

基于区间模糊匹配函数的数据清洗算法研究 及其在问卷调查中的应用

米允龙¹, 李金海², 米春桥^{1,3}, 刘文奇², 刘佳¹, 王添³

(1.怀化学院计算机科学与工程学院, 湖南 怀化 418000)

(2.昆明理工大学理学院, 云南 昆明 650500)

(3.武陵山片区生态农业智能控制技术湖南省重点实验室, 湖南 怀化 418000)

[摘要] 数据清洗是保证数据质量的重要步骤. 由于人类的活动通常带有一定的主观性与情绪性, 因此现实中部分数据往往存在不合理性甚至错误. 而此类不合理数据常具有不确定性、模糊性与隐藏性, 这给数据清洗带来了困难. 传统的数据清洗方法对此类数据难以充分发挥作用. 结合区间值模糊集理论与匹配函数提出一种区间模糊匹配函数方法, 构建区间模糊匹配算法来清洗数据、提高数据质量, 并将其应用在问卷调查数据中. 实验结果表明本算法具有较高的准确度及运行效率, 适应处理数据中的不合理数据.

[关键词] 数据清洗, 匹配函数, 区间模糊集, 区间模糊匹配函数, 问卷调查数据

[中图分类号] TP311 **[文献标志码]** A **[文章编号]** 1672-1292(2017)03-0070-10

Research into Data Cleaning Algorithm Based on Interval Fuzzy Matching Functions and Its Application to Questionnaire Data

Mi Yunlong¹, Li Jinhai², Mi Chunqiao^{1,3}, Liu Wenqi², Liu Jia¹, Wang Tian³

(1.School of Computer Science and Engineering, Huaihua University, Huaihua 418000, China)

(2.Faculty of Science, Kunming University of Science and Technology, Kunming 650500, China)

(3.Hunan Provincial Key Laboratory of Ecological Agriculture Intelligent Control Technology, Huaihua 418000, China)

Abstract: Data cleaning is a very important step to ensure data quality. The real-world data often has some unreasonable data even error because of human activities usually with subjectivity and emotionality, such as the questionnaire data. However, there are some difficulties to process data cleaning due to these unreasonable data often being uncertainty, ambiguity and hiding. For this type of data, the traditional data cleaning methods have difficulty in handling the unreasonable data. Therefore, by combining the basic theories of interval-valued fuzzy set and matching function, we propose an interval fuzzy matching function method. Based on this method we construct a new algorithm to clean data and improve data quality, and then apply it to questionnaire data. Experiments show that our algorithm have a good precision and running efficiency, and that it is adaptable to process the unreasonable data.

Key words: data cleaning, matching function, interval-valued fuzzy set, interval-valued fuzzy matching function, questionnaire data

数据清洗是数据质量的重要保证, 它通过检测并消除数据中的错误、不一致等问题来提高数据质量^[1-2]. 由于各种缺失、错误、重复及不合理数据存在, 数据质量在决策中尤为重要. 实际上, 在质量欠佳的数据上是很难得做出有意义的决策. 为了提高数据质量, 已有很多学者对数据清洗进行了研究. 例如, 文献[3-7]主要研究重复数据信息的问题, 文献[8]主要关注的是缺失数据问题, 而文献[9]强调直接从不确定性数据中进行挖掘等.

收稿日期: 2017-05-23.

基金项目: 湖南省教育科学规划课题(XJK016QXX003)、湖南省自然科学基金项目(2017JJ3252)、国家自然科学基金项目(41301084)、怀化学院一般项目(HHUY2016-05).

通讯联系人: 米春桥, 博士, 副教授, 研究方向: 数据挖掘与分析、地理信息系统、农业与教育信息化. E-mail: michunqiao@163.com

问卷调查是进行调查研究的一种重要方法. 目前,其应用领域主要有市场客户关系管理^[10-11]、学生对教学水平的评价^[12]及旅游景点与服务等方面. 例如,文献[13]中将问卷调查的问题分为开放式的与封闭式两种,其问卷调查的数据则由开放式与封闭式的数据组成. 在文献[12]中,其将用户回答数据划分为5种类型:类别数据、列表数据、数值数据、等级数据、多语义等级数据. 其中,对于后面两类数据,存在很大的主观性与不确定性. 例如,“此景点是很好、好、一般还是很差?”这类问题. 为了处理此类问题,文献[12, 14]强调挖掘问卷调查的关联规则,文献[15]则直接挖掘调查数据的开放式回答.

毋庸置疑,数据质量是数据挖掘的基础. 在问卷调查中,我们发现存在一些不合理数据影响到数据质量. 虽然已有学者提及到问卷调查的数据质量的重要性^[16-17],但尚未提出一种处理问卷调查中不合理数据的方法与算法. 如表1表示,来自2014年“昆明市建设世界知名旅游城市对策研究”中的部分问卷调查数据.

表1 景区满意度问卷调查

Table 1 Scenic spot satisfaction questionnaire survey

交通状况	景区景观	餐饮条件	工作人员服务态度	旅行社	安全性问题	环境卫生	基础设施	总体满意度
C	C	B	C	C	C	C	C	C
C	C	D	C	D	C	C	C	B
B	C	D	C	C	B	B	B	C
B	A	B	A	B、C	A	A	A	C
B	A	B	A	C	A	A	B	A

注:A表示很好;B表示较好;C表示一般;D表示差;E表示很差

从表1可知,由于被调查者的主观性与自身对问题理解程度不同,数据难免会存在的随意性与矛盾性. 例如,虽然已对“旅行社”选项标注是单选项,但仍有一部分游客给出的答案是多选项. 然而,不能随意将其清除. 如表1中的第2行与第4行所示:

第2行:条件(C,C,D,C,D,C,C,C) \Rightarrow 整体满意度(B);

第4行:条件(B,A,B,A,B,C,A,A) \Rightarrow 整体满意度(C).

