计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2014.20130869
51(11):2518-2527, 2014

不确定度模型下数据流自适应网格密度聚类算法

刘 卓[1] 杨 悦[2] 张健沛[2] 杨 静[2] 初 妍[2] 张泽宝[2]
[1](哈尔滨工程大学自动化学院 哈尔滨 150001)
[2](哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)
(liuzhuo@hrbeu.edu.cn)

An Adaptive Grid-Density Based Data Stream Clustering Algorithm Based on Uncertainty Model

Liu Zhuo[1], Yang Yue[2], Zhang Jianpei[2], Yang Jing[2], Chu Yan[2], and Zhang Zebao[2]
[1](*College of Automation*, *Harbin Engineering University*, *Harbin* 150001)
[2](*College of Computer Science and Technology*, *Harbin Engineering University*, *Harbin* 150001)

Abstract Uncertain data stream, a new widespread data form which is emerging in many application fields with the development of computer and sensing technology. The research of data analysis and processing of uncertain data stream has attracted the attention of many researchers. Existing data stream clustering techniques generally ignored uncertainty characteristics. It often makes the clustering results unreasonable even unavailable. The two aspects of uncertain character, existence-uncertainty and attributive-uncertainty, can affect the clustering process and results significantly. But they can't be considered at same time in existing relevant work. The lately reported clustering algorithms are all based on *K*-Means algorithm with inherent shortage. In order to solve this problem, a data stream adaptive grid-density based algorithm, ADC-UStream, is proposed under the uncertainty of model. For the uncertainty characteristic, with the unified strategy of the presence and properties uncertainty, the algorithm builds the entropy uncertainty model to measure the uncertainty. With the comprehensive survey of uncertainty, the grid-density based clustering algorithm over attenuation window model is adopted to design the temporal and spatial adaptive density threshold, to adapt to the temporal and non-uniform distribution characteristics of the uncertainty data flow. The experimental results show that the ADC-UStream algorithm under the uncertainty model has good performance both in clustering quality and clustering efficiency.

Key words uncertain character; data stream; clustering; grid-density; adaptive density threshold; uncertainty model

摘 要 随着计算机技术及感知技术的发展及应用,各个领域普遍出现不确定性数据流形态的新型数据,吸引了众多研究者的关注.现有的数据流聚类技术普遍忽略不确定性特征,常导致聚类结果的不合理甚至不可用.为数不多的针对不确定性特征的聚类方法片面考察不确定性,且大多基于 *K*-Means 算法,具有先天缺陷.针对这一问题展开研究,提出了不确定度模型下数据流自适应网格密度聚类算法

收稿日期:2013-06-24;**修回日期**:2013-09-30
基金项目:国家自然科学基金项目(61202274);中国博士后科学基金项目(2012M510927);黑龙江省博士后科学基金项目(LBH-Z12066);中央高校基本科研业务费专项资金项目(HEUCF100602)
通信作者:杨 悦(yangyue@hrbeu.edu.cn)

(adaptive density-based clustering algorithm over uncertain data stream, ADC-UStream). 对于不确定性特征,该算法在存在级和属性级不确定性统一策略下,构建熵不确定度模型进行不确定性度量,综合考察不确定性. 采用网格-密度的聚类算法,基于衰减窗口模型设计时态和空间的自适应密度阈值,以适应不确定性数据流的时态性和非均匀分布特征. 实验结果表明,不确定模型下的数据流网格密度自适应聚类算法 ADC-UStream 在聚类结果质量和聚类效率方面都具有较好的性能.

关键词　不确定性;数据流;聚类;网格-密度;自适应密度阈值;不确定度模型

中图法分类号　TP311

随着数据获取技术和数据处理技术的不断发展,人们认识到数据不确定性的存在具有普遍性,越来越多学者的关注点由确定性转向不确定性数据的处理技术研究,面向不确定性的数据分析和挖掘技术已经成为新近的研究热点[1].

实际应用中数据采集设备及采集技术收集到的数据均具有不确定性特征,其中最典型的物联网数据感知层中的主要生力军射频识别(RFID)技术和无线传感器网络技术由于物理设备本身及外界环境等主客观因素的影响,其获取的数据具有以下不确定性的特征:1)本体因素. 传感器节点受其体积、能源、成本等因素限制,采集数据精度难以保证,RFID 的阅读器在实际应用中经常发生误读、漏读、多读现象,错误率高达 30%～40%[2]. 2)外界因素. 工作环境的复杂多变使得物理设备数据采集精度下降、质量不稳定,如无线传感器网络中数据的传输会受到带宽、传输延时、能量、外磁场干扰等因素的影响. 3)预处理因素. 无线传感器网络中各网络簇节点被设计对所采集数据进行一定的预处理,如数据集成、异构融合、缺失值处理等,在为后续数据处理过程作必要准备的同时,引入了新的不确定因素. 4)隐私保护因素. 某些应用出于隐私保护的目的,通过一系列方法对数据进行了相应处理,不能获得准确的细节化的原始数据[1]. 以上因素所引发的数据的不确定性特征严重影响数据处理和分析结果的准确性,甚至使得结果不可接受,不仅造成时间和资源的浪费,更会使数据所反映知识的时效性降低为零,实际应用中很可能会造成无法挽回的社会经济损失. 同时,物联网的感知层会产生大规模具有实时性的流式数据,而 RFID 及无线传感器网络本身仅具有十分有限的数据处理能力和计算资源,迫切需要有针对性地研究强有力的数据流处理、分析及数据挖掘方法,使得所获得的数据能被有效利用,真正实现物联网环境下的智慧感知. 在物联网技术方兴未艾的大背景下,研究其感知层获取数据的不确定数据流聚类技术具有非常重要的理论意义和应用价值. 基于聚类分析技术在数据处理领域的重要性及其与其他学科交叉的特性,本文重点研究针对不确定性数据流的聚类技术.

