一种新颖的基于密度的祛噪声方法

王 扬 [1]

摘 要 由于采集技术和设备的局限, 以及外界的各种干扰, 采集得到的数据中常常参杂着噪声, 直接影响着后续数据分析的结果. 传统的祛噪声方法, 或是依赖于数据服从某一特定分布的假设, 或是只能对服从单一分布的数据进行祛噪声处理, 这些固有的缺陷大大降低了处理后数据的可信度. 本文提出了一种新颖的基于密度的祛噪声方法, 并应用在实际交通数据的处理中. 通过与传统方法的实验比较, 结果表明该方法摒除了传统方法的缺陷, 能够对服从多个相异分布的数据进行有效的祛噪声处理, 且处理后的数据能够很好地保留系统本质的特征.

关键词 祛噪声, 密度估算, 期望值, 噪声识别, 行驶速度

DOI 10.3724/SP.J.1004.2010.00343

A Novel Algorithm for Outlier Removal Based on Density

WANG Yang[1]

Abstract Due to the limitation of the present techniques and facilities for data collection and various interferences, the data obtained are often distorted and noised, directly influencing the result of subsequent data analysis. The conventional approaches to outlier removal either assume that the data follow a certain known distribution or deal with the data that are from a single distribution, resulting in a reduced credibility of the data processed. This paper proposes a novel method to remove outliers based on density estimation and it has been applied to real-world traffic data. By comparison with the conventional approach, the experimental results indicate that the proposed algorithm is capable of detecting and removing outliers effectively for the data that may follow different unknown distributions, and the processed data retain the original and significant characteristics possessed by the system.

Key words Outlier removal, density estimation, expectation, outlier detection, travel speed

在数据采集过程中, 由于采集技术与设备的局限以及各种外界干扰, 所采集到的数据往往包含许多不真实的数据. 这些统称为噪声的数据, 很大程度上掩盖了所要研究系统的真实本质, 使测量得到的期望值远远偏离真实期望值. 为了祛除噪声, 以便更好地揭示系统的本质特征, 各种祛除噪声的方法孕育而生. 目前使用比较普遍的两种祛噪声方法是: 1) 假设数据是服从单一高斯分布, 先对其分布的期望值进行估算, 然后把那些偏离期望值超过一定程度的数据 (通常偏离期望值超过标准偏差的 2～3 倍的数据) 归为噪声并把这些噪声剔除[1−2]. 因此, 该方法属于参数统计法. 由于该方法是建立在数据服从高斯分布这个假设的基础之上, 因此, 对于服从其他分布的数据, 该方法不能准确有效地识别出噪声; 2) 依据四分位数法, 将从小到大排列而成的数据分为四等份, 那些偏离上四分位数和下四分位数超过一定程度 (通常偏离上四分位数和下四分位数 1.5 倍的上下四分位数范围) 的数

收稿日期 2008-12-29 录用日期 2009-09-17
Manuscript received December 29, 2008; accepted September 17, 2009
1. 北京工业大学交通工程北京市重点实验室 北京 100124
1. Beijing Key Laboratory of Transportation Engineering, Beijing University of Technology, Beijing 100124

据被认定为噪声[3−4]. 该方法属于非参数方法. 上述第一种方法, 由于噪声的影响导致期望值与标准偏差的计算结果不同程度地偏离真实值. 另外, 以上两种方法均局限于单个分布的情形. 对于服从多个相同或不同分布的数据, 以上两种方法的祛噪效果明显不佳, 甚至有时把那些能够体现实质的数据也当作噪声过滤掉了. 近几年, 有些研究者[5−6] 通过聚类来识别并祛除噪声, 但是聚类本身就是一个复杂耗时的过程, 而且聚类结果的好坏直接影响着噪声的识别. 本文所提出的基于密度的祛噪声方法已经在作者参加并完成的科研课题 "避免拥堵的动态导航系统 (Congestion avoidance dynamic routing engine, CADRE)" 中得到实验验证, 并与传统基于四分位数的方法进行了仿真比较. 实验结果表明本文所提出的祛噪声方法不依赖于数据服从某一特定分布的假设, 且能够从服从多个相同或不同分布的数据中有效地祛除噪声, 并同时较好地保留了系统的本质特征.

1 基于密度的祛噪声方法

1.1 方法描述

该方法首先对数据密度进行估算, 然后依据所得的密度进行噪声识别并作祛噪处理.