然而,由于这种不合理数据带有隐藏性,传统的数据清洗方法难以充分发挥作用. 为了更好地清洗这些不合理数据,通过分析问卷调查数据中的不合理数据特征,本文提出一种区间模糊匹配算法来进行处理.

区间值模糊集理论是一种能够处理不确定与不准确信息的有力数学工具^[18-19]. 目前,已经被广泛应用于近似推理、信号传输及模糊控制等领域^[20-25]. 值得注意的是,区间值模糊集理论也是一种能够处理等级与语言信息的数学工具. 匹配函数与函数依赖是一种在重复数据检测与对象识别有着重要作用的函数模式,已经有很多学者将其应用于数据清洗与数据挖掘领域(见文献[5-7, 26-27]). 本文通过运用区间值模糊集理论与匹配函数的思想,提出一种区间模糊函数算法,并将其直接应用到旅游问卷调查中.

1 相关理论

1.1 区间值模糊理论

定义1^[18-19] 设 X 为非空集合,即论域. A 为 X 到 $I_{[0,1]}$ 的映射,即:

$$A: X \rightarrow I_{[0,1]}.$$

式中, $I_{[0,1]} = \{a = [\underline{a}, \bar{a}] \mid 0 \leq \underline{a} \leq \bar{a} \leq 1, \underline{a}, \bar{a} \in \mathbf{R}\}$.

此时,称 A 为区间值模糊集. 且将基于论域 X 的区间值模糊集称为 $F_1(X)$.

推论1^[24] 设 $A \in F_1(X)$, $A(x) \triangleq [\underline{A}(x), \bar{A}(x)]$, 其中 $\underline{A}(x)$ 与 $\bar{A}(x)$ 分别为 A 的下确界与上确界. 则有:

$$(A \cup B)(x) = [\underline{A}(x) \vee \underline{B}(x), \bar{A}(x) \vee \bar{B}(x)], \forall x \in X,$$

$$(A \cap B)(x) = [\underline{A}(x) \wedge \underline{B}(x), \bar{A}(x) \wedge \bar{B}(x)], \forall x \in X.$$

式中, \wedge 和 \vee 均为Zadeh算子^[28].

定义2 设 T 为任意指标集且有 $A_t \in F(X) (t \in T)$, 则 $\cup_{t \in T} A_t$ 与 $\cap_{t \in T} A_t$ 可以表述为:

$$(\cup_{t \in T} A_t)(x) = \sup_{t \in T} A_t(x) = \vee_{t \in T} A_t(x),$$

$$(\cap_{t \in T} A_t)(x) = \inf_{t \in T} A_t(x) = \wedge_{t \in T} A_t(x).$$

其中, $F(X)$ 为基于论域 X 的所有模糊集合, 则 $\vee_{t \in T} A_t(x)$ 与 $\wedge_{t \in T} A_t(x)$ 分别是其上确界与下确界.

推论 2 设 A_i 为区间值模糊集, T 为任意指标集. 由可扩展性原理有:

$$\begin{aligned} (\cup_{i \in T} A_i)(x) &= [\sup_{i \in T} \underline{A}_i(x), \sup_{i \in T} \bar{A}_i(x)] = [\vee_{i \in T} \underline{A}_i(x), \vee_{i \in T} \bar{A}_i(x)], \\ (\cap_{i \in T} A_i)(x) &= [\inf_{i \in T} \underline{A}_i(x), \inf_{i \in T} \bar{A}_i(x)] = [\wedge_{i \in T} \underline{A}_i(x), \wedge_{i \in T} \bar{A}_i(x)]. \end{aligned}$$

由推论 1 与定义 2 可直接证明.

1.2 匹配函数

设 A 为属性集的关系模式, 记为 $R_i (i=1, 2, \dots, n)$. D_{A_j} 为相对应属性 $A_j (j=1, 2, \dots, m)$ 的域值, R 为数据库 D 属性的模式集, 记为 $\{R_1, R_2, \dots, R_n\}$. 设 t^D 为任意属性 A_j 到关系模式 R_i 的值映射, 并记 $t^D[A_j]$ 为相对应的属性值. 若 X 为属性集合, 则用 $t^D[X]$ 来表示相对应的属性值集合.

定义 3 匹配依赖度^[5-6] 设 R_1 与 R_2 分别为两不同的关系模式, X_1, X_2 与 Y_1, Y_2 分别来自于 R_1 与 R_2 的属性集. 则对应于 R_1 与 R_2 的匹配依赖度 φ 表示为:

$$\varphi: R_1[X_1] \approx R_2[X_2] \rightarrow R_1[Y_1] \rightleftharpoons R_2[Y_2].$$

式中, \approx 表示相似算子, \rightleftharpoons 表示匹配算子.