不确定性数据流聚类技术的主要挑战表现为:1)数据随时间推移高速到达,具有实时性,要求聚类算法应具备较低的时空复杂度,节省存储空间和处理时间,且尽量采用单次扫描的方式(重复读取数据的代价高);2)数据流具有时态特征,数据点对聚类形成簇的贡献权值不断衰减,要求算法反映数据点随时间蜕化的特点;3)不确定性问题引入了新的概率维,数据点存在级不确定性(数据点存在与否的可能性)以及数据点属性级不确定性(数据点特定属性取某值的可能性)往往同时并存且相互作用,对聚类结果造成影响,迫切要求算法对于不确定性建立适合的概率模型且在聚类过程中反映该概率维;4)数据规模及流速率不可预知,加之概率维的参与,难以获取簇个数及簇形状等先验知识,对阈值参数设定造成困难,要求算法善于挖掘任意形状的簇并涉及尽量少的人工设定参数,以降低人为干预使得结果不准确的风险.

本文提出一种不确定性数据流自适应密度聚类算法(adaptive density-based clustering algorithm over uncertain data stream, ADC-UStream). 该算法的主要贡献包括:1)对于不确定性数据流数据对象,针对目前为数不多的同类算法中仅对存在级或属性级不确定性片面考虑进行聚类的问题,提出了存在级和属性级属性综合统一的策略,并在此基础上构建了基于超椭球体指标的熵不确定度模型,进行不确定性度量;2)基于网格密度的不确定性数据流聚类,改善现有同领域算法的仅能发现球形聚类、人工参数设定等造成的聚类结果不准确的问题,算法构建适合的衰减窗口模型,设计具有时态和空间自适应能力的密度阈值,以适合数据对象的流式特征和空间非均匀分布特征,取得良好的聚类效果.

1 已有相关工作分析

目前国内外相关研究工作大多集中在确定性数据流聚类分析方法上[3-5],而对于实际应用中普遍存在的不确定性数据,聚类方法讨论仅局限在静态数据[6-8],而非数据流上.公认的传统的聚类方法很难直接应用到数据流上,加之不确定因素就更加难以胜任.

针对具有不确定性特征的数据流聚类方法,国内外相关的研究成果并不多.Thanh 等人[9]对不确定性数据流这种新的数据形式提出了基于自然提取的连续性随机变量的 CLARO 模型,以支持该数据形式的流式计算;Aggarwal 等人[10]提出了 UMicro 算法,将聚类特征(clustering feature, CF)扩展为扩展性聚类特征(entended clustering feature, ECF)结构,在新到样本点与簇中心之间距离以及簇半径的计算过程中扩展考虑了不确定因素;Aggarwal 等人[11]又针对高维数据"维度灾难"问题提出了投影空间下的不确定数据流聚类算法;张晨等人[12]采用信息熵描述元组的不确定性信息提出基于信息熵的不确定性数据流聚类算法.以上方法针对属性级不确定性进行描述,而对 RFID 及无线传感器网络应用中经常出现的数据存在级不确定性并未提及.为此,张晨等人[13]又提出 EMicro 算法,改进 UMicro 算法中的 ECF 为不确定性聚类特征(uncertian clustering feature, UCF),对存在级不确定性元组进行聚类.近期,Aggarwal 等人[14]、Chen 等人[15]又展开了对于不确定性数据流的图形态数据的聚类方法及技术的讨论和研究.

对于近期为数较少的不确定性数据流聚类研究成果,仅是对该领域的初涉,不确定数据流聚类问题尚有相当大的研究空间亟待探索.当前研究工作存在的主要问题包括:1)对不确定性特征的关注和描述片面化,只关注存在级不确定性或只关注属性级不确定性,实际应用中存在级和属性级不确定性同时并存且具有一定相关性,将二者割裂开来讨论难以得到有说服力的结果;2)就聚类方法本身而言未形成切实可行的方案,现有相关算法均以欧氏距离作为考察依据,脱离不了此类传统算法的先天不足,即形成球形簇而非任意形状簇,不能很好地刻画真实聚类情况,样本数量增加时导致时空复杂度过高,需预先设定簇数目等关键参数值,聚类结果人为干预明显,不容易进行高维扩展等;3)忽视数据流的时效性,未考虑数据点聚类贡献随时间蜕化的问题,而是新到样本点与陈旧样本点(到达后经过一段时间的样本点)在聚类过程中同等对待,显然与现实应用中人们对数据流的关注情况不符.

本文提出的 ADC-UStream 算法综合考察了数据的存在级和属性级不确定性特征,采用基于网格密度的聚类算法,并设计时间空间自适应密度阈值避免结果人为干预,实现不确定性数据流的自适应网格密度聚类分析.

2 不确定性数据流多维变量不确定度模型

2.1 不确定性数据流

定义 1. 不确定性数据流可以描述为相互独立的具有不确定性的元组构成序列,即

$$S=\{(\bar{\boldsymbol{X}}_1,p_1),(\bar{\boldsymbol{X}}_2,p_2),\cdots,(\bar{\boldsymbol{X}}_n,p_n)\},\quad(1)$$

其中,$\bar{\boldsymbol{X}}_i$ 是 d 维的不确定性元组,$\bar{\boldsymbol{X}}_i=(x_i^{(1)},x_i^{(2)},\cdots x_i^{(d)})$,$x_i^{(j)}(i=1,2,\cdots,n),(j=1,2,\cdots,d)$,$p_i(i=1,2,\cdots,n)$为元组 $\bar{\boldsymbol{X}}_i$ 的存在概率.不失一般性,p_i 取值为 $p_i\in[\theta,1](0\leqslant\theta\leqslant1)$.

2.2 存在级和属性级不确定性

在不确定性数据库中,元组不是以单个确定的值出现,而是一个可能取值的集合,集合中的每个元组都附带一个概率值以标识其出现的可能性,即元组级不确定性,也称作存在级不确定性;同时元组的属性也可能附带相应的概率值表示其属性取值的可能性,即属性级不确定性.也就是说,在不确定性数据库中,对于数据不确定性的描述可分为描述元组存在可能性的存在级不确定性、描述元组各构成属性的取值不确定性的属性级不确定性.在现有的不确定性数据流分析方法中,只考虑存在级与属性级不确定性中的一种,大多采用点概率估计模型描述存在级不确定性,采用概率密度函数描述元组的属性级不确定性.而实际上两种不确定性是普遍存在于不确定性数据流环境中的,不确定性数据流的数据分析处理过程中只考察其中某一种数据不确定性显然不具有说服力,不完备,难以得到可信的数据分析结果.因此,本文在研究不确定性数据流聚类算法之前先提出存在级及属性级不确定性统一的不确定性数据流模型.

