 2 个问题的依据.

本文第 2 节简单介绍本研究涉及的相关理论; 第 3 节结合关系数据库, 给出一种基于证据理论的相似关系模型及相似类构建方法; 第 4 节依据第 3 节中的相似关系模型, 对数据库的分类不一致程度评价进行讨论, 并给出相关计算方法. 最后给出相关评价方法的应用实例.

2 相关理论简介

2.1 证据理论^[23-27]

证据理论是在概率论的基础上进一步扩充而产生的一种新的数学理论, 在缺失数据的处理中有着普遍的应用. 其基本理论构架如下.

定义 1(mass 函数). 设 X 是有限集, 有 X 的幂集 2^X . 称 $m: 2^X \rightarrow [0, 1]$ 是一 mass 函数, 若满足:

- ① $m(\emptyset) = 0$;
- ② $\sum_{A \subseteq X} m(A) = 1$.

mass 函数在一些研究中也称为基本指派函数、基本概率指派等, 表示变量取值在集合 $A \subseteq X$ 中的可信程度, 但其中不反映变量在 A 任何子集中取值的可信程度. 可以看出, 在 mass 函数中考虑了两种不确定, 第一种是取值为集合 A 的不确定, 即 $m(A)$, 我们称之为变量具有不确定性; 第二种是当 $m(A)$ 已知时, 变量取值为 A 的某个具体子集的不确定, 我们称之为变量具有不精确性. 显然, 当 mass 函数为一概率函数, 即对于任一 $m(A) > 0$, 有 $|A| = 1$ 时, 变量只有不确定性而不存在不精确性, 其中 $| \cdot |$ 表示集合的秩.

为对不精确性进行处理以得到 X 的任一子集 $B \in 2^X$ 的可信程度, 证据理论引入两个模糊测度, 并证明了其可由 mass 函数通过以下关系获得.

定义 2(信任测度). 设 m 是 X 上的 mass 函数, 则

$$Bel(L) = \sum_{A \subseteq L} m(A)$$

为一信任测度.

定义 3(似然测度). 设 m 是 X 上的 mass 函数, 则

$$Pls(L) = \sum_{L \cap A \neq \emptyset} m(A)$$

为一似然测度.

信任测度 Bel 和似然测度 Pls 表示了对数据不精确性的不同处理态度, 信任测度相对保守, 其认为只有必然属于 L 的集合才支持变量在 L 中取值; 而似然测度相对乐观, 其认为只要与 L 相似(即 $L \cap A \neq \emptyset$), 则变量在 A 中取值就支持变量在 L 中取值. 不难看出, 对于任意 $L \subseteq X$, 有 $Bel(L) \leq Pls(L)$.

另外, 在文献中, 将由如下 mass 函数构成的信任测度和似然测度也称为 μ_* 和 μ^* 测度:

$$m(A) = \begin{cases} 1, & A = X \\ 0, & \text{其它} \end{cases}$$

此种 mass 函数表示系统不存在对变量取值分布的任何信息, 也称为完全混沌状态.

类似于信息熵, 利用信任测度和似然测度, 分别有系统的混淆度和不协调度.

定义 4(混淆度). 设 m 是 X 上的 mass 函数, Bel 为由 m 生成的信任测度, 即 $Bel(L) = \sum_{A \subseteq L} m(A)$, 称

$$C(m) = - \sum_{A \subseteq X} m(A) \log Bel(A)$$

为 $[X, m]$ 的混淆度.

定义 5(不协调度). 设 m 是 X 上的 mass 函数, Bel 为由 m 生成的信任测度, 即 $Pls(L) = \sum_{L \cap A \neq \emptyset} m(A)$, 称

$$E(m) = - \sum_{A \subseteq X} m(A) \log Pls(A)$$

为 $[X, m]$ 的不协调度.

2.2 等价关系^[6]与粗糙分类熵

对于属性值均为有限集的数据库系统, 粗糙分类熵是一种常用的分类不一致评价方法, 其理论基础是粗集理论和等价关系. 为进一步分析不完备数据库下分类不一致评价问题, 我们先对粗糙分类熵进行简介.

定义 6(知识表示系统). 设 $S = \langle U, R, V, f \rangle$ 为一知识表示系统, 其中 U 为论域; R 为属性集合; $V = \bigcup_{r \in R} V_r$ 是属性值集合, V_r 表示属性 $r \in R$ 的属性值范围; $f: U \times R \rightarrow V$ 是一个信息函数, 它指定 U 中每一对象 x 的属性值.

以上方法定义了一关系数据库, 其中论域 U 表示数据库的所有对象集合, R 则为数据库的属性集合.

在以上知识表示系统中, 若有 $R = (C \cup D)$, 其中 $C = \{c_1, c_2, \dots, c_N\}$ 为条件属性集, $D = \{d_1, d_2, \dots, d_M\}$ 为决策属性集, 则称 S 为一决策表, 简记为 $S = \langle U, R = (C \cup D) \rangle$.

定义 7(等价类). 设有知识表示系统 S (见定义 6), 令有 U 上的等价关系 $I_B (B \subseteq R)$: $x I_B y \Leftrightarrow \forall r \in B, f(x, r) = f(y, r)$, 其中 $x, y \in U$. 由等价关系 I_B 可得 U 上的划分 U/I_B , 或简记为 U/B , 有

$$U/B = \bigcup [x]_B,$$

称 U/B 为论域 U 上的一个等价类, 其中 $[x]_B$ 称为论域 U 上的一个范畴.

由文献[30]可知, 此时 $[x]_B$ 为一信息粒.

定义 8(粗糙分类熵)^[7]. 设有知识表示系统 S 及 U 上的等价关系 $I_C, I_D, (C \cup D = R, C \cap D = \emptyset)$, 则称条件信息熵

$$I(D|C) = - \sum_{[x]_C \in U/C} \frac{|[x]_C|}{|U|} \sum_{[x]_D \in U/D} \frac{|[x]_C \cap [x]_D|}{|[x]_C|} \log \frac{|[x]_C \cap [x]_D|}{|[x]_C|}$$

为知识表示系统 S 上的粗糙分类熵.

可以看出, 当数据库中不存在缺失数据时, 粗糙

分类熵中各部分符合概率分布,从而满足条件信息熵要求.由条件信息熵的性质可知,当系统 $|V_C|$, $|V_D|$ 及 $\frac{|[x]_C|}{|U|}$ 确定时,有 $I(D|C)$ 最大当且仅当 $\frac{|[x]_C \cap [x]_D|}{|[x]_C|} = \frac{1}{|V_D|}$,即系统中所有规则的决策值服从等概分布,其中 V_C, V_D 分别为属性集 C, D 的值域空间. $I(D|C)=0$ 当且仅当对于任一 $\frac{|[x]_C|}{|U|} > 0$ 有 $\frac{|[x]_C \cap [x]_D|}{|[x]_C|} = 1$,即系统中所有决策规则均为确定性规则.

2.3 相似关系与相似类^[20-21,28]

当数据库存在缺失数据时,由于缺失数据取值的不确定,一般不再构成等价关系.虽然可以通过估值或将缺失数据作为一个特殊值进而构造等价关系,但在估值过程中需要较多的先验信息和计算量,而将缺失数据作为特殊值则无法反映缺失数据与其它数据间的相互关系.因此在当前粗集理论研究和数据库检索中,利用相似关系对缺失数据进行处理是一种常用方法.

定义 9(相似关系). 设有知识表示系统 S, U 上的相似关系 $S_B (B \subseteq R)$ 定义为

$$xS_By \Leftrightarrow \forall r \in B, f(x, r) = f(y, r) \text{ 或}$$

$$f(x, r) = \text{"null"} \text{ 或 } f(y, r) = \text{"null"}, x, y \in U \quad (1)$$

定义 10(相似类). 设有知识表示系统 S 及 U 上的相似关系 $S_B (B \subseteq R)$,则由相似关系 S_B 可得 U 上的覆盖

$$U/S_B = \bigcup S_B(x) \quad (2)$$

称 U/S_B 为论域 U 上的一个相似类, $S_B(x)$ 为论域 U 上的相似范畴.