定义 4 匹配函数^[5-7] 设任意具有可比性的属性集对 A_1, A_2 都有与之相对应的域值集 D_A . 则其匹配函数如下:

$$m_A: D_A \times D_A \rightarrow D_A.$$

通过匹配函数发现, 若需使 a 与 b 相等 ($a, b \in D_A$), 则用 $m_A(a, b)$ 进行相应的替代. 同时, 匹配函数存在以下性质^[7]:

$$\begin{aligned} \text{等幂律: } m_A(a, a) &= a, \\ \text{交换律: } m_A(a, b) &= m_A(b, a), \\ \text{结合律: } m_A(m_A(a, b), c) &= m_A(a, m_A(b, c)). \end{aligned}$$

式中, $a, b, c \in D_A$, 若符合以上 3 条性质, 则 (D_A, m_A) 形成一个偏序关系.

定义 5^[7] 设 (D_A, m_A) 是偏序关系, 对任意 $a, b \in D_A$, 若 D_A 都存在上确界 (lub) 与下确界 (glb), 则 (D_A, m_A) 构一个格. 则有:

- (1) 对偏序 \leq_A 可定义为: 对 $\forall a, b \in D_A$, 若 $m_A(a, b) = b$, 则有 $a \leq_A b$;
- (2) 若 $L_A = (D_A, m_A)$ 是格, 可定义匹配函数 m_A 为: $m_A(a, b) := \text{lub}_{\leq_A} \{a, b\}$, 或 $m_A(a, b) := \text{glb}_{\geq_A} \{a, b\}$;
- (3) 若 $L_A = (D_A, m_A)$ 是格, 且 L_A 存在最小值 (N) 与最大值 (M), 则对 $\forall a \in D_A$ 有: $m_A(a, N) = a, m_A(a, M) = M, N \leq_A a \leq_A M$.

2 区间模糊匹配函数

下文将提出一种区间模糊匹配函数来处理问卷调查中的不合理数据.

2.1 问卷调查数据预处理

在数据库 D 中, 设 $I_i (i=1, 2, \dots, n)$ 为任意属性项, $X = (I_1, I_2, \dots, I_n)$ 为属性集合, $T(t_1, t_2, \dots, t_m)$ 为对象 (或称元组) 集合, 则 $t_j^D[I_i]$ 为对应对象 $t_j (j=1, 2, \dots, m)$ 的值. 根据表 2, 对 $I_i \in X, t_j \in T$ (其中, $i = \{1, 2, \dots, n\}, j = \{1, 2, \dots, m\}$), 则有:

$$\begin{aligned} t_j^D[I_i] \{A\} &:= [0.9, 1], & t_j^D[I_i] \{B\} &:= [0.8, 0.9], \\ t_j^D[I_i] \{C\} &:= [0.6, 0.8], & t_j^D[I_i] \{D\} &:= [0.4, 0.6], \\ t_j^D[I_i] \{E\} &:= [0, 0.4). \end{aligned}$$

将表 1 中的数据重新描述为表 3.

将表 3 中的属性从左至右分别记 I_1, I_2, \dots, I_9 . 则由定义 3~5, 可得简单的匹配函数如下:

$$m: (I_1, I_2, \dots, I_8) \Rightarrow I_9.$$

若 $m(I_9, I_9') := I_9'$ (或 $I_9 \leq I_9'$), 则 I_9 与 I_9' 匹配. 其中, I_9' 是根据匹配函数推导出的, I_9 为真实数据值. 因此, 为了获取合理的匹配规则, 我们对提出了区间模糊匹配函数.

表 2 等级语义、等级数据的区间值模糊描述

Table 2 The interval-valued descriptions of questionnaire data with ranking and linguistic ranking

等级	语义等级	区间值
A	很好	[0.9, 1)
B	好	[0.8, 0.9)
C	一般	[0.6, 0.8)
D	差	[0.4, 0.6)
E	很差	[0, 0.4)

表 3 景区满意度问卷调查(区间值描述)

Table 3 Scenic spot satisfaction questionnaire survey(interval-value descriptions)

交通状况	景区景观	餐饮条件	工作人员服务态度	旅行社	安全性问题	环境卫生	基础设施	总体满意度
[0.6,0.8)	[0.6,0.8)	[0.8,0.9)	[0.6,0.8)	[0.6,0.8)	[0.6,0.8)	[0.6,0.8)	[0.6,0.8)	[0.6,0.8)
[0.6,0.8)	[0.6,0.8)	[0.4,0.6)	[0.6,0.8)	[0.4,0.6)	[0.6,0.8)	[0.6,0.8)	[0.6,0.8)	[0.8,0.9)
[0.8,0.9)	[0.6,0.8)	[0.4,0.6)	[0.6,0.8)	[0.6,0.8)	[0.8,0.9)	[0.8,0.9)	[0.8,0.9)	[0.6,0.8)
[0.8,0.9)	[0.9,1]	[0.8,0.9)	[0.9,1]	[0.6,0.9)	[0.9,1]	[0.9,1]	[0.9,1]	[0.6,0.8)
[0.8,0.9)	[0.9,1]	[0.8,0.9)	[0.9,1]	[0.6,0.8)	[0.9,1]	[0.9,1]	[0.8,0.9)	[0.9,1]

2.2 区间模糊匹配函数相关理论

为了与区间值模理论表述一致,以下相关理论仍以闭区间来描述.

