

基于不完整数据的异常信号检测方法

马 捷^a, 钟子发^b, 史英春^a

(解放军电子工程学院 a. 309 研究室; b. 信息系, 合肥 230037)

摘 要: 针对异常电磁信号检测中常见的输入数据存在参数缺失的问题, 提出一种基于不完整数据的异常信号检测方法。该方法借鉴几何数学的思想, 通过将缺失数据与正常数据进行比对, 分析出缺失数据异常的可能性, 给出该数据的异常概率计算方法。通过该异常概率能直接检测出部分异常信号, 并给出剩余不完整数据的异常可能性的排序, 从而有利于在资源有限时优先处理异常概率高的信号, 达到处理资源优化配置的目的。实验结果表明, 该方法能给出缺失数据点的异常概率。

关键词: 不完整数据; 缺失数据处理; 异常信号检测; 异常概率

Abnormal Signal Detection Method Based on Incomplete Data

MA Jie^a, ZHONG Zi-fa^b, SHI Ying-chun^a

(a. 309 Research Room; b. Information Department, Electronic Engineering Institute of PLA, Hefei 230037, China)

【Abstract】 A new technique for the problem of incomplete data in abnormal signal detection system is proposed. Getting inspiration from the geometry, the new method compares the incomplete date with normal data, it presents a computation method of abnormal probability. With the abnormal probability, some abnormal signals can be detected directly, and the other incomplete data can be arranged. The algorithm decreases the workload and makes good use of calculation resources. Experimental result shows that when some parameters are lost, the method can get the reasonable abnormal probability of the incomplete data.

【Key words】 incomplete data; missing data processing; abnormal signal detection; abnormal probability

DOI: 10.3969/j.issn.1000-3428.2011.14.028

1 概述

异常信号检测是电磁频谱检测中的重要内容。然而, 由于电磁环境的复杂性, 信号检测的输入参数往往存在一定程度的缺失, 因此研究基于不完整数据的异常信号检测方法具有现实意义。现有的不完整数据的处理方法可以分为删除元组、数据填充和不处理三大类。删除元组法主要是将存在缺失数据的记录删除, 从而得到一个没有含缺失数据的数据集。这种方法简单易行, 其适用于数据中仅含有少量的不完整数据, 并且默认这些少量的不完整数据对所求的结果影响不大的领域。由于不完整的数据在电磁频谱检测中经常出现, 异常信号在一般情况下在数量上占少数, 但重要性很大, 因此随意丢失数据的代价太大。

数据填充方法的主要思想是根据没有缺失的记录, 对有缺失的位置估计出一个数据去代替缺失的数据, 从而使数据集完备化。这种方法在统计学和数据挖掘领域都十分流行, 研究人员提出了非常多的缺失数据填充方法^[1-3]。常见的方法有人工填写法、特殊值填充法、平均值填充法、热卡填充法、回归法、期望值最大化方法以及近几年提出的粗糙集数据补缺法和神经网络数据补缺法等。这些方法的提出, 极大地丰富了缺失数据处理的理论, 也解决了大多数情况下的数据不完整问题。但这些方法最核心的思想不外乎于“平滑”或“就近”的原则, 即将缺失数据要么用最常见的未缺失数据补齐, 要么用可能与之最接近的未缺失数据补齐。这一思路可以较好地解决一般问题, 但是用于本领域则不太合适。首先, 异常检测的目的就是试图在正常数据中找出异常信号, 即找出“奇异点”, 如果用大多数正常点来“平滑”, 则很可能将“奇异点”抹去; 同时, 异常信号其本质上就是与其

他信号不同, 若试图用“就近”原则找到相似点, 同样可能将其“正常化”。单独处理的方法是一些研究人员结合上述 2 种处理方法的不足, 提出直接在含有缺失数据的数据集上进行数据挖掘的方法^[4-5]。但是, 由于电磁频谱检测中不完整数据常与一些毫无规律的干扰或噪声混在一起, 因此不完整数据集的规律性较难发掘。所以, 如果完全忽略含有丰富背景信息的无缺失数据, 单纯从缺失项中挖掘信息, 很容易陷入毫无规律的噪声中, 产生奇异的结果。