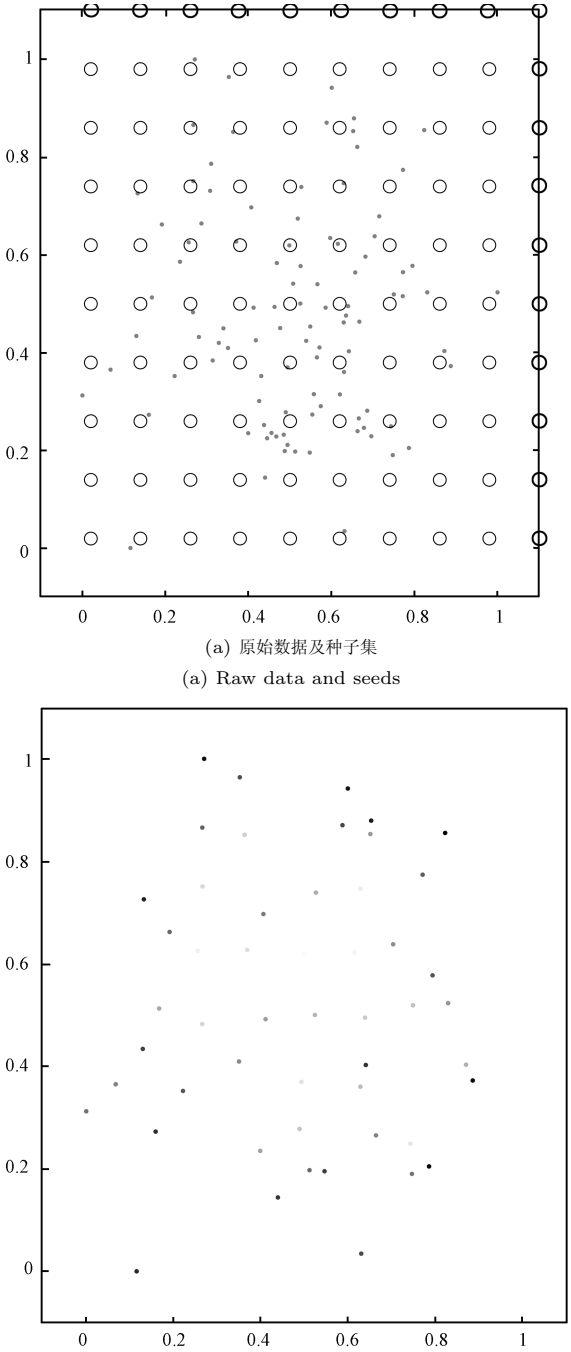
假设采集得到一个总数为 N 的二维数据集 Z (如图 1 (a) 中的圆点所示), 并产生一个称为种子群的数据集 S (如图 1 (a) 中的圆圈所示), 种子群 S 所含的种子个数 M 需事先指定, 并保证各个种子点与其相邻种子之间的距离恒等, 此外还应保证种子群的范围能够包含采集所得的数据. 每个数据点 $\boldsymbol{z}_k$ ($k \in \{1, 2, \cdots, N\}$) 均附有一个初始值为 0 的种子吸附计数器 c_k, 该种子吸附计数器用来累计该数据点可以吸附种子的数目, 计算吸附种子数是通过计算数据点与种子之间的距离实现的, 具体计算方法如下:

对于每个种子 $\boldsymbol{s}_j$ ($j \in \{1, 2, \cdots, M\}$) 分别计算它与数据集 Z 的各个数据点之间的距离. 本文中的事例与实验均采用欧式 (Euclidean) 距离[7] 计算, 但该方法同样适用于其他距离度量, 例如: 曼哈顿 (Manhattan) 距离[8]、汉明 (Hamming) 距离[9]、编辑 (Levenshtein) 距离[10] 等.

$$i = \arg\min\left(\left\|\boldsymbol{z}_k - \boldsymbol{s}_j\right\|^2\right), \quad k \in \{1, 2, \cdots, N\} \tag{1}$$

依据式 (1) 求得距离该种子 $\boldsymbol{s}_j$ 最近的数据点, 并将该数据点 $\boldsymbol{z}_i$ 所附带的种子吸附计数器 c_i 增 1. 如果存在多个数据点与该种子 $\boldsymbol{s}_j$ 距离相等且为最近, 则等比份的分配给每个数据点, 即距离最近的每个数据点的种子吸附计数器累积 $1/p$ (p 为距离该种子点最近的数据点的个数). 以此类推, 对于每个种子点均按上述方法计算距离, 并按上述原则更新相应数据点种子吸附计数器的值. 因此, 该方法的计算复杂度即为 $\mathrm{O}(M \times N)$.

