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　工　大学　　学与　学　　哈尔

摘 要: 对不 定　对　于　似　，　出一　参 估 与　合　。利　参 估 对不 定　对　，后利　对　，似 ，　对　与 对判　似 。仿 实　了　不仅可以准 地　不 定对　似 ，且在对　大　况下，依 可以　和 定　。

关键词: 不 定　参 估　似

中图法分类号:　　**文献标识码:**　　**文章编号:**

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

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

Abstract

Keywords

信　、存储　与　不 发展，信　以及属　变　为 大，不仅使信　学　工作变　加困 、复 ，也同　了 多　和乐 。例如，传 器　、　、　、对　别以及 动对　。

为了对不 定　与分 ，关　大多先对不 定　型以保原始　。　型如可 世　型 ，于对　存在 不 定　；以及密 函 ，于　中属 值 不 定　。基于 些不 定　型，一 对不 定　分 ，如不 定　存储与　，以及　不 定　中　。前，不 定对　基 一　基于不 定对　从　分　，典型 密 分 函 代 了不 定对　主　，　对　不 定，但不 定信 却　到 好 保 ，且　单　分 假 ，对实大 偏 ，导　型与实　型　大。

对上对于不 定对　与 似 存在　，　出了一 基于 参 估 与　合　。先　参　密 估　对不 定　对　，后利　卡　对　密 函　对　与 对　，两个分　定不 定对　似 。在 二　了基于 参 估 不 定　对　型以及 似　，从 上 出了　，后在 三　仿 实　了　与可　。

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

参 估　在不假　分　前 下，基于　对不 定对　密 函

收稿日期:　　**修回日期:**

作者简介:　华　博士　主　向为不 定　分 及人工

合 。其中 密 估 一 基于 ， 于估 密 函 参 ，不 先假 变 函 关 ，其估 如 （ ） 。

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

其中 K 为一个 函 ， 如 函 、三 函 、均匀 函 。 密 估 密 函 够在 与 够大 ， 到任 一 密 函 ，因 在 分 况下，为了 取 个不 定对 分 ， 利 参 估 取不 定 对 分 。

假 不 定 U，包含 n 个 不 定对 ，任一个不 定对 u_i 含 l 个 ， 以 不 定 如 $U = \{u_{1,l},...,u_{i,l}\}, 0 \le i \le n$。 合上 参 估 密 估 ，

U 中 不 定 对 u_i 密 可 为 $\hat{f}_{u_i}(x) = \frac{1}{nh}\sum_{j=1}^{l} K\left(\frac{x-u_{i,j}}{h}\right)$， 公 一个 列中 一个不 定 值 密 估 函 ， 个不 定对 包含了多少个 ，密 函 就 多少个 函 。其中 h 函 宽，一 均 定。

1.2 相似性度量

假 两个不 定 对 X Y，其中 U_x 代 不 定 对 $X, U_x = \{u_{1,l},...,u_{i,l}\}, 0 \le i \le N$，$u_i$ 于不 定 产 到 分 合，大小为 l，因 存在不 定 $|u_i| > 1$，即 元 取值不唯一。在 不 定 一 中，假 不 定 X,Y 定义为 $dists(X,Y) = \{L(x,y) | x \in U_x, y \in U_y\}$，$L$ 一 ， 么 X,Y 定 值 为：$PR\left(distance(X,Y)\right) \le \varepsilon) = \frac{|\{d \in dists(X,Y) | d \le \varepsilon\}|}{|dists(X,Y)|}$

其中，d X 与 Y 之 任 元 之 ， 假 可 ， 为任 元 之 ，即 在固定 匹 中 也可 Nl 。从中不 出， 一 够合 不 定 似 ， 一 合 似 ，但 元 中 大 历 ， 则 低。

借 了 ， 合 参 估 ，将不 定 似 化为 密 似 。 密 取值区域不同，但密 中 似 ，也 好 匹 。假 两个不 定 对 $U = \{u_{i,l}\}$ 和 $V = \{v_{i,l}\}$，可 各 密 函 $F_U = \{\hat{f}_{u_i}\}$ 和 $F_V = \{\hat{f}_{v_i}\}$，$0 \le i \le N$，因 任 两个不 定 中 元 似 可 为：