综上所述, 用于电磁异常信号检测中的不完整数据处理方法具有以下特点: (1)需要算法从大量正常信号中发掘少量的异常信号。(2)异常信号一般没有明显的规律性, 可能没有可供类比的相似信号。(3)缺失参数的信号往往混在复杂的干扰或噪声中, 难以单独发掘。(4)不能轻易放弃某一未确定的信号, 宁可加大虚警概率以换取漏警概率的降低。

本文提出一种通过将不完整数据中的未缺失参数与具有完整数据的正常信号数据集进行比较的方法, 计算不完整数据的异常概率, 通过该异常概率能直接检测出部分异常信号, 并给出剩余不完整数据的异常可能性的排序, 从而有利于在资源有限时优先处理异常概率高的信号, 达到处理资源优化配置的目的。

2 适用于电磁频谱监控的不完整数据处理算法

在提出算法之前, 首先考虑以下三维信号的情况。

一般情况下, 正常信号点往往能够按某种规律聚成一类,

作者简介: 马 捷(1986—), 男, 硕士, 主研方向: 通信与信号处理, 智能信息处理; 钟子发, 教授、博士生导师; 史英春, 博士
收稿日期: 2010-12-13 **E-mail:** majierex@126.com

或者说正常点能够分布于某个“规律面”的附近。

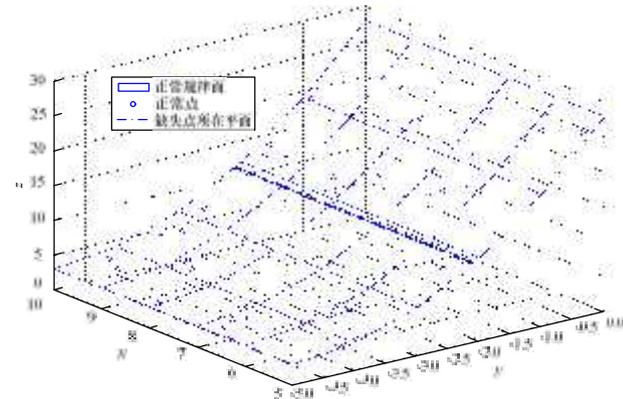
假设信号的三维参数为 (x,y,z) , 正常的没有缺失信号参数的数据点为 $P_1(x_1, y_1, z_1), P_2(x_2, y_2, z_2), P_3(x_3, y_3, z_3), \dots$, 能够按某种规律聚在某个规律面上, 即:

$$P_1(x_1, y_1, z_1), P_2(x_2, y_2, z_2), P_3(x_3, y_3, z_3) \dots \in \text{规律面}\Sigma$$

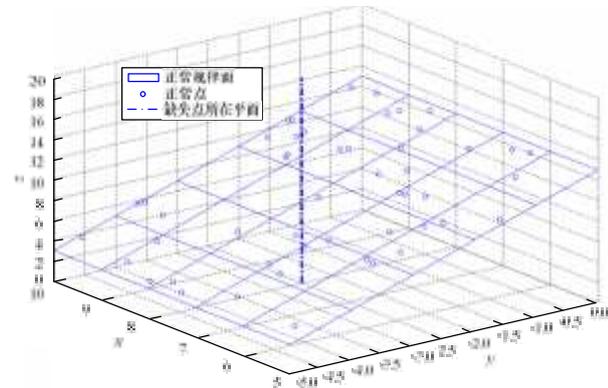
某一不完整信号点 $P_0(x_0, y_0, z_0)$ 有一参数缺失, 为方便起见, 不妨设 $y_0 = \emptyset$ (或 $z_0 = \emptyset$)。

根据 P_0 剩下的参数 x_0, z_0 (y_0), 可知 P_0 只可能分布在三维空间的一条直线上, 即 $P_0(x_0, y_0, z_0) \in \text{直线}l$ 。

