文章编号：1000-5013(2016)03-0308-04　　**doi**：10.11830/ISSN.1000-5013.2016.03.0308

考虑数据不确定性的非均匀挖掘算法

刘竹松，陈洁

（广东工业大学 计算机学院，广东 广州 510006）

摘要：针对高维大数据不确定性的非均匀挖掘问题，提出一种基于不确定频繁模式树的模糊逻辑非均匀数据挖掘算法. 首先，在考虑数据不确定性的前提下建立高维数据的区域连接演算(RCC)模型，并基于数据集合组元定义分析不确定数据集合的模糊距离；然后，采用不确定模式树对数据的非均匀特性进行均匀泛化处理，并给出了具体的实现步骤. 仿真结果表明：文中方法有效地提升不确定非均匀数据集合在不同支持度情况下的挖掘效率.

关键词：高维大数据；数据挖掘；模糊逻辑；不确定频繁模式树；区域连接演算

中图分类号：TP 311.13　　**文献标志码**：A

数据挖掘算法已经成为大数据领域的热点. 通过有效的算法分析，能够及时有效地发现海量数据的价值信息，增强目标预测的有效性和准确性[1-2]. 为了提升数据的保密功能，大部分数据都具有人为的不确定性及集合区分的非均匀特性，而现有的挖掘算法针对这类数据的挖掘效率太低甚至失效[3]. 2000 年，Eliseo 等[4]提出采用多层集合分界进行高维大数据的均匀规则挖掘. 但是，目前不确定非均匀数据集合的研究还非常少，特别是高维数据的分集及集合大小的选择问题[5]. Shifei 等[6]在现有的挖掘思想中引入智能推理的思想，将定性空间推理(qualitative spatial reasoning，QSR)技术引入到高维数据的不确定性挖掘算法. 张继福等[7]实现了一种基于相关子空间的局部离群数据挖掘算法，并有效改善离群程度较大的局部离群数据的挖掘效率. 基于此，本文提出一种考虑数据不确定性的非均匀挖掘算法.

1　高维大数据模型

1.1　高维数据的建模

通常针对大数据集合建模的主要思想就是将数据集合看作空间目标进行整体建模. 针对确定数据建模的代表性工作是文献[8]提出的区域连接演算(region connection calculus，RCC)理论，该理论通过连接算子 $C(x,y)$ 先后分析了 RCC-5，RCC-8，RCC-15 等多种数据拓扑结构. 文中在 RCC-5 模型的基础上，通过融入模糊逻辑构建基元数据关系和不确定模式树的方法，针对不确定数据集合进行建模. 为了便于分析，采用三个基元进行简化模型分析，其三基元等价模型为

$$(\alpha_1,\alpha_2,\alpha_3)=(x\wedge y\neq\varnothing, x\wedge -y\neq\varnothing, -x\wedge y\neq\varnothing). \quad (1)$$

式(1)中：各基元数据的取值范围都是{0，1}；具有实际物理意义的基元数据有 5 种. 相应的 RCC-5 模型的对应关系如表 1 所示.

表 1　RCC-5 模型的对应关系

Tab. 1　Corresponding relationship for RCC-5 model

α_1	α_2	α_3	RCC-5
0	1	1	DC
1	1	1	PO
1	0	1	PP
1	1	0	PPI
1	0	0	EQ

1.2　高维数据集合模糊距离分析

数据的确定通常都是一种理想的假设，在实际的大数据信息中，由于保密和应用范围的需求，通常都会人为地加入不确定信息，以保持数据集合的唯一性[9]. 数据的不确定性表现在集合分割的特征方面

收稿日期：2016-03-18

通信作者：刘竹松(1979-)，男，副教授，博士，主要从事云计算、大数据的研究. E-mail：liuzs@gdut.edu.cn.

基金项目：国家自然科学基金资助项目(61572144)；广东省科技计划项目(2013B090200006)；广东省现代信息服务业发展专项基金资助项目(GDEID2011IS022)

就是区域边界的模糊性，在分析研究中通常将这种模糊性称为近似分割集合. 近似分割思想是 QSR 理论近年来的研究热点，目前还没有形成统一的理论框架.