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

在公 中，$I_{u_i} \cup I_{v_i}$ 元 u_i 与 v_i 大取值 ，$\int_{I_{u_i} \cup I_{v_i}} \left(\hat{f}_{u_i}(x) - \hat{f}_{v_i}(x)\right)dx$ 密 分， 在 $I_{u_i} \cup I_{v_i}$ 区域中，一 义上 $\int_{I_{u_i} \cup I_{v_i}} \left(\hat{f}_{u_i}(x)\right)dx = 1$、$\int_{I_{u_i} \cup I_{v_i}} \left(\hat{f}_{v_i}(x)\right)dx = 1$，为了 ，公 可 化为：

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

公 可以 出， 似 包含三个主 因 ，取值 、 对 以及 对 。$I_{u_i} \cap I_{v_i}$ 代 了取值 ， 交 为 Φ 似 ，$\int_{I_{u_i} \cap I_{v_i}} \left(\hat{f}_{u_i}(x) - \hat{f}_{v_i}(x)\right)dx \le \varepsilon$ 对 ， $\hat{f}_{u_i}(x)$ 与 $\hat{f}_{v_i}(x)$ 取值 完全 同 ， 对 似为 。$\int_{I_{u_i} \cap I_{v_i}} |\left(\hat{f}_{u_i}(x) - \hat{f}_{v_i}(x)\right)|dx \le \delta |)$ 对 ， 了分 密 几何 上 不同，两 可以互 不 。取值 决定了在允 取值 ， 大 上决定 和 。因 先 定取值 。 对 一个不 定 中 元 ，假 取值 $I = [a,b]$，ε，因 ：

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

含义 参 密 函 信区 。一 况下， 假 密 函 对 ，因 可 $\frac{\varepsilon}{2} =$

$\frac{1}{nh}\sum\int_{-\infty}^{a}K(\cdot)$，$\frac{\varepsilon}{2}=1-\frac{1}{nh}\sum\int_{-\infty}^{b}K(\cdot)$。 于，$x_1\le x_2\le\ldots\le x_n$，可 小值与 大值 取值区域 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$\frac{\varepsilon}{2}=1-\frac{1}{nh}\sum\int_{-\infty}^{b}K\left(\frac{x-x_n}{h}\right)$，$\varepsilon=2-\frac{2}{h}\int_{-\infty}^{b}K\left(\frac{x-x_n}{h}\right)$

复合函 导 ，$\varepsilon=\frac{2}{\sqrt{2\pi}h\sigma}\int_{-\infty}^{a}e^{-\frac{1}{2\sigma}\left(\frac{x-x_1}{h}\right)^2}dx$

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 $\varepsilon=2\phi$（$\frac{a-x_1}{h\sigma}$

，可 到a，同 可 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 $\varepsilon_a,\varepsilon_b\le\varepsilon$，因 可 少 估 密 函 取值区 I。假 不 定对 元 u与v，其 实 密 函 与估 密 函 分别为f_u，f_v与$\widehat{f_u}$，$\widehat{f_v}$，在 ε 件可 U与V 信区 分别为I_u和I_v。 可 对 与 对 分别为：

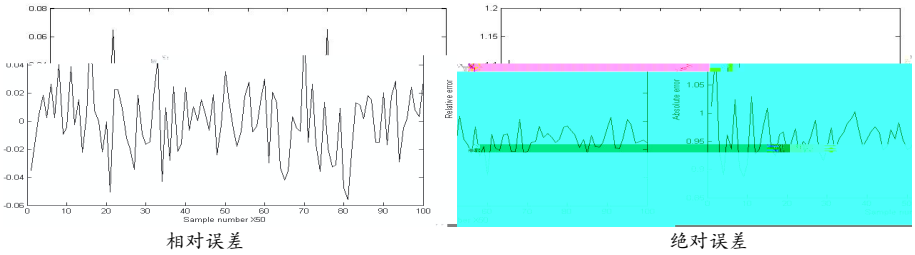
$$\int_{I_u\cap I_v}(f_u-\widehat{f_v})dx$$