定义 2. 设给定不确定性数据流的时间序列:$S=\{(\bar{\boldsymbol{X}}_1,p_1),(\bar{\boldsymbol{X}}_2,p_2),\cdots,(\bar{\boldsymbol{X}}_n,p_n)\}$,由定义 1 易知,$p_i(i=1,2,\cdots,n)$描述元组 $\bar{\boldsymbol{X}}_i$ 的存在级不确定

性;对于属性级不确定性,元组具有相应的概率密度函数 pdf,但实际应用中,由于概率密度函数很难获得,一般采用连续模型的一种离散形式进行采集简化,即每个属性级不确定性元组 $\bar{\boldsymbol{X}}_i=((x_i^{(1)},k_{i1}),(x_i^{(2)},k_{i2}),\cdots(x_i^{(d)},k_{id}))$,其中,$d=1,2,\cdots,\bar{\boldsymbol{X}}_i$ 表示元组 $\bar{\boldsymbol{X}}_i$ 的 d 维度的可能值,k_{id} 则描述取这一可能值的可能性,由此构成不确定性数据流的可能世界模型:

$$\begin{aligned}S=\{&((x_1^{(1)},k_{11}),(x_1^{(2)},k_{12}),\cdots,(x_1^{(d)},k_{1d}),p_1),\\&((x_i^{(1)},k_{21}),(x_i^{(2)},k_{22}),\cdots,(x_2^{(d)},k_{2d}),p_2),\cdots,\\&((x_n^{(1)},k_{n1}),(x_n^{(2)},k_{n2}),\cdots,(x_n^{(d)},k_{nd}),p_n)\}.\end{aligned}\tag{2}$$

对该可能世界模型首先采取属性级不确定性的概率密度函数拟合,得到其对应的连续 pdf 模型,并进行归一化处理,再将其各个数据所对应的存在级不确定性,即元组的存在概率与得到的 pdf 模型相整合,用简单的方法实现不确定性数据流数据元组级不确定性与属性级不确定性的统一.按照前文所述思路,本文提出的存在级与属性级不确定性统一下的不确定性数据流模型为

$$\begin{aligned}S=\{&((x_1^{(1)},p_1k_{11}),(x_1^{(2)},p_1k_{12}),\cdots,(x_1^{(d)},p_1k_{1d})),\\&((x_i^{(1)},p_2k_{21}),(x_i^{(2)},p_2k_{22}),\cdots,(x_2^{(d)},p_2k_{2d})),\cdots,\\&((x_n^{(1)},p_nk_{n1}),(x_n^{(2)},p_nk_{n2}),\cdots,(x_n^{(d)},p_nk_{nd}))\}.\end{aligned}\tag{3}$$

令 $P_{ij}=p_ik_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,d)$,则式(3)变形为

$$\begin{aligned}S=\{&((x_1^{(1)},P_{11}),(x_1^{(2)},P_{12}),\cdots,(x_1^{(d)},P_{1d})),\\&((x_i^{(1)},P_{21}),(x_i^{(2)},P_{22}),\cdots,(x_2^{(d)},P_{2d})),\cdots,\\&((x_n^{(1)},P_{n1}),(x_n^{(2)},P_{n2}),\cdots,(x_n^{(d)},P_{nd}))\}.\end{aligned}\tag{4}$$

在该不确定性数据流的可能世界模型下,参与数据处理阶段的可能世界实例的规模远大于不确定性数据库的规模,甚至是其指数倍,在相关的工作中必须研究不确定性的量化机制来描述数据的不确定性,即进行不确定性度量.本文引入在空间数据不确定性研究方面已取得不错效果的熵指标进行不确定性度量.熵指标下存在级与属性级不确定性统一的数据流模型被视为多维随机变量.对于多维随机变量,概率密度函数为 $p(x_i^1,x_i^2,\cdots,x_i^d)$,则其联合熵 $H(X)$定义为

$$H(X)=-\int_R\cdots\int_R p(x_i^1,x_i^2,\cdots,x_i^d)\cdot\ln p(x_i^1,x_i^2,\cdots,x_i^d)\mathrm{d}x_i^1\mathrm{d}x_i^2\cdots\mathrm{d}x_i^d.\tag{5}$$

通常情况下,实际应用中采集数据大多呈现正态分布.对于 d 维正态随机变量 X,其概率密度为

$$p(X)=\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\mathrm{e}^{\left|-\frac{1}{2}(X-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(X-\boldsymbol{\mu})\right|},\tag{6}$$

其中,$\boldsymbol{\mu}$ 为 d 维均值向量,$\boldsymbol{\Sigma}$ 是 $d\times d$ 维协方差矩阵,$|\boldsymbol{\Sigma}|$ 为 $\boldsymbol{\Sigma}$ 的行列式值.若令 $(X-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(X-\boldsymbol{\mu})=k^2$,则式(6)为一个正定二次型,从数理角度分析具有明显的几何意义,即中心为 $(\mu_1,\mu_2,\cdots,\mu_d)$ 超椭球体,其大小是观测向量对均值向量的离散度度量[16].

根据上述分析与推导,在 $\mathbb{R}^n$ 空间中服从正态分布的 n 维随机变量 $\bar{\boldsymbol{X}}_i=(x_i^1,x_i^2,\cdots,x_i^d)$,概率密度函数为式(6),则熵不确定度为

$$H(X)=\ln\{2\pi\mathrm{e}^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\}.\tag{7}$$

该熵不确定度作为具有不确定性的数据流样本的不确定性度量模型,量化不确定性参与后续的自适应性聚类算法.