文献[10-11]分别提出了近似区域和宽边界区域的“蛋黄”模型，陈爱东等[12]提出了一种应用高维数据的交集扩展模型，其主要事项都是针对均匀不确定数据的聚类处理. 采用模糊逻辑的思想进行数据不确定关系的理论建模分析，假设给定的近似点集为 $P^*=(P,\varepsilon)$，$P(a,b)$表示给定集合的确定点，$\varepsilon>0$ 为 P 的有效延伸区域，则 P^* 的隶属度函数可以表示为

$$\mu_{P^*}(x,y)=k(1-\frac{d^{00}((x,y),(a,b))}{\varepsilon}). \tag{2}$$

式(2)中：$k(x)=\max\{\min\{x,1\},0\}$；$d^{00}$为基元点集之间的欧式距离.

近似点集构成的封闭集合区域 $R^*=(R,\varepsilon)$表示为

$$\mu_{R^*}(x,y)=k(1-\frac{d^{02}((x,y),R)}{\varepsilon}). \tag{3}$$

式(3)中：R 为 R^* 的核，是数据分割集合的确定区域.

同样，点与分割集合区域之间的距离函数表示为

$$d^{02}((x,y),R)=\begin{cases}0, & (x,y)\in R,\\ \min\limits_{(a,b)\in R}(d^{00}((x,y),(a,b))), & (x,y)\notin R.\end{cases} \tag{4}$$

1.3　高维不确定关系计算

通过前面的分析，该部分主要基于元素的基元进行数据集合区域的近似计算分析，近似区域的总体表示为

$$R_1^*=(R_1,\varepsilon_1),\qquad R_2^*=(R_2,\varepsilon_2). \tag{5}$$

由于这种逻辑数学的表达形式不能直接判断集合的关系，文中主要通过集合基元数据(STUPLE)进行近似区域的表达，相互关系可以表示为

$$\mathrm{STUPLE}(R_1^*,R_2^*)=(\mathrm{TAG},d_1,d_2,s_1,s_2). \tag{6}$$

式(6)中：TAG 为 RCC-5 模型的数据核；d_1 为两个圆心为 R_1 和 R_2 的圆集合的边缘距离，其中，正数为在两集合之外，负数为两集合之内，当两个集合重叠的时候，取集合边界的最大值；d_2 是 R_2 相对于 R_1 的距离，其中，最大值分别表示为 s_1 和 s_2.

根据给定数据集合，初始化点集区域 R_1 和 R_2，并结合式(1)计算不确定区域 R_1^* 和 R_2^* 的隶属度函数，即

$$\mu_{a_1}=\bigvee_{(x,y)}[\mu_{R_1}^*(x,y)\wedge\mu_{R_2}^*(x,y)], \tag{7}$$

$$\mu_{a_2}=\bigvee_{(x,y)}[\mu_{R_1}^*(x,y)\wedge(1-\mu_{R_2}^*(x,y))], \tag{8}$$

$$\mu_{a_3}=\bigvee_{(x,y)}[(1-\mu_{R_1}^*(x,y))\wedge\mu_{R_2}^*(x,y)]. \tag{9}$$

在数据集合分布较密集的情况下，如果满足分析条件的最大距离为 $s_1\gg\varepsilon_1$，$s_2\gg\varepsilon_2$，可以将文中模型的三元隶属度函数表示为

$$\mu_{a_1}=\begin{cases}k[1-\dfrac{d_1}{\varepsilon_1+\varepsilon_2}], & \mathrm{TAG}=1,\\ 1, & \mathrm{TAG}\neq 1.\end{cases}$$

$$\mu_{a_2}=\begin{cases}1, & \mathrm{TAG}=1,\\ k[\dfrac{\varepsilon_1+d_1}{\varepsilon_1+\varepsilon 2}], & \mathrm{TAG}=2,\\ k[\dfrac{\varepsilon_1-d_1}{\varepsilon_1+\varepsilon 2}], & \text{其他}.\end{cases}\qquad \mu_{a_3}=\begin{cases}1, & \mathrm{TAG}=1,\\ k[\dfrac{\varepsilon_2+d_2}{\varepsilon_1+\varepsilon 2}], & \mathrm{TAG}=2,\\ k[\dfrac{\varepsilon_2-d_1}{\varepsilon_1+\varepsilon 2}], & \text{其他}.\end{cases}$$

