

海洋数据仓库的数据清洗规则应用研究

李郅明^{1,2}, 余荣君^{1,2}, 叶馥榕^{1,2}

(1. 国家海洋局厦门海洋预报台 厦门 361000; 2. 国家海洋局宁德海洋环境监测中心站 宁德 352000)

摘要: 建立海洋数据仓库以实现海洋多源信息的汇集融合已成为趋势, 数据清洗是数据仓库建立过程中的重要环节。文章介绍基于滤波和离差法的数据清洗规则, 并以厦门市海洋水温数据为例, 采用该规则方案对异常值进行批量探测和修正, 得到较好效果, 可为数据挖掘技术的研究和应用提供支持。

关键词: 滤波; 离差法; 数据仓库; 数据清洗; 数据挖掘

中图分类号: P7

文献标志码: A

文章编号: 1005-9857(2018)02-0094-04

Application of Data Cleaning Rules of Marine Data Warehouse

LI Zhiming^{1,2}, YU Rongjun^{1,2}, YE Furong^{1,2}

(1. Xiamen Marine Forecast Station, SOA, Xiamen 361000, China;

2. Ningde Marine Environmental Monitoring Center, SOA, Ningde 352000, China)

Abstract: Building marine data warehouse to realize multi-source marine information fusion has become the developing trend. Data cleaning is an important issue for the whole marine data warehouse. In this paper, a processing method based on filter and interpolation to clean data was introduced. With the benchmark of water temperature in Xiamen, the processing method was applied to detect outliers. The results proved the efficiency of the proposed method, which would provide support for the research and application of marine data in the future.

Key words: Filter, Interpolation, Data warehouse, Data cleaning, Data mining

经过多年的海洋调查和信息收集, 我国已拥有大量珍贵的海洋科学数据, 涉及海洋水文、海洋气象、海洋生物、海洋化学、海洋资源、海洋地球物理、海洋基础地理以及海洋航空和卫星遥感等学科领域。这些海洋数据具有类型多样、获取手段多样、存储介质多样、空间特征强、尺度丰富、数据量大和动态更新频繁等特点, 建立海洋数据仓库以实现多源信息的汇集融合是大势所趋。

然而海洋环境的复杂性和观测仪器设备的不稳定性常导致出现数据异常值, 面对种类繁多、结构复杂、表现形式和存储格式各异以及质量问题严重的大量海洋数据, 为保证进入数据仓库的数据质量, 有必要建立数据清洗规则, 以过滤和修正不符合要求的异常数据^[1]。

1 海洋数据仓库的建立

以国家海洋局厦门海洋预报台建立的海洋数

收稿日期: 2017-05-24; 修订日期: 2018-02-03

基金项目: 国家海洋局东海分局青年科技基金项目(201525)。

作者简介: 李郅明, 工程师, 研究方向为海洋观测数据通信与应用

通信作者: 余荣君, 助理工程师, 研究方向为海洋水文气象

据仓库为例,其总体思路为:以厦门市海域周边的多年海洋观测数据为基础,整理种类繁杂的现有原始数据,根据数据的不同特点和情况,利用不同的数据挖掘方法对同一数据集进行处理,对比选取最优结果,从而构建分类、分级的海洋数据仓库。

具体来说,将现有原始数据分为 2 个部分,分别做定型集和测试集;分析现有原始数据,根据分析结果和挖掘目标对各种数据挖掘方法进行评估,选择合适的方法对定型集数据进行数据处理;利用测试集数据对采用的数据挖掘方法进行验证,经过对比分析获取最优结果。其技术路线如图 1 所示。

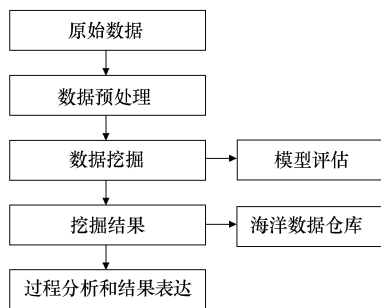


图 1 海洋数据仓库建立的技术路线

2 数据预处理——数据清洗

数据清洗是数据预处理过程的第 1 步,即清除错误和不一致的数据,同时需要解决孤立点和元组重复问题,主要包括范围检验和重新格式化,以去除源数据中不规范的部分。

数据清洗的主要任务是过滤或修正不符合要求的数据。不符合要求的数据主要包括不完整数据、错误数据和重复数据 3 大类。数据清洗不仅检查字段或字段组的存储格式,而且检查字段中数据的有效值。国外对数据清洗的研究起源于 20 世纪 50 年代的美国,研究主要集中在 5 个方面:①异常数据的检测和处理;②重复数据的检测和处理;③面向特定领域的的数据清洗;④与领域无关的数据清洗;⑤数据集成。我国开展相关研究较晚,主要研究面向特定领域的的数据清洗。