显然,此时 $S_B(x)$ 构成信息粒.

然而,当数据库存在缺失数据时,由于 U/S_C 和 U/S_D 不构成论域 U 的划分, $\frac{|S_C(x)|}{|U|}$ 和 $\frac{|S_C(x) \cap S_D(x)|}{|S_D(x)|}$ 不再满足概率分布,因此对其分类不一致程度的评价无法采用定义 8 中的粗糙信息熵方法,而需要采取新的测度类型和信息量测量方法.

3 基于证据理论的相似关系与相似类

3.1 基于证据理论的相似关系

由于定义 9 及定义 10 中,并未包括对缺失数据的取值状况和测度类型的说明,因而不便于进一步

分析.为此,我们首先对其进行改写.

实际上,当属性 $r \in R$ 的值域 V_r 为有限集,在没有任何先验信息时,对于缺失数据,我们显然可以判定取值必然属于 V_r ,但是我们无法判断其取值属于 V_r 任一真子集的可能性大小.

据此,由证据理论及信息检索中的相关方法^[23-25],对于缺失数据,即 $f(x, r) = \text{"null"}$,可构建如下 mass 函数:

$$m(f(x, r) = A) = \begin{cases} 0, & A \neq V_r \\ 1, & A = V_r \end{cases}, A \subseteq V_r \quad (3)$$

$m(f(x, r) = A)$ 表示记录 x 对应的属性 r 的属性值 $f(x, r)$ 在集合 A 中取值的可能性.由式(3)可以看出,当且仅当 $A = V_r$ 时该可能性为 1,否则为 0,即无法判断其在 V_r 某个真子集中取值的可能性大小.因此,可将 $f(x, r) = \text{"null"}$ 替换为相应属性的值域 V_r .

同样,对于非缺失数据,若有 $f(x, r) = k$,则有如下 mass 函数:

$$m(f(x, r) = A) = \begin{cases} 0, & A \neq k \\ 1, & A = k \end{cases}, A \subseteq V_r \quad (4)$$

即对于非缺失数据,其保持原有取值不变.

由 $m(f(x, r) = A)$,可分别计算 $f(x, r) = A$ 的似然测度 $Pls(f(x, r) = A)$ 和信任测度 $Bel(f(x, r) = A)$:

$$Pls(f(x, r) = A) = \sum_{L \cap A \neq \emptyset} m(f(x, r) = L) \quad (5)$$

$$Bel(f(x, r) = A) = \sum_{L \subseteq A} m(f(x, r) = L) \quad (6)$$

比较式(5),(6)可以看出,当 $m(f(x, r) = V_r) = 1$ 时(由 mass 函数性质可知,此时必有 $m(f(x, r) = A) = 0, A \subset V_r$ 且 $A \neq V_r$)有

$$Pls(f(x, r) = A) = 1,$$

$$Bel(f(x, r) = A) = 0.$$

不能看出,对于缺失数据取值可能性的态度,采用似然测度和信任测度是不同的.其中似然测度表示了相对乐观的处理态度,即认为可以完全判定缺失数据为任一值,而信任测度则表示了相对保守的处理态度,即认为缺失数据完全无法判定其取值.这也是与证据理论完全相符的.

由 $f(x, r)$ 的可能取值,可以构建相似关系.

定义 11(基于似然测度的相似关系). 设有知识表示系统 $S, x, y \in U$.称在属性集 $B \subseteq R$ 上记录 y 是基于似然测度与 x 相似的,当且仅当

$$\forall r \in B, Pls(f(y, r) = f(x, r)) = 1 \quad (7)$$

并记为 $xS_B y$.

定义 12(基于信任测度的相似关系). 设有知识表示系统 $S, x, y \in U$. 称在属性集 $B \subseteq R$ 上记录 y 是基于信任测度与 x 相似的, 当且仅当:

$$\forall r \in B, Bel(f(y, r) = f(x, r)) = 1 \quad (8)$$

并记为 $xS'_B y$.

可以看出, 定义 11、12 的区别在于缺失数据取值可能性的判定态度.

定理 1. 设有知识表示系统 S , 对于属性集合 $B \subseteq R$ 若有对于任意 $x \in U$ 有 $\forall r \in B, f(x, r) \in V_r \cup \{\text{"null"}\}$, 即 S 在属性集合 B 上为非多值数据库, 则在属性集 B 上定义 9 与定义 11 相同.

证明. 只需证 $Pls(f(y, r) = f(x, r)) = 1$ 与 $f(x, r) = f(y, r)$ 或 $f(x, r) = \text{"null"}$ 或 $f(y, r) = \text{"null"}$ 等价.

充分性:

① $f(x, r) = f(y, r)$, 则 $m(f(y, r) = f(x, r)) = 1$, 有 $Pls(f(y, r) = f(x, r)) = 1$.

② $f(x, r) = \text{"null"}$, 则 $m(f(x, r) = V_r) = 1$, 有 $Pls(f(y, r) = f(x, r)) = m(f(y, r) = f(y, r)) = 1$.

③ $f(y, r) = \text{"null"}$, 则 $m(f(y, r) = V_r) = 1$, 有 $Pls(f(y, r) = f(x, r)) = m(f(y, r) = V_r) = 1$.

充分性得证.

必要性:

利用反证法, 令 $Pls(f(y, r) = f(x, r)) = 1$ 时有 $f(x, r) \neq f(y, r)$ 且 $f(x, r) \neq \text{"null"}$ 且 $f(y, r) \neq \text{"null"}$.

不失一般性, 令 $f(y, r) = k, f(x, r) = l, k, l \in V_r, k \neq l$. 显然此时有

$$m(f(y, r) = A) = 0, A \neq \{k\},$$

因此

$$Pls(f(y, r) = f(x, r)) = \sum_{L \cap \{l\} \neq \emptyset} m(f(y, r) = L) = 0,$$

与条件相反. 因此, 有 $f(x, r) = f(y, r)$ 或 $f(x, r) = \text{"null"}$ 或 $f(y, r) = \text{"null"}$ 成立.

必要性得证.

证毕.

由定理 1 可知, 定义 11 实际是定义 9 在利用证据理论基础上的改写.

同样, 对定义 10 也可进行类似定义 9 的表述, 形式如下.

定义 13(相似关系 S'_B). 设有知识表示系统 S , 令有 U 上的相似关系 $S'_B (B \subseteq R)$:

$$xS'_B y \Leftrightarrow \forall r \in B, f(x, r) = f(y, r)$$

或 $f(x, r) = \text{"null"}, x, y \in U \quad (9)$

定理 2. 设有知识表示系统 S , 对于属性集合 $B \subseteq R$ 若有对于任意 $x \in U$ 有 $\forall r \in B, f(x, r) \in V_r \cup \{\text{"null"}\}$, 即 S 在属性集合 B 上为非多值数据库, 则在属性集 B 上定义 12 与定义 13 相同.

证明. 只需证 $Bel(f(y, r) = f(x, r)) = 1$ 与 $f(x, r) = f(y, r)$ 或 $f(x, r) = \text{"null"}$ 等价.

充分性:

① $f(x, r) = f(y, r)$, 则 $m(f(y, r) = f(x, r)) = 1$, 有 $Bel(f(y, r) = f(x, r)) = 1$.

② $f(x, r) = \text{"null"}$, 则 $m(f(x, r) = V_r) = 1$, 有 $Bel(f(y, r) = f(x, r)) = Bel(f(y, r) = V_r) = \sum_{L \subseteq V_r} m(f(y, r) = L) = 1$.

充分性得证.

必要性:

利用反证法, 令 $Bel(f(y, r) = f(x, r)) = 1$ 时有 $f(x, r) \neq f(y, r)$ 和 $f(x, r) \neq \text{"null"}$.

