山东农业大学学报(自然科学版),2017,48(4):521-524 VOL.48 NO.4 2017
Journal of Shandong Agricultural University (Natural Science Edition) doi:10.3969/j.issn.1000-2324.2017.04.008

基于非参数估计与随机模拟的不确定数据流相似性度量方法

迟荣华,黄少滨,李熔盛

哈尔滨工程大学 计算机科学与技术学院, 黑龙江 哈尔滨 150001

摘 要: 针对不确定数据流对象难于度量相似性的问题，本文提出一种非参数估计与随机模拟相结合的方法。本方法利用非参数估计对不确定数据流对象建模，然后利用随机模拟计算对象间的误差相似性，通过相对距离与绝对距离判断相似度。仿真实验验证了本方法不仅可以准确地度量不确定对象间的相似性，而且在对象规模较大的情况下，依然可以获得较快速和稳定的计算结果。

关键词: 不确定数据流; 非参数估计; 随机模拟; 相似性

中图法分类号: TP274+.2 **文献标识码**: A **文章编号**: 1000-2324(2017)04-0521-04

An Uncertain Data Stream Similarity Measurement Method Based on Nonparametric Estimation and Stochastic Simulation

CHI Rong-hua, HUANG Shao-bin, LI Rong-sheng

College of Computer Science and Technology/Harbin Engineering University, Harbin 150001, *China*

Abstract. To solve the problem that the current uncertain data stream is difficult to measure the similarity, this paper proposes a method combining non-parametric estimation with stochastic simulation. The method used the non-parametric estimation to model the uncertain data stream objects, and then used stochastic simulation to calculate the error similarity between objects, judged the similarity by relative distance and absolute distance. Simulation experiment verified this method can not only measure the similarity between the uncertain objects accurately, but also can obtain fast and stable results when the object scale is large.

Keywords: Uncertain data stream; non-parametric estimation; stochastic simulation; similarity

随着信息收集、存储等技术与手段的不断发展，信息的规模以及属性都变的极为庞大，这不仅使信息科学的研究工作变得更加困难、复杂，也同时带来了更多方面的挑战和乐趣。例如，传感器网络、物联网、RFID 网络、对象识别以及移动对象搜索等[1-5]。

为了对不确定性数据进行有效管理与分析，相关研究大多先对不确定性数据构建模型以保留数据原始的自身特性。常见的数据模型如可能世界模型[6]，用于对数据存在的不确定性建模；以及概率密度函数，用于描述数据中属性值的不确定性[7]。基于这些不确定性数据模型，进一步对不确定性数据进行管理分析，如不确定性数据的存储与查询，以及挖掘不确定性数据中蕴涵知识的数据挖掘算法等[8-11]。目前，不确定对象建模方法的理论基础一般是基于不确定对象的数据服从某种理论分布[12,13]，典型的密度分布函数代表了不确定对象的主要特征，这类方法虽然没有忽略对象的不确定性，但不确定信息却没有得到很好的保留，而且这种简单的数据分布假设，相对实际的数据来说都有较大的偏差，导致理论模型与实际模型的误差较大。

针对上述对于不确定对象建模与相似性度量存在的问题，本文提出了一种基于非参数估计与随机模拟相结合的方法。该方法首先采用非参数的概率密度估计方法对不确定数据流对象进行建模，然后利用蒙特卡洛随机模拟方法对概率密度函数的相对距离与绝对距离进行计算，通过两个指标的分析确定不确定对象的相似性。文本在第二节论述了基于非参数估计的不确定数据流对象模型的建模以及相似性度量方法，并从理论上给出了方法的有效性证明，然后在第三节通过仿真实验的方法验证了本方法的有效性与可靠性。

1 基于非参数估计的不确定数据流对象建模以及距离度量

1.1 不确定数据流对象建模

非参数估计方法是在不假设数据分布的前提下，基于样本点数据对不确定对象构建概率密度函

收稿日期: 2015-01-05 **修回日期**: 2015-03-06

作者简介: 迟荣华(1981-),男,博士研究生.主要研究方向为不确定数据分析及人工智能. E-mail:chironghua@hrbeu.edu.cn

数字优先出版:2017-06-13 http://www.cnki.net

数的较合适的方法。其中核密度估计是一种基于样本数据特征，用于估计未知密度函数的非参数统计方法，不需要预先假设变量间的函数关系，其估计形式如式（1）所示。

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

其中 K 为一个核函数，常见的形式如高斯核函数、三角核函数、均匀核函数等。核密度估计方法求得的概率密度函数能够在样本点与目标数据量足够大时，收敛到任意一种密度函数，因此在数据分布未知的情况下，为了获取每个不确定对象的分布特征，本文利用非参数估计方法获取不确定性对象的是分布特征。