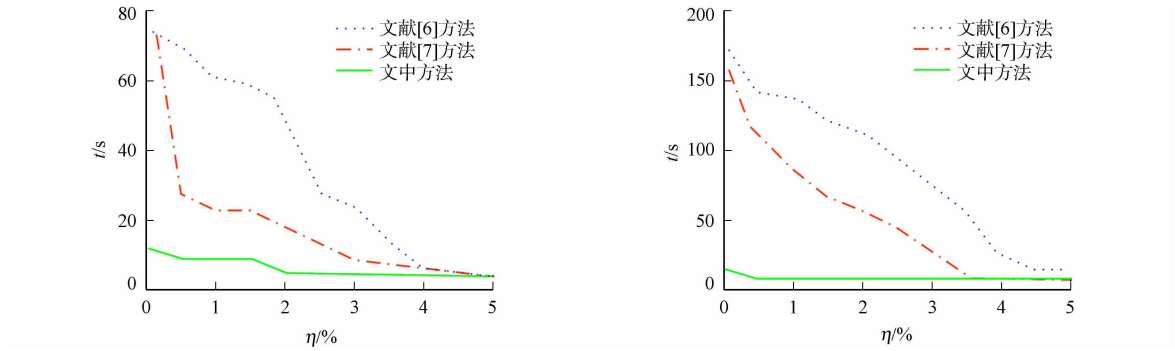
2　仿真与结果分析

为充分分析文中方法的可行性，对基于确定均匀数据和非确定非均匀两类数据库进行仿真分析. 仿真采用 Intel Core 2，CPU 为 2.4 GHz，RAM 内存为 2 GB，操作系统为 Window XP，软件为 Matlab

2012. 实验选择了两组数据集合,基于 UCI 机器学习数据库的 Adult 数据[13-14],该数据共 48 842 条记录,属确定均匀分布的数据集合,仿真中取前 10 000 条数据记为 D_1. 另一条数据是 IBM 数据集生成器生成的数据,该数据集总记录数据为 100 万条. 由于数据太长,根据实验分析需要,将数据集合进行了分割处理,仅提取需要的 5 000 条数据进行分析,记录为 D_2. 为定量说明方法的性能,在相同的实验条件下,选择了文献[6-7]的方法进行对比分析.

2.1 不同支持度情况下的效率分析

针对两组不同数据集合,在支持度从 5%降到 0.1%的情况下,不同挖掘算法的运行时间,如图 1 所示. 图 1 中:t 为运行时间;η 为支持度. 由图 1 可知:相对于确定、均匀数据集合而言,文献[6-7]的方法在非确定、非均匀数据集合上的运行时间大幅提升,已经无法满足数据挖掘的实时性需求,效率低下;但是文中方法在两种数据集合情况下均保持了较好的挖掘效率,虽然在非确定、非均匀情况下的运行时间有所增加,但是仍然处于有效的挖掘效率范围内.



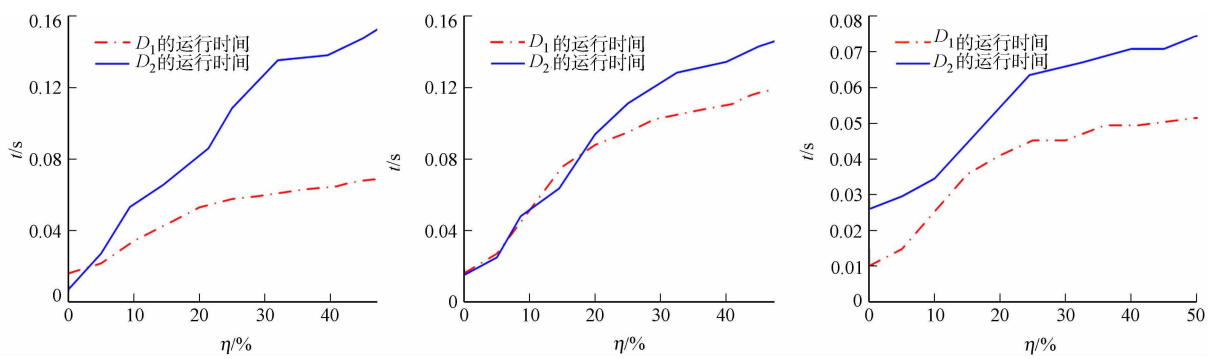
(a) D_1 (b) D_2

图 1 不同支持度的运行效率比较

Fig. 1 Running efficiency comparison of different support degree

2.2 生成一个频繁树的时间消耗分析

为进一步定量分析文中挖掘算法在相同条件下的挖掘性能,对比分析两个不同的数据库进行了生成一个频繁树需要的时间消耗,仿真结果如图 2 所示. 由图 2 可知:文中方法针对两个数据库的运行时间消耗都控制在 0.08 s 以内;而文献[6-7]的方法的运行时间均高出文中方法一个数量级,这一点的主要原因是文中方法采用了模糊逻辑进行了模型的建模,在理论上克服了边界模糊效应的影响.