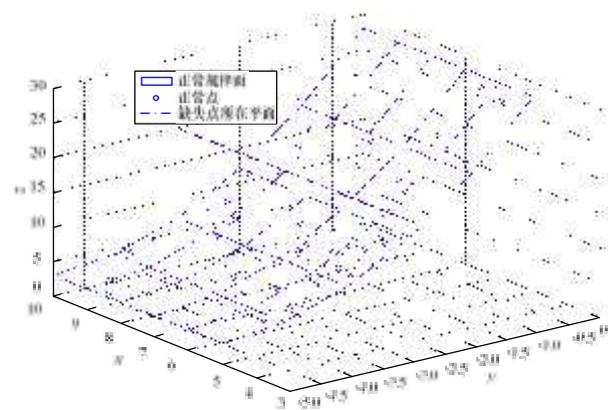
由图1可知, 缺失参数信号点所在直线与正常规律面存在如下可能的关系: 直线 l 在规律面内, 即 $l \subset \Sigma$ (图1(a)); 直线 l 与规律面相交 (图1(b)); 直线 l 不在规律面内, 即 $l \cap \Sigma = \emptyset$ (图1(c))。



(a) 缺失点所在直线在规律面上



(b) 缺失点所在直线与规律面相交 p



(c) 缺失点所在直线与规律面异面

图1 缺失点所在直线与规律面之间的关系

当 $l \subset \Sigma$ 时, 说明无论 $P_0(x_0, y_0, z_0)$ 缺失的参数 y_0 为何值, 该点必然为正常点, 即该信号正常, 不必对其进行处理。

当直线 l 与规律面相交时, 只有一部分点在规律面附近, 若直线与规律面夹角较大, 在规律面附近的点较少, 即信号异常的概率较小; 若直线与规律面夹角较小, 则在规律面附近的点较多, 即信号异常的概率较大。

当 $l \cap \Sigma = \emptyset$ 时, 可以根据直线到面的距离判断出该点异常的概率: 当距离小于某阈值时, 认为距离与该信号点异常的概率呈正比; 而当距离大于该阈值时, 则直接判断该信号点为异常信号。

根据上述分析, 本算法不试图填补不完整数据的缺失参数项, 也不删除该不完整数据, 而是通过比较分析, 给出该缺失数据异常的概率。具体算法如下:

假设正常接收的信号数据为 r 维, 即正常数据为:

$$P_1(k_{1,1}, k_{1,2}, \dots, k_{1,r}), P_2(k_{2,1}, k_{2,2}, \dots, k_{2,r}), P_3(k_{3,1}, k_{3,2}, \dots, k_{3,r}) \dots$$

存在一个缺失第 s ($1 \leq s < r$) 维参数的缺失数据 $P_0(k_{0,1}, k_{0,2}, \dots, k_{0,s-1}, k_{0,s+1}, \dots, k_{0,r})$, 定义与该点的异常概率相关的以下特征量:

(1) 正常数据中心在 $(r-1)$ 维空间的投影点到缺失点所在向量的距离 d :

1) 计算出正常数据的各维参数均值 (m_1, m_2, \dots, m_r) 。

2) 求出正常数据与缺失数据之间在除第 s 维之外 $(r-1)$ 维空间的欧氏距离:

$$d = \sqrt{(m_1 - k_{01})^2 + (m_2 - k_{02})^2 + \dots + (m_{s-1} - k_{0,s-1})^2 + (m_{s+1} - k_{0,s+1})^2 + \dots + (m_r - k_{0r})^2}$$

(2) 正常数据参数的 r 维主分量向量 A 与缺失点所在向量之间的夹角 α :

1) 通过主分量分析(PCA)方法找出正常数据参数的 r 维主分量向量 A 。

2) 计算出正常数据 r 维主分量向量 A 与缺失数据所在 s 轴之间的夹角 α ($\alpha \in [0, \pi]$)。

(3) 正常点最大阈值 R

根据主分量分析理论, 选取的特征面为方差最大的几个方向, 所以, 在其他方向的正常点到中心点距离不会超过正常规律面上距中心最远的点, 所以选取正常点阈值 R 为规律面上距离中心点最远的距离, 即:

$$R = \max(\sqrt{(k_{i,1} - m_1)^2 + (k_{i,2} - m_2)^2 \dots + (k_{i,r} - m_r)^2}), \quad i = 1, 2, 3, L$$

当某点距离正常点中心的距离大于 R 时, 直接认为其为异常点。因此, 以正常点中心为球心, 以 R 为半径的球就是正常点的最大范围, 落入球外点的异常概率为1, 即其必然是异常点。

