

用 Shannon 熵度量两个数据集的一致性

车晓雅 米据生

(河北师范大学数学与信息科学学院 石家庄 050024)

摘要 粗糙集理论的基本思想是根据已知数据自身的不可分辨关系,通过一对近似算子,对某一给定概念进行近似表示。这种思想被应用在研究一个数据集对于另一个数据集的分类一致性上。提出了一种测量两个数据集一致性的新方法,并用 Shannon 熵定义了分类一致性。考虑到不同数据临近关系的影响,引入了模糊概念将测量对象由清晰分类转化为模糊分类,进而构造了一个广义的一致性度量,这种方法可以产生稳定的可判结果,有效地阻止建模技术中常出现的“黑箱”现象。

关键词 一致性程度,不可辨识关系,模糊划分,Shannon 熵

中图分类号 O236 文献标识码 A DOI 10.11896/j.issn.1002-137X.2016.1.014

Measuring Consistency of Two Datasets Using Shannon Entropy

CHE Xiao-ya MI Ju-sheng

(College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang 050024, China)

Abstract The basic idea of rough set theory is based on an indiscernibility relation, and through a pair of approximate operators, it can approximatively represent a given concept. It is used in the study of a data set for classification consistency to another data set. This paper presented a new approach to measure consistency degree of two datasets, and defined classification consistency by Shannon entropy. Taking the influence of neighborhood relations of different data into account, a general consistency measure was defined by introducing the expert knowledge into a fuzzy inference system, then we constructed a consistent generalized metric. Moreover, this method can prevent the “black box” phenomenon encountered in many modeling techniques and produce robust and interpretable results.

Keywords Consistency degree, Indiscernibility relation, Fuzzy partition, Shannon entropy

1 引言

粗糙集理论是波兰数学家 Pawlak 教授于 1982 年提出的一种数学方法,这种方法可以有效地处理不一致、不完整、不精确的信息^[7]。经过三十多年的发展,该理论已经成功地应用到数据挖掘、知识发现、机器学习、模式识别、决策分析和专家系统等领域。

近年来,科学工作者已经开发出许多利用多数据集间复杂关系的数据挖掘方法,基于统计学的方法是最常见的,包括:线性回归分析^[12]、主成分分析^[5]、多维尺度^[4]和多因素分析^[6]等。上述方法因为能够很好地识别不同来源的线性模式并在大量数值数据中发现数据属性间的相关性,所以被广泛地应用在经济、医学、生物学、化学和工程学^[1]等研究领域。

实际生活中,在不同数据集之间的关系模型中通常会遇到不确定性和不精确性问题,经典方法在解决此类问题时存在很大弊端。基于此情况,各学科工作者将模糊概念引入智能计算工具中并成功应用^[4,12]。粗糙集理论已经广泛地应用于测量基于分类的一致性和两数据集的包含度。实践中,相较于经典方法,粗糙集理论更加适合处理小型数据集。

本文在传统的包含度概念中,通过引入模糊技术来测量对象的相关性,将测量对象由清晰分类转化为模糊分类。这种方法能够防止在建模技术中常遇到的“黑箱”问题。

我们利用 Shannon 熵代替文献^[13]中的信息度量,从而得到另外一种分类一致性的定义,进而重新度量两数据集的一致性。

2 两数据集模糊分类的一致性

1982 年波兰数学家 Pawlak 用粗糙集理论定义了一个数据集对另一个数据集基于一致性的分类,这是一个分析解决不完整数据的新计算工具。不可辨识关系和描述不同对象集的等价关系是粗糙集理论的核心概念,也是本文所提方法的主要理论基础^[7]。

粗糙集的一个基本假设是论域中每个对象均与知识(或数据)有关,以相同信息为特征的有效知识是不可辨识的(相似的),以此种方式产生的不可辨识关系是粗糙集理论的数学基础^[8]。一个确定集合表示所有知识能被划分为可辨识粒,粗糙集所有的知识不能被细分和辨识,因此粗糙集理论中知识粒的信息由不可辨识关系表示。

到稿日期:2015-04-30 返修日期:2015-06-09 本文受国家自然科学基金(61170107,61300153,61300121,61573127,61502144),河北省高校创新团队领军人才培养计划项目(LJRC022),河北省自然科学基金(A2014205157, A2013208175)资助。