不失一般性, 可令 $f(x, r) = k, k \in V_r$, 则

$$Bel(f(y, r) = f(x, r)) = \sum_{L \subseteq \{k\}} m(f(y, r) = L) = m(f(y, r) = k).$$

由 $f(x, r) \neq f(y, r)$, 有 $m(f(y, r) = k) = 0$.

因此 $Bel(f(y, r) = f(x, r)) = 0$, 与条件矛盾.

因此, 必有 $f(x, r) = f(y, r)$ 或 $f(x, r) = \text{"null"}$ 成立.

必要性得证.

证毕.

由此可以看出, 定义 12 是定义 13 在证据理论基础上的改写.

对于相似关系 S_B, S'_B 的两种定义方式各有利弊, 其中定义 9、13 相对直观且便于计算, 而定义 11、12 则有利于分析和深入理解. 在本文中, 分析过程将采用定义 11、12 的形式, 而仿真实现则更多地采用定义 9、13 的方法.

定理 3. 设有知识表示系统 $S, x, y \in U$, 则 $xS_B y \Leftrightarrow yS_B x, xS'_B y \not\Leftrightarrow yS'_B x$.

证明. 由 \cap 运算的对称性易证 $xS_B y \Leftrightarrow yS_B x$.

由 \subseteq 的非对称性易证 $xS'_B y \not\Leftrightarrow yS'_B x$. 证毕.

由定理 3 可知, 基于似然测度的相似关系具有对称性, 而基于信任测度的相似关系则不具备对称性. 从而构建基于信任测度的相似关系时需要注意先后顺序.

3.2 基于证据理论的相似类构建

由相似关系, 可得以下相似类定义.

定义 14(基于似然测度的相似类). 设有知识

表示系统 S 及 U 上的基于似然测度的相似关系 S_B ($B \subseteq R$), 则由相似关系 S_B 可得 U 上的覆盖

$$U/S_B = \bigcup S_B(x),$$

称 U/S_B 为论域 U 上基于似然测度的相似类, $S_B(x) = \{y \in U \mid x S_B y, x \in U\}$ 为论域 U 上基于似然测度的相似范畴。

由定理 1 易证定义 10 与定义 14 相同。

定义 15(基于信任测度的相似类). 设有知识表示系统 S 及 U 上的基于信任测度的相似关系 S'_B ($B \subseteq R$), 则由相似关系 S'_B 可得 U 上的覆盖

$$U/S'_B = \bigcup S'_B(x),$$

称 U/S'_B 为论域 U 上基于信任测度的相似类, $S'_B(x) = \{y \in U \mid x S'_B y, x \in U\}$ 为论域 U 上基于信任测度的相似范畴。

由定理 3 易证有 $S_B(x) = S_B(y)$, 然而一般有 $S'_B(x) \neq S'_B(y)$ 。

由此, 有如下相似类计算方法:

① 对所有缺失值进行替换, 令有

$$f(x, r) = \text{"null"} \rightarrow f(x, r) = V_r,$$

即根据缺失数据所对应属性, 将空值替换为该属性值域。

② 根据等价关系 B 构建等价类, 即有

$$x I_B y \Leftrightarrow \forall r \in B, f(x, r) = f(y, r),$$

其中, 若存在 $f(x, r) = V_r$, 则 $f(x, r) = f(y, r)$ 当且仅当 $f(y, r) = V_r$ 。

记所得划分为 U/B 。

③ 按下述公式分别计算 U 上关于 $B \subseteq R$ 的相似范畴:

$$S_B([x]_B) = \bigcup_{\substack{x \in [x]_B, y \in U \\ (Pls(f(y, r) = f(x, r)) = 1)}} y = \bigcup_{\substack{x \in [x]_B, y \in [y]_B \\ f(y, B) \cap f(x, B) \neq \emptyset \\ p.s.}} [y]_B, \\ [x]_B, [y]_B \in U/B \quad (10)$$

$$S'_B([x]_B) = \bigcup_{\substack{x \in [x]_B, y \in U \\ Bel(f(y, r) = f(x, r)) = 1}} y = \bigcup_{\substack{x \in [x]_B, y \in [y]_B \\ \forall r \in B, f(y, r) \subseteq f(x, r)}} [y]_B, \\ [x]_B, [y]_B \in U/B \quad (11)$$

其中 $S_B([x]_B), S'_B([x]_B)$ 分别表示依据相似关系 S_B, S'_B 构建的包含范畴 $[x]_B$ 的相似范畴。其中 $f(x, B) \bigcap_{p.s.} f(y, B) \neq \emptyset \Leftrightarrow \forall r \in B, f(x, r) \cap f(y, r) \neq \emptyset$ 。

记 $U/S_B = \bigcup S_B([x]_B), U/S'_B = \bigcup S'_B([x]_B)$ 。

特别地, 当 B 为条件属性集 C 时, 有条件相似类 $U/S_C = \bigcup S_C([x]_C)$;

当 B 为决策属性集 D 时, 有决策相似类 $U/S_D = \bigcup S_D([x]_D)$ 。

4 缺失环境下分类不一致程度评价

4.1 基于相似关系的分类规则精度计算

为评价系统的分类不一致程度, 首先需要计算缺失环境下中规则的分类精度:

根据分类精度定义, 分别有分类精度^[29] $\tau(Y|X) = \frac{|X \cap Y|}{|X|}$, $X, Y \subseteq U$, 则有

$$\tau(S_D([x]_D) | S_C([x]_C)) = \frac{|S_C([x]_C) \cap S_D([x]_D)|}{|S_C([x]_C)|} \quad (12)$$

$$\tau(S'_D([x]_D) | S'_C([x]_C)) = \frac{|S'_C([x]_C) \cap S'_D([x]_D)|}{|S'_C([x]_C)|} \quad (13)$$

可以看出, 当决策属性存在缺失时, 由于 $S_D([x]_D)$ 及 $S'_D([x]_D)$ 为相似范畴, 可能不再构成 $S_C([x]_C)$ 及 $S'_C([x]_C)$ 上的划分, 且同时 U/S_C 及 U/S'_C 也不再构成论域 U 上的划分, 均不符合文献[28]中的概率测度要求, 为此需进行改进。

首先设 $U/C = \{C_i \mid 1 \leq i \leq m\}$, $U/D = \{D_j \mid 1 \leq j \leq n\}$ 。

定义如下函数:

$$m(D_i | Pls(C_i)) = \frac{\sum_{\substack{f(C_s) \cap f(C_i) \neq \emptyset \\ f(C_s) \cap f(C_i) \neq \emptyset}} |C_s \cap D_i|}{\sum_{\substack{f(C_s) \cap f(C_i) \neq \emptyset \\ f(C_s) \cap f(C_i) \neq \emptyset}} |C_s|} \\ = \frac{|S_C(C_i) \cap D_i|}{|S_C(C_i)|} \quad (14)$$

$$m(D_i | Bel(C_i)) = \frac{\sum_{\substack{f(C_s) \subseteq f(C_i) \\ f(C_s) \subseteq f(C_i)}} |C_s \cap D_i|}{\sum_{\substack{f(C_s) \subseteq f(C_i) \\ f(C_s) \subseteq f(C_i)}} |C_s|} \\ = \frac{|S'_C(C_i) \cap D_i|}{|S'_C(C_i)|} \quad (15)$$

其中 $f(C_s) \cap f(C_i) \neq \emptyset$ 表示 $\forall x \in C_i, y \in C_s$, 有 $f(x, C) \bigcap_{p.s.} f(y, C) \neq \emptyset$. $f(C_s) \subseteq f(C_i)$ 表示 $\forall x \in C_i, y \in C_s$, 有 $\forall c \in C, f(y, c) \subseteq f(x, c)$ 。

