doi: 10.7690/bgzdh.2013.04.019

基于软加权 k-均距异常因子的雷达数据剔野方法

胡奎
（中国人民解放军 92941 部队 96 分队，辽宁 葫芦岛 125000）

摘要：针对雷达测量数据的特点，提出一种基于软加权 k-均值距离异常因子的雷达数据剔野方法。对测量序列进行软性加权，再将序列映射到特征空间通过计算异常因子来对野值进行检测，并以某型雷达在某次任务中的一段测量数据进行实验。实验结果表明：该方法能很好地识别野值，最大限度地保持轨迹测量数据的完整性。

关键词：软加权；k-均值距离异常因子；雷达数据；剔野
中图分类号：TJ02 **文献标志码**：A

Radar Data Outlier Elimination Method Based on Soft-Weighted k-Mean Distance Outlier Factor

Hu Kui
(*No*. 96 *Team*, *No*. 92941 *Unit of PLA*, *Huludao* 125000, *China*)

Abstract: Aiming at the features of radar measurement data, put forwards the radar data outlier elimination method based on soft-weighted K-mean distance outlier factor. At first, carry out soft weight for measurement sequence, then, maps them into feature space and compute the k-mean distance outlier factor to detect the outliers. Take one of the task measurement data of certain type radar to carry out experiment. The experimental results show that the method can effectively recognize the outlier data, and protect race measurement data with a maximum of integrity.

Key words: soft weight; k-mean distance outlier factor; radar data; outlier elimination

0 引言

雷达作为靶场的重要测量装备，其测量数据能为研制部门提供飞行器轨迹的基本情况，对飞行器的性能分析和精度评定起着重要的作用。但在实际的目标跟踪中，即使是高精度的测量雷达，在测量中受外来干扰，如宇宙噪声、海面反射和工业干扰等[1]，也会造成测量数据与真实轨迹之间的误差。特别是当雷达处于复杂电磁环境[2]时，有的跟踪数据会严重偏离真值，形成明显不连续的异常数据。对于异常数据，具有代表性的定义是 Barnett 在统计学研究领域中给出的：一个异常点是这样的数据点，基于某种度量而言，该数据点与数据集中的其他数据有着显著的不同[3]。在雷达数据处理中这样的异常点被称为野值[4]。野值的出现导致了测量值严重失真，严重影响了数据的质量，进而影响飞行器轨迹处理精度；因此，在进行数据处理前，对野值进行剔除就成为一项十分重要的工作。

围绕如何识别和剔除野值，人们进行了大量的研究，提出了外推预报识别[5]、差分检测法、似然比检测法[6]等一系列算法，但这些算法都有一些不足：外推预报和差分检测法不能处理成串的野值；似然比检测需要知道数据的概率密度函数，但是实际中要求得测量数据的密度函数往往很困难。近年来，随着数据挖掘研究的日益进展，异常点（即野值）检测受到了越来越多的关注，出现了许多新的算法，比如基于聚类的算法[7]、基于距离的算法[8]以及基于密度的算法[9]等。这些算法在很多场合都能够较好地检测到数据中的野值，但是它们处理的对象大多是无序数据集，在面对雷达测量数据这样的有序数据集时，检测效果就没有那么明显。詹艳艳等提出了一种针对有序数据集的异常点检测方法 K-MODF[10]，该算法首先利用边缘权重因子提取序列的边缘点，再将子序列的一些特征值映射到特征空间并计算 k-均距异常因子，然后利用异常因子对野点进行剔除。

K-MODF 从模式的角度检测序列的异常行为，为序列剔野提供了一个新的思路。但是，该算法在计算边缘权重因子时，将检测窗口的最大最小值作为参考，将那些等于检测窗口最大值或最小值的点在计算权重时硬性地定义为“1”或“-1”，忽略检测窗口中的其他点。但飞行器的轨迹是连续的，雷达对其轨迹的测量数据序列各点之间应有很强的相

收稿日期：2012-10-10；**修回日期**：2012-11-27
作者简介：胡奎（1985—），男，四川人，硕士，助理工程师，从事数据处理工作研究。

关性，都带有属于整个轨迹的信息，硬性加权忽略了中间点，必然造成轨迹信息的丢失，因此在处理轨迹序列的异常时往往效果不太理想。鉴于此，笔者进行了改进，提出一种“软”加权法，这种加权法能利用所有数据的信息，最大限度地保持轨迹测量数据的完整性。

1 “软”加权法

设 $S=\{(t_1,s_1),(t_2,s_2),\cdots,(t_n,s_n)\}$ 是雷达测量数据序列，其中 $t_1<t_2<\cdots<t_n$ 是测量数据的时间码。一般地，应有 $t_2-t_1=t_3-t_2=\cdots=t_n-t_{n-1}=\Delta t$，$\Delta t=1/f$ 是测量雷达的采样间隔，这里 f 代表雷达的采样频率。定义如下的软性边缘权重因子：

$$W(i)=\left|\sum_{j=1}^{v}w_j(i)\right|,\ w_j(i)=\frac{s_i-E_j}{\sigma_j+\varepsilon\delta(\sigma_j)} \tag{1}$$