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

于$f_u,f_v>0$，且$\int f_u=1$，$\int f_v=1$，则可 $0\le\int_{I_u\cap I_v}(f_u-f_v)dx\le1$

$\int_{I_u\cap I_v}|(f_u-f_v)|dx\le\int_{I_u\cap I_v}|(f_u+f_v)|dx\le2$，则可 $0\le\int_{I_u\cap I_v}|(f_u-f_v)|dx\le2$

公 可 ， 两个不 定对 取值区域完全不同 ， 对 和 对 分别为 和 ，取值区域完全 同 ， 对 为 ， 对 可取 任何值。因 ，可以 即使两个对在取值 完全 同 ，仍可 于分 不同 产 。



相对误差 绝对误差

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 不 定对 密 对 与 对 ， 了两个 同取值 ，但分 完全 ， 且不同 不 定对 对 与 对 变化，从图 可以 出，在 下， 对 基 定在 密 ，但 对 可体 出不同 变化。 在可以 ，不 定对 可以 密 对 与 对 准 ，但如何 同 一个复 。

2 仿真实验分析

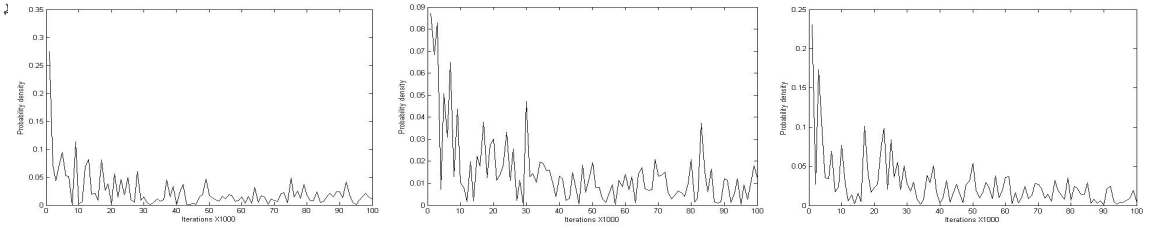
为了从实 可 ， 仿 实 。 先，分别 三 仿 ，对 分 分别为 分 、均匀分 和 分 ， 个 值， 值取值 为 ， 分 均值为 为 ， 分 均值为 。 ，分别对三 不 定对 。 ， 密 函 宽 参 估 一个 ，宽 合 与否决定了 密 函 。 前 为 基于均 ，假 不 定对 实 密 函 $f(x)$与估 密 函 $\widehat{f(x)}$ 均 ：

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

可 $h_{opt}=1.06\times\sigma\times n^{-\frac{1}{5}}$（ 前 假 实 密 函 ）。

先，为了分 在不 定对 属于 同区域 分 密 不同 ， 了属于两个不同分 型 不 定对 分属于不同 别 ， 对 与 对 变化 况。如图 可 ， 两个对 在 同区域内 密 几乎完全 同 ，于分 型不同， 对

仍可导 两个对 具 大 别。 后， 不同 代 ， 察分 三 分
变化 况如图 。



（ ） 均匀分布 （ ） 正态分布 （ ） 指数分布

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

仿 实 可 均匀分 ， 分 之， 分 小， 原因可 与 函 假 关，但 ， 可以 制在 以下，实 ， 且 多 实 ， 依 定。

3 总 结

向 属 存在 不 定 ， 对不 定对 以及不 定对 似 中 临主 ， 出基于 参估 及 合 似 。 利 参 估 对不定对 ，可以 因为属 以及存在 导 不 定 ，同 又 决了 参 估 函 复 于 似 。 后， 仿 实 对 了 ，实 不仅 够 地对 ，同 又 在 下 地 不 定对 似 ， 和实 了 与准 。

参考文献

et al

et al

et al

et al