车晓雅(1991-),女,硕士生,主要研究方向为人工智能的数学基础,E-mail:chexiaoya@163.com;米据生(1966-),男,教授,主要研究方向为粗糙集、概念格、近似推理和随机集等,E-mail:mijsh@263.net。

2.1 基本概念

由上述分析可知,粗糙集理论事实上就是用于处理信息、数据和知识的分类问题。比较两个数据集时,它们不一定只有隶属关系,很多时候需要处理其包含关系^[10]。先给出如下概念。

定义 1 称 $S=(U, C \cup D, V, f)$ 是决策表,其中 $U=\{x_1, x_2, x_3, \dots, x_n\}$ 是非空有限对象集; C 是条件属性集合, D 是决策属性集合, V 是属性值的集合, $f: U \times C \cup D \rightarrow V$ 称为信息函数,它指定 U 中每一个对象的属性值。

任意属性子集 $B \subseteq C$ 决定一个二元不可辨识关系 R_B , 定义为:

$$R_B = \{(x_i, x_j) \in U \times U \mid \forall a \in B, f(x_i, a) = f(x_j, a)\}$$

显然, R_B 是集合 U 上的一个等价关系,对于 $B \subseteq C$, 关系 R_B 产生 U 的一个划分,记作 U/R_B 或 U/B 。以这种方法定义的不可辨识关系是等价关系。

定义 2 设 (X, \leq) 是一个偏序集,若对于任意 $(x, y) \in X^2$, 存在一个实数 $P(Y/X)$, 满足以下条件:

- (1) $0 \leq P(y/x) \leq 1, (x, y) \in X^2$;
- (2) $x \leq y \Rightarrow P(y/x) = 1, (x, y) \in X^2$;
- (3) $z \leq x \leq y \Rightarrow P(z/y) \leq P(z/x), (x, y, z) \in X^3$;
- (4) $x \leq y \Rightarrow \forall z \in X, P(x/z) \leq P(y/z), (x, y) \in X^2$ 。则称 P 为 X 上的包含度^[13,14]。

由于 $(P(U), \subseteq)$ 是偏序集,我们可以定义 $P(U)$ 上的包含度 $P: \forall X, Y \in P(U)$,

$$P(Y/X) = \begin{cases} \frac{|X \cap Y|}{|X|}, & X \neq \emptyset \\ 1, & X = \emptyset \end{cases}$$

若 $X \subseteq Y$, 则称 X 包含在 Y 中,或 X 对于 Y 是协调的。

定义 3 设 P 是 $X = \{x_1, x_2, \dots, x_n\}$ 上的概率测度,则

$$H(X) = - \sum_{i=1}^n P(x_i) \log_2 P(x_i)$$

称为 $[X, P]$ 上的不确定度或 Shannon 熵^[2]。

2.2 分类一致性

利用 Shannon 熵和包含度,我们可以定义两个数据集之间的一致性。

定义 4 设 $S=(U, C \cup D, V, f)$ 是决策表,记 $U/C = \{X_0, X_1, \dots, X_m\}$, $U/D = \{Y_0, Y_1, \dots, Y_s\}$ 。则每一个 X_i 关于 D 的分类一致性定义为:

$$Cons(X_i, D) = 1 + \frac{e}{s \times \log_2 e} \sum_{j=0}^s P(Y_j/X_i) \log_2 P(Y_j/X_i),$$

$i=0, \dots, m$

记 $G(Y_j, X_i, D) = -P(Y_j/X_i) \log_2 P(Y_j/X_i)$, 可以推导出以下结论:

- (1) $0 \leq G(Y_j, X_i, D) \leq \frac{1}{e} \log_2 e$;
- (2) 当 $P(Y_j/X_i) \rightarrow 0$ 时, $G(Y_j, X_i, D) \rightarrow 0$, 当 $P(Y_j/X_i) = 1$ 时, $G(Y_j, X_i, D) = 0$, 均对应 X_i 和 Y_j 之间分类一致性的最好情况;
- (3) 当 $P(Y_j/X_i) = \frac{1}{e}$ 时, $G(Y_j, X_i, D) = \frac{1}{e} \log_2 e$, 对应 X_i 和 Y_j 之间分类一致性的最差情况。