3　不确定性数据流自适应网格密度聚类算法

针对目前对不确定性数据流聚类问题的研究不多,且多基于距离考察相似度,具有先天缺陷(详见引言部分)的现状,本文将擅长进行时态空间数据聚类处理的网格密度聚类算法引入不确定性数据流聚类分析工作中.在大多数的数据流数据聚类分析算法中采用滑动窗口技术,可以有效解决数据的时序性问题以及过期数据处理问题,但对于本文的聚类分析对象的不确定性特征,单纯的滑动窗口技术不能很好地处理过期数据聚类贡献程度不一致的问题,因此需进一步研究适合的衰减窗口模型;同时,网格密度算法自身存在的一个弊端是密度阈值的设定,人工方法设定的方式在复杂的不确定性数据流环境中显然难以实行,复杂的数据流环境更容易出现数据空间分布不均匀的状态,在这个数据聚类过程中采用统一的密度阈值进行分析,很可能不能反映出数据流真实的聚类情况.因此,在引入网格密度聚类算法进行不确定性数据流聚类分析时,需要研究适合的针对数据不确定性、时态性和数据分布不均匀性质的自适应密度阈值调整方法,这些都是本文提出的 ADC-UStream 算法中重点解决的难题.

3.1　ADC-UStream 算法涉及的基本概念

定义 3. 假定给定的数据集空间 D 有 n 个 d 维属性:$S=\{(\bar{\boldsymbol{X}}_1,p_1),(\bar{\boldsymbol{X}}_2,p_2),\cdots,(\bar{\boldsymbol{X}}_n,p_n)\}$,其中,$\bar{\boldsymbol{X}}_i$ 为 d 维向量 $\bar{\boldsymbol{X}}_i=(x_i^1,x_i^2,\cdots,x_i^d)$,并且每个属性的属性值是有界的.为了不失一般性,设任意一个属性

区间是一个半开半闭的区间$[m_j,n_j)$,$j=1,2,\cdots,d$,则 d 维空间数据可表示为

$$SpaceX=[m_1,n_1)\cdot[m_2,n_2)\cdot\cdots\cdot[m_d,n_d). \quad (8)$$

这种划分方法为标准的 k-正规划分.该方法将 d 维数据集空间 D 用相同的尺度划分为 $K(K=kd)$ 个网格,每一个属性的范围为 $\delta_i=(n_i-m_i)/k$,属性 x_i^j 便是所有满足式(9)左闭右开区间的集合:

$$I_{i,j}=[m_i+(j-1)\delta_i,m_i+j\delta_i),\ j=1,2,\cdots,k. \quad (9)$$

每个网格均有一个独立的索引,第$(i_1,i_2,\cdots,i_d)$个网格,也就是 $I_{1,i_1}\cdot I_{2,i_2}\cdot\cdots\cdot I_{d,i_d}$.既然每一维的数据都有相同的划分尺度,可以假定每个网格中样本点的个数可以代表数据点的密度.这样,通过 k-正规划分将数据集空间 D 划分为相同尺度的网格(grid).如果一个数据点 $\bar{\boldsymbol{X}}_i$ 落入某个单元网格 G,则称点 $\bar{\boldsymbol{X}}_i$ 属于 G,记作 $\bar{\boldsymbol{X}}_i\in G$.

定义 4. 密度阈值判断当前考察网格的类型的临界值 $MinPts$.网格中数据点数目 Num 若大于等于 $MinPts$,则当前网格为稠密网格;若网格中数据点数目 Num 小于 $MinPts$,则当前网格为稀疏网格.一般的网格密度算法中会采用人为设定的方式,本文的 ADC-UStream 算法中将采取自适应调整的方式,以适应研究对象不确定性数据流的时序性、不确定性和非均匀分布性,详见 3.3 节.

定义 5. 相邻网格一般分为两类:1)与当前考察网格 G 有公共边的网格;2)与当前考察网格 G 有公共顶点的网格.

定义 6. 若当前考察网格 G 是稠密网格,且是某稠密网格 G'的相邻网格,则称 G 直接密度可达 G'.若给定一系列空间相邻网格 $G_1,G_2,\cdots,G_n$ 且均为稠密网格,则有 G_i 直接密度可达于 $G_{i+1}(i=1,2,\cdots,n-1)$.若 $G=G_1,G'=G_n$,则空间网格 G 密度可达 G'.

定义 7. 若空间网格 G 与 G'同时密度可达空间网格 G^o,则 G 密度相连 G'.由该定义可知,密度相连是对称的.

定义 8. 将数据集空间 D 中所有密度相连的空间网格聚集成多个子空间,每一个非空子空间中的所有数据点形成一个聚类簇 C.簇 C 由其中某个稠密网格及与该稠密网格密度可达的所有网格中的数据点构成.

3.2 聚类衰减窗口模型

由于数据流的无限性,不可能将所有的数据全部存储后再进行计算和处理,需要采用一定的窗口模型体现对不同时间点到达数据的重视程度(即聚类贡献度).衰减窗口由于能够体现陈旧数据对聚类结果的一定影响.而不同于滑动窗口将陈旧数据一律抛弃的做法,能够较好地模拟对数据流的关注特点.然而,对于不确定性数据流,除考虑时态因素引起的聚类贡献衰减外,不确定性因素(特别是存在级不确定性)也要加以考虑.同样是陈旧数据,不确定性小的数据对聚类结果的正面贡献度理论上讲要大于不确定性大的数据,应对其赋予较大的贡献度权重.对于不确定性数据流数据,其聚类贡献程度表现出如下特征:1)数据点离当前时刻越近其聚类贡献程度越大;2)数据点的数据质量越好,即不确定度越小,其聚类贡献程度越大;3)数据点的聚类贡献程度既与自身有关,又与簇中其他数据点(特别是新到数据点)有关;4)随着数据点的到达,数据分布可能呈现不均匀趋势,即所形成簇的密度不同也会随时间变化.由以上分析,同时借鉴杨宁等人[5]对于不考虑不确定性因素进行的数据流聚类算法中的衰减窗口设计方案,本文将 ADC-UStream 算法中的聚类衰减窗口定义如下:

定义 9. 设 $X=(\bar{\boldsymbol{X}},P_X,t_X)$是不确定性数据流中的一个数据点,其中 t_X 为 X 的到达时间,$P_X(P_X\in[0,\delta],0<\delta\leqslant1)$为其不确定度指标,$P_X$ 的确定方式详见 2.2 节的熵不确定度模型,以此实现对于参与聚类的数据流样本点的不确定性的度量,在聚类过程中考虑并重视不确定性这一普遍特性,获得更为准确的聚类结果.由此,X 在时刻 $t(t>t_X)$的权值为函数 $UTW(X,t)=2^{-\lambda(t-t_X)}P_X,\lambda\geqslant1$.其中,$\lambda$ 为衰减系数,代表衰减过程的速度.

