·先进计算与数据处理· 文章编号：1000—3428(2013)05—0061—04 文献标识码：A 中图分类号：TP391.5

基于聚类的多传感器数据融合方法研究

黎 亮[1]，谭世海[2]，师 伟[3]

(1. 电子科技大学经管学院信息系统研究所，成都 610054；
2. 成都市经信委，成都 610041；3. 中国燃气涡轮研究院，四川 江油 621703)

摘 要：在复杂环境下进行多传感器测试，其数据分布往往不规则和不一致。针对该情况，提出一种基于聚类的多传感器数据融合方法。该方法不按权重相加，侧重于分析数据整体分布状况。采用模糊梯形函数对数据进行一致度量化，使用聚类算法对数据分布进行聚类分析，按照最大支持度原则寻找最优点。实验结果表明，该方法能得到较精确的融合值，并可以查找在测试过程中可能出现的故障。

关键词：数据融合；多传感器；一致度；聚类；支持度

Research on Multi-sensor Data Fusion Method Based on Clustering

LI Liang[1], TAN Shi-hai[2], SHI Wei[3]

(1. Institute on Information System, School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 610054, China; 2. Board of Chengdu Municipal Economic and Informatization, Chengdu 610041, China; 3. China Gas Turbine Establishment, Jiangyou 621703, China)

【Abstract】Multi-sensors are often used when testing in a complex environment, however the results sometimes are difficult to interpret especially when data distribution is irregular and even inconsistent. This paper presents a new analysis and fusion method which focuses on the distribution analysis for all the data rather than just making the weights for each single sensor in the traditional ways. It uses a fuzzy gradient function to quantify the consistent degree of the sensor data, and makes an algorithm for cluster analysis. The method integrates data by using the supportive degree. Experimental results show that the method can make a better integration and also can help to find faults.

【Key words】data fusion; multi-sensor; consistency; clustering; support degree

DOI: 10.3969/j.issn.1000-3428.2013.05.012

1 概述

利用多传感器的信息冗余进行数据融合对于提高系统测试的准确性和鲁棒性有很大帮助[1]。在高空飞行中会遇到大量不稳定的因素，其测试建模往往非常复杂，仿真效果较差，常规方法多停留在理论研究层面上，在实际测试中由于实时性要求高，其效果极为有限。许多融合方法是建立在信号较为稳定的前提下[2]，而高空环境往往不具备这些特征。

高空台上的实验如飞机发动机测试，由于其环境复杂往往采用多个传感器，如在同一截面可能会排放 21 个及其以上的同质传感器。测试出的数据有时会在一个较大的范围内，分布极不均匀，用简单的方法进行加权相加不能很好地解释其意义。不同于静态环境下可以用传感器反复测量，在高空台的动态实验中，数据处于连续变化之中，一般的融合方法如主成分分析法、方差分析法等都不适用。按照测量理论，若传感器的测量值 V_i=真实值+正态分布噪声 σ_i，则当多个传感器数据相加后，噪声部分将相互抵消，融合值的偏离部分将由原来的 σ 逐步下降为 0，这就是多传感器融合的目的所在。但大多数文献把融合方法的重点变成对各个传感器数据的赋权值过程，简单地对不同的传感器给予大小不同的权重。这样做的后果是权重大的传感器噪声将在融合时占据主导地位，因此，融合后的噪声部分将难以消除。本文分析近年来国内外数据融合理论，提出一套适合于高空台实验的数据融合方法，并探讨一种不是基于按权重大小进行融合的方法。

2 相关研究

在相关研究中，为刻画传感器数据之间的差异，采用许多方式，有绝对差异[3-4]但更普遍则偏重于相对差异[5-6]，

基金项目：航空科学基金资助项目(20101024006)