易知 $0 \leq Cons(X_i, D) \leq 1$ 。且当 $Cons(X_i, D) = 1$ 时, X_i 对于所有 Y_j 是协调的(最好情况); 当 $Cons(X_i, D) = 0$ 时, X_i 对于所有 Y_j 是不协调的(最差情况)。

在此基础上, C 对于 D 的基本分类的一致度定义为:

$$CCons(C, D) = \sum_{i=0}^m \frac{|X_i|}{s} \left(1 + \frac{e}{s \times \log_2 e} \times \sum_{j=0}^s P(Y_j/X_i) \cdot \log_2 P(Y_j/X_i) \right)$$

$CCons(C, D)$ 是所有条件划分 U/C 类对于决策 D 一致度的聚合。

2.3 模糊分类的一致性

事实上,基于前面样本所给出的划分,任何样本都可以明显地被分辨出是否为某一类成员,这可能会导致一些重要信息的流失。但许多现实问题中,是与不是、属于与不属于之间的区别不是突变的,而是有一个边缘地带、量变的过渡关系^[9]。一个对象是否属于某个集合,不能简单地用“是”或“否”来回答。因而有必要允许隶属度取 0 与 1 之间的其他实数值,从而用隶属函数来表示模糊概念。

根据上述思想,通过引入模糊划分的概念,我们改进前面定义的分类一致性的概念^[3]。

首先定义模糊集之间的一种包含度。

定义 5 设 $U = \{e_1, e_2, \dots, e_n\}$, 记 $F(U)$ 为 U 上的模糊集合全体, $(F(U), \subseteq)$ 为偏序集, 定义 $F(U)$ 上的包含度 FP : 对 $\forall X, Y \in F(U)$, $FP(Y/X) = \frac{\sum_{k=1}^n \min(X(e_k), Y(e_k))}{\sum_{k=1}^n X(e_k)}$ 。

定义 6 设 $\{X_p, p=0, 1, \dots, m\}$ 是属性 C 产生的模糊划分, $\{Y_q, q=0, 1, \dots, s\}$ 是属性 D 产生的模糊划分, 对于 $\forall e_k \in U$, 其模糊隶属度 $\mu_{X_p}(e_k)$ 和 $\mu_{Y_q}(e_k)$ 的定义如下:

$$\mu_{X_p}(e_k) = \max\{1 - h|p - c(e_k)|, 0\} \quad (1)$$

$$\mu_{Y_q}(e_k) = \max\{1 - h|q - d(e_k)|, 0\} \quad (2)$$

它们是以 p 和 q 为中心的三角形隶属函数。其中 h 是控制函数灵敏度的系数, $c(e_k)$ 和 $d(e_k)$ 是样本 e_k 在属性 C 和 D 中的属性值。

在模糊集中,根据样本 e_k 对于属性 D 的隶属函数来定义 $[e_k]_D$, 若 $\max\{\mu_{Y_0}(e_k), \dots, \mu_{Y_m}(e_k)\} = \mu_{Y_j}(e_k)$, 则令 $[e_k]_D = Y_j$ 。

值得注意的是:(1)根据最大隶属原则,当 e_k 的模糊隶属度 $\mu_{Y_j}(e_k)$ 在所有决策类的隶属函数中取到最大值时,认为 e_k 属于决策集 Y_j ; (2)对 $\forall e_k \in U$, 总存在一个决策类 Y_j , 其值 j 最接近 $d(e_k)$ (差值最大为 0.5)。

定义 7 X_i 对于属性 D 的分类一致性可定义为:

$$Cons^*(X_i, D) = 1 + \frac{e}{\left(\sum_{i=0}^m \sum_{k=0}^s \mu_{X_i}(e_k)\right) \times \log_2 e} \times \sum_{j=0}^s FP$$

$([e_k]_D/X_i) \log_2 FP([e_k]_D/X_i)$, 易知 $0 < Cons^*(X_i, D) < 1$, 且 $0 \leq -FP([e_k]_D/X_i) \log_2 FP([e_k]_D/X_i) \leq \frac{1}{e} \log_2 e$ 对任意的 X_i 和 $[e_k]_D$ 成立。