由定义 9 给出的权值直接反映了数据点的聚类贡献程度,可以看出数据点的权值与其到达时间距当前时间的间隔有关,间隔越长(即数据点越陈旧),其权值越小,聚类贡献程度越小;同时,权值又与其不确定度指标有关,不确定度越大数据质量越差,其聚类贡献程度越小.

命题 1. 新到数据点的聚类贡献程度由其不确定度指标决定,即对于新到数据点 X,其权值 $UTW(X,t)=P_X$.

证明. 由定义 9 可知,$UTW(X,t)=2^{-\lambda(t-t_X)}P_X$,对于新到样本点 $X=(\bar{\boldsymbol{X}},P_X,t_X)$,

因为:$t_X=t$,

所以:$UTW(X,t)=2^{-\lambda(t-t_X)}P_X=2^{-\lambda(t-t)}P_X=2^0P_X=P_X$. 证毕.

定义 10. 数据空间的单元网格 G 中所有数据点在时刻 t 的权值之和 $Den(G,t)=\sum_{X\in G} UTW(X,t)$,为网格 G 在时刻 t 的密度.

命题 2. 给定时刻 t_0 与 t_1,$t_0<t_1$,且 t_0 为时刻 t_1 之前最近的一个考察时刻,时刻 t_1 到达单元网格 G 中的数据点数目为 ΔNum,则 G 在时刻 t_1 的密度为 G 在时刻 t_0 网格密度衰减后的值加上新增数据点的不确定度指标之和,即

$$Den(G,t_1)=2^{-\lambda(t_1-t_0)}Den(G,t_0)+\sum_{i\in\Delta Num}P_{X_i}.$$

证明. 设 Num_0 为时刻 t_0 网格 G 中的数据点数目,由定义 10 可知,网格 G 在时刻 t_0 的密度为 $Den(G,t_0)=\sum_{X\in G}UTW(X,t_0)$,即有

$$Den(G,t_0)=\sum_{i=1}^{Num_0}UTW(X_i,t_0).$$

对于 $\forall X_i, X_i\in G, 0\leqslant i\leqslant Num_0$,数据点 X_i 在时刻 t_1 的权值为

$$UTW(X_i,t_1)=2^{-\lambda(t_1-t_s)}P_i=$$
$$2^{-\lambda(t_1-t_0)}2^{-\lambda(t_0-t_s)}P_i=2^{-\lambda(t_1-t_0)}UTW(X_i,t_0),$$

其中,$t_s(t_s\leqslant t_0<t_1)$为数据点 X_i 的到达时间. 对于 $\forall X_j, X_j\in G, 0\leqslant j\leqslant\Delta Num$,数据点 X_j 为时刻 t_1 的新到样本点,由命题 1 可知,X_j 在时刻 t_1 的权值为 $UTW(X_j,t_1)=P_{X_j}$. 可得:

$$Den(G,t_1)=\sum_{i=1}^{Num_0+\Delta Num}UTW(X_i,t_1)=$$
$$\sum_{i=1}^{Num_0}UTW(X_i,t_1)+\sum_{j=1}^{\Delta Num}UTW(X_j,t_1)=$$
$$\sum_{i=1}^{Num_0}2^{-\lambda(t_1-t_0)}UTW(X_i,t_0)+\sum_{j=1}^{\Delta Num}P_{X_j}=$$
$$2^{-\lambda(t_1-t_0)}\sum_{i=1}^{Num_0}UTW(X_i,t_0)+\sum_{j=1}^{\Delta Num}P_{X_j}=$$
$$2^{-\lambda(t_1-t_0)}Den(G,t_0)+\sum_{j=1}^{\Delta Num}P_{X_j}.$$

证毕.

显然,命题 2 提供了一种对于网格 G 的网格密度的增量式计算方式,即当前时刻 G 的网格密度为其之前某时刻的网格密度随时间过程衰减与当前时刻所到数据点的不确定度指标之和. 以此方式实现网格 G 的网格密度的增量计算,避免在每一个考察时刻内对 G 中所有数据点权值的重复计算,考虑到权值计算以及网格密度计算在基于网格密度聚类算法中的基础地位和核心地位,其计算量的大小会直接影响到整体算法的执行效率,网格密度的增量式计算有助于算法执行效率的提高.

3.3　密度阈值自适应调整策略

为摆脱基于网格密度聚类算法人为参与因素的限制,保证聚类结果的准确性和可靠性,需要研究密度阈值自适应调整的方法. 对于不确定性数据流聚类过程中,除了考虑时态因素和不确定性因素,数据样本点还极有可能出现空间分布不均匀的情况. 因此,本文提出的 ADC-UStream 算法中采用时空自适应密度阈值的主动调节方法,根据时间推移和空间分布情况自动设定密度阈值,减少基于密度聚类算法的人工干预程度,保证聚类结果的准确性.

命题 3. 设 S 为起始时刻 t_0 到当前时刻 t_c 的时间间隔内到达数据集空间 D 的数据集,则 D 中所有网格单元的密度之和满足:

$$\lim_{t\to\infty}\sum_{X\in S}UTW(X,t)=\frac{1}{2-2^{1-\lambda}}. \tag{10}$$

证明. 由前述的定义和命题可得:

$$\lim_{t\to\infty}\sum_{X\in S}UTW(X,t)=\lim_{t\to\infty}\sum_{X\in S}2^{-\lambda(t-t_s)}P_{X_i}=$$
$$\lim_{t\to\infty}(2^{-\lambda(t-0)}+2^{-\lambda(t-1)}+\cdots+2^{-\lambda(t-1)}+2^{-\lambda(t-t)})\bar{P}=$$
$$\lim_{t\to\infty}\sum_{\tau=0}^{\tau=t}2^{-\lambda\tau}\bar{p}=\lim_{t\to\infty}\frac{1-2^{-\lambda\tau}}{1-2^{-\lambda}}\bar{P}=\frac{1}{2-2^{1-\lambda}}.$$

证毕.

由命题 3 可知,数据集空间 D 中所有数据点的权值之和不会超过 $\frac{1}{2-2^{1-\lambda}}$,而对于 D 通过 k-正规划分(见定义 3)可划分为 K 个网格单元 G,可知 D 中网格 G 的平均密度不会超过 $\frac{1}{K(2-2^{1-\lambda})}$.