每个数据点附带一个种子吸附计数器, 用来累计该数据点所吸附种子的数目, 如果某个数据点的种子吸附计数器值高, 则表明该数据点吸附的种子多, 也就是说该数据点的邻域没有很多的数据点与其竞争分享这些种子, 因此表明该数据点密度低. 反之, 若一个数据点的邻域存在许多数据点, 那么就意味着该数据点与其周围的数据点在吸附种子时存在着较为激烈的竞争, 因此每个数据点所吸附获得的种子数目必定减少. 以图 1 (b) 为例, 颜色较浅的点代表密度较大的数据点, 而颜色较深的那些点则表示它们的密度相对较低. 具有较低密度的数据点表明测量结果在其邻域出现的概率较小, 即该数据点表征系统本质的置信度较低, 因此可把种子吸附计数器值高于某个设定值 (种子吸附阈值) 的数据点归为噪声并祛除.



(a) 原始数据及种子集

(a) Raw data and seeds

(b) 噪声识别之后的数据

(b) Data after outlier detection

图 1 噪声识别事例

Fig. 1 An example of outlier detection

1.2 参数确定

该算法有两个参数需要确定: 一是种子数目; 二是种子吸附阈值.

对于种子数目的确定, 这里提出了一种较为简便的启发式方法. 先按照式 (2) 计算每一个数据点与其他数据点之间

最短距离; 再利用式 (3) 求出所有数据点与其他数据点间最短距离值的均值, 并将其作为种子点与其邻近种子之间的距离;

$$d_i = \min(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2), \quad j \neq i, j \in \{1, 2, \cdots, N\} \tag{2}$$

$$\bar{d} = E(d_i) = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{3}$$

在确定种子范围时, 首先确定采集所得数据的范围 (对每一维即为 $z_{\max}$, $z_{\min}$); 其次, 为了确保种子的范围能够包括所有的数据点, 种子每一维的上下边界应满足:

$$s_{\max} - z_{\max} > \bar{d}, \quad z_{\min} - s_{\min} > \bar{d} \tag{4}$$

在确定种子之间距离和种子范围之后, 便可直接计算出种子的数目. 当然也可由其他方法确定种子数目. 例如, 在计算出每个数据点与其他数据点之间最短距离后, 可以取其中最小距离值作为种子之间的距离, 但是由此而得的种子数目较大, 从而增加了后续计算处理的负荷并延长了计算时间, 导致实时性变差, 因此不适于实时性要求较高的场合.

对于种子吸附阈值, 可依据所得全部种子吸附值的总体分布来确定. 本文中采用的方法是: 先把种子吸附值按从大到小进行排列, 并平均分为 5 等份, 取第一与第二等份的分界吸附值作为种子吸附阈值. 在实际应用中, 视具体情况灵活调整参数确定方法, 以便取得最佳效果.

参见图 2, 该祛噪声方法的流程可归纳为如下:

步骤 1. 确定参数;

步骤 2. 生成一个恒等间距的种子群;

步骤 3. 初始化数据点的种子吸附计数器;

步骤 4. 计算当前种子与所有数据点间的距离, 找到距离该种子最近的数据点, 并更新该数据点的种子吸附计数器的值;

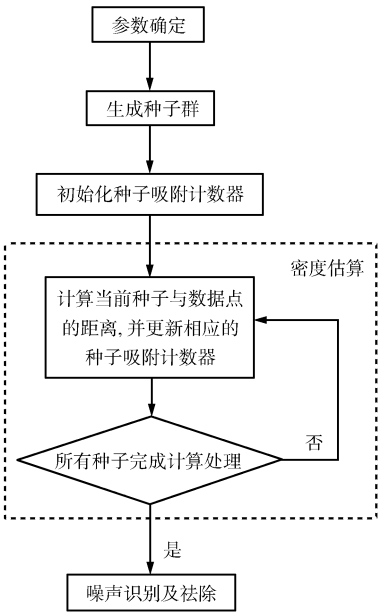


图 2　基于密度估算的祛噪声方法流程图

Fig. 2　The flow chart of the outlier removal algorithm based on density estimation

步骤 5. 重复步骤 4, 直至完成对所有种子的计算处理;

步骤 6. 依据给定密度阈值识别并祛除噪声.

