基于 K-Means 动态数据中值估算的研究

林金田　　王梦娣

【摘要】本文介绍了一种基于推广型 K-Means 分类决策的动态数据中心值估算方法，能够有效实时准确地计算出具有幅度波动的动态称重系统中的正确称重重量，准确度高，实时性好，鲁棒性高。

【关键词】 动态称重　中值估算　动态仿真

一、技术背景

动态称重是指在测量数据不稳定时，通过一定的算法估算出真实的称重重量。此类系统常用于汽车称重、牲畜称重等动态称重领域。

目前常用的动态数据求中心值算法主要为滑动平均算法。滑动平均算法根据滑动窗口的大小可分为两种，固定窗口长度和可变窗口程度，其具体如下：

$$\bar{d} = \mathrm{h}\left(d_1, d_2, ..., d_K\right) = \frac{\sum_{i=1}^{K} d_i}{K}$$

其中，$d_1, d_2, ..., d_K$ 为原始数据值，K 为数据个数。

对于固定窗口长度算法，K 为固定值。此算法适用于波动幅度较小的动态数据，实时性好，但准确度低。对于可变窗口长度算法，K 是可变的。每测量到一个数据，K 的值增加 1。随着 K 数值的增加，数据接近稳定，求得动态数据中心值。此算法可用于波动幅度较大的动态数据，准确度高，但延时较长，实时性差。

二、中值估算算法的研究

K-Means 决策算法是基于原型的目标函数聚类方法，算法利用相似度度量将一组数据通过不断的迭代得到 K 个聚类中心，以此作为动态数据的中心值。该算法的具体步骤如下：

步骤 1、算法数据初始化，包括：

1) 建立长度为K的数据队列；

2) 获取特定宽度窗口N_0内动态数据求取均值$\bar{u}_1$，作为初始中心点，并加入队列；

$$\bar{u}_1 = \frac{d_1 + d_2 + ... + d_{N_0}}{N_0}$$

步骤 2、窗口依次扩大，加入新的原始数据d_i，求取新的平均值$\bar{u}_j$并加入队列$\{\bar{u}_1, \bar{u}_2, ..., \bar{u}_{j-1}\}$，若队列已满，则遵循先进先出的原则覆盖已有值，求取新的平均值的具体算法为：

$$\bar{u}_{N+1} = \mathrm{f}\left(N, \bar{u}_N, d_{N+1}\right) = \frac{\bar{u}_N * N + d_{N+1}}{N+1}$$

其中，d_{N+1}为新加入的原始数据，$\bar{u}_N$为上一步已求得的N个数据的均值。

步骤 3、队列内数据均值$\bar{\bar{u}}$作为新的类中心点，并求队列内数据相似度$distance$；

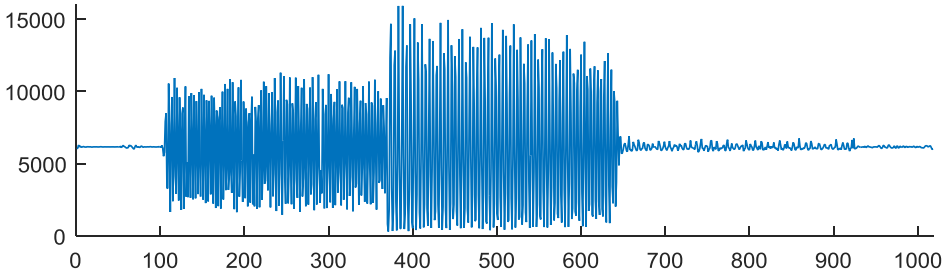
$$\bar{\bar{u}} = \mathrm{h}\left(\bar{u}_1, \bar{u}_2, ..., \bar{u}_K\right) = \frac{\sum_{i=1}^{K} \bar{u}_i}{K}$$

$$distance = \mathrm{g}\left(\bar{u}_1, \bar{u}_2, ..., \bar{u}_K\right) = \sqrt{\frac{\sum_{i=1}^{K} \left(\bar{u}_i - \bar{\bar{u}}\right)^2}{K-1}}$$

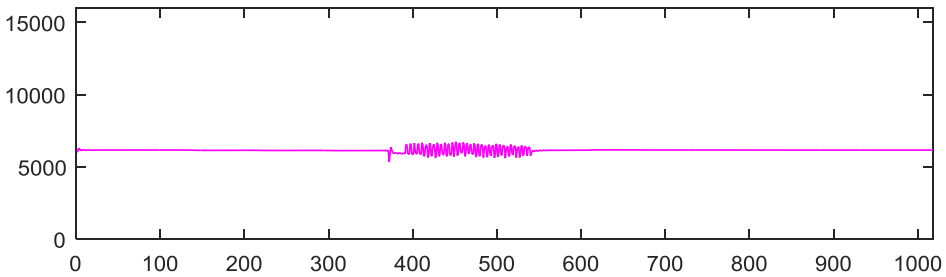
其中，K 为队列长度，$\bar{u}_1,\bar{u}_2,...,\bar{u}_K$ 为队列内数据。