假设有不确定数据流 U，包含 n 个有序的不确定对象，任一个不确定对象 u_i 含有 l 个样本点，所以有不确定数据流形如 $U = \{u_{1,l},...,u_{i,l}\}, 0 \le i \le n$。结合上述非参数估计的概率密度估计方法，数据流 U 中的不确定观测对象u_i的概率密度可表示为$\hat{f}_{u_i}(x) = \frac{1}{nh}\sum_{j=1}^{l} K\left(\frac{x-u_{i,j}}{h}\right)$，该公式表示一个数据流序列中的某一个不确定观测值的概率密度估计函数，这个不确定对象包含了多少个样本，密度函数就有多少个核函数累积计算。其中 h 是核函数的带宽，一般根据样本的误差均方确定。

1.2　相似性度量

现假设两个不确定数据流对象 X,Y，其中U_x代表不确定数据流对象 X，$U_x = \{u_{1,l},...,u_{i,l}\}, 0 \le i \le N$，$u_i$是所有由于不确定性产生或观测到的分量集合，大小为 l，因此当存在不确定性现象时$|u_i| > 1$，即数据流元素的取值不唯一。在计算不确定数据流的一般方法中，假设不确定数据流 X,Y 的距离定义为 $dists(X,Y) = \{L(x,y) | x \in U_x, y \in U_y\}$，$L$ 一般采用欧式距离的计算方法，那么 X,Y 满足某特定距离阈值的概率为：$PR\,(distance(X,Y)) \le \varepsilon) = \frac{|\{d \in dists(X,Y) | d \le \varepsilon\}|}{|dists(X,Y)|}$　　(2)

其中，d 是 X 与 Y 之间的任意元素之间的距离，根据假设可推得，这种距离计算方式的计算量为任意元素之间的距离，即是在固定长度的模式匹配问题中计算总量也可达$(Nl)^2$。从中不难看出，这种一般方法能够合理的解释不确定的相似性结果，是一种合理的相似性度量指标，但当元素中样本量较大或有繁琐的遍历需求时，该方法的时效性则较低。

本文借鉴了这种方法的概率解释的思想，并结合非参数估计的方法，将不确定的相似性问题转化为概率密度的相似性问题。当概率密度的取值区域不同，但密度中心相似时，也能有较好的匹配结果。假设有两个不确定数据流对象 $U = \{u_{i,l}\}$和 $V = \{v_{i,l}\}$，可根据样本获得各自的概率密度函数 $F_U = \{\hat{f}_{u_i}\}$和$F_V = \{\hat{f}_{v_i}\}$，$0 \le i \le N$，因此任意两个不确定数据流中的元素的相似性问题可表示为：

$$distance(X,Y) = \frac{1}{N}\sum (\left| \int_{I_{u_i} \cup I_{v_i}} \left(\hat{f}_{u_i}(x) - \hat{f}_{v_i}(x)\right)dx \le \varepsilon \cap \int_{I_{u_i} \cup I_{v_i}} \left|\left(\hat{f}_{u_i}(x) - \hat{f}_{v_i}(x)\right)\right|dx \le \delta \right|)$$

在公式中，$I_{u_i} \cup I_{v_i}$表示元素u_i与v_i的最大取值空间，$\int_{I_{u_i} \cup I_{v_i}} \left(\hat{f}_{u_i}(x) - \hat{f}_{v_i}(x)\right)dx$表示概率密度的误差的积分，然而在$I_{u_i} \cup I_{v_i}$的区域中，一般意义上$\int_{I_{u_i} \cup I_{v_i}} \left(\hat{f}_{u_i}(x)\right)dx = 1$、$\int_{I_{u_i} \cup I_{v_i}} \left(\hat{f}_{v_i}(x)\right)dx = 1$，所以为了约简计算，公式可转化为：

$$distance(X,Y) = \frac{1}{N}\sum (\left| \int_{I_{u_i} \cap I_{v_i}} \left(\hat{f}_{u_i}(x) - \hat{f}_{v_i}(x)\right)dx \le \varepsilon \cap \int_{I_{u_i} \cap I_{v_i}} \left|\left(\hat{f}_{u_i}(x) - \hat{f}_{v_i}(x)\right)\right|dx \le \delta \right|)$$