通常可用一些预先定义的规则(算法)对数据进行清洗,清洗规则将保存在数据仓库中由 ETL 系统调用。ETL 系统首先调出指定的清洗规则,再由规则解析模块进行解析并加入清洗队列,抽取的数

据经过队列中的每个规则进行清洗,最后输出干净的数据集。对海洋数据仓库的数据清洗动作主要包括探测异常值和修正异常值 2 个部分。

2.1 探测异常值——滤波

由于海洋数据探测周期短和时间序列长,而异常值的出现常常没有规律可循,人工探测存在极大困难。近年来滤波被广泛应用,其在时间序列平滑和异常值探测中有显著效果。海洋数据庞大,信息细节较多,且由于天气和周边环境等因素影响,时而会出现时段性的探测误差,产生具有连续性的异常值。传统的线性滤波在处理异常值的过程中难以保留细节信息,对于具有连续性的异常值信息呈现不敏感性,不能满足海洋数据的异常值探测要求。目前常用的非线性滤波主要包括中值滤波、最小值滤波和基于序统计的 L 滤波等^[2-3],本研究主要考虑使用中值滤波模型和最小值滤波模型。

中值滤波模型的公式为:

$$\hat{d}(k) = \text{mediumn}\{d(k-N), \dots, d(k), \dots, d(k+N)\}$$

式中: $d(k)$ 表示探测序列中的 k 时刻的探测值; $\{-N, \dots, 0, \dots, N\}$ 表示滤波窗口,窗口长度 $L = 2N+1$; mediumn 函数表示取序列中值。

最小值滤波模型的公式为:

$$\hat{d}(k) = \text{miniumn}\{d(k-N), \dots, d(k), \dots, d(k+N)\}$$

式中: miniumn 函数表示取序列中的最小值。

2.2 修正异常值——离差法

海洋数据探测记录的异常值常表现为异常大值或异常小值,而在实际记录中常无法确定,因此在判定异常数据时选择离差法^[4-5]。

如果某个时刻的探测记录值相较前一时刻或后一时刻有较大偏差时,则认为该时刻的记录值有较大概率为异常值。利用 Δd 来判断记录值是否为异常值,离差定义为:

$$\text{左离差 } \Delta d_L(k) = |d(k) - d(k-1)|$$

$$\text{右离差 } \Delta d_R(k) = |d(k) - d(k+1)|$$

$$\text{离差 } \Delta d(k) = \Delta d_L(k) + \Delta d_R(k)$$

已利用 Δd 来判断记录值是否为异常值。本研究给定的判定方法如下。

令窗口标准差为:

$$\hat{\sigma} = \frac{1}{2N_j} \sum_{j=-N}^N \sqrt{[d(k-j) - \bar{d}(k)]^2}$$

$$\bar{d}(k) = \frac{1}{2N+1} \sum_{j=-N}^N d(k-j)$$

当 $\Delta d(k) > 4\hat{\sigma}$ 时, 则 k 时刻的记录值为异常值。

3 实例研究

以厦门市海洋水温数据为例, 对异常值进行探测和修正(本实验将异常值修正为前一个非异常时刻的记录值)。在处理水温数据的过程中遵循 2 个原则: ①保留正常的探测值, 对于正常范围内的水温波动情况, 不应进行处理; ②对于连续的异常值, 应保留数据并进行人工处理。

在对每个时刻进行异常值验证的过程中, 考虑到之前验证过的时刻的数据已经过滤波处理, 具有较高的可靠性, 可利用优化后的数据进行离差计算。经处理的 2016 年 1 月 1 日 0 时至 2016 年 1 月 31 日 0 时的厦门市海域水温数据序列如图 2 所示。

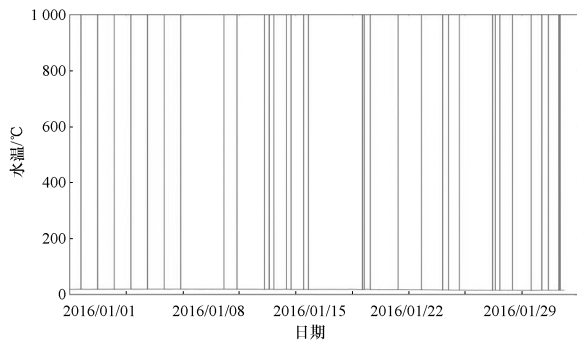


图2 厦门市海域水温数据序列

经过采用上述离差法探测异常值并修正后, 得到时间序列如图 3 所示。由图 3 可以看出, 大多数异常值都可被探测并修正。

经过观察数据, 由于数据中包含较长序列的异常值(连续 10 次)以及在计算窗口标准差时窗口设置过大, 导致窗口标准差无法准确区分异常值和正常值。而对于连续的异常值应保留数据并进行人工处理, 因此在采用离差法发现异常后应采用中值滤波, 滤波窗口的长度可影响要保留的连续长度。