(4) 正常点分布密集度 *density*:

$$density = \frac{\text{正常点总个数}N}{\text{正常点最大阈值}R}$$

正常点分布密度与异常概率呈反比, 当正常点越密集地分布于中心点附近时, 离正常中心点一定距离的点异常概率越大。

(5) 异常概率系数 k

由于实际情况不同, 误判的代价不同, 可按实际情况通过调节异常概率系数来对计算所得的结果做出一定调整。例如, 如果漏检的代价较大, 则可相应增大 k 值; 而少量漏检的代价不大, 且希望在尽量少的时间内判断出可能最大的异常目标, 则相应减小 k 值。 k 的具体取值可依照专家知识或通过实际情况使用样本训练获得。

三维空间下各特征量的示意图如图2所示。

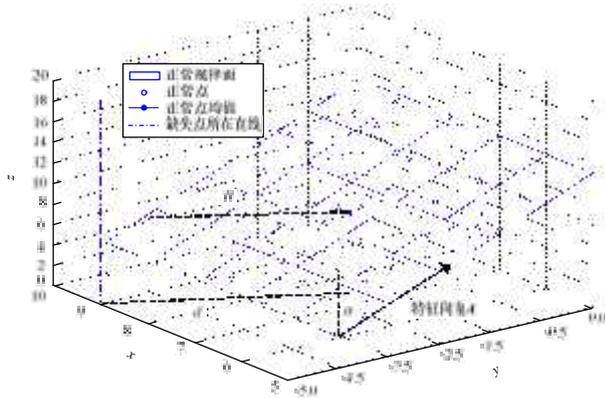


图2 三维空间下各特征量的示意图

下面,用上述特征量表示出异常概率:

为方便起见,假设正常点的特征向量和缺失点所在向量呈 α 角,正常点的中心与缺失点所在向量距离为 d ,正常最大阈值为 R .缺失向量落入正常阈值的部分被认为是正常.所关心的异常概率就与异常线段的长度和上面提到的正常点分布密集度 $density$ 有关.

这里用角度关系定义异常区域占总体的比率:

$$\frac{\text{异常区域}}{\text{缺失点分布区域}} = \begin{cases} \frac{\pi - \alpha}{\pi} & d \leq R \\ 1 & d > R \end{cases} = \begin{cases} \frac{d}{R} & d \leq R \\ 1 & d > R \end{cases}$$

对于缺失点可能的直线与正常点所在规律面的异面距离 D 有: $D = R \cdot \sin \alpha$. 易见,如果缺失点可能的直线与正常点所在规律面在正常点最大阈值 R 范围内无交点,则缺失点必为异常,即异常概率为 1;若有交点,则可根据焦点距中心点远近判断异常概率的大小:

$$P1 = \begin{cases} \frac{R \cdot \sin \alpha}{R} & d \leq R \cdot \sin \alpha \\ 1 & d > R \cdot \sin \alpha \end{cases} = \begin{cases} \sin \alpha & d \leq R \cdot \sin \alpha \\ 1 & d > R \cdot \sin \alpha \end{cases}$$

因此,总的异常概率:

$$P = k \cdot \frac{\text{异常区域}}{\text{缺失点分布区域}} \cdot \frac{1}{R \cdot \text{density}} \cdot p1 = \begin{cases} k \cdot \frac{\pi - \alpha}{\pi} \cdot \frac{d}{R} \cdot \frac{1}{R \cdot \text{density}} \cdot \sin \alpha & d \leq R \cdot \sin \alpha \\ 1 & d > R \cdot \sin \alpha \end{cases}$$

3 仿真实验

采用仿真实验对以上提出的缺失数据处理方法进行验证.仍然用三维数据来仿真,如图3所示.