定理 4. 设有 $m(D_i | Pls(C_i)), m(D_i | Bel(C_i))$ 定义如前, 则 $m(D_i | Pls(C_i)), m(D_i | Bel(C_i))$ 分别为 $S_C(C_i)$ 及 $S'_C(C_i)$ 上关于 U/D 的 mass 函数。

证明. 显然有 $0 \leq m(D_i | Pls(C_i)) \leq 1, 0 \leq m(D_i | Bel(C_i)) \leq 1$ 。

根据 D_i 定义, 有

$$\begin{aligned} \sum_{\mathcal{D}_i \in U/D} m(\mathcal{D}_i | Pls(\mathcal{C}_i)) &= \sum_{\mathcal{D}_i \in U/D} \frac{\sum_{f(\mathcal{C}_s) \cap f(\mathcal{C}_i) \neq \emptyset} |\mathcal{C}_s \cap \mathcal{D}_i|}{\sum_{f(\mathcal{C}_s) \cap f(\mathcal{C}_i) \neq \emptyset} |\mathcal{C}_s|} \\ &= \frac{\sum_{f(\mathcal{C}_s) \cap f(\mathcal{C}_i) \neq \emptyset} |\mathcal{C}_s \cap U|}{\sum_{f(\mathcal{C}_s) \cap f(\mathcal{C}_i) \neq \emptyset} |\mathcal{C}_s|} = 1, \end{aligned}$$

因此 $m(\mathcal{D}_i | Pls(\mathcal{C}_i))$ 为 $S_C(\mathcal{C}_i)$ 上关于 U/D 上的 mass 函数。

同理可证 $m(\mathcal{D}_i | Bel(\mathcal{C}_i))$ 为 $S'_C(\mathcal{C}_i)$ 上关于 U/D 上的 mass 函数。证毕。

定理 5. $\tau(S_D(\mathcal{D}_j) | S_C(\mathcal{C}_i))$ 及 $\tau(S'_D(\mathcal{D}_j) | S'_C(\mathcal{C}_i))$ 分别为 $S_C(\mathcal{C}_i)$ 和 $S'_C(\mathcal{C}_i)$ 上关于 U/D 上的似然测度及信任测度。即

$$\begin{aligned} \tau(S_D(\mathcal{D}_j) | S_C(\mathcal{C}_i)) &= Pls(\mathcal{D}_j | Pls(\mathcal{C}_i)) \\ &= \sum_{f(\mathcal{D}_i) \cap f(\mathcal{D}_j) \neq \emptyset} m(\mathcal{D}_i | Pls(\mathcal{C}_i)) \end{aligned} \quad (16)$$

$$\begin{aligned} \tau(S'_D(\mathcal{D}_j) | S'_C(\mathcal{C}_i)) &= Bel(\mathcal{D}_j | Bel(\mathcal{C}_i)) \\ &= \sum_{f(\mathcal{D}_i) \subseteq f(\mathcal{D}_j)} m(\mathcal{D}_i | Bel(\mathcal{C}_i)) \end{aligned} \quad (17)$$

证明。由式(12)有

$$\begin{aligned} \tau(S_D(\mathcal{D}_j) | S_C(\mathcal{C}_i)) &= \frac{|S_C(\mathcal{C}_i) \cap S_D(\mathcal{D}_j)|}{|S_C([x]_C)|} \\ &= \frac{\sum_{f(\mathcal{D}_i) \cap f(\mathcal{D}_j) \neq \emptyset} |S_C(\mathcal{C}_i) \cap \mathcal{D}_i|}{\sum_{f(\mathcal{C}_s) \cap f(\mathcal{C}_i) \neq \emptyset} |\mathcal{C}_s|} \\ &= \sum_{f(\mathcal{D}_i) \cap f(\mathcal{D}_j) \neq \emptyset} \frac{\sum_{f(\mathcal{C}_s) \cap f(\mathcal{C}_i) \neq \emptyset} |\mathcal{C}_s \cap \mathcal{D}_i|}{\sum_{f(\mathcal{C}_s) \cap f(\mathcal{C}_i) \neq \emptyset} |\mathcal{C}_s|} \\ &= \sum_{f(\mathcal{D}_i) \cap f(\mathcal{D}_j) \neq \emptyset} m(\mathcal{D}_i | Pls(\mathcal{C}_i)). \end{aligned}$$

由定理 4 及定义 3 可知 $\tau(S_D(\mathcal{D}_j) | S_C(\mathcal{C}_i))$ 为 $S_C(\mathcal{C}_i)$ 关于 U/D 上的似然测度。

同理, 由式(13)可证 $\tau(S'_D(\mathcal{D}_j) | S'_C(\mathcal{C}_i)) = \sum_{f(\mathcal{D}_i) \subseteq f(\mathcal{D}_j)} m(\mathcal{D}_i | Bel(\mathcal{C}_i))$ 。

由定理 4 及定义 2 可知 $\tau(S'_D(\mathcal{D}_j) | S'_C(\mathcal{C}_i))$ 为 $S'_C(\mathcal{C}_i)$ 上关于 U/D 上的信任测度。证毕。

4.2 缺失环境下的分类不一致程度评价

根据以上的规则分类精度计算方法, 由定理 4、5 及混淆度及不协调度定义, 可对相似范畴 $S_C(\mathcal{C}_i)$ 及 $S'_C(\mathcal{C}_i)$ 上的分类不一致程度描述如下。

定义 16(相似范畴 $S_C(\mathcal{C}_i)$ 的不协调度)。设有知识表示系统 S 及各定义如前, 称

$$\begin{aligned} E(D | S_C(\mathcal{C}_i)) &= - \sum_{\mathcal{D}_j \in U/D} m(\mathcal{D}_j | S_C(\mathcal{C}_i)) \log Pls(\mathcal{D}_j | Pls(\mathcal{C}_i)) \\ &= - \sum_{\mathcal{D}_j \in U/D} \frac{|S_C(\mathcal{C}_i) \cap \mathcal{D}_j|}{|S_C(\mathcal{C}_i)|} \log \tau(S_D(\mathcal{D}_j) | S_C(\mathcal{C}_i)) \end{aligned} \quad (18)$$

为相似范畴 $S_C(\mathcal{C}_i)$ 的不协调度。

定义 17(相似范畴 $S'_C(\mathcal{C}_i)$ 的混淆度)。设有知识表示系统 S 及各定义如前, 称

$$\begin{aligned} C(D | S'_C(\mathcal{C}_i)) &= - \sum_{\mathcal{D}_j \in U/D} m(\mathcal{D}_j | S'_C(\mathcal{C}_i)) \log Bel(\mathcal{D}_j | Bel(\mathcal{C}_i)) \\ &= - \sum_{\mathcal{D}_j \in U/D} \frac{|S'_C(\mathcal{C}_i) \cap \mathcal{D}_j|}{|S'_C(\mathcal{C}_i)|} \log \tau(S'_D(\mathcal{D}_j) | S'_C(\mathcal{C}_i)) \end{aligned} \quad (19)$$

为相似范畴 $S'_C(\mathcal{C}_i)$ 的混淆度。

定义 16、17 分别反映了 $S_C(\mathcal{C}_i)$ 及 $S'_C(\mathcal{C}_i)$ 上分类的不一致程度, 然而由于 $S_C(\mathcal{C}_i)$ 及 $S'_C(\mathcal{C}_i)$ 为相似范畴, 不构成 U 上的划分。为计算系统 S 的分类不一致程度, 需对等价类 \mathcal{C}_i 的分类不一致程度进行描述。为此定义

$$m(\mathcal{D}_j | \mathcal{C}_i) = \frac{|\mathcal{C}_i \cap \mathcal{D}_j|}{|\mathcal{C}_i|} \quad (20)$$