作者简介：黎 亮(1968－)，男，副教授、博士，主研方向：数据挖掘，人工智能；谭世海，工程师、硕士；师 伟，高级工程师、硕士

收稿日期：2012-05-29 **修回日期：**2012-08-20 **E-mail：**liliang@uestc.edu.cn

并且基本使用的是线性或类似线性的方式来描述[7]，即相同为 1、差异太大则为 0。这些方法存在一个问题，其忽略了传感器之间正常的差异，即当传感器之间有微小的差异时不应认为传感器之间有不一致；而当传感器之间有较大的差异时，其“一致性”应该迅速下降到 0，此时认为它们之间完全没有相互支持。也就是说传感器之间的“一致”的程度不应该是线性的，而是一条非线性的曲线，随着传感器间数据值差异的增加，其刚开始时下降缓慢，而后则迅速下降到 0。一些文献对于线性下降的不足主要是采用截断法[6]，即根据数据本身的情况，人工设置一个阈值，把低于此阈值的一致度全部变为 0，这样做的缺点是阈值不易掌握，如果过大则不容易判断数据的总体结构(本文进行总体排序)。文献[8]使用了 arccot 函数来量化一致度，虽然具有较好的连续性，然而该函数的形状并不能够反映出上述传感器数据间的物理意义。本文部分借鉴了文献[8]的思想，将探讨通过多段直线来模拟非线性下降的一致度趋势。

在多数文献中，把一致度简单地转变为融合的权值，然而假如测试传感器中有 2 组以上传感器，组内一致度高而组间一致度低，当这 2 组传感器数据值相差很远，是否能够简单地融合它们，如文献[9]所述“对高冲突证据合成会得到有悖常理的结果”。既然 2 组值之间不相互支持，就不能简单地按一致度权值相加。这种情况的发生在实际实验中多是由传感器排放位置不合理、没有相互校准、或故障，如漏油、干扰等引起，也就是说严重的不一致往往会有很重要的提示，可以对系统进行故障排查。因此，不一致现象是进行检查的信号而不是进行融合的信号。不一致的分析对于系统检测有重要的意义，本文提出了根据一致度进行聚类的算法，其目的在于发现数据的分布结构，而不是简单的赋权相加。在已有文献中也有使用聚类进行融合的例子[10]，只是在本文中目标和模型完全不同。

本文在数据正式融合前首先进行一致性计算，然后以聚类分析来判断数据的整体分布状态，根据分布情况来判断系统是否存在可能的故障或干扰，再在此基础上进行最后的融合。

3 一致度计算

探讨一致度的目的在于为聚类提供基础。通过聚类分析，可以把一致性高的数据进行融合，对“不一致”的个别数据则可能认为是在恶劣环境下偶然造成的“奇异点”。如果在聚类分析中发现出现了多个距离分布较远的类，则说明整体数据具有高度的不一致性，这些数据反映了不同的系统特性，不能进行简单的融合。一致度应该是一种势函数 $K(X,Y)$，它的选择应该满足以下条件：

(1)函数的输出为 0~1 之间，当 2 点 X、Y 距离很近时输出值大，当 2 点距离很远时输出值小。

(2)函数的主体趋势是逐步下降，即随着 2 点 X、Y 间距离的增大，函数的输出逐渐减小。

(3)当 X、$Y2$ 点重合时(距离为 0)，输出最大为 1。当 2 点距离无穷大时，输出最大为 0。

(4) $K(X,Y)=K(Y,X)$，$K(X,Y)$ 为一连续的函数。

在分析了多篇文献的基础上，本文选择了一种模糊梯形函数为一致度函数：

定义 1 两传感器数据的一致度为：

$$K=1-\frac{1-VA}{\alpha}\times dif \quad \text{当 } dif\leqslant\alpha \text{ 时}$$

$$K=\frac{VA\times(\beta-dif)}{\beta-\alpha} \quad \text{当 } \alpha\leqslant dif\leqslant\beta \text{ 时}$$

$$K=0 \quad \text{当 } dif\geqslant\beta \text{ 时}$$

其中，dif 代表 2 点 X、Y 间距离，上面函数被分为 3 段。0 到 α 为第 1 段，α 到 β 为第 2 段，β 以外为第 3 段，如图 1 所示。VA 是第 1 段函数的输出下限和第 2 段函数输出的上限，通过对 VA 值的设置可以调整这 2 段的斜率。在一般情况下，函数在第 1 段下降慢，在第 2 段下降快。上面 K 函数的输出，其值总体来说随着 dif 的增大而减小，不仅满足上面对势函数的基本要求，而且也符合现场测量的实际情况。即用多传感器对同一目标测量时，由于排放位置的不同和各种噪声的干扰，传感器之间有一点差异是完全正常的，不过相互间误差不能太大。函数输出随着 dif 的增加，一开始时下降很慢(对小差异不在乎)，但到了后面就衰减得很快(对大差异不容忍)。当两传感器的测量值差异很小时，认为是正常的(即相互之间一致度很强)，函数的输出很接近 1；当两传感器的测量值差异很大时，认为是不正常的(即相互之间没有一致性)，函数的输出非常迅速地远离 1。