其中：v 是软性边缘权重因子的窗口宽度；E_j、σ_j 分别为每一个检测窗口的均值和标准差；$\varepsilon>0$ 是一个常值小量；$\delta(\sigma_j)=\begin{cases}1 & \sigma_j=0\\ 0 & \sigma_j\neq 0\end{cases}$。将测量数据序列的边缘点定义为那些软性边缘权重因子较大的点。

设 $S_i=\left\{(t_{i_1},s_{i_1}),(t_{i_2},s_{i_2}),\cdots,(t_{i_s},s_{i_s})\right\}$ 是测量数据序列的一个子列。定义有 4 个指标：

子列高度：$h_i=s_{i_s}-s_{i_1}$ (2)

子列长度：$l_i=i_s-i_1+1$ (3)

子列均值：$\overline{S_i}=\dfrac{\sum_{j=1}^{i_s}s_{i_j}}{l_i}$ (4)

子列标准差：$\sigma_i=\sqrt{\dfrac{1}{l_i-1}\sum_{j=1}^{i_s}(s_{i_j}-\overline{S_i})^2}$ (5)

为了将上述 4 个指标放到同一个空间，需先对他们进行无量纲化。设 $c_i=\left\{c_{i_1},c_{i_2},\cdots,c_{i_s}\right\}$ 是其中一组指标，按式 (6)：

$$\begin{aligned}&d(c_{i_j})=\frac{c_{i_j}-\min(c_i)}{\max(c_i)-\min(c_i)+\varepsilon\delta(c_i)},\quad \varepsilon>0\\ &\delta(c_i)=\begin{cases}1 & \max(c_i)=\min(c_i)\\ 0 & \max(c_i)\neq\min(c_i)\end{cases}\end{aligned} \tag{6}$$

对 c_i 进行无量纲化后，每个指标的值都将介于 0 和 1 间。用无量纲化后的 4 个指标构成特征空间：

$$C=C(d(h),d(l),d(\overline{S}),d(\sigma))$$

有了特征空间后，给出如下定义：

点 p 的 k 距离、k 平均距离：给定 $k\in N^+$，对 $p,q\in C(d(h),d(l),d(\overline{S}),d(\sigma))$，称 p、q 之间的距离是点 p 的 k 距离 $\mathrm{dist}_k(p)$ 当且仅当同时满足以下 2 个条件：

1） $\left|\{r\mid r\in C(d(h),d(l),d(\overline{S}),d(\sigma))\backslash\{p\},\mathrm{dist}(p,r)\leqslant\mathrm{dist}(p,q)\}\right|\geqslant k$；

2） $\left|\{r\mid r\in C(d(h),d(l),d(\overline{S}),d(\sigma))\backslash\{p\},\mathrm{dist}(p,r)<\mathrm{dist}(p,q)\}\right|\leqslant k-1$。

其中 $|\cdot|$ 是集合的势；$\mathrm{dist}(\cdot,\cdot)$ 是特征空间 C 中的欧式距离。称 $\overline{\mathrm{dist}_k(p)}=\dfrac{1}{k}\sum_{j=1}^{k}\mathrm{dist}_j(p)$ 为点 p 的 k 平均距离。

对于特征空间中的每一个点 $p\in C(d(h),d(l),d(\overline{S}),d(\sigma))$，分别在特征空间 C 和其子空间 $C_1(h_1,h_2,\cdots,h_m)$、$C_2(l_1,l_2,\cdots,l_m)$、$C_3(\overline{S}_1,\overline{S}_2,\cdots,\overline{S}_m)$ 及 $C_4(\sigma_1,\sigma_2,\cdots,\sigma_m)$ 上计算其 k 平均距离，将这些距离进行无量纲化后记为 $\mathrm{dist}(p)$，$\mathrm{dist}_1(p)$，$\mathrm{dist}_2(p)$，$\mathrm{dist}_3(p)$ 及 $\mathrm{dist}_4(p)$。定义如下的异常因子：

$$OF(p)=\mathrm{dist}(p)+\max\{\mathrm{dist}_i(p),i=1,2,3,4\} \tag{7}$$

显然，$OF(p)$ 越大，其对应的子列是野值的可能性就越大。

2 实验分析

采用某型雷达在某次任务中的一段测量数据进行实验，数据信息如表 1 所示。

表 1 测量数据信息

数据量	起始时间/s	终止时间/s	时间氏度/s	采样率/(帧/s)
987	0	98.60	98.60	10

图 1 是飞行器方位角 A(t(i))理论轨迹（真值）曲线图，该飞行器在 43.2 s 附近有一个加速转弯过程；图 2 是该雷达对飞行器的实测数据。在实验中分别用文中的方法及 K-MODF 对测量数据进行野值剔除，所得结果如表 2，表中“+”表示剔除，“-”表示未剔除，括号内的“T”表示剔野正确，“F”表示剔野不正确。

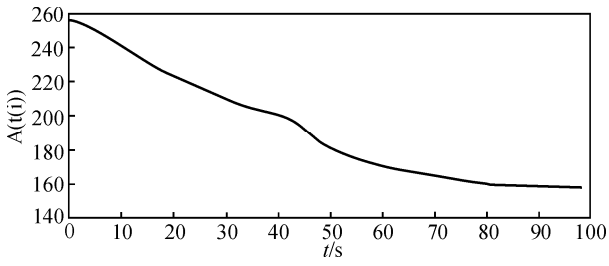

图 1 飞行器方位角真值

（下转第 77 页）