定义 11. 设与当前考察时刻 t 最邻近的前一次考察时刻为 t_u,则网格单元 G 在当前时刻 t 的密度阈值 $MinPts$ 为

$$MinPts(t_u,t)=\frac{Den(G,t_u)\sum_{\tau=0}^{t-t_u}2^{-\lambda\tau}}{K(1-2^{-\lambda})}=$$
$$\frac{Den(G,t_u)(1-2^{-\lambda(t-t_u+1)})}{K(1-2^{-\lambda})}. \tag{11}$$

式(11)中定义的密度阈值函数可以满足不确定性数据流空间和时间的变化要求,能够进行自适应调整,通过衰减速度参数 λ 反映密度随时间衰减变化的因素,通过当前考察时刻的前一次考察时刻 t_u 的网格 G 的密度值 $Den(G,t_u)$反映数据的不确定性特征对密度阈值的影响,对于非均匀分布状态下的数据点能够进行自适应调整密度阈值,有效避免传统的基于网格密度聚类算法中数据集全局采用统一的密度阈值造成的聚类结果不准确和偏差问题.

3.4 ADC-UStream 算法

ADC-UStream 算法是针对于不确定性数据流这一数据对象,利用基于网格密度的聚类算法的思想,在一定的网格划分下,采用聚类衰减窗口的方式计算数据点的权值,并以加权计和方式计算网格单元的密度,根据时空自适应调整的密度阈值,对数据集空间的网格进行类型判断,之后进行密度相连,最终形成聚类结果.

ADC-UStream 算法关键过程描述如下:

过程 1. 一般聚类过程.

```
Initial Clustering ()
delay the main thread Span second;
/* 对数据流样本点延迟 Span 时间 */
divide the current data space D into several d-
dimension hypercubes as grids;
judge the attributive grid of data sample and
computer its uncertain tense weight;
reject the empty grids;
/* 网格单元初始化 */
for (i=1,i≤Nonempty,i++)
/* Nonempty 为当前数据空间中非空网格的
   数目 */
{
compute the uncertain tense density Den;
  if (Den≥MinPts)
/* MinPts 由参数 t=t_c 自适应计算 */
     stamp the grid with dense-grid;
     else
     stamp the grid with spare-grid;
      find the neighbor grid NeiG of every
      dense-grid;  /* 网格密度相连 */
  initial cluster ID=0;
     case NeiG is a dense-grid
{find the core dense-grid to begin a new
cluster; stamp all the neighbor dense-grid
with a new cluster ID; cluster ID ++;}
/* 确定稠密网格 */
  case NeiG is a spare-grid
  {performance the boundary point process;}
find the neighbor grid NeiG of every spare-grid;
  while (NeiG is a spare-grid)
    {performance the noise point process;}
    /* 确定稀疏网格 */
  }
```

过程 2. 增量式聚类过程.

```
for new arriving data sample X_i^(j)
  update T=current time t;
    if X_i^(j) ∈ g_i
    X_i^(j) is already belonging to any grid
    then determine the attributive grid of it;
    calculate the uncertain tense weight of X_i^(j);
     update uncertain tense density of the
     grid;  /* 数据点归于已有类簇,进行密
     度增量式计算 */
        else
      X_i^(j) is not belongs to any current grid;
        create a new grid for it;
        /* 数据点归于新簇 */
    endif
endfor
```

4 实验及结果分析

4.1 实验平台及数据集

本文 ADC-UStream 算法的实验平台为 Intel Core2 双核 CPU,主频 2.0 GHz,内存 2 GB 的 PC 机,Windows XP Professional 操作系统,文中算法由 C++实现.

为了能够与目前为数不多的同类算法进行实验数据比较,实验数据集采用同类算法中使用的真实数据集 UCI Forest Covertype,该数据集中包含 581 012 个实例,具有 54 个属性列,为了单纯验证本文算法的聚类性能,抛开非数值属性处理的问题,仅选取其中的 10 个数值型属性列.该数据集为确定性数据集,为适应本文算法的不确定性数据流数据对象特征,首先按固定时间间隔加入时间维属性值,模拟数据在一定时序下到达的流式特征,然后需将其转化成相应的不确定性数据集.具体做法为数据集中的每个实例项都随机增加一个介于(0,1]之间的概率代表存在级不确定性;保持元组的期望值 X 不变,令其方差为 $(1-p)X$(p 为(0,1]的随机值)表示属性级不确定性.综合上述两种方式构造出具有存在级和属性级不确定性的数据集.

4.2 实验结果及分析

实验分为两部分:1)在同时具有存在级和属性级不确定性的数据集下,验证本文 ADC-UStream 的有效性,并考察其衰减速度参数 λ 值调整对算法结果和效率的影响;2)分别与 UMicro 算法和 EMicro

算法进行比较,考察 ADC-UStream 算法仅处理存在级不确定性与仅处理属性级不确定性的聚类能力和算法性能.除特殊说明外,根据多次实验结果取网格 k 正规划分的参数 $k=20$,以数据流流速 $v=10$ KBps 输入真实数据模拟数据流过程.

实验 1. 衰减速度参数设置对聚类质量的影响.

该实验主要是通过改变算法涉及到的重要参数衰减速度 λ 的值,考察其对聚类结果质量的影响.将 ADC-UStream 算法作用于数据集,经过多次实验,以 λ 分别取值 1,2,3 时的聚类结果进行质量情况分析.对于聚类结果质量评价,目前未见针对于不确定性数据流的聚类评价标准,本文对实验结果的分析采用确定性数据流聚类算法的评价方法,鉴于平均距离和 SSQ 方式已不作为优秀的评价方法,本文采用聚类精度 CP 评价法.具体的聚类结果质量情况如图 1 所示:

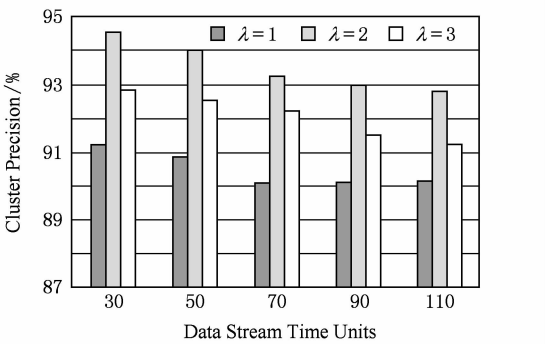


Fig. 1　Influence of attenuation rate λ on clustering quality.