易证 $m(\mathcal{D}_j | \mathcal{C}_i)$ 为等价范畴 \mathcal{C}_i 上关于决策属性的 mass 函数。

下面定义 $\mathcal{C}_i \in U/C$ 相对相似关系 S_R 及 S'_R 的分类不一致程度。

定义 18(等价范畴 \mathcal{C}_i 的条件不协调度)。设有知识表示系统 S 及各定义如前, 称

$$E(D | \mathcal{C}_i) = - \sum_{\mathcal{D}_j \in U/D} m(\mathcal{D}_j | \mathcal{C}_i) \log Pls(\mathcal{D}_j | Pls(\mathcal{C}_i)) \quad (21)$$

为等价范畴 \mathcal{C}_i 的不协调度。

定义 19(等价范畴 \mathcal{C}_i 的条件混淆度)。设有知识表示系统 S 及各定义如前, 称

$$C(D | \mathcal{C}_i) = - \sum_{\mathcal{D}_j \in U/D} m(\mathcal{D}_j | \mathcal{C}_i) \log Bel(\mathcal{D}_j | Bel(\mathcal{C}_i)) \quad (22)$$

为等价范畴 \mathcal{C}_i 的混淆度。

可以看出定义 18、19 与不协调度、混淆度定义并不完全一致。然而其基本构成仍由 mass 函数和似然、信任测度构成, 且相较于条件信息熵, 该定义

具有类似性质(见 4.3 节),为方便计,本文称其为等价范畴 \mathcal{C}_i 的条件不协调度、条件混淆度。

由 $Pls(\mathcal{D}_j | Pls(\mathcal{C}_i))$ 和 $Bel(\mathcal{D}_j | Bel(\mathcal{C}_i))$ 定义可以看出, $Pls(\mathcal{D}_j | Pls(\mathcal{C}_i))$ 表示了对缺失数据采取相对乐观的处理态度时分类规则精度,因此 $E(D | \mathcal{C}_i)$ 表示了对缺失数据采取相对乐观的处理态度时等价范畴 \mathcal{C}_i 的分类不一致程度;同理可知 $C(D | \mathcal{C}_i)$ 表示了对缺失数据采取相对保守的处理态度时等价范畴 \mathcal{C}_i 的分类不一致程度。

为计算系统分类不一致程度,我们采用了对等价范畴 \mathcal{C}_i 的条件不协调度、条件混淆度进行加权平均的方法,由此进一步定义知识表示系统 S 的加权平均分类不一致程度,令有

$$m(\mathcal{C}_i) = \frac{|\mathcal{C}_i|}{|U|}, \mathcal{C}_i \in U/C \quad (23)$$

则对应于相似关系 S_R 及 S'_R , 有系统 S 的加权平均分类不一致程度 $E(D | C)$ 及 $C(D | C)$ 。

定义 20(系统 S 的条件不协调度). 设有知识表示系统 S 及各定义如前,称

$$E(D | C) = \sum_{\mathcal{C}_i \in U/C} m(\mathcal{C}_i) E(D | \mathcal{C}_i) \quad (24)$$

为系统 S 的条件不协调度。

定义 21(系统 S 的条件混淆度). 设有知识表示系统 S 及各定义如前,称

$$C(D | C) = \sum_{\mathcal{C}_i \in U/C} m(\mathcal{C}_i) C(D | \mathcal{C}_i) \quad (25)$$

为系统 S 的条件混淆度。

4.3 几个基本性质

定理 6. $E(D | C) \geq 0$ 且等号成立当且仅当 $\forall \mathcal{C}_i \in U/C, \mathcal{D}_j \in U/D$, 若 $m(\mathcal{D}_j | \mathcal{C}_i) > 0$, 则必有 $\tau(S_D(\mathcal{D}_j) | S_C(\mathcal{C}_i)) = 1$; $C(D | C) \geq 0$ 且等号成立当且仅当 $\forall \mathcal{C}_i \in U/C, \mathcal{D}_j \in U/D$, 若 $m(\mathcal{D}_j | \mathcal{C}_i) > 0$, 则必有 $\tau(S'_D(\mathcal{D}_j) | S'_C(\mathcal{C}_i)) = 1$ 。

证明. 由 mass 函数性质及式 (19)、(20)、(22)、(23) 易证 $E(D | C) \geq 0$ 及 $C(D | C) \geq 0$ 。

由式 (21) 易知 $\forall \mathcal{C}_i \in U/C$, 有 $m(\mathcal{C}_i) > 0$ 。

由式 (22) 易知 $E(D | C) = 0$ 当且仅当 $\forall m(\mathcal{C}_i) > 0$, 必有 $E(D | \mathcal{C}_i) = 0$ 。

由式 (20) 易知 $E(D | \mathcal{C}_i) = 0$ 当且仅当 $\forall m(\mathcal{D}_j | \mathcal{C}_i) > 0$, 必有 $Pls(\mathcal{D}_j | Pls(\mathcal{C}_i)) = 1$ 。

由此可证 $E(D | C) = 0$ 当且仅当 $\forall \mathcal{C}_i \in U/C$, 若 $m(\mathcal{D}_j | \mathcal{C}_i) > 0$, 必有 $Pls(\mathcal{D}_j | Pls(\mathcal{C}_i)) = 1$ 。

同理可证 $C(D | C) = 0$ 的情况。证毕。

由定理 6 可见, 当且仅当 $\tau(S_D(\mathcal{D}_j) | S_C(\mathcal{C}_i)) =$

1 及 $\tau(S'_D(\mathcal{D}_j) | S'_C(\mathcal{C}_i)) = 1$ 时, $E(D | C)$ 及 $C(D | C)$ 达到最小, 即 $E(D | C) = 0, C(D | C) = 0$ 当且仅当相对应的分类规则均为确定性规则, 显然与分类不一致程度评价的要求相吻合。

定理 7. 当且仅当 $m(\mathcal{D}_j | \mathcal{C}_i) = Pls(\mathcal{D}_j | Pls(\mathcal{C}_i)) = \frac{1}{|V_D|}$, $E(D | C)$ 达到最大; 当且仅当 $m(\mathcal{D}_j | \mathcal{C}_i) = Bel(\mathcal{D}_j | Bel(\mathcal{C}_i)) = \frac{1}{|V_D|}$, $C(D | C)$ 达到最大。

证明. 由 mass 函数及 $Pls(\mathcal{D}_j | Pls(\mathcal{C}_i))$, $Bel(\mathcal{D}_j | Bel(\mathcal{C}_i))$ 性质即证。证毕。

由定理 7 可以看出, 当且仅当对于任一 \mathcal{C}_i , 其分类规则精度服从等概分布时, $E(D | C)$ 和 $C(D | C)$ 达到最大。

由定理 6, 7 可以看出, $E(D | C)$ 及 $C(D | C)$ 对分类不一致程度进行了较好描述。

定义 21(联合 mass 函数). 设有知识表示系统 S 及各部分定义如前. 称

$$m(\mathcal{C}_i, \mathcal{D}_j) = \frac{|\mathcal{C}_i \cap \mathcal{D}_j|}{|U|}, [x]_C \in U/C, [y]_D \in U/D \quad (26)$$

为 U 上关于属性集 $(C \cup D)$ 上的联合 mass 函数。

由 mass 函数定义, 易证 $m(\mathcal{C}_i, \mathcal{D}_j)$ 为一 mass 函数。

依据该 mass 函数和不协调度、混淆度定义有

$$\begin{aligned} E(D, C) &= - \sum_{\substack{\mathcal{C}_i \in U/C \\ \mathcal{D}_j \in U/D}} m(\mathcal{C}_i, \mathcal{D}_j) \log Pls(\mathcal{C}_i, \mathcal{D}_j) \\ &= - \sum_{\substack{\mathcal{C}_i \in U/C \\ \mathcal{D}_j \in U/D}} m(\mathcal{C}_i, \mathcal{D}_j) \log \sum_{\substack{f(\mathcal{C}_i) \cap f(\mathcal{C}_i) \neq \emptyset \\ f(\mathcal{D}_i) \cap f(\mathcal{D}_j) \neq \emptyset}} m(\mathcal{C}_s, \mathcal{D}_t) \end{aligned} \quad (27)$$

$$\begin{aligned} C(D, C) &= - \sum_{\substack{\mathcal{C}_i \in U/C \\ \mathcal{D}_j \in U/D}} m(\mathcal{C}_i, \mathcal{D}_j) \log Bel(\mathcal{C}_i, \mathcal{D}_j) \\ &= - \sum_{\substack{\mathcal{C}_i \in U/C \\ \mathcal{D}_j \in U/D}} m(\mathcal{C}_i, \mathcal{D}_j) \log \sum_{\substack{f(\mathcal{C}_i) \subseteq f(\mathcal{C}_i) \\ f(\mathcal{D}_i) \subseteq f(\mathcal{D}_j)}} m(\mathcal{C}_s, \mathcal{D}_t) \end{aligned} \quad (28)$$