定义 6 在数据库 D 中,对任意属性项 $I_i(i=1,2,\dots,n)$,对象 $t_j(j=1,2,\dots,m)$ 对应的值为 $t_j^D[I_i]$. 从区间模糊集角度可描述为:

$$t_j^D[I_i]\{V_k\}=[\underline{V}_k(I_i),\bar{V}_k(I_i)].$$

对于任意属性项 I_i ,则 $\underline{V}_k(I_i)$ 与 $\bar{V}_k(I_i)$ 分别为 V_k 的最小值与最大值. 其中, $V_k=\{A,B,C,D,E\}$, $k=1,2,3,4,5$. 如,对 $k=1$ 时,则有: $t_j^D[I_i]\{V_1\}=[\underline{V}_1(I_i),\bar{V}_1(I_i)]=[\underline{Z},\bar{A}]$.

定理 1 对任意属性项 I_x 与 $I_y(x,y=1,2,\dots,n$ 且 $x\neq y)$,对象 $t_j(j=1,2,\dots,m)$ 对应的值分别为 $t_j^D[I_x]$ 与 $t_j^D[I_y]$. 则有:

$$\begin{aligned} t_j^D[I_x]\{V_{k1}\}\cup t_j^D[I_y]\{V_{k2}\}&=[\underline{V}_{k1}(I_x)\cup\underline{V}_{k2}(I_y),\bar{V}_{k1}(I_x)\cup\bar{V}_{k2}(I_y)]=[\underline{V}_{k1}(I_x)\vee\underline{V}_{k2}(I_y),\bar{V}_{k1}(I_x)\vee\bar{V}_{k2}(I_y)], \\ t_j^D[I_x]\{V_{k1}\}\cap t_j^D[I_y]\{V_{k2}\}&=[\underline{V}_{k1}(I_x)\cap\underline{V}_{k2}(I_y),\bar{V}_{k1}(I_x)\cap\bar{V}_{k2}(I_y)]=[\underline{V}_{k1}(I_x)\wedge\underline{V}_{k2}(I_y),\bar{V}_{k1}(I_x)\wedge\bar{V}_{k2}(I_y)]. \end{aligned}$$

且当 $i=1,2,\dots,n$ 时,有:

$$\begin{aligned} U_{i=1}^n t_j^D[I_i]\{V_k\}&=[U_{i=1}^n \underline{V}_k(I_i),U_{i=1}^n \bar{V}_k(I_i)]=[\vee_{i=1}^n \underline{V}_k(I_i),\vee_{i=1}^n \bar{V}_k(I_i)], \\ \cap_{i=1}^n t_j^D[I_i]\{V_k\}&=[\cap_{i=1}^n \underline{V}_k(I_i),\cap_{i=1}^n \bar{V}_k(I_i)]=[\wedge_{i=1}^n \underline{V}_k(I_i),\wedge_{i=1}^n \bar{V}_k(I_i)]. \end{aligned}$$

式中, $k,k1,k2=1,2,3,4,5$.

证明 由推论 1 与定义 6 易证.

推论 3 设 T 为任意指标集且 $A_t\in F_1(X)(t\in T)$,则有:

$$\begin{aligned} \inf_{t\in T}\{\inf_{t\in T}(t_j^D[A_t]\{V_k\})\}&=\wedge_{t\in T}[\wedge_{t\in T}\underline{V}_k(A_t),\wedge_{t\in T}\bar{V}_k(A_t)]=[\wedge_{t\in T}\{\wedge_{t\in T}\underline{V}_k(A_t)\},\wedge_{t\in T}\{\wedge_{t\in T}\bar{V}_k(A_t)\}], \\ \sup_{t\in T}\{\sup_{t\in T}(t_j^D[A_t]\{V_k\})\}&=\vee_{t\in T}[\vee_{t\in T}\underline{V}_k(A_t),\vee_{t\in T}\bar{V}_k(A_t)]=[\vee_{t\in T}\{\vee_{t\in T}\underline{V}_k(A_t)\},\vee_{t\in T}\{\vee_{t\in T}\bar{V}_k(A_t)\}]. \end{aligned}$$

为了方便,简记:

$$NN=\inf_{t\in T}\{\inf_{t\in T}(t_j^D[A_t]\{V_k\})\},$$

$$MM=\sup_{t\in T}\{\sup_{t\in T}(t_j^D[A_t]\{V_k\})\}.$$

由定义 5,若 $(D_\alpha,m_\alpha)(N\leq\alpha\leq M)$ 是格,则对任意 $NN,MM\in D_\alpha$,则可得一种匹配函数如下:

$$m_M(NN,MM):=[\min\{NN\},\max\{MM\}],$$

$$m_N(NN,MM):=[\max\{NN\}\wedge\min\{MM\},\max\{NN\}\vee\min\{MM\}].$$

例 1 如表 3 中的第 4 条元组,设 $A_1=\{I_1,I_2,I_3\}$, $A_2=\{I_4,I_5,I_6,I_7,I_8\}$,则由推论 3 可得:

$$NN=\inf_{t\in T}\{\inf_{t\in T}(t_4^D[A_t]\{V_k\})\}=[0.6,0.9],$$

$$MM=\sup_{t\in T}\{\sup_{t\in T}(t_4^D[A_t]\{V_k\})\}=[0.9,1],$$

$$m_M(NN,MM):=[0.6,1],m_N(NN,MM):=[0.9,0.9].$$

正如文献[29]所指,数据清洗应尽量合理,最大可能保持数据原有信息. 清洗力度过大,将会清洗掉有用的信息. 而力度过小,将达不到清洗的效果. 例 1 中 $m_M(NN,MM):=[0.6,1]$ 表现出力度过小, $m_N(NN,MM):=[0.9,0.9]$ 呈显出过大. 然而,根据统计学原理,样本的均值能够反映出数据的集中趋势水平. 因此,我们的匹配函数以均值为中心进行上下合理的幅度匹配.