通过公式可以看出，相似性的计算包含三个主要因素，取值空间、相对误差以及绝对误差。$I_{u_i} \cap I_{v_i}$代表了取值空间，当交集为Φ时表示没有相似性，$\int_{I_{u_i} \cap I_{v_i}} \left(\hat{f}_{u_i}(x) - \hat{f}_{v_i}(x)\right)dx \le \varepsilon$表示的相对误差，当$\hat{f}_{u_i}(x)$与$\hat{f}_{v_i}(x)$取值空间完全相同时，相对误差近似为 0。$\int_{I_{u_i} \cap I_{v_i}} \left|\left(\hat{f}_{u_i}(x) - \hat{f}_{v_i}(x)\right)\right|dx \le \delta|)$表示绝对误差，描述了分布密度几何结构上的不同，两种误差可以互相弥补不足。取值空间决定了在允许精度下的取值空间，很大程度上决定计算的规模和效率。因此首先确定取值空间的计算方法。针对一个不确定数据流中的元素，假设取值空间 $I = [a,b]$，误差ε，因此有：

$$1 - \left[\frac{1}{nh}\sum \int_{-\infty}^{a} K(\cdot) - \frac{1}{nh}\sum \int_{-\infty}^{b} K(\cdot)\right] = \varepsilon \tag{3}$$

含义是非参数密度函数置信区间。一般情况下，本文假设密度函数的边缘是对称的，因此可有$\frac{\varepsilon}{2} =$

$\frac{1}{nh}\sum\int_{-\infty}^{a}K(\cdot)$，$\frac{\varepsilon}{2}=1-\frac{1}{nh}\sum\int_{-\infty}^{b}K(\cdot)$。由于，$x_1\le x_2\le\ldots\le x_n$，可根据样本最小值与最大值计算取值区域的近似值，令$\frac{\varepsilon}{2}=\frac{1}{nh}\sum\int_{-\infty}^{a}K\left(\frac{x-x_1}{h}\right)$，$\varepsilon=\frac{2}{h}\int_{-\infty}^{a}K\left(\frac{x-x_1}{h}\right)$以及$\frac{\varepsilon}{2}=1-\frac{1}{nh}\sum\int_{-\infty}^{b}K\left(\frac{x-x_n}{h}\right)$，$\varepsilon=2-\frac{2}{h}\int_{-\infty}^{b}K\left(\frac{x-x_n}{h}\right)$

根据复合函数的求导方法，有$\varepsilon=\frac{2}{\sqrt{2\pi}h\sigma}\int_{-\infty}^{a}e^{-\frac{1}{2\sigma}\left(\frac{x-x_1}{h}\right)^2}dx$

令，$Z=\frac{x-x_1}{h}$，则$\varepsilon=\frac{2}{\sqrt{2\pi}h\sigma}\int_{-\infty}^{\frac{a-x_1}{h}}e^{-\frac{Z^2}{2\sigma}}hdx$，可得$\varepsilon=2\phi\left(\frac{a-x_1}{h\sigma}\right)$

此时，可通过查表得到 a，同理可得 b。且，$\varepsilon_a=\frac{2}{nh}\sum\int_{-\infty}^{a}K(\cdot)$，$\varepsilon_b=2-\frac{2}{nh}\sum\int_{-\infty}^{b}K(\cdot)$

可推得$\varepsilon_a,\varepsilon_b\le\varepsilon$，因此可根据少量样本快速推算估计密度函数的取值区间 I。假设不确定对象的元素 u 与 v，其真实的密度函数与估计密度函数分别为f_u，f_v与$\widehat{f_u}$，$\widehat{f_v}$，在误差ε的条件可得 U 与 V 的置信区间分别为I_u和I_v。现可得相对误差与绝对误差分别为：