图 1　衰减速度 λ 对聚类质量的影响

从实验结果可以看出,λ 取值对聚类结果影响可以接受,整体的聚类结果的精度保持在 91%～94%,并未因 λ 值的调整出现较大浮动.其中,$\lambda=2$ 时聚类精度略高.证明了针对一定的数据流流速,在某一个 λ 值下会收到较好的聚类效果.主要是因为:1)λ 取值过小,会使得旧数据点的聚类贡献度衰减过慢,造成新到数据点的聚类贡献影响不足;2)λ 取值过大,对旧数据点的聚类贡献度衰减过快,其对当前聚类结果的影响不够,过分强调新到数据点的聚类贡献.由此看出,衰减速度参数 λ 的调整会对不确定性数据流的聚类结果造成一定程度的影响,但总体浮动在可接受的范围.为全面考察算法的聚类性能,本文修改了 ADC-UStream 算法中对于不确定性的处理环节,将其修改为针对确定性数据流的自适应网格密度算法,在同等条件下对未作不确定性转化的该数据集 Forest Covertype 进行聚类,也取得了较高的聚类精度(篇幅所限不作详解),从另外的角度验证了 ADC-UStream 算法中所基于的聚类算法主体-自适应网格密度聚类算法对数据流聚类的有效性,但结果中的聚类精度要比上述结果略高,主要是因为 Forest Covertype 数据集本身是确定性的数据,人为进行的不确定性转化本身就引入了不准确因素.

实验 2. 与 UMicro 算法和 EMicro 算法的性能比较.

目前已有的同类算法中未见类似的能够同时针对存在级及属性级不确定性进行统一描述处理的算法,相对权威的 UMicro 算法[11]仅考虑属性级不确定性,EMicro 算法[13]仅考虑存在级不确定性,这也为实验中的同类算法比较造成了一定困难.在实验中,将本文提出的 ADC-UStream 算法分别与 EMicro 算法和 UMicro 算法进行性能比较.在多次实验进行两种算法的参数调整后,选择其中表现均较好的结果,3 种算法比较的实验结果如图 2 和图 3 所示:

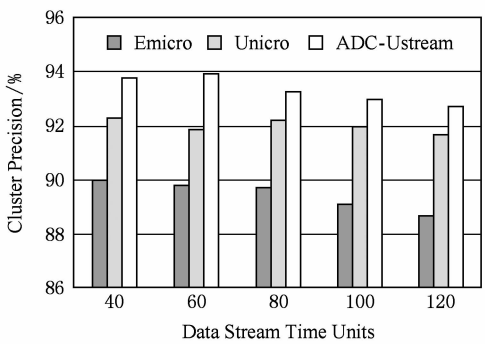


Fig. 2　Cluster precision of different uncertain data stream clustering algorithms.

图 2　不确定性数据流聚类算法聚类精度比较

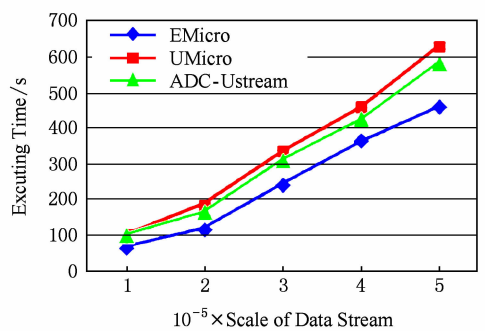


Fig. 3　Performance efficiency of different uncertain data stream clustering algorithms.

图 3　不确定性数据流聚类算法执行效率比较

实验结果显示,与 EMicro 算法相比,ADC-UStream 有着更高的聚类精度,但算法执行效率略逊一筹,分析其原因是由于 EMicro 算法仅处理存在

级不确定性,相比于综合处理不确定性特征的 ADC-UStream 算法,其聚类的准确度显然会较低,并且 EMicro 算法基于 K-Means 聚类算法,用距离进行相似度度量,仅能形成球形簇,且参数 k 的设定会引入不准确因素,ADC-UStream 算法基于网格密度聚类算法,参数依赖小,且能够发现任意形状的簇,因此在聚类结果的质量上,ADC-UStream 算法表现更好.但同时,ADC-UStream 算法需要对存在级和属性级不确定性统一并进行不确定度建模,耗费了一定的计算时间,并且 EMicro 算法基于的 K-Means 算法在执行效率上优于 ADC-UStream 算法基于的网格密度聚类方法,因此 EMicro 算法的执行效率明显高于本文的 ADC-UStream 算法.但鉴于目前的计算机硬件的支持和计算处理速度的提高,牺牲少量时间换取针对不确定性特征下的数据流这一特殊数据对象的聚类精度是值得的.

可以看出,与 UMicro 算法相比,ADC-UStream 的聚类结果精度更高,算法执行效率二者相差不多,随着数据流的向前推进,ADC-UStream 算法的执行时间略少于 UMicro 算法.主要是因为 UMicro 算法仅针对属性级不确定性进行处理,并且也同样是基于 K-Means 算法进行聚类,ADC-UStream 算法具有更高的聚类精度.UMicro 算法对于属性级不确定性的处理的计算复杂度也与数据集本身的维度有关,这一点与 ADC-UStream 算法中不确定度的计算方式是类似的,使得两种算法在执行效率上差别不大,但 UMicro 算法中对过期数据的处理需要一定的计算时间,随着数据流数据的不断到达,本文的 ADC-UStream 算法逐渐略显优势.实验结果表明 ADC-UStream 算法对不确定性数据流的聚类分析是有效的,能够形成任意形状的簇,参数依赖小,能够获得相对准确度聚类结果,算法的执行效率也是可接受的.