定义 7 区间模糊匹配函数 设 $m_M(NN,MM)$ 与 $m_N(NN,MM)$ 分别是任意对象的上确界集合与下确

界集合, $AVG(w_i * \prod_{i=1}^n Dom I_i)$ (其中, $w_i=1, (i=1,2,\dots,n)$ 且 $\vee_{i=1}^n w_i=1$) 为平均权重,则有:

$$m_{I_d}(NN, AVG, MM) := [AVG(w_i * \prod_{i=1}^n \underline{V}_k(I_i)) - dis_1, AVG(w_i * \prod_{i=1}^n \bar{V}_k(I_i)) + dis_2].$$

式中,

$$\begin{aligned} AVG(w_i * \prod_{i=1}^n DomI_i) &= AVG(w_i * \prod_{i=1}^n t_j^D[I_i] \{V_k\}) = [AVG(w_i * \prod_{i=1}^n \underline{V}_k(I_i)), AVG(w_i * \prod_{i=1}^n \bar{V}_k(I_i))], \\ dis_1 &= \min \{ |AVG(w_i * \prod_{i=1}^n \underline{V}_k(I_i)) - \inf m_M(NN, MM)|, |AVG(w_i * \prod_{i=1}^n \underline{V}_k(I_i)) - \inf m_N(NN, MM)| \}, \\ dis_2 &= \min \{ |AVG(w_i * \prod_{i=1}^n \bar{V}_k(I_i)) - \sup m_M(NN, MM)|, |AVG(w_i * \prod_{i=1}^n \bar{V}_k(I_i)) - \sup m_N(NN, MM)| \}. \end{aligned}$$

例 2(续例 1) 由定义 7 可得:

$$m_{I_d}(NN, AVG, MM) := [AVG(w_i * \prod_{i=1}^n \underline{V}_k(I_i)) - dis_1, AVG(w_i * \prod_{i=1}^n \bar{V}_k(I_i)) + dis_2] = [0.8, 1].$$

式中,

$$\begin{aligned} AVG(w_i * \prod_{i=1}^n t_j^D[I_i] \{V_k\}) &= [AVG(w_i * \prod_{i=1}^n \underline{V}_k(I_i)), AVG(w_i * \prod_{i=1}^n \bar{V}_k(I_i))] = [0.8375, 0.9625], \\ dis_1 &= \min \{ |0.8375 - 0.6|, |0.8375 - 0.8| \} = \min \{ 0.2375, 0.0375 \} = 0.0375, \\ dis_2 &= \min \{ |0.9625 - 0.9|, |0.9625 - 1| \} = \min \{ 0.0625, 0.0375 \} = 0.0375. \end{aligned}$$

由例 2 可知,元组 4 的属性项 I_9 合理值应是 $[0.8, 1]$, 即 A 或 B .

由表 1 可知,元组 4 的总体满意度为 C . 这与所有前提条件都为 B 以上的评价相矛盾. 主要是由于评价者的随意性,或者还有其它更为重要的因素未考虑到. 本文只考虑前者的情况下,本文算法能匹配出更为合理的结果 $[A \text{ 或 } B]$. 即文献[26-27]采用函数依赖关系进行数据修复与检测的出发点.

3 基于区间模糊匹配函数的数据清洗算法

算法 1 区间值模糊集的上确界与下确界算法.

Input:

D : 已经进行预处理的数据库.

$A_i: A_i = \{I_1, I_2, \dots, I_n\}$.

Output:

$ISIS$: 区间模糊集的上确界与下确界. (即, $m_M(NN, MM)$ 与 $m_N(NN, MM)$).

```

1. for  $j=0$  to (row.length-1) do
2.   for  $i=0$  to (column.length-1) do
3.      $lowerNN = \inf_{t \in T} \{ \inf_{t \in T} (t_j^D[A_i] \{ \underline{V}_k \}) \}$ ;
4.      $upperNN = \inf_{t \in T} \{ \inf_{t \in T} (t_j^D[A_i] \{ \bar{V}_k \}) \}$ ;
5.      $lowerMM = \sup_{t \in T} \{ \sup_{t \in T} (t_j^D[A_i] \{ \underline{V}_k \}) \}$ ;
6.      $upperMM = \sup_{t \in T} \{ \sup_{t \in T} (t_j^D[A_i] \{ \bar{V}_k \}) \}$ ;
7.   if  $k = (\text{column.length}-1)$  then
8.      $m_M(NN, MM) = \text{matchFunction\_}m_M(lowerNN, upperNN, lowerMM, upperMM)$ ;
9.      $m_N(NN, MM) = \text{matchFunction\_}m_N(lowerNN, upperNN, lowerMM, upperMM)$ ;
10.    output( $m_M(NN, MM), m_N(NN, MM)$ );
11.  end if
12. end for
13. end for
14. % the method of matchFunction_ $m_M()$ 
15. matchFunction_ $m_M(lowerNN, upperNN, lowerMM, upperMM)$ ;
16. return  $[\min \{ lowerNN, upperNN \}, \max \{ lowerMM, upperMM \}]$ ;
17. % the method of matchFunction_ $m_N()$ 
18. matchFunction_ $m_N(lowerNN, upperNN, lowerMM, upperMM)$ 
19. return  $[\max \{ lowerNN \} \wedge \min \{ lowerMM \}, \max \{ upperNN \} \vee \min \{ upperMM \}]$ ;

```

算法 2 区间值模糊集的样本均值算法.**Input:**

D:已经进行预处理的数据库.