为 U 上关于属性集 $(C \cup D)$ 上的联合不协调度与联合混淆度。

不难看出, 联合不协调度 $E(C, D)$ 与联合混淆度 $C(C, D)$ 实际即是依据属性集 R 上的 mass 函数 $m(R_i)$ 构成的不协调度和混淆度。

同理, 由 $m(\mathcal{C}_i)$, 有条件属性的混淆度及不协调度:

$$\begin{aligned}
E(C) &= - \sum_{C_i \in U/C} m(C_i) \log Pls(C_i) \\
&= - \sum_{C_i \in U/C} m(C_i) \log \sum_{f(C_s) \cap f(C_i) \neq \emptyset} m(C_s) \quad (29)
\end{aligned}$$

$$\begin{aligned}
C(C) &= - \sum_{C_i \in U/C} m(C_i) \log Bel(C_i) \\
&= - \sum_{C_i \in U/C} m(C_i) \log \sum_{f(C_s) \subseteq f(C_i)} m(C_s) \quad (30)
\end{aligned}$$

定理 8. 设有 $E(D|C), E(D, C), E(C)$ 及 $C(D|C), C(D, C), C(C)$ 定义如上, 则

$$E(D|C) = E(D, C) - E(C) \quad (31)$$

$$C(D|C) = C(D, C) - C(C) \quad (32)$$

证明.

$$\begin{aligned}
E(D, C) &= - \sum_{C_i \in U/C} m(C_i, D_j) \log Pls(C_i, D_j) \\
&= - \sum_{C_i \in U/C} \sum_{D_j \in U/D} m(C_i, D_j) \log Pls(C_i, D_j) \\
&= - \sum_{C_i \in U/C} \sum_{D_j \in U/D} [m(D_j|C_i) m(C_i)] \cdot \\
&\quad \log [Pls(D_j|C_i) Pls(C_i)] \\
&= - \sum_{C_i \in U/C} m(C_i) \sum_{D_j \in U/D} m(D_j|C_i) \cdot \\
&\quad [\log Pls(D_j|C_i) + \log Pls(C_i)] \\
&= - \sum_{C_i \in U/C} m(C_i) \sum_{D_j \in U/D} m(D_j|C_i) \cdot \\
&\quad \log Pls(D_j|C_i) - \\
&\quad \sum_{C_i \in U/C} m(C_i) \sum_{D_j \in U/D} m(D_j|C_i) \log Pls(C_i) \\
&= \sum_{C_i \in U/C} m(C_i) E(D|C_i) - \\
&\quad \sum_{C_i \in U/C} m(C_i) \log Pls(C_i) \\
&= E(D|C) + E(C).
\end{aligned}$$

同理可证 $C(D|C) = C(D, C) - C(C)$. 证毕.

由定理 8 可以看出, $E(D|C)$ 可表示为联合不协调度与条件属性集的不协调度之差, 类似于条件信息熵可表示为联合信息熵与条件属性信息熵之差, 这也是本文称 $E(D|C)$ 为条件不协调度的一个原因. 同样, $C(D|C)$ 也具有类似性质.

由定理 8 有如下 $E(D|C)$ 及 $C(D|C)$ 的计算方法:

① 根据 3.2 节, 分别计算 $U/R, U/S_R, U/S'_R$, $R = (C, D)$ 及 $U/C, U/S_C, U/S'_C$;

② 分别按下式计算 $E(D, C), C(D, C)$ 及 $E(C)$,

$C(C)$:

$$E(D, C) = - \sum_{[x]_R \in U/R} \frac{|[x]_R|}{|U|} \log \frac{|S_R([x]_R)|}{|U|},$$

$$C(D, C) = - \sum_{[x]_R \in U/R} \frac{|[x]_R|}{|U|} \log \frac{|S'_R([x]_R)|}{|U|},$$

$$E(C) = - \sum_{[x]_C \in U/C} \frac{|[x]_C|}{|U|} \log \frac{|S_C([x]_C)|}{|U|},$$

$$C(C) = - \sum_{[x]_C \in U/C} \frac{|[x]_C|}{|U|} \log \frac{|S'_C([x]_C)|}{|U|}.$$

③ 依据定理 8, 计算 $E(D|C) = E(D, C) - E(C), C(D|C) = C(D, C) - C(C)$.

5 算例分析

为对本文方法进行分析, 我们首先选取 UCI (<http://www.ics.uci.edu/~mlearn>) 的 mushroom 数据集进行测试.

mushroom 数据库共有 23 个属性, $C = \{\#1, \#2, \dots, \#22\}$ 为条件属性, $D = \{\#23\}$ 为决策属性; 8124 条记录, 2480 条记录存在属性值缺失, 且全部集中在属性 #11.

此时有缺失纪录百分比近似为 30.53%.

由式 (31)、(32), 不一致程度如分别为表 1 所示.

表 1 对 mushroom 数据库的计算结果

	$C(X)$	$E(X)$
$X = (C, D)$	12.9879	12.9879
$X = (C)$	12.9879	12.9879
$X = (D C)$	0	0

可以看出, mushroom 数据集的分类不一致程度在两种状况下均为 0.

通过对 mushroom 数据集进一步分析, 可以发现条件属性 #11 是 mushroom 数据集中可省略的, 即删除属性 #11 不影响记录间的相似性, 因此 mushroom 数据集中缺失不影响分类不一致程度. 而当删除 mushroom 数据集中属性 #11 时, 其分类不一致程度为 0. 因而, mushroom 数据集的分类不一致程度均为 0.

为对存在不可省略属性的缺失环境下的分类不一致程度进行评价, 我们进一步选取 UCI 的 car 数据集进行测试.

car 数据集共包含 7 个属性 (#1: buying; #2: maint; #3: doors; #4: persons; #5: lug_boot; #6: safety; #7: class), $C = \{\#1, \#2, \#3, \#4, \#5, \#6\}$

分别为条件属性,均不可省略; $D=\{\#7\}$ 为决策属性.所有属性均为名义尺度,其中条件属性 $\#1\sim\#3$ 均有 4 种属性值,条件 $\#4\sim\#6$ 各有 3 种属性值,决策属性 $\#7$ 有 4 种属性值;整个数据集共包含 1728 条记录,不存在缺失.

(1) 原分类不一致程度为 0 的情况

为便于分析,分别对 car 数据库中的记录,由第一条记录起,按照步长(*step*)分别删除相应记录的属性值作为缺失数据,然后对 $E(D|C),C(D|C),I(D|C)$ 分别进行计算.

对 car 数据库的分类不一致程度的计算结果如表 2 所示.

表 2 对 car 数据库的计算结果

	$C(X)$	$E(X)$
$X=(C,D)$	10.7548875	10.7548875
$X=(C)$	10.7548875	10.7548875
$X=(D C)$	0	0

显然,由于原数据库中所有记录均不相似,且不存在缺失,因此分类不一致程度为 0.

属性 $\#1$ 在不同程度缺失下的分类不一致程度的计算结果如表 3 所示.