条件属性 C 对于决策属性 D 的分类一致性可定义为:

$$CCons^*(C, D) = \sum_{i=0}^m \frac{\sum_{k=0}^s \mu_{X_i}(e_k)}{\sum_{i=0}^m \sum_{k=0}^s \mu_{X_i}(e_k)} Cons^*(X_i, D)$$

3 说明性例子

表 1 是一个决策表,它包含 6 个样本, C 为条件属性, D 为决策属性。

我们以表 1 为例展示模糊分类一致性的计算步骤,求条件属性 C 对应数据集对决策属性 D 对应数据集的分类一致性。取相应参数为: $h=0.2, m=s=10$ 。

表1 一个决策表

样本	C(条件变量)	D(决策变量)
e ₁	2.786	2.643
e ₂	4.071	4.357
e ₃	1.875	1.429
e ₄	8.071	8.429
e ₅	7.286	6.214
e ₆	8.786	9.214

根据根定义6,则 $X_2 = \{e_3\}, X_3 = \{e_1\}, X_4 = \{e_2\}, X_7 = \{e_5\}, X_8 = \{e_4\}, X_9 = \{e_6\}, Y_3 = [e_1]_D = \{e_1\}, Y_4 = [e_2]_D = \{e_2\}, Y_6 = [e_5]_D = \{e_5\}, Y_8 = [e_4]_D = \{e_4\}, Y_9 = [e_6]_D = \{e_6\}$, 其它 X_i 和 Y_j 均为空集。

应用式(1)和式(2),计算所有6个样本对于条件类 $\{X_i\}$ ($i=1,2,\dots,m$)和决策类 $\{Y_j\}$ ($j=1,2,\dots,s$)的模糊隶属值,并分别列在表2、表3中。

现计算 X_i 对于所有类 $\{Y_j\}$ ($j=0,\dots,s$)的模糊包含度。以 Y_3 和 X_0 为例,有:

$$FP(Y_3/X_0) = \sum_{k=1}^6 \min(\mu_{X_0}(e_k), \mu_{Y_3}(e_k)) / \sum_{k=1}^6 \mu_{X_0}(e_k) = (\min(0.443, 0.929) + \min(0.186, 0.729) + \min(0.629, 0.686) + \min(0, 0) + \min(0, 0.357) + \min(0, 0)) / (0.443 + 0.186 + 0.629 + 0 + 0 + 0) = 1$$

其它模糊包含度同理可得,相应结果在表4中列出。

表2 条件类 X_i 的模糊隶属值

样本	$\mu_{X_0}(e_k)$	$\mu_{X_1}(e_k)$	$\mu_{X_2}(e_k)$	$\mu_{X_3}(e_k)$	$\mu_{X_4}(e_k)$	$\mu_{X_5}(e_k)$	$\mu_{X_6}(e_k)$	$\mu_{X_7}(e_k)$	$\mu_{X_8}(e_k)$	$\mu_{X_9}(e_k)$	$\mu_{X_{10}}(e_k)$
#1	0.471	0.671	0.871	0.929	0.729	0.529	0.329	0.129	0.000	0.000	0.000
#2	0.129	0.329	0.529	0.729	0.929	0.871	0.671	0.471	0.271	0.071	0.000
#3	0.714	0.914	0.886	0.686	0.486	0.286	0.086	0.000	0.000	0.000	0.000
#4	0.000	0.000	0.000	0.000	0.114	0.314	0.514	0.714	0.914	0.886	0.686
#5	0.000	0.000	0.157	0.357	0.557	0.757	0.957	0.843	0.643	0.443	0.243
#6	0.000	0.000	0.000	0.000	0.157	0.357	0.557	0.757	0.957	0.843	0.000