 $A_i: A_i = \{I_1, I_2, \dots, I_n\}$. $w_i (i = 1, 2, \dots, n)$:权重.**Output:**

AIS:区间值模糊集的样本均值集合.

```

1. lowerSum←0, upperSum←0;
2. for j=0 to (row.length-1) do
3.   for i=0 to (column.length-1) do
4.     lowerSum =  $w_i * \prod_{k=1}^n \underline{V}_k(I_i)$ ;
5.     upperSum =  $w_i * \prod_{k=1}^n \bar{V}_k(I_i)$ ;
6.     if k = (column.length-1) then
7.       lowerAverage = lowerSum / column.length;
8.       upperAverage = upperSum / column.length;
9.       output( lowerAverage, upperAverage );
10.    end if
11.  end for
12. end for

```

算法 2 中的权重需要根据具体属性项在数据集中的比重来进行取值. 在此,我们取每个属性项的权重相等. 同时,以算法 1 与算法 2 的输出,作为算法 3 的输入,则有区间模糊匹配函数的数据清洗算法如算法 3 所示.

算法 3 区间模糊匹配函数的数据清洗算法.**Input:**

ISIS:区间模糊集的上确界与下确界.

AIS:区间值模糊集的数据均值集合.

Output:

RRS:合理值集合.

```

1. for j=0 to (row.length-1) do
2.   for i=0 to (ISIScolumn.length-1) do
3.     lowerBottom =  $M_N(NN, MM).Bottom$ ;
4.     lowerTop =  $m_N(NN, MM).Top$ ;
5.     upperBottom =  $m_M(NN, MM).Bottom$ ;
6.     upperTop =  $m_M(NN, MM).Top$ ;
7.   end for
8.   for k=0 to (AIScolumn.length-1) do
9.      $dis_1 = \min\{|lowerAverage - lowerBottom|, |lowerAverage - upperBottom|\}$ ;
10.     $dis_2 = \min\{|upperAverage - lowerTop|, |upperAverage - upperTop|\}$ ;
11.    RL = reasonableLower( lowerAverage,  $dis_1$  );
12.    RU = reasonableUpper( upperAverage,  $dis_2$  );
13.    output( RL, RU );
14.  end for
15. end for
16. % the method of reasonableLower()
17. reasonableLower( lowerAverage,  $dis_1$  );
18. return lowerAverage -  $dis_1$ ;

```

```

19. % the method of reasonableUpper()
20. reasonableUpper( upperAverage, dis2 );
21.      return upperAverage+dis2;

```

算法 4 匹配结果(RCF)

Input:

D: 已经进行预处理的数据库.

RRS: 合理值集合.

ItemNumber: 数据库中对象行号.

Output:

MRS: 匹配结果, 将不合理数据存储下来.

```

1. for j = 0 to ( row.length-1 ) do
2.   originalData = getValue[ itemNumber ];
3.   if ! matchCompare( originalData, [ RL, RU ] ) then
4.     sum ← sum + 1;
5.     output( j, sum );
6.   end if
7. end for
8. % the method of matchCompare()
9. matchCompare( originalData, [ RL, RU ] )
10.   originalLower = originalData.Bottom;
11.   originalUpper = originalData.Top;
12.   if ( originalLower >= RL ) && ( originalUpper <= RU ) then
13.     return true;
14.   else
15.     return false;
16.   end if

```

算法 4 是将算法 3 所得合理区间值与实际值进行区间匹配, 当与匹配结果不一致时, 则记录下不一致的数据. 为后续进行进一步处理做好准备. 在此, 我们不强调直接进行删除操作.

算法分析 算法 1 的时间复杂度主要表现在第 1~13 行的循环中, 而循环过程与数据集的属性项数目、对象数有直接关联, 经分析, 其时间复杂度为: $O(|row.length| * |column.length|)$. 容易得算法 2 的时间复杂度也为 $O(|row.length| * |column.length|)$. 以 (ISIS, AIS) 作为算法 3 的输入, 算法 3 的时间复杂度主要由第 1-15 行决定, 则为 $O(|row.length| * |ISISColumn.length| + |AISColoumn.length|)$. 若 p 设为匹配函数 $matchCompare()$ 的运行时间, 则易得算法 4 的时间复杂度为 $O(p * |row.length|)$.

4 实验

4.1 实验环境

实验中采用的配置如下: CPU 为 Intel(R) Pentium(R) Dual(2.16 GHz), 2 GB 内存, 250 GB 硬盘; 开发平台为 Eclipse 4.4, JDK 为 Java jdk 1.8.0_20. 数据集源于 2014 年“昆明市建设世界知名旅游城市对策研究”(countermeasures for constructing the world famous tourism city in Kunming(2014), China. 记为 CCWFCK) 的真实问卷调查数据集, 原始数据集为 610 条记录, 每一条记录有 18 项属性. 18 项属性中, 其中有 5 项属性可作为决策属性, 其余为条件属性. 同时, 为了更好的测试运行时间, 对 CCWFCK 进行复制 10 次、50 次及 100 次, 具体如表 4 所示.