5 总结与展望

本文针对不确定性数据流提出了一种具有时空自适应能力的基于网格密度的聚类算法 ADC-UStream,该算法对于不确定性数据流数据的存在级不确定性和属性级不确定性进行了统一,并构建了基于超椭球体指标的信息熵不确定度模型,相比于同类算法中片面针对存在级和属性级某一种不确定性进行考察的做法,聚类结果更具说服力;ADC-UStream 算法在聚类衰减窗口模型下计算数据点的权值及网格的密度,符合不确定性数据流数据的时态性和不确定性的特征要求;同时,传统网格密度聚类算法中人为设定的密度阈值影响聚类准确性,是该类算法的瓶颈,鉴于本文聚类对象不确定性数据流的复杂程度,设计时态和空间自适应的密度阈值函数,使其能够满足时态性、不确定性以及数据空间非均匀分布的特性.实验结果表明,ADC-UStream 算法相对于目前为数不多的不确定性数据流聚类算法,能够产生较为准确的聚类结果,其聚类效率也是可接受的.在后续的研究工作中,将着力研究算法中的重要参数的调整策略、不确定性数据流聚类质量评价及其可视化表示的方法和技术、不确定性数据流聚类动态演化分析等.

参 考 文 献

[1] Zhou Aoying, Jin Cheqing, Wang Guoren, et al. A survey on the management of uncertain data [J]. Chinese Journal of Computers, 2009, 32(1): 1-16 (in Chinese)
(周傲英, 金澈清, 王国仁, 等. 不确定性数据管理技术研究综述[J]. 计算机学报, 2009, 32(1): 1-16)

[2] Shawn R J, Minos G, Michael F. Adaptive cleaning for RFID data streams [C] //Proc of the 32nd Int Conf on Very Large Data Bases. San Fransisco, CA: Morgan Kaufmann, 2006: 163-174

[3] Aggarwal C C, Han J W, Yu P S, et al. A framework for clustering evolving data streams [C] //Proc of the 29th Int Conf on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann, 2003: 81-92

[4] Chang Jianlong, Cao Feng, Zhou Aoying. Clustering evolving data streams over sliding windows [J]. Journal of Software, 2007, 18(4): 905-918 (in Chinese)
(常建龙, 曹锋, 周傲英. 基于滑动窗口的进化数据流聚类算法[J]. 软件学报, 2007, 18(4): 905-918)

[5] Yang Ning, Tang Changjie, Wang Yue, et al. Clustering algorithm on data stream with skew distribution based on temporal density [J]. Journal of Software, 2010, 21(5): 1031-1041 (in Chinese)
(杨宁, 唐常杰, 王悦, 等. 一种基于时态密度的倾斜分布数据流聚类算法[J]. 软件学报, 2010, 21(5): 1031-1041)

[6] Kriegel H P, Pfeifle M. Hierarchical density-based clustering of uncertain data [C] //Proc of the 5th Int Conf on Data Mining. Los Alamitos, CA: IEEE Computer Society, 2005: 689-692

[7] Cormode G, McGregor A. Approximation algorithms for clustering uncertain data [C] //Proc of the 27th ACM SIGMOD-SIGACT-SIGART Symp on Principles of Database Systems. New York: ACM, 2008: 191-200

[8] Chen Jianmei, Lu Hu, Song Yuqing, et al. A possibility fuzzy clustering algorithm based on the uncertainty membership [J]. Journal of Computer Research and Development, 2008, 45(9): 1486-1492 (in Chinese)
(陈健美,陆虎,宋余庆,等. 一种隶属关系不确定的可能性模糊聚类方法[J]. 计算机研究与发展, 2008, 45(9): 1486-1492)

[9] Thanh T L T, Peng L P, Diao Y L, et al. CLARO: Modeling and processing uncertain data streams [J]. The VLDB Journal, 2012, 21(5): 651-676

[10] Aggarwal C C, Yu P S. A framework for clustering uncertain data streams [C] //Proc of the 24th Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2008: 150-159

[11] Aggarwal C C. On high dimension projected clustering of uncertain data streams [C] //Proc of the 25th Int Conf on Data Engineering. Los Alamitos, CA: IEEE Computer Society, 2009: 1152-1154

[12] Zhang C, Gao M, Zhou A Y. Tracking high quality clusters over uncertain data streams [C] //Proc of the 1st Workshop on Management and Mining of Uncertain Data. Los Alamitos, CA: IEEE Computer Society, 2009: 1641-1648

[13] Zhang Chen, Jin Cheqing, Zhou Aoying. Clustering algorithm over uncertain data streams [J]. Journal of Software, 2010, 21(9): 2173-2182 (in Chinese)
(张晨,金澈清,周傲英. 一种不确定数据流聚类算法[J]. 软件学报, 2010, 21(9): 2173-2182)

[14] Aggarwal C C, Zhao Y C, Philip S Y. On clustering graph streams [C] //Proc of the 10th SIAM Int Conf on Data Mining. New York: ACM, 2010: 478-489

[15] Chen L, Wang C L. Continuous sugraph pattern search over certain and uncertain graph streams [J]. IEEE Trans on Knowledge and Data Engineering, 2010, 22(8): 1093-1109

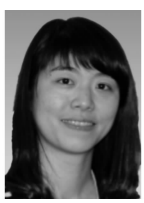
[16] Li Dajun, Cheng Penggen, Gong Jianya, et al. Entropy uncertainty of multi-dimensional random variable [J]. Acta Metrologica Sinica, 2006, 27(3): 290-293 (in Chinese)
(李大军,程朋根,龚健雅,等. 多维随机变量的熵不确定度[J]. 计量学报, 2006, 27(3): 290-293)

Liu Zhuo, born in 1979. PhD candidate. His main research interests include data fusion, information analysis and data stream mining.

Yang Yue, born in 1980. Received her PhD degree in computer application in 2008. Her main research interests include data mining, uncertainty analysis and clustering processing.

Zhang Jianpei, born in 1956. Professor and PhD supervisor. His main research interests include database and knowledge base, data mining, social network and social computing.

Yang Jing, born in 1962. Professor and PhD supervisor. Her main research interests include data stream analysis and mining, database management and data privacy protection.



Chu Yan, born in 1979. Received her PhD degree in computer application in 2008. Her main research interests include data mining, mobile data management and wireless sensor network.

Zhang Zebao, born in 1978. Received his PhD degree in computer application in 2010. His main research interests database management, spatial database and spatial data index.