表3 决策类 Y_j 的模糊隶属值

样本	$\mu_{Y_0}(e_k)$	$\mu_{Y_1}(e_k)$	$\mu_{Y_2}(e_k)$	$\mu_{Y_3}(e_k)$	$\mu_{Y_4}(e_k)$	$\mu_{Y_5}(e_k)$	$\mu_{Y_6}(e_k)$	$\mu_{Y_7}(e_k)$	$\mu_{Y_8}(e_k)$	$\mu_{Y_9}(e_k)$	$\mu_{Y_{10}}(e_k)$
#1	0.471	0.671	0.871	0.929	0.729	0.529	0.329	0.129	0.000	0.000	0.000
#2	0.129	0.329	0.529	0.729	0.929	0.871	0.671	0.471	0.271	0.071	0.000
#3	0.714	0.914	0.886	0.686	0.486	0.286	0.086	0.000	0.000	0.000	0.000
#4	0.000	0.000	0.000	0.000	0.114	0.314	0.514	0.714	0.914	0.886	0.686
#5	0.000	0.000	0.157	0.357	0.557	0.757	0.957	0.843	0.643	0.443	0.243
#6	0.000	0.000	0.000	0.000	0.157	0.357	0.557	0.757	0.957	0.843	0.000

表4 所有6个样品的模糊包含度

样本	FP([e _k] _D /X ₀)	FP([e _k] _D /X ₁)	FP([e _k] _D /X ₂)	FP([e _k] _D /X ₃)	FP([e _k] _D /X ₄)	FP([e _k] _D /X ₅)	FP([e _k] _D /X ₆)	FP([e _k] _D /X ₇)	FP([e _k] _D /X ₈)	FP([e _k] _D /X ₉)	FP([e _k] _D /X ₁₀)
#1	1.000	0.923	0.881	0.935	0.832	0.691	0.515	0.361	0.197	0.152	0.195
#2	0.886	0.815	0.750	0.806	0.901	0.824	0.623	0.422	0.305	0.281	0.313
#3	1.000	0.969	0.798	0.667	0.545	0.431	0.294	0.165	0.074	0.006	0.000
#4	0.148	0.146	0.113	0.156	0.292	0.495	0.667	0.796	0.872	0.912	1.000
#5	0.477	0.431	0.417	0.462	0.574	0.775	0.907	0.811	0.670	0.632	0.727
#6	0.057	0.038	0.030	0.081	0.223	0.392	0.529	0.660	0.773	0.912	0.992

3.1 以 Shannon 熵定义分类一致性所得数据结果

基于前面的结果, X_i 对于属性 D 的分类一致性计算如下:

$$Cons^*(X_i, D)$$

$$= 1 + \frac{e}{(\sum_{i=0}^{10} \sum_{k=0}^6 \mu_{X_i}(e_k)) \times \log_2 e} \times \sum_{j=0}^6 FP([e_k]_D/X_i) \log_2 FP([e_k]_D/X_i)$$

$$= 1 + \frac{e}{27} \sum_{k=1}^6 FP([e_k]_D/X_i) \ln FP([e_k]_D/X_i)$$

$$= 1 + \frac{e}{27} (0 - 0.1072 - 0 - 0.2828 - 0.3531 - 0.1633)$$

$$= 0.9087$$

同理得: $Cons^*(X_1, D) = 0.8954, Cons^*(X_2, D) = 0.8768, Cons^*(X_3, D) = 0.8634, Cons^*(X_4, D) = 0.8399, Cons^*(X_5, D) = 0.8298, Cons^*(X_6, D) = 0.8297, Cons^*(X_7, D) = 0.8334, Cons^*(X_8, D) = 0.8528, Cons^*(X_9, D) = 0.8661, Cons^*(X_{10}, D) = 0.9072$ 。

最后,对于此决策表,条件属性 C 对于决策属性 D 的分类一致性可被计算为:

$$Cons^*(C, D)$$

$$= \sum_{i=0}^{10} \frac{\sum_{k=1}^6 \mu_{X_i}(e_k)}{\sum_{i=0}^{10} \sum_{k=1}^6 \mu_{X_i}(e_k)} Cons^*(X_i, D)$$

$$= \frac{1.257}{27} \times 0.804 + \frac{1.857}{27} \times 0.777 + \frac{1.774}{27} \times 0.739 + \frac{2.657}{27} \times 0.710 + \frac{2.886}{27} \times 0.660 + \frac{2.914}{27} \times 0.639 + \frac{2.914}{27} \times 0.665 + \frac{2.934}{27} \times 0.647 + \frac{2.900}{27} \times 0.688 + \frac{2.443}{27} \times 0.758 + \frac{1.829}{27} \times 0.803 = 0.8394$$