2　实验结果

本文所提出的基于密度的祛噪声方法已在具体项目中得到实验验证, 并且与传统基于四分位数的方法, 在 Matlab 环境下进行了仿真比较. 该项目是 CADRE, 由英格兰东南发展委员会 (South East England Development Agency) 设立, 前期总计投资 80 万英镑, 旨在研发智能车载导航系统. 作者在该项目中主要承担数据分析处理以及模糊预测模型建模的工作. 如图 3 所示, 本文所述实验中的原始数据是由英国公路局 (British Highways Agency) 统计而得, 该数据包含从 2006 年 12 月 4 日到 2006 年 12 月 10 日一周内英格兰汉普郡 (Hampshire) 某段高速公路的行驶速度.

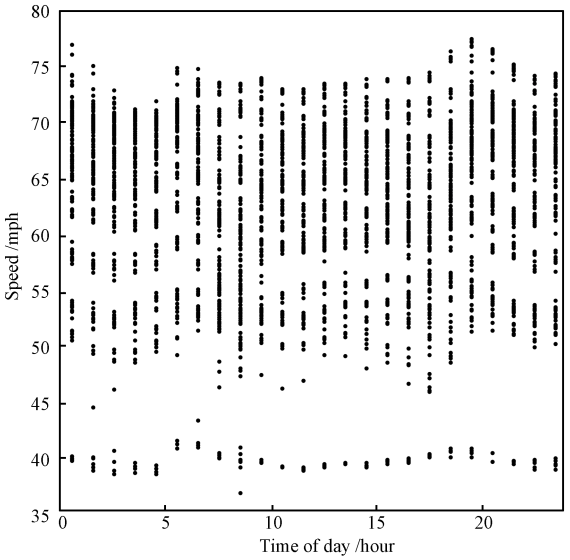


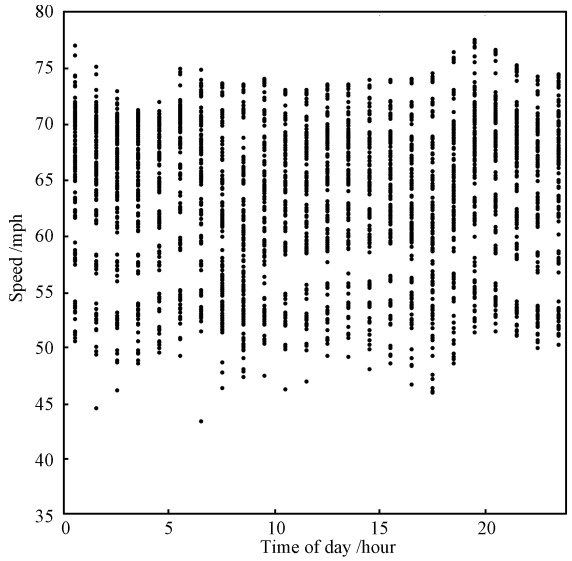
图 3　原始数据

Fig. 3　Raw data

图 4 (a) 和 (b) 分别表示基于四分位数法和本文所提出方法的祛噪声结果. 对比两者可以发现: 基于四分位数的方法基本上滤除了所有低于 45 mph 的数据, 而本文所提出的基于密度的祛噪声方法仅仅祛除小于 45 mph 以下的一小部分数据. 此外, 从图 3 所示的原始数据中还可看出, 虽然低于 45 mph 的数据距离样本均值较远, 但是这些数据相对稠密, 且持续出现在一天内所有的时刻, 暗示着一种较强的关联性. 这种关联性表明这些低于 45 mph 的数据不容忽视; 也就是说, 该数据集里很有可能包含着两个或两个以上服从相同或不同分布的数据子集. 基于四分位数的方法在进行噪声识别时, 将所有数据看作为服从单个分布的数据, 显然此传统方法不能从含有多个分布的数据中正确地检测出噪声. 然而, 采用本文提出的方法祛噪之后 (参见图 4 (b) 所示), 不仅保留了低于 45 mph 数据中蕴含的系统特征, 而且通过滤除其周围可信度较差的数据, 更加突现了系统的本质特征.