(a) 文献[6]方法 (b) 文献[7]方法 (c) 文中方法

图 2 生成一个频繁树的时间消耗对比

Fig. 2 Time consumption comparison of generating a frequent tree

3 结束语

针对高维大数据的不确定性非均匀挖掘问题,文中提出了一种基于不确定频繁模式树的模糊逻辑非均匀数据挖掘算法. 在考虑数据不确定性的前提下,建立了高维数据的 RCC 模型,并基于数据集合组元定义分析了不确定数据集合的模糊距离. 由于采用模糊逻辑的思想进行建模,在理论上消除了数据集合的边缘效应,增强了算法的运行效率;同时,由于不确定模式树对数据的非均匀特性进行均匀泛化处理,进一步增强了算法对无序大数据的处理能力. 最后的仿真结果表明:文中方法为处理非确定和非均

匀大数据提供一种可行的思路，

参考文献：

[1] 喻小光，陈维斌，陈荣鑫. 一种数据规约的近似挖掘方法的实现[J]. 华侨大学学报(自然科学版)，2008，29(3)：370-374.
[2] 朱龙. 利润约束的关联规则挖掘算法[J]. 华侨大学学报(自然科学版)，2015，36(5)：522-526.
[3] 吴章玲，金培全，岳华丽，等. 基于 PCM 的大数据存储与管理研究[J]. 计算机研究与发展，2015，52(2)：343-361.
[4] ELISEO C，FELICE P D. Mining multiple-level spatial association rules for objects with a broad boundary[J]. Data and Knowledge Engineering，2000，34(3)：251-270.
[5] 王珊，王会举，覃雄派. 架构大数据：挑战、现状与展望[J]. 计算机学报，2011，34(10)：174-181.
[6] SHIFEI D，FULIN W，JUN Q，et al. Research on data stream clustering algorithms [J]. Artificial Intelligence Review，2013，43(4)：593-600
[7] 张继福，李永红，秦啸，等. 基于 MapReduce 与相关子空间的局部离群数据挖掘算法[J]. 软件学报，2015，26(5)：1079-1095.
[8] 刘大有，王生生，虞强源. 基于定性空间推理的多层空间关联规则挖掘算法[J]. 计算机研究与发展，2004，41(4)：565-570.
[9] JONATHAN A S，ELAINE R F，RODRIGO C B，et al. Data stream clustering：A survey[J]. ACM Computing Surveys，2013，46(1)：1-13，31.
[10] 李洁，高新波，焦李成. 一种基于 GA 的混合属性特征大数据集聚类算法[J]. 电子与信息学报，2004，26(8)：1203-1209.
[11] 任家东，王倩，王蒙. 一种基于频繁模式有向无环图的数据流频繁模式挖掘算法[J]. 燕山大学学报(自然科学版)，2011，35(2)：115-120.
[12] 陈爱东，刘国华，费凡，等. 满足均匀分布的不确定数据关联规则挖掘算法[J]. 计算机研究与发展，2013，50(增刊 1)：186-195.
[13] 雷向欣，杨智应，黄少寅，等. XML 数据流分页频繁子树挖掘研究[J]. 计算机研究与发展，2012，49(9)：1926-1936.
[14] 孙力娟，陈小东，韩崇，等. 一种新的数据流模糊聚类方法[J]. 电子与信息学报，2015，37(7)：1620-1625.

Non-Uniform Mining Algorithm for Considering Data Uncertainty

LIU Zhusong，CHEN Jie

(School of Computer Science and Technology，Guangdong University of Technology，Guangzhou 510006，China)

Abstract：In order to solve high-dimensional large data uncertainty and non-uniform mining problems，this paper proposed a new kind of non-uniform data mining algorithm based on the fuzzy logic and uncertain frequent pattern tree. Firstly，the high-dimensional region connection calculus (RCC) data model is established under the premise of considering the data uncertainty. The uncertain fuzzy distance of data sets is defined and analyzed based on the data sets elements. Secondly，the non-uniform data is generalized by the uncertain frequent pattern tree，and the specific implementation steps is given. Simulation results show that the proposed method effectively improved the mining efficiency of the uncertain heterogeneous data sets in different support conditions.

Keywords：high dimensional data；data mining；fuzzy logic；uncertain frequent pattern tree；region connection calculus

(**责任编辑**：陈志贤　　**英文审校**：吴逢铁)