意味着属性 C 的分类结果和属性 D 的分类结果是相当协调的,可认为属性 C 很强地包含在 D 中。

3.2 文献[13]中的数据结果

X_i 对属性 D 的分类一致性计算如下:

$$Cons^*(X_i, D)$$

$$= 1 + \frac{e}{(\sum_{i=0}^{10} \sum_{k=0}^6 \mu_{X_i}(e_k)) \times \log_2 e} \times \sum_{j=0}^6 FP([e_k]_D/X_i) \log_2 FP([e_k]_D/X_i)$$

$$= 1 - \frac{4}{27} (0 + 0.101 + 0 + 0.126 + 0.249 + 0.054)$$

$$= 0.922$$

同理得: $Cons^*(X_1, D) = 0.902, Cons^*(X_2, D) = 0.878, Cons^*(X_3, D) = 0.868, Cons^*(X_4, D) = 0.837, Cons^*(X_5, D) = 0.812, Cons^*(X_6, D) = 0.815, Cons^*(X_7, D) = 0.831$,

(下转第80页)

sion Theory and Applications[M]. Nanjing: Nanjing University Press, 2012; 1-16(in Chinese)

贾修一, 商琳, 周献中, 等. 三支决策理论与应用[M]. 南京: 南京大学出版社, 2012; 1-16

[6] Zhong Jin-yi, Ye Dong-yi. Extended Decision-theoretic Rough Set Models Based on Fuzzy Minimum Cost[J]. Computer Science, 2014, 41(3): 50-54, 75(in Chinese)

衷锦仪, 叶东毅. 基于模糊数风险最小化的拓展决策粗糙集模型[J]. 计算机科学, 2014, 41(3): 50-54, 75

[7] Liu Dun, Li Tian-rui, Li Hua-xiong. Rough set theory: A three-way decisions perspective[J]. Journal of Nanjing University (Natural Sciences), 2013, 49(5): 574-581(in Chinese)

刘盾, 李天瑞, 李华雄. 粗糙集理论: 基于三支决策视角[J]. 南京大学学报(自然科学版), 2013, 49(5): 574-581

[8] Zhi Hui-lai. Knowledge Representation on Incomplete Formal Context[J]. Computer Science, 2015, 42(1): 276-278(in Chinese)

智慧来. 不完备形式背景上的知识表示[J]. 计算机科学, 2015, 42(1): 276-278

[9] Xie Cheng, Shang Lin. Detection of abnormal behavior in video using three-way decision rough sets[J]. Journal of Nanjing Uni-

versity(Natural Sciences), 2013, 49(4): 475-482(in Chinese)

谢骋, 商琳. 基于三支决策粗糙集的視頻异常行为检测[J]. 南京大学学报(自然科学版), 2013, 49(4): 475-482

[10] Li Jian-lin, Huang Shun-liang. Multistage Three-Way Decisions of Span SMS Filtering Model[J]. Journal of Frontiers of Computer Science and Technology, 2014(2): 226-233(in Chinese)

李建林, 黄顺亮. 多阶段三支决策垃圾短信过滤模型[J]. 计算机科学与探索, 2014(2): 226-233

[11] Zhang Li-bo, Li Hua-xiong, Zhou Xian-zhong, et al. Multi-granularity cost-sensitive three-way decision for face recognition[J]. Journal of Shandong University(Natural Science), 2014, 49(8): 48-57(in Chinese)

张里博, 李华雄, 周献中, 等. 人脸识别中的多粒度代价敏感三支决策[J]. 山东大学学报(理学版), 2014, 49(8): 48-57

[12] Du Li-na, Xu Jiu-cheng, Liu Yang-yang, et al. Research on the evaluation of venture investment based on the risk minimization of three-way decision[J]. Journal of Shangdong University(Natural Science), 2014(8): 66-72(in Chinese)

杜丽娜, 徐久成, 刘洋洋, 等. 基于三支决策风险最小化的风险投资评估应用研究[J]. 山东大学学报(理学版), 2014, 49(8): 66-72

(上接第 63 页)

$Cons^*(X_8, D) = 0.860$, $Cons^*(X_9, D) = 0.892$, $Cons^*(X_{10}, D) = 0.914$.