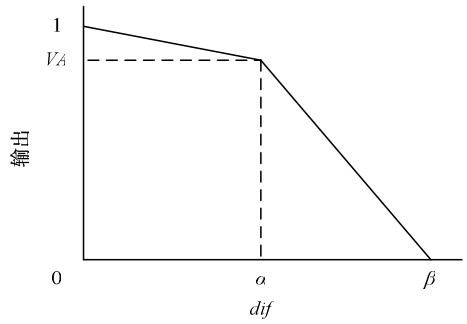

图 1 两传感器数据的一致度

对于上式中的距离 dif，不同的学者有不同诸多的定义，有的采用绝对距离，有的采用相对距离。本文参考了文献[6]定义的相对距离：

定义 2 数据间的距离为：

$$dif=1-\frac{\min(X,Y)}{\max(X,Y)}$$

通过设置适当的 α、β 和 VA 值，可以按需定制势函数对不同距离输入的响应，即哪些情况下的距离其相互一致度大，哪些情况下的距离其相互一致度小。从定义 1 中可知，当距离为 α 时，一致度的大小是 VA，表示基本支持；当超过 α 时函数值将加速下降。具体的参数值要根据传感

器精度和用户要求的精度来调整，精度要求越高，VA 值越小。α、β 可以看成是不同下降速度的界限。假如测量对象的在 1%的差异内是正常的，超过 1%的差异表示两传感器测试值有不可忽略的不同，若差异超过 25%的不一致就完全不能够接受，则可以有：α=0.01，VA=0.95，β=0.25。

本文使用了二段法分别来描述不同斜率一致度下降的情况，理论上还可以把一致度的下降过程分成更多的线段。线段越多越能够精确地描述，在测试中发现二段法基本能够满足要求。

4　聚类分析

在得到各个传感器数据的一致度后，很容易根据一致度来进行数据点的聚类，即把相似相近的点构成一个个集合。从上面一致度的公式定义可知，距离小于阈值的点其一致度大，距离大的点其一致度小。可以利用上面的公式计算出描述点与点之间的一致度矩阵，矩阵中的每个元素 K_{ij} 就代表第 i 个传感器数据和第 j 个传感器数据的一致度。把矩阵的各行数据分别相加，可以得到各个传感器的总体一致度:

$$TK_i = \sum_{j=1}^{n} K_{ij}$$

由上式可知，在密集区域的 TK_i 值大，孤独点的 TK_i 值小；相同密集程度的点，靠近中心的 TK_i 值大。在图 2 中 4 个传感器的测试数据分布于 A、B、C、D 4 点，如果传感器的精度较高，其覆盖半径较小，图中点 AB 与点 C 与点 D 所得到的数据值明显的不一致。图中的 C 点虽然处于中心位置，但总体一致度最高的却是 B 点。因此，总体一致度反映的是“重心”，而不是“中心”的理念。由于差异太大，图中测试数据分布在 3 个不同的区域：AB 区，C 区，D 区。D 传感器的数据得不到任何其他传感器的支持，同其他传感器明显不一致，没有可信度；同样 C 传感器也一样没有可信度。A、B 两传感器数据值接近，计算后存在一致性，即它们相互支持，故 AB 区的测试结果相对可信度最大，可以只用 AB 进行融合。强调重心(密集点处)而不是的中心(平均处)来进行数据融合是本文的核心思想之一。其优越性是可以有效地排除奇异点对中心位置(平均数)带来的不利影响。

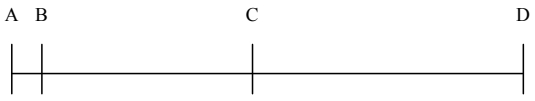

图 2　分散在 3 个区域的 4 个数据

本文进行聚类分析的方法是，先把各个传感器数据按照 TK_i 值从大到小进行排序；设置 n 个空的集合 G_1~G_n，其排号为 1 到 n。对排序好的数据使用如下算法:

(1)设置初始使用的集合号 S=1。

(2)从排序的数据点中选择最大者(排在最前面的)，即最大的 TK_i 的对应点，放入集合 G_S 中。

(3)把步骤(2)中选取点的所有相临近点都放入集合 G_S 中。对于临近点的定义是指 K_{ij}<阈值 γ 的那些点，γ 相当于聚类半径。γ 值的含义同上面的 VA 值，在 0 和 1 之间，可以像上面一样取 0.95 左右或略小一点如 0.85，但不宜太小。