步骤 4、若数据相似度小于阈值则可锁定数据中心值，否则重复进行步骤 2、3。

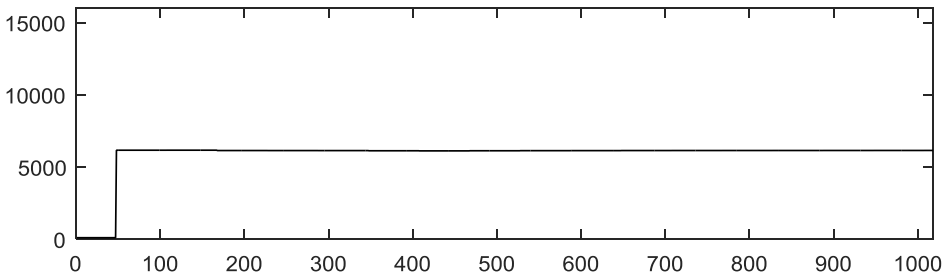
图 1-2 是本文中心值估算算法与现有技术估算算法的仿真效果对比图。



(a) 原始数据

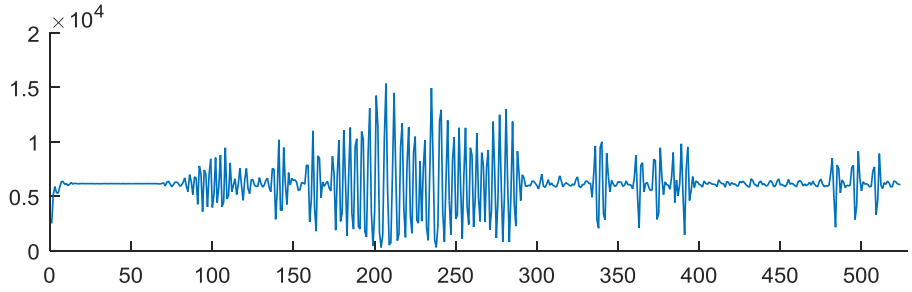


(b) 滑动平均算法

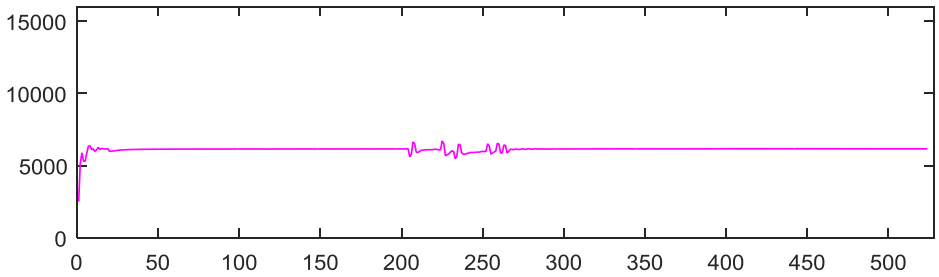


(c) 本文算法

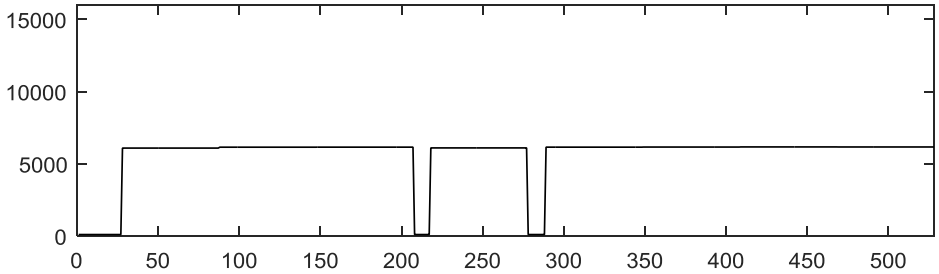
图 1



(d) 原始数据



(e) 滑动平均算法



(f) 本文算法

图 2

如图 1-2 为两组数据的仿真结果，如图示所述，(a)为原始动态数据，(c)为本文算法求取的中心值数据。根据称重系统（一般 1s 获取 10 个数据）的实际延时需求，这里取滑动窗口初始大小 $N_0 = 20$，队列长度 $K = 10$。本仿真所用数据波动较大，即实际中心值为 D 时，波动数据可在 $[0, 2D]$ 的范围内波动。如图，对比可见，滑动平均算法只能获取有一定误差的中心值数据，且在动态数据波动较大时，滑动平

均算法所估测的中心值也具有一定的波动，而本文算法则可获取较为稳定的中心值数据。由图可见，在波动较大时，本文可能出现数据无法锁定的状态（图示显示为恢复零值）但本文算法可在 2s 时间范围内重新锁定中心值数据，满足实际应用即时稳定的需求。

三、小结

本文基于推广型 K-Means 决策算法的动态数据中心值估算算法，相比较于传统的滑动平均估算算法，抛弃传统的滤波思想，从 K-Means 分类决策算法思想出发，将所有数据分为一类，算法的目标由分类伸展为求解类中心。本算法兼顾了时间效率和准确度两个方面，可以用较短的数据估算出准确的中心值数据，可用于具有大幅度波动的动态数据的中心值估算。

参考文献

[1] 张健沛，杨悦，杨静,等. 基于最优划分的 K-Means 初始聚类中心选取算法[J]. 系统仿真学报，2009，第 9 期:2586-2590.

[2] 杨善林，李永森，胡笑旋,等. K-means 算法中的 k 值优化问题研究[J]. 系统工程理论与实践，2006，第 2 期(2):97-101.

[3] 冀素琴,石洪波.面向海量数据的K-means聚类优化算法[J].计算机工程与应用,2014,50(14)：143-147.

作者简介

林金田，男，1972 年 7 月，工程师，经济师，现任锐马（福建）

电气制造有限公司董事长兼总经理，主要从事称重测力传感器及仪表的研发工作。

王梦娣，女，1992 年 1 月，清华大学电子工程硕士学位，主要从事算法优化，系统设计的研发工作。