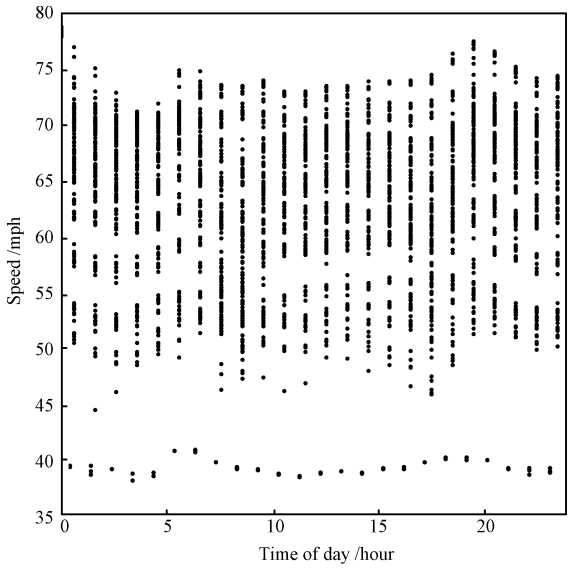
对实验中的行驶速度数据作进一步分析可知: 影响行驶速度的因素很多 (诸如天气、能见度、人们出行的时间等), 所有这些因素可看作为速度这个应变量的自变量; 也就是说该数据实质上是一个高维数据, 每一维是其中的一个影响因素, 但在实际采集中, 不太可能对每一维数据做到较为精确地采

集, 对于本文中所采用的数据, 只对时间和速度进行了统计记录, 也就是说把一个高维数据简化为二维数据, 由于忽略了其他影响因素, 这是可能造成多个分布的主要原因之一. 例如图 3 和图 4 (b) 中低于 45 mph 的数据可能是由于天气恶劣能见度低导致的行驶缓慢.



(a) 基于四分位数的方法

(a) Quartile-based approach



(b) 基于密度的祛噪声方法

(b) Density-based approach

图 4 祛噪声结果

Fig. 4 The results of outlier removal

3 结论

本文阐述了一种新颖的基于密度的祛噪声方法, 并通过仿真对比实验验证了该方法的有效性. 实验结果表明该方法不依赖于数据服从某个特定分布的假设, 而且可以处理服从多个相同或不同分布的杂合数据. 该方法有两个重要参数: 种子数目和种子吸附阈值, 需要针对实际情况灵活确定. 此外, 虽然本文只对二维数据进行了实验, 但该方法同样适用于多维数据的祛噪声处理.

References

1 Taylor J R. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements (Second Edition)*. Sausolito: University Science Books, 1997. 120−187

2 Bevington P R, Robinson D K. *Data Reduction and Error Analysis for the Physical Sciences (Third Edition)*. New York: McGraw Hill, 2002

3 Rousseeuw P J, Ruts I, Tukey J W. The bagplot: a bivariate boxplot. *The American Statistician*, 1999, **53**(4): 382−387

4 Fornasini P. *The Uncertainty in Physical Measurements: An Introduction to Data Analysis in the Physics Laboratory*. New York: Springer, 2008

5 He Z Y, Xu X F, Deng S C. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 2003, **24**(9-10): 1641−1650

6 Fu L, Medico E. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics*, 2007, **8**(3): 1−15

7 Ferrer-i-Cancho R. The Euclidean distance between syntactically linked words. *Physical Review E*, 2004, **70**(5): 1−5

8 Krause E F. *Taxicab Geometry*. New York: Dover, 1987. 63−89

9 Huang W, Shi Y Y, Zhang S Y, Zhu Y F. The communication complexity of the Hamming distance problem. *Information Processing Letters*, 2006, **99**(4): 149−153

10 Li Y J, Liu B. A normalized Levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, **29**(6): 1091−1095

王 扬 北京工业大学交通研究中心讲师. 分别于 2004 年及 2007 年获得英国约克大学和英国拉夫堡大学的硕士和博士学位, 并在英国朴茨茅斯大学从事博士后研究工作. 主要研究方向为模式识别、机器学习、交通信息处理与控制. E-mail: hiscott@126.com

(**WANG Yang** Lecturer at the Transportation Research Center, Beijing University of Technology. He received his master and Ph. D. degrees from University of York and Loughborough University, UK, respectively, and worked as a postdoctoral researcher in University of Portsmouth, UK. His research interest covers pattern recognition, machine learning, traffic information processing and control.)