(4)把步骤(2)、步骤(3)选出的点从排序数据中删掉。

(5)若排序数据没有剩余的点则算法结束。

(6)S=S+1。

(7)转到步骤(2)对剩余的点继续处理。

该算法的主要工作集中在排序和临近点的选择两方面上。若有 n 个传感器，则排序复杂度为 $O(n\text{lb}n)$ 。对于临近点的选择可考虑 2 种极端的情况: (1)各个结点完全独立，此时算法要把序列上的所有点扫描一遍；(2)所有的结点都相邻，即测试数据几乎一样。在这 2 种情况下，算法都需把各个传感器查看一遍，计算的复杂度都是 $O(n)$。因此，综合考虑算法的总体复杂度与传感器的数量有关，为 $O(n\text{lb}n)$ 。

算法执行完后，如果所有的点都在一个集合中(情况 1)，说明所有数据点的一致性很好，可以直接融合。如果每个点都处在许多不同的集合中(情况 2)，说明各个数据点完全不能相互支持，无法直接融合。出现这种情况的原因可能是外部环境的突发事件产生，也可能说明传感器的布点不够合理或传感器输出需要校正等。如果是前一种情况，则不一致，可以帮助发现追踪系统测试过程中的偶发事件，进而可能找到系统设计制造上的缺陷。如果传感器布点有问题但又无法改变，则在预处理时可以对各个数据点进行校准；如 D 点的温度总是比 C 点高 2°C 左右，则处理时可把 D 数据减 2 或 C 点数据加 2。在高空台实验中，由于安装等原因，有时会出现一个传感器总是比其他传感器高(低)几个数值的情况，使用上面的聚类方法可以有效地找出问题，在后面处理时提前校正。

在算法执行完后，遇到上述 2 种情况之外的情形可以看成是这 2 种情况的组合。如果有多个传感器数据都分布在一个大集合中，个别少数传感器数据零星分散在多个小集合中，说明这些小集合数据缺乏其他传感器的相互支持而不足以采用；这时只需对大集合中数据进行融合即可。如果好几个集合的大小都一样，即每个集合中含的传感器个数相同，则说明了数据之间的严重不一致，这种情况同上段的第 2 种情形一样，其处理方法也一样。就一般情况而言，首先查看哪个集合中的传感器个数多，以最多者表示其相互获得的支持多，将用它来进行下一步的融合。在传感器数量很少时如果有多个集合大小一样，则集合编号小的那个可信度更高些，因为其主要点的 TK_i 值大些，所以排号靠前。如在 3 个传感器情形下，数据值在中间的那个传感器的 TK_i 值会最大，其最应该被采纳。在这种情况下，本文认为应选取各个数据点的中位值。

通过上述对各种可能状态的分析，清理出了在什么情

况下数据无法融合，什么情况下可以融合，用哪些数据融合的方法。在对飞机发动机的实测中，全部传感器集结在一个大集合中的情况最多。然后是少数个别传感器数据分散于若干小集合，而多数传感器集中于一个大集合的情况次之(这时可排出掉奇异出错点)，其他情况很少且多是系统故障造成(因此，这为系统故障的追踪分析提供了依据)，或传感器本身需要校准。

5 基于支持度的融合

当得到了一组一致性较强的数据后，本文的工作就是完成最后一步的最终融合。在数据融合的研究中，其基本思路是找出各个传感器的不同重要程度，然后使用重要程度为权重进行数据相加。本文的思路不在于找传感器的重要度，而在于找所有传感器对所有可能数据候选点的支持程度，得到支持程度最大的数据点即可视为最佳融合点。

首先要确定范围，最终的融合点应该是介于融合传感器数据的最小值 *low* 与最大值 *high* 之间。一般文献中对支持的定义多是基于加法的，即传感器数值的加权相加。其特点是可以以肥遮丑，即某些传感器的支持可以弥补另外一些传感器的不支持。然而，本文认为在数据融合中，能够获得众多传感器一致支持的点才能够代表充分的融合。因此，定义的支持度使用乘法。不失一般性，假设第 i 个传感器的测试数据符合正态分布，如图 3 所示，其密度函数为 $P_i(x)=N(a_i,\sigma_i)$；对单个而言在 a_i 处密度最大。