4.2 实验结果

(1) 精度比 (Precision): 设数据集 CCWFCK 的 3 模式分别为: R_1, R_2, R_3 . 其中, R_1 共 13 项属

表 4 CCWFCK 景区满意度问卷调查

Table 4 Scenic spot satisfaction questionnaire survey of CCWFCK

数据集	对象数	条件属性	决策属性
CCWFCK	610	13	5
CCWFCK10	6 100	13	5
CCWFCK50	30 500	13	5
CCWFCK100	61 000	13	5

性(12 项条件属性,1 项决策属性); R_2 共 6 项属性(5 项条件属性,1 项决策属性); R_3 共 3 项属性(2 项条件属性,1 项决策属性).

清洗度 (Cleaning degree,CD):表示整体数据集 CCWFCK 中不合理数据占的比例,具体如下:

$$CD(data,f)=\frac{N_f}{N}.$$

式中, N_f 表示不合理数据的对象数目, N 表示总体数据集的对象数目.

精确度 (Precision degree,PD):表示不合理数据的对象数中确实有问题的数据,具体如下:

$$PD(data,f,i)=\frac{N_f-N_i}{N_f}.$$

式中, N_f 表示不合理数据的对象数目, N_i 表示清洗出不合理数据中确实没有问题的对象数目.

表 5 显示的是区间模糊匹配函数与均值区间匹配函数对比结果. 表 5 显示出与本文算法相比,均值区间匹配函数的清洗力度大,但实际清洗效果有待提高.

表 5 CCWFCK 不合理数据
Table 5 The unreasonable data of CCWFCK

模式	对象数	本文算法				均值区间匹配函数			
		N_f	N_i	CD	PD	N_f	N_i	CD	PD
R_1	610	88	12	14.4%	86.4%	149	73	24.4%	51.0%
R_2	610	120	19	19.7%	84.2%	181	80	29.6%	55.8%
R_3	610	123	20	20.2%	82.1%	241	141	39.5%	41.5%

其中,均值区间匹配函数直接源于文献[5-7] 匹配函数与函数依赖思想.

图 1 展示了随着模式中的属性集的减少,则清洗度越高,清洗出不合理数据越多. 图 2 描述了随着模式中的属性集的减少,则精确度越来越低. 同时,图 2 显示出在模式 R_1 精确度达到 86%,则表示区间模糊匹配函数数据清洗算法对这类型数据清洗的效果不错.

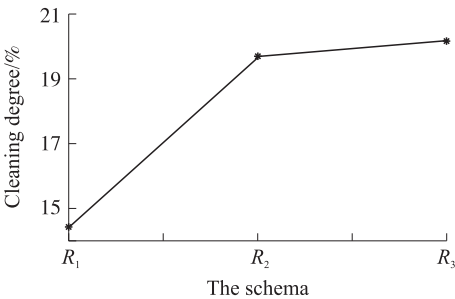


图 1 清洗度
Fig. 1 Cleaning degree

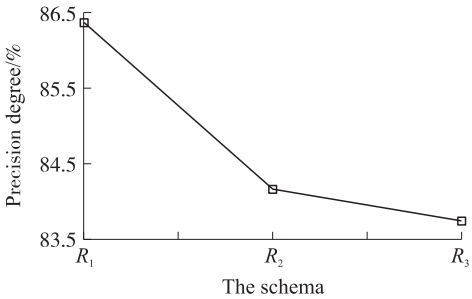


图 2 精确度
Fig. 2 Precision degree

(2) 运行时间 (Runtime): 我们选取模式 R_1, R_2 来测试时间效率,运行结果如图 3 所示.

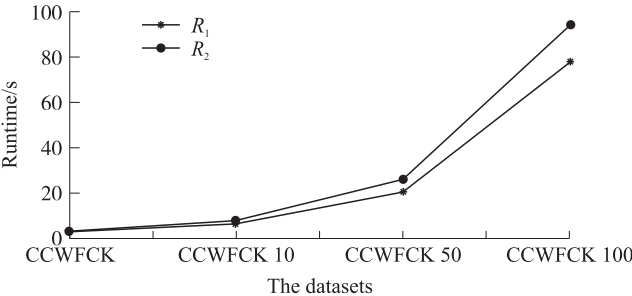


图 3 CCWFCK 运行时间
Fig. 3 The runtime of CCWFCK

在图 3 中,显示出随着数据量的增加其执行时间不断增加. 同时,相同数据集的模式属性集小的运行时间小于模式属性集多的.

5 结语

数据清洗是保证数据质量的重要保证,也是决策的基石. 对于问卷调查数据集亦是如此. 在此文中,为处理具有隐藏性的不合理数据,我们提出了一种区间模糊匹配函数算法. 通过实验评估,该算法能处理此种不合理数据. 但我们也再次强调,这种不合理数据不宜轻易删除,更重要的是进行修复、重新评估属性集的设置问题及调查的人群问题等.