表 3 属性 #1 的计算结果

<i>Step</i>	缺失百分比	$E(D C)$	$C(D C)$	$I(D C)$
1	100.00	0.3933601	0.3933601	0.3933601
5	19.97	0.2046000	0.0793468	0.0115741
10	9.95	0.1163516	0.0399628	0
20	4.98	0.0592131	0.0204400	0
40	2.49	0.0310533	0.0107986	0
60	1.62	0.0203965	0.0070863	0
80	1.22	0.0161054	0.0056887	0
100	0.98	0.0123930	0.0042911	0
200	0.46	0.0071846	0.0025550	0

从表 3 可以看出,当所有记录属性 $\#1$ 全部丢失时($step=1$),分类不一致程度达到最大,且有 $E(D|C)=C(D|C)=I(D|C)$,此时相当于删除属性 $\#1$.由于属性 $\#1$ 不可省略,从而造成部分具有不同决策属性值的纪录合并,因此分类不一致程度大于 0.又由于其它属性不存在缺失,故有 $S_{(R-\#1)}=S'_{(R-\#1)}=I_{(R-\#1)}$,即与等价关系相同.

同时还可以发现,随着缺失百分比的降低, $E(D|C),C(D|C),I(D|C)$ 均有不同程度的下降,且越来越趋近于原数据集的分类不一致程度.当 $step=10$ 后,由 $I(D|C)=0,E(D|C)>C(D|C)>0$,可以看出等价关系往往难以真实反映缺失数据集的分类不一致程度,而要结合实际挖掘需要选取相应评价方法.

属性 $\#2$ 在不同程度缺失时的分类不一致程度的计算结果如表 4 所示.

表 4 属性 #2 的计算结果

<i>Step</i>	缺失百分比	$E(D C)$	$C(D C)$	$I(D C)$
1	100.00	0.3438074	0.3438074	0.3438074
5	19.97	0.1799882	0.0741272	0.0205496
10	9.95	0.1016438	0.0385103	0.0066607
20	4.98	0.0602722	0.0214007	0.0011574
40	2.49	0.0309550	0.0107004	0
60	1.62	0.0208768	0.0069880	0
80	1.22	0.0140307	0.0047715	0
100	0.98	0.0155267	0.0056887	0
200	0.46	0.0083420	0.0031337	0

可以看出,当所有记录属性 $\#2$ 全部丢失时($step=1$),与属性 $\#1$ 相似,有分类不一致程度达到最大,随着缺失百分比的降低, $E(D|C),C(D|C),I(D|C)$ 也均有不同程度的下降,且越来越趋近于原数据集的分类不一致程度.然而其中当 $step=100$ 时,分类不一致程度较 $step=80$ 时相对增加.由抽样方法来看,有 $step=100$ 时的缺失记录与 $step=80$ 时不同,因此可以看出,除了缺失百分比对分类不一致程度评价存在影响外,缺失的位置也对分类不一致程度评价存在影响.

属性 $\#5$ 在不同程度缺失时的分类不一致程度的计算结果如表 5 所示.

表 5 属性 #5 的计算结果

<i>Step</i>	缺失百分比	$E(D C)$	$C(D C)$	$I(D C)$
1	100.00	0.4428990	0.4428990	0.4428990
5	19.97	0.1953025	0.0960814	0.0295252
10	9.95	0.1017022	0.0473911	0.0115741
20	4.98	0.0560377	0.0251697	0.0046296
40	2.49	0.0277568	0.0128524	0.0034722
60	1.62	0.0234545	0.0095656	0
80	1.22	0.0100895	0.0037237	0
100	0.98	0.0110502	0.0046845	0
200	0.46	0.0039091	0.0015943	0

分别考虑属性 $\#1$ 、属性 $\#2$ 、属性 $\#5$ 的不同缺失百分比下的分类不一致程度,可以发现虽然每个属性重要程度($I(D|C),step=1$)不同,然而当其不存在明显差异时,属性重要程度与分类不一致程度并没有绝对关系,分类不一致程度仅与数据的缺失程度和具体缺失位置相关.

(2) 原分类不一致程度大于 0 的情况

由于 car 数据集本身分类不一致程度为 0,为对原有分类不一致程度大于 0 的情况分析,我们将原有 car 数据集中属性 $\#1$ 删除,并记删除后的数据集为 car' .

有 car' 的分类不一致程度 $E(D|C)=C(D|C)=I(D|C)=0.3933601$.

car' 在属性 # 2 不同程度缺失时的分类不一致程度的计算结果如表 6 所示.

表 6 car' 在属性 # 2 不同缺失时的计算结果

Step	缺失百分比	$E(D C)$	$C(D C)$	$I(D C)$
1	100.00	0.5438998	0.5438998	0.5438998
5	19.97	0.4912157	0.3832346	0.3300474
10	9.95	0.4506710	0.3968317	0.3619448
20	4.98	0.4268846	0.3959622	0.3725094
40	2.49	0.4096396	0.3937350	0.3809802
60	1.62	0.4037782	0.3933491	0.3837093
80	1.22	0.4020456	0.3931188	0.3854340
100	0.98	0.4057802	0.3956813	0.3875957
200	0.46	0.4004236	0.3943425	0.3897574

car' 在属性 # 5 不同程度缺失时的分类不一致程度的计算结果如表 7 所示.

表 7 car' 在属性 # 5 不同缺失时的计算结果

Step	缺失百分比	$E(D C)$	$C(D C)$	$I(D C)$
1	100.00	0.7553367	0.7553367	0.7553367
5	19.97	0.5587774	0.4281899	0.3669704
10	9.95	0.4870195	0.4154460	0.3731899
20	4.98	0.4467943	0.4050601	0.3740810
40	2.49	0.4169508	0.3971903	0.3817008
60	1.62	0.4182403	0.3992423	0.3837093
80	1.22	0.4018699	0.3945174	0.3880326
100	0.98	0.4027757	0.3945797	0.3875957
200	0.46	0.3952566	0.3923227	0.3897573

综合分析表 6 和表 7,可以看出,当原有分类不一致程度大于 0 时,数据集的分类不一致程度在原 car' 表的分类不一致程度附近呈现随机波动,且随着缺失程度的降低, $E(D|C),C(D|C),I(D|C)$ 均一致趋近于 car' 的分类不一致程度.然而,当属性 # 2 缺失时, $C(D|C)$ 较 car' 的分类不一致程度低,而当属性 # 5 缺失时, $C(D|C)$ 较 car' 的分类不一致程度高.这也进一步说明了缺失对分类不一致程度的影响,不但与缺失数量相关,也同时与具体的缺失位置相关.

由以上分析可见,本文方法较好地评价了存在缺失数据时的分类不一致程度,同时当数据集不存在缺失时,本文评价方法与文献[7]相同.

6 结 语

现实世界中,由于各种原因数据集往往存在着不同程度的缺失,在先验知识相对较少时,利用相似关系代替等价关系是当前缺失数据集学习中的一种

常用方法.因此,本文对基于相似关系的缺失数据集的分类不一致程度评价进行了研究.

由于当前数据分析通常是反映信息粒之间的关系,因此我们首先构建了缺失环境的信息粒.由常用相似关系,我们根据证据理论进行了分析,并给出了基于信任度与似然度的信息粒构建方法.在此基础上,我们利用相似范畴对分类精度的计算进行分析,为分类不一致程度计算奠定了基础.由 mass 函数构建可知,此时分类精度可表示为信任度和似然度.由模糊熵理论,我们给出类似于不协调度和混淆度的系统分类不一致程度评价方法.

进一步可见,本文的评价方法可看作是条件属性与决策属性的联合不协调度及混淆度中去除条件属性的不协调度及混淆度,从而具有类似于信息熵的性质,同时也简化了系统分类不一致程度的计算.

由算例分析可以看出,本文方法较好地描述了对缺失环境下的系统分类不一致程度,同时当数据集不存在缺失时,本文方法与以往研究具有相同结果.