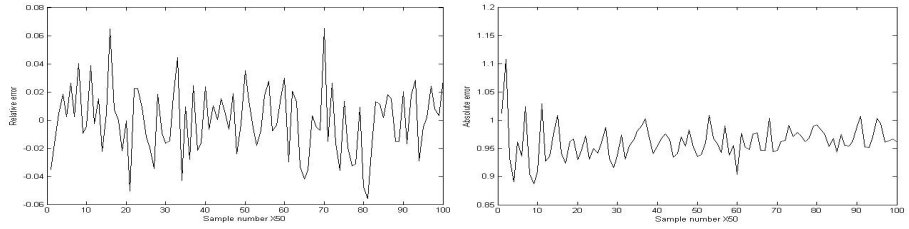
$$\int_{I_u\cap I_v}(f_u-f_v)dx \tag{4}$$

$$\int_{I_u\cap I_v}|(f_u-f_v)|dx \tag{5}$$

由于$f_u,f_v>0$，且$\int f_u=1$，$\int f_v=1$，则可知 $0\le\int_{I_u\cap I_v}(f_u-f_v)dx\le 1$

而$\int_{I_u\cap I_v}|(f_u-f_v)|dx\le\int_{I_u\cap I_v}|(f_u+f_v)|dx\le 2$，则可知 $0\le\int_{I_u\cap I_v}|(f_u-f_v)|dx\le 2$

根据公式可知，当两个不确定对象的取值区域完全不同时，相对误差和绝对误差分别为 1 和 2，而当取值区域完全相同时，相对误差为 0，绝对误差可取[0,2]的任何值。因此，可以想象即使两个对象在取值空间完全相同时，仍可能由于分布形式的不同而产生很高的误差。



(a) 相对误差 Relative error　　(b) 绝对误差 Absolute error

图 1 不同样本规模下的相对误差（a）与绝对误差（b）比较

Fig.1 Comparison between relative error (a) and absolute error (b) at different sample size

为了分析不确定对象的概率密度的相对误差与绝对误差，本文模拟了两个相同取值空间，但分布特征完全随机的，并且不同样本数的不确定对象的相对误差与绝对误差变化，从图 1 可以看出，在随机模拟的方法下，相对误差基本稳定在 99%的概率密度上，但绝对误差可体现出不同程度的变化。现在可以知道，不确定对象可以通过概率密度的相对误差与绝对误差进行准确描述，但如何进行这种计算同样是一个复杂的问题。

2　仿真实验分析

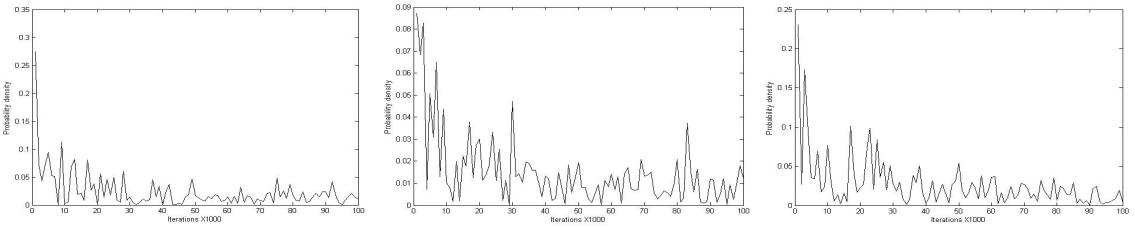
为了从实际应用的角度验证本文所提算法的可靠性，本文采用仿真的方法进行实验。首先，本文分别生成三组仿真数据，对应分布分别为正态分布、均匀分布和指数分布，每组 50 个数值，数值的取值空间为[0,1]，正态分布的均值为 0 方差为 1，指数分布的均值为 0.5。根据算法流程，分别对三组不确定对象进行建模。然而，核概率密度函数的带宽选择是非参数估计方法的一个核心问题，带宽选择的合适与否决定了构造密度函数的精度。本文选择当前较为普遍采用的基于均方误差的计算方法，假设不确定对象的真实概率密度函数 $f(x)$与估计的概率密度函数$\widehat{f(x)}$间的均方误差：

$$\mathrm{MSE}(x;h)=E\{\widehat{f(x)}-f(x)\}^2 \tag{6}$$

可知$h_{opt}=1.06\times\sigma\times n^{-\frac{1}{5}}$（该结果的前提假设是真实概率密度函数接近高斯）。

首先，为了分析在不确定对象属于相同区域而分布密度不同时的现象，本文模拟了属于两个不同分布类型的不确定对象分属于不同级别的样本数量时，相对误差与绝对误差的变化情况。如图 2 可知，当两个对象在相同区域内的概率密度几乎完全相同时，由于分布类型的不同，绝对误差较高