[参考文献] (References)

- [1] KUMAR R, CHADRASEKARAN D R. Attribute correction-data cleaning using association rule and clustering methods[J]. International journal of data mining and knowledge management process, 2011, 1(2): 22-32.
- [2] RAHM E, HONG H D. Data cleaning: problems and current approaches[J]. IEEE data engineering bulletin, 2000, 23(4): 3-13.
- [3] GARDEZI J, BERTOSSI L, KIRINGA I. Matching dependencies: semantics and query answering[J]. Frontiers of computer science, 2012, 6(3): 278-292.
- [4] LOW W L, LEE M L, LING T W. A knowledge-based approach for duplicate elimination in data cleaning[J]. Information systems, 2001, 26(8): 585-606.
- [5] FAN W, JIA X, LI J, et al. Reasoning about record matching rules[J]. Proceedings of the VLDB endowment, 2010, 2(1): 407-418.
- [6] FAN W, MA S, TANG N, et al. Interaction between record matching and data repairing[J]. Journal of data and information quality, 2014, 4(4): 1-38.
- [7] BERTOSSI L, KOLAH S, LAKSHMANAN L V S. Data cleaning and query answering with matching dependencies and matching functions[J]. Theory of computing systems, 2013, 52(3): 441-482.
- [8] GRAHAM J W. Missing data analysis: making it work in the real world[J]. Annual review of psychology, 2009, 60: 549-576.
- [9] WENG C H, CHEN Y L. Mining fuzzy association rules from uncertain data[J]. Knowledge and information systems, 2010, 23(2): 129-152.
- [10] CHANG S E, CHANGCHIEN S W, HUANG R H. Assessing users' product-specific knowledge for personalization in electronic commerce[J]. Expert systems with applications, 2006, 30(4): 682-693.
- [11] DOHERTY N, ELLIS-CHADWICK C F, HART C. An analysis of the factors affecting the adoption of the Internet in the UK retail sector[J]. Journal of business research, 2003, 56(11): 887-897.
- [12] CHEN Y L, WENG C H. Mining fuzzy association rules from questionnaire data[J]. Knowledge-based systems, 2009, 22(1): 46-56.
- [13] MARSHALL G. The purpose, design and administration of a questionnaire for data collection[J]. Radiography, 2005, 11(2): 131-136.
- [14] BURTON S H, MORRIS R G, GIRAUD-CARRIER C G, et al. Mining useful association rules from questionnaire data[J]. Intelligent data analysis, 2014, 18(3): 479-494.
- [15] YAMANISHI K, LI H. Mining open answers in questionnaire data[J]. IEEE intelligent systems, 2002, 17(5): 58-63.
- [16] BROECK J V D, CUNNINGHAM S A, EECKELS R, et al. Data cleaning: detecting, diagnosing, and editing data abnormalities[J]. Plos medicine, 2005, 2(10): e267.
- [17] BOYNTON P M. Administering, analysing, and reporting your questionnaire[J]. BMJ, 2004, 328(7452): 1372-1375.
- [18] SAMBUC R. Fonctions and floues: application a l'aide au diagnostic en pathologie thyroïdienne[D]. Marseille: University of Marseille, 1975.
- [19] ZADEH L A. The concept of a linguistic variable and its application to approximate reasoning[J]. Information sciences, 1975, 8(3): 199-249.
- [20] SANZ J, FERNÁNDEZ A, BUSTINCE H, et al. A genetic tuning to improve the performance of Fuzzy Rule-Based Classification Systems with Interval-Valued Fuzzy Sets: Degree of ignorance and lateral position[J]. International journal of approximate reasoning, 2011, 52(6): 751-766.
- [21] DESCHRIJVER G. Triangular norms which are meet-morphisms in interval-valued fuzzy set theory[J]. Fuzzy sets and systems, 2008, 181(1): 88-101.

- [22] WU Z G, SHI P, SU H, et al. Network-based robust passive control for fuzzy systems with randomly occurring uncertainties[J]. IEEE transactions on fuzzy systems, 2013, 21(5): 966–971.
- [23] ZHANG H, YAN H, YANG F, et al. Quantized control design for impulsive fuzzy networked systems[J]. IEEE transactions on fuzzy systems, 2011, 19(6): 1 153–1 162.
- [24] ATANASSOV K. Interval valued intuitionistic fuzzy sets[J]. Fuzzy sets and systems, 1989, 31(3): 343–349.
- [25] 曾文艺, 李洪兴, 施煜. 区间值模糊集合的分解定理[J]. 北京师范大学学报(自然科学版), 2003, 39(2): 171–177.
ZENG W Y, LI H X, SHI Y. Decomposition theorem of interval-value fuzzy sets[J]. Journal of Beijing normal university(natural science), 2003, 39(2): 171–177.(in Chinese)
- [26] 金澈清, 刘辉平, 周傲英. 基于函数依赖与条件约束的数据修复方法[J]. 软件学报, 2016, 27(7): 1 671–1 684.
JIN C Q, LIU H P, ZHOU A Y. Functional dependency and conditional constraints based data repair[J]. Journal of software, 2016, 27(7): 1 671–1 684.(in Chinese)
- [27] 钟评, 李战怀, 陈群. 关系数据中函数依赖检测方法[J]. 计算机学报, 2017, 40(1): 207–222.
ZHONG P, LI Z H, CHEN Q. A functional dependencies checking method in relational data[J]. Chinese journal of computers, 2017, 40(1): 207–222.(in Chinese)
- [28] ZADEH L A. Fuzzy sets[J]. Information and control, 1965, 8(3): 338–353.
- [29] 刘文奇. 中国公共数据库数据质量控制模型体系及实证[J]. 中国科学:信息科学, 2014, 44(7): 836–856.
LIU W Q. Modeling data quality control system for Chinese public database and its empirical analysis[J]. Scientia sinica (informationis), 2014, 44(7): 836–856.(in Chinese)

[责任编辑:陈 庆]