条件属性 C 对于决策属性 D 的分类一致性计算如下:

$Cons^*(C, D)$

$$\begin{aligned}
 &= \frac{\sum_{i=0}^{10} \sum_{k=1}^6 \mu_{X_i}(e_k)}{\sum_{i=0}^{10} \sum_{k=1}^6 \mu_{X_i}(e_k)} Cons^*(X_i, D) \\
 &= \frac{1.257}{27} \times 0.902 + \frac{1.857}{27} \times 0.878 + \frac{1.774}{27} \times 0.868 + \\
 &\quad \frac{2.657}{27} \times 0.837 + \frac{2.886}{27} \times 0.812 + \frac{2.914}{27} \times 0.815 + \\
 &\quad \frac{2.914}{27} \times 0.831 + \frac{2.934}{27} \times 0.860 + \frac{2.900}{27} \times 0.892 + \\
 &\quad \frac{2.443}{27} \times 0.902 + \frac{1.829}{27} \times 0.914 = 0.859
 \end{aligned}$$

从以上计算可以看出, 利用这两种方法计算得到的数据集的分类一致性结果相近。但 Shannon 熵更为常用, 物理意义更为明确。

结束语 文献[13]提出了一种计算一致度的方法, 用以测量两个数据集的一致性, 从而判断一个物理测量是否能被另一个特定数据集所替代。粗糙集中用包含度来测量两个数据集的分类一致性, 由于经典分类方法在考虑不同数据集的相邻关系时常常不够准确, 文献[13]利用包含度的思想, 通过引入模糊分类, 对数据集的一致性定义进行分类, 从而得到更为精准的分类结果。本文利用更为常用的 Shannon 熵, 得到了一种新的分类一致性的定义并利用例子计算得到了相关数据结果。

参考文献

[1] Agresti A, Finlay B. Statistical Methods for the Social Sciences (third edition)[M]. Prentice Hall, New Jersey, 1997

[2] Chen C B, Wang L Y. Rough set-based clusing with refinement using Shannon's entropy theory[J]. Computer, Mathematics

with Applications, 2006, 52(10/11): 1563-1576

[3] Dubois D, Prade H, Yager R R. Fuzzy Information Engineering: A Guide Tour of Application[M]. Wiley, New York, 1997

[4] Hollins M, Faldowski R, Rao S, et al. Perceptual dimensions of tactile surface texture: a multidimensional scaling analysis[J]. Perception and Psychophys, 1997, 54(6): 697-705

[5] Jolliffe I T. Principal Component Analysis(2rd edition)[M]. Information Publisher Science, New York, 2002

[6] Le Dien S. Hierarchical multiple factoranalysis: application to the comparison of sensory profiles[J]. Food Quality Preference, 2003, 14(5/6): 397-403

[7] Pawlak Z. Rough sets[J]. International Journal Computer and Information Sciences, 1982, 11(5): 341-356

[8] Pawlak Z. Rough set theory and its applications in data analysis [J]. Cybernet System, 1998, 29(7): 661-688

[9] Polkowski L, Skowron A. Rough mereology and analytical morphology: new developments in rough set Theory[C]//De Glass M, Pawlak Z. eds. Proceedings of WOCFAI-95, Second World Conference on Fundamntals of Artificial Intelligence. Angkor, Paris, 1995: 343-354

[10] Qian Y H, Liang J Y, Dang C Y. Consistency measure, inclusion degree and fuzzy measure in decision tables[J]. Fuzzy Sets and Systems, 2008, 159(18): 2353-2377

[11] Tripathy B C, Ray G C. On mixed fuzzy topological spaces and countability[J]. Soft Computing, 2012, 16(10): 1691-1695

[12] Weisberg S. Applied Linear Regression(third edition)[M]. John & Sons, New York, 2005

[13] Xue Z, Zeng X, Koehl L, et al. Measuring consistency of two datasets using fuzzy techniques and the concept of indiscernibility[J]. Engineering Applications of Artificial Intelligence, 2014, 36: 54-63

[14] Yao Y Q, Mi J S. Hybrid Monotonic Measure on Intuitionistic Fuzzy Sets[J]. Computer Science, 2010, 37(1): 255-257(in Chinese)

姚燕青, 米据生. 直觉模糊集上的混合单调包含度[J]. 计算机科学, 2010, 37(1): 255-257