仍可导致两个对象具有较大的差别。然后，通过不同的迭代次数，观察分析三组分布时误差的距离误差的变化情况如图 2。



（a） 均匀分布 Uniform distribution　（b） 正态分布 Normal distribution　（c） 指数分布 Exponential distribution

图 2 三种典型分布下的模拟效果示意图

Fig.2 Simulation performances of three typical distributions

通过仿真实验可知均匀分布的误差最高，指数分布的误差次之，正态分布的误差最小，考虑原因可能与核函数的假设有关，但当迭代次数趋近 40000 次时，误差率普遍可以控制在 0.5%以下，实验效果明显，而且当进行多组实验时，结果依然稳定。

3 总 结

本文面向数据属性存在的不确定性，针对不确定对象建模以及不确定对象间相似性度量中面临的主要问题，提出基于非参估计及随机模拟相结合的相似性计算方法。本方法利用非参数估计对不确定对象建模，可以有效描述因为属性以及存在所导致的不确定性问题，同时又采用随机模拟方法解决了非参数估计函数复杂难于计算相似性的问题。最后，本文通过仿真实验的方式对所提方法进行了验证，实验结果表明本方法不仅能够有效地对建模，同时又能在有限次计算数量下较高精度地计算不确定对象间的相似度，理论和实验都验证了本文所提方法的有效性与准确性。

参考文献

[1] Yang Z, Liu Y. Quality of trilateration: Confidence-based iterative localization[J]. IEEE Transactions on Parallel and Distributed Systems, 2010,21(5):631-640

[2] Mokbel MF, Chow CY, Aref WG. The new casper: Query processing for location services without compromising privacy[J]. Proceedings of the 32nd international conference on very large data bases, 2006,34(4): 763-774

[3] Jeffery SR, Franklin MJ, Garofalakis M. An adaptive RFID middle ware for supporting metaphysical data independence[J]. The International Journal on Very Large Data Bases, 2008,17(2):265-289

[4] Bohm C, Pryakhin A, Schubert M. The gauss-tree: Efficient object identification in databases of probabilistic feature vectors[C]. Proceedings of the 22nd International Conference on IEEE, 2006:9

[5] Chen L, Özsu MT, Oria V. Robust and fast similarity search for moving object trajectories[C]. USA: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, ACM, 2005:491-502

[6] Muzammal M, Raman R. Mining sequential patterns from probabilistic databases[C]. Springer Heidelberg Berlin: Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2011:210-221

[7] Soliman MA, Ilyas IF, Chang KCC. Top-k query processing in uncertain databases[C]. IEEE 23rd International Conference on Data Engineering, 2007:896-905

[8] Barbará D, Garcia-Molina H, Porter D. The management of probabilistic data[J]. IEEE Transactions on knowledge and data engineering, 1992,4(5):487-502

[9] Antova L, Jansen T, Koch C, *et al*. Fast and simple relational processing of uncertain data[C]. IEEE 24th International Conference on Data Engineering, 2008:983-992.

[10] Tang R, Cheng R, Wu H, *et al*. A framework for conditioning uncertain relational data[C]. Springer Heidelberg Berlin: International Conference on Database and Expert Systems Applications, 2012:71-87

[11] Taskar B, Segal E, Koller D. Probabilistic classification and clustering in relational data[C]. Lawrence Erlbaum Associates Ltd.: International Joint Conference on Artificial Intelligence, 2001,17(1):870-878

[12] Kriegel HP, Kunath P, Pfeifle M, *et al*. Probabilistic similarity join on uncertain data[C]. Springer Heidelberg Berlin: International Conference on Database Systems for Advanced Applications, 2006:295-309

[13] Agarwal PK, Aronov B, Har-Peled S, *et al*. Nearest-neighbor searching under uncertainty II[J]. Symposium on Principles of Database Systems, 2013,13(1):115-126