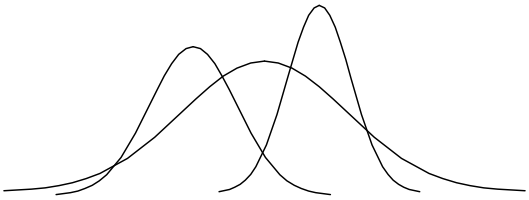

图 3 具有不同方差和均值中心的数据

定义 3 某数据 X 得到的各个传感器的支持度为:

$$TP(x)=\prod_{i=1}^{n}p_i(x)=\left(\frac{1}{\sqrt{2\pi}}\right)^n\times\frac{1}{\prod_{i=1}^{n}\sigma_i}\times e^{-\left[\sum_{i=1}^{n}\frac{(x-a_i)^2}{2\sigma_i^2}\right]}$$

其中，$X\in[low,\ high]$。为了求得上式的极值，求其导数和二阶导数，在其一阶导数等于 0 处其二阶导数小于 0，故有极大值。进一步可知，$TP(x)$在$\sum_{i=1}^{n}\frac{a_i}{\sigma_i^2}\Big/\sum_{i=1}^{n}\frac{1}{\sigma_i^2}$处存在最大值，因此，认为此时的数据值代表各个传感器的最大“联合”支持度，反映了获得各个传感器一致支持的点的位置。该结论是从 min 到 max 之间找一个点其联合概率密度最大的角度出发推算出来的，此结论的结果同基于“求最小方差”法的结果是一样的。值得注意的是，两者虽然结论相同，但目标不同，解决问题的方法也不相同。后者以“最小方差”为目标，使用拉格朗日法求解。而本文则是以“最大联合支持度”为目标，使用二阶导数法求解。这样就避免了个别权重大的传感器的噪声强烈影响融合结果的问题。

6 实验结果分析

本文方法总结起来其过程就是找密度最大的点集合，然后进行融合。如果遇到密度大小相等的情况，则找其中获得全局支持度最高的数据(按算法一般会是中位值的数据或偏向中位值的数据)。其特点是宏观上以重心法为主，微观上以中心法为主，结合了两者的特色。对于方法的效果选用了文献[5]中数据进行验证，此数据以 900°为标准值，并曾被大量文献所引用，见表 1。

表 1 3 个传感器的 6 次测量值

传感器	第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	第 6 次
s1	899.5	905.3	901.9	900.6	889.9	899.4
s2	898.3	875.9	888.1	886.2	907.5	904.4
s1	896.7	906.8	898.2	904.0	896.4	891.6

考虑到温度传感器的精度一般在 1/1 000 左右，偏离 2/1 000 属于完全正常，偏离 5%以上则完全不可接受(有故障或需校正)。故参数选择为 α=0.002，VA=0.95，β=0.05，γ=0.9。最后的总绝对误差为 14.4，好于其他方法[5]。从表 2 可知，不同情况下参与融合的传感器数量是不一样的。

表 2 6 次测量中参与融合的传感器与最终融合值

第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	第 6 次
s1+s2+s3	s1+s3	s1+s3	s1+s3	s3	s1
898.2	906.1	900.1	902.3	896.4	899.4

值得注意的是上面的参数并非最佳参数，即如果在本文算法中再调整此参数，完全可以获得更小的绝对误差。在实际的系统中，参数一旦确定不会轻易改变，不会根据数据反过来改参数以证明本文方法的优势。

以上的数据都是正常数据，即没有传感器出问题时的情况。现在假设 3 个传感器中的 1 个出了问题，这在实际系统中是常见的，如漏油、排线等。仍然使用上面数据，但假设其中偏离 900°最大的那个传感器的偏离程度分别增加 5%、10%和 20%。从融合结果可以看出，本文方法融合值的绝对误差始终保持在 13.5，不受奇异值的干扰。而其他方法融合值随着奇异值的增大而加速偏离 900°，完全没有抗野值的能力。这是因为其他方法中心值(平均值)的位置会严重影响融合后的结果，而如果不进行人工调整则奇异点会影响中心值的位置。使用本文方法不需要根据数据的不同情况来人工干预参数值，自动化智能化程度高，实时处理能力得到了保障。

7 结束语

本文提出一套在复杂环境下实时处理的数据融合方法，包括使用二阶段模糊梯形函数进行一致度的计算、使用聚类算法进行数据的分布分析和使用支持度最大化进行融合。实验结果表明，该方法具有很好的抗野值能力，能帮助发现系统故障。

(下转第 68 页)