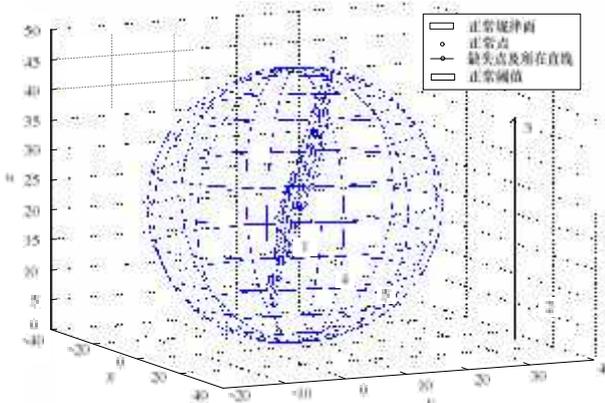


图3 三维仿真实验

现有5个数据缺失点,图中标出了实际缺失点的位置,而缺失了 z 坐标后,其可能分布的向量如图中线条所示,根

据第2节的算法,求出各点的异常概率如表1所示.

表1 三维缺失数据分析

缺失数据点	异常概率
1	0.152
2	1.000
3	1.000
4	0.257
5	0.837

通过从图中观察,上述结果较为合理,点2和点3落在正常阈值之外,可直接认为是异常信号,点1和点4离正常规律面较近,异常可能性较小,而点5距离正常规律面较远,异常概率较大,因此,可以根据实际需求进行按异常概率高低有选择性地后续处理.

下面用接收的信号数据对该算法进行验证.本实验数据采用某一体化分析接收机接收的无线电监测数据,实验平台为 Pentium 4 2.6 GHz 的 PC 机,1 GB 内存、Window XP 专业版操作系统.使用通过信号处理得到的信号的持续时间、中心频率、来波方位、信号带宽、信号场强作为模型的 5 维输入参数.将空间接收的无线电背景信号作为正常信号;将使用放置在不同方位的 5 台某综合信号源发射的信号作为异常信号源.实验共采集 1 000 组参数完整的信号.前 800 个信号只有正常信号作为正常信号的训练样本,后 200 个混有异常信号作为测试信号.在实验中,从后 200 个测试信号中随机抽取 5 个信号,分别拿去这 5 个信号的某一维输入参数,将这 5 个只有四维输入参数的信号作为不完整的待测信号,使用上述算法对其进行判断,判别结果如表 2、表 3 所示.

表2 实际三维信号数据实验 1

参数	正常数据										
	1	2	3	4	5	6	7	8	9	10	...
持续时间/s	4.35	0.72	4.99	3.22	2.14	1.79	4.82	3.09	3.74	2.27	...
中心频率/MHz	2.704	37.846	69.467	28.261	13.643	56.009	66.626	27.358	17.728	81.778	5...
来波方位/(°)	13.5	6.9	4.5	19.6	18.7	3.4	1.7	15.9	18.9	15.7	...
信号带宽/kHz	3.3	7.7	5.8	7.3	6.6	9.4	5.6	14.8	6.6	4.4	...
信号场强/dBm	-55.7	-48.0	-63.1	-84.3	-63.5	-41.5	-47.9	-81.1	-82.6	-50.1	...

表3 实际三维信号数据实验 2

参数	缺失数据(括号中为缺失参数的真实数据)				
	1	2	3	4	5
持续时间/s	3.47	0.57	2.89	5.16	缺失(5.73)
中心频率/MHz	16.582 4	10.757 2	18.136 2	缺失(9.936 4)	11.181 1
来波方位/(°)	27.7	14.0	缺失(18.7)	19.8	38.1
信号带宽/kHz	缺失(17.5)	10.5	9.1	2.7	19.3
信号场强/dBm	-48.2	缺失(-52.4)	-80.8	-105.1	-35.1
信号的实际异常状况	异常	正常	异常	异常	正常
判断的异常概率	1.000	0.237	1.000	0.910	0.490

从表 2、表 3 的数据中可以看出,通过采用此算法,可以较合理地给出各点的异常概率.通过多次实验的结果来看,异常概率为 1 的点其真实数据也肯定异常;同时,在一般情况下,异常概率大的点,其真实值是异常的概率也大于异常概率低的点.因此,认为该异常概率的定义可以较准确地反映缺失数据点的异常可能性.

4 结束语

本文通过分析现有数据补缺方法运用于异常信号检测领域的不足,结合实际提出异常信号领域对于数据补缺的特殊需求,提出了一种新的缺失数据处理方法.该方法通过将正常数据与缺失数据的未缺失项进行联合分析处理,得出两者之间的关系,从而给出缺失数据的异常概率.

(下转第 93 页)