参 考 文 献

[1] Quinlan J R. Induction of decision trees. Machine Learning, 1986, 1: 81-106

[2] Lior R, Oded M. Top-down induction of decision trees classifiers—A survey. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 2005, 35(5): 476-487

[3] Salvatore G, Benedetto M, Roman S. Extension of the rough set approach to multicriteria decision support. INFOR Journal, 2000, 38: 161-195

[4] Wojciech Z. Variable precision rough set model. Journal of Computer and System Sciences, 1993, 46(2): 39-59

[5] Zhang Wen-Xiu, Wu Wei-Zhi et al. The Theory and Method of Rough Set. Beijing: Science Press, 2001(in Chinese)
(张文修, 吴伟志等. 粗糙集理论与方法. 北京: 科学出版社, 2001)

[6] Silva L M, Alexandre L A, de Sá J M. Neural network classification: Maximizing zero-error density//Proceedings of the 3rd International Conference on Advances in Pattern Recognition—ICAPR 2005. Bath, United Kingdom, 2005: 127-135

[7] Duntsch I, Gediga G. Uncertainty measures of rough set prediction. Artificial Intelligence, 1998, 106: 109-137

[8] Chen Xiang-Hui, Zhu Shan-Jun, Ji Yin-Dong. Entropy based uncertainty measures for classification rules with inconsistency tolerance//Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. Nashville, TN, USA, 2000, 4: 2816-2821

- [9] Chen Xiang-Hui, Zhu Shang-Jun, Ji Ning-Dong. Uncertain measure of rules based on entropy and Variable precision rough set model. *Journal of Tsinghua University (Science and Technology)*, 2001, 41(3): 109-112(in Chinese)
(陈湘晖, 朱善君, 吉吟东. 基于熵和变精度粗糙集的规则不确定性量度. *清华大学学报*, 2001, 41(3): 109-112)
- [10] Yu Hong, Wang Guo-Yin et al. Knowledge reduction algorithms based on rough set and conditional information entropy//*Proceedings of the International Society for Optical Engineering*. Orlando, USA, 2002, 4730: 422-431
- [11] Roman S, Jerzy S. Rough classification in incomplete information systems. *Mathematical and Computer Modelling (Oxford)*, 1989, 12(10-11): 1347-1357
- [12] Wang Shou-Hong. Classification with incomplete survey data: A Hopfield neural network approach. *Computers and Operations Research*, 2005, 32(10): 2583-2594
- [13] Hisao I, Akihiro M et al. Learning from incomplete training data with missing values and medical application//*Proceedings of the International Joint Conference on Neural Networks*. Nagoya, Japan, 1993: 1871-1874
- [14] Zhang Min, Cheng Jia-Xing. The research on the classification of the incomplete information system//*Proceedings of the 2004 International Conference on Machine Learning and Cybernetics*. Shanghai, China, 2004, 7: 3781-3786
- [15] Kryszkiewicz M. Rough set approach to incomplete information systems. *Information Sciences*, 1998, 112(1-4): 39-49
- [16] Hisao I, Akihiro M, Hideo T. Neural-network-based diagnosis systems for incomplete data with missing inputs//*Proceedings of the IEEE International Conference on Neural Networks*. Orlando, USA, 1994: 3457-3460
- [17] Bogdan G. Pattern classification for incomplete data//*Proceedings of the International Conference on Knowledge-Based Intelligent Electronic Systems*. Brighton, England, 2000: 454-457
- [18] Zhang Mei, Xu Li-Da. A rough set approach to knowledge reduction based on inclusion degree and evidence reasoning theory. *Expert Systems*, 2003, 20(5): 298-304
- [19] Li Yu-He, Dai Hai-Hong. Reducing uncertainties in data mining//*Proceedings of the APSEC'97 and ICSC'97, Software Engineering Conference*. Hong Kong, 1997: 97-105
- [20] Liang Ji-Ye, Xu Zhong-Ben et al. Reduction of knowledge in incomplete information systems//*Proceedings of the 16th World Computer Congress*. Beijing, China, 2000
- [21] Liang Ji-Ye, Xu Zhong-Ben. Uncertainty measures of roughness of knowledge and rough sets in incomplete. *Information Systems//Proceedings of the 3rd World Congress on Intelligent Control and Automation*. Hefei, China, 2000, 4: 2526-2529
- [22] McClean S, Scotney B et al. Aggregation of imprecise and uncertain information in databases. *IEEE Transactions on Knowledge and Data Engineering*, 2001, 13(6): 902-912
- [23] Lee S K. Imprecise and uncertain information in databases; An evidential approach//*Proceedings of the 8th International Conference on Data Engineering*. Tempe, USA, 1992: 614-621
- [24] Bell D A, Guan J W et al. Generalized union and project operations for pooling uncertain and imprecise information. *Data and Knowledge Engineering*, 1996, 18: 89-117
- [25] Vladarean C-M. About aggregation of imprecise evidence in database//*Proceedings of the 25th International Conference Information Technology Interfaces*. Cavtat, Croatia, 2003: 167-172
- [26] Yager R R. Measuring the information and character of a fuzzy measure//*Proceedings of the IFSA World Congress and the 20th NAFIPS International Conference*. Vancouver, BC, 2001, 3: 1718-1722
- [27] Zhang Wen-Xiu, Liang Yi. *Theory of uncertain reasoning*. Xi'an: Xi'an Jiaotong University Press, 1996(in Chinese)
(张文修, 梁怡. 不确定性推理原理. 西安: 西安交通大学出版社, 1996)
- [28] Peters J F, Pawlak Z et al. A rough set approach to measuring information granules//*Proceedings of the 26th Annual International Computer Software and Applications Conference*. Oxford, England, 2002: 1135-1139
- [29] Pawlak Z. Why rough sets. *Fuzzy Systems//Proceedings of the 5th IEEE International Conference*. New Orleans, USA, 1996, 2: 738-743
- [30] Ruan Da, Huang Chong-Fu. *Fuzzy Sets and Fuzzy Information-Granulation Theory: Key Selected Papers by Lotfi A Zadeh*. Beijing: Beijing Normal University Press, 2000



ZHANG Wei-Gang, born in 1975, Ph. D., lecturer. His research interesting includes data mining, rough set, automatic reasoning, etc.

PAN Quan, born in 1961, professor, Ph. D. supervisor. His research interesting includes object identification and tracking, data fusion, etc.

ZHANG Hong-Cai, born in 1938, professor, Ph. D. supervisor. His research interesting includes cybernetics, nonlinear system, etc.

Background

This is partially supported by the National Natural Science Foundation of China under grant No. 60172037 and Science Research Foundation of Civil Aviation University of China under grant No. 05qd08q.

ence Research Foundation of Civil Aviation University of China under grant No. 05qd08q.

In the field of data mining and machine learning, how to appraise an algorithm in theory research and select a proper one for an application in practice is crucial. And the inconsistency measure, one of the most important characters of database for learning classification rules, is utilized popularly in such fields. However, most of the tradition measures are based on rough entropy theory, which requires no data missing be present. Since the databases in real life are often incomplete and learning with such database has received many attentions in recent years, estimating inconsistency with incomplete database is significative.

In this paper, the authors consider this problem with two steps: matching of missing data and formulizing of measures. In the first step, the authors utilize two different similar relations that are based on evidence theory and be used widely in field of fuzzy query of information to reflect optim-

ism and conservative attitudes with missing data instead of equivalence relation, the basis of tradition measure based on rough entropy. In the second step, the authors extend the tradition measure with these two similar relations and give two measures by use of evidence theory and fuzzy entropy theory. The measures can be seen as a nature extent of tradition measure based rough entropy with evidence theory and fuzzy entropy and have reasonable constructions in mathematics and well understandability in intuition. At the same time, the difference between these two measures makes a opportunity to estimate the influence of different learning algorithms than just using one measure and reflects the essence of missing data. It may be helpful to evaluate and select learning algorithm, and also can be used to modify learning algorithm based on information entropy to learn with incomplete database.