在确定窗口长度时, 当 7 个数据中有 4 个及以上的连续异常值时, 中位数则取异常值; 当数据中

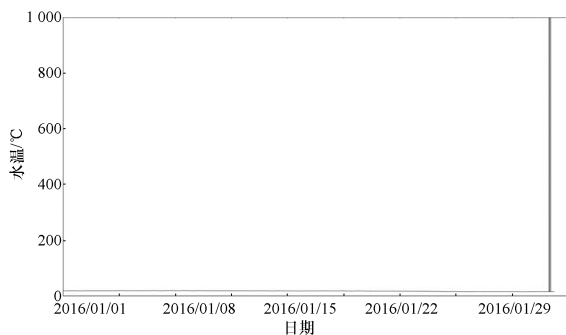


图3 经探测并修正后的水温数据序列

只有 3 个连续异常值时, 中位数则取 1 个正常值。同样, 在中值滤波模型中仍采用之前优化后的值, 模型为:

$$\hat{d}(k) = \min\{\hat{d}(k-3), \hat{d}(k-2), \hat{d}(k-1), d(k), d(k+1), d(k+2), d(k+3)\}$$

对应离差计算公式为:

$$\Delta d_L(k) = |\hat{d}(k) - \hat{d}(k-1)|$$

$$\Delta d_R(k) = |\hat{d}(k) - d(k+1)|$$

$$\Delta d(k) = \Delta d_L(k) + \Delta d_R(k)$$

综上所述, 探测和修正海洋水温数据时间序列, 并保留连续异常值的方案分为 2 个步骤。

(1) 计算离差 $\Delta d(k)$ 和窗口标准差 $\hat{\sigma}$ 。如果 $\Delta d(k) < 4\hat{\sigma}$, 则认为 $d(k)$ 是正常值, $\hat{d}(k) = d(k)$ (滤波窗口长度 $L = 1$), 继续步骤(1), 探测 $d(k+1)$; 否则, 执行步骤(2)。

(2) 取 $\hat{d}(k) = \min\{\hat{d}(k-3), \hat{d}(k-2), \hat{d}(k-1), d(k), d(k+1), d(k+2), d(k+3)\}$, 重新计算离差 $\Delta d(k)$ 和窗口标准差 $\hat{\sigma}$ 。如果 $\Delta d(k) < 4\hat{\sigma}$, 则认为 $d(k)$ 是正常值, $\hat{d}(k) = d(k)$; 否则, 修正 $\hat{d}(k) = \hat{d}(k-1)$ 。

通过上述方案可得到时间序列如图 4 所示, 可保留连续的异常值。

由于水温值一般在 $0 \sim 100^\circ\text{C}$, 可预设所有不在该区间内的记录值都为异常值并将其修正。对于首个时刻只取其右离差, 对于末个时刻只取其左离差, 对于第 2 个和倒数第 2 个时刻取窗口长度为 3 的窗口标准差进行计算。通过该方案可得到时间序列如图 5 所示, 该时间序列值表现较好。

由于原时间序列的异常值是 999.8 或 999.9 这

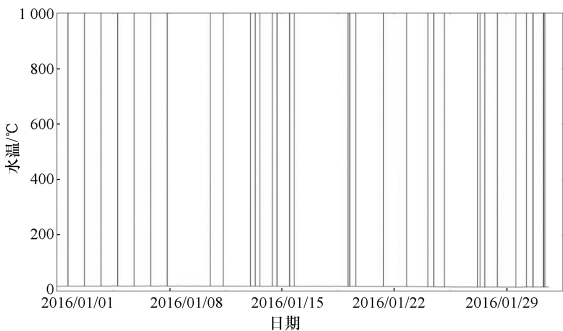


图 4 保留连续异常值的水温数据序列

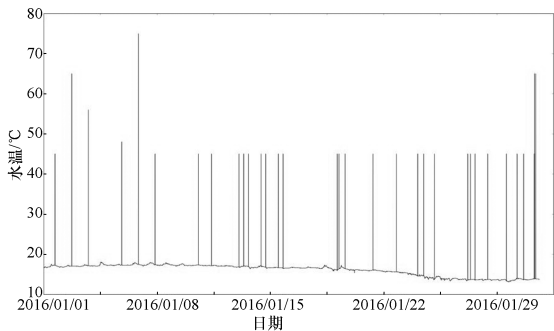


图 7 经探测和修正保留连续出现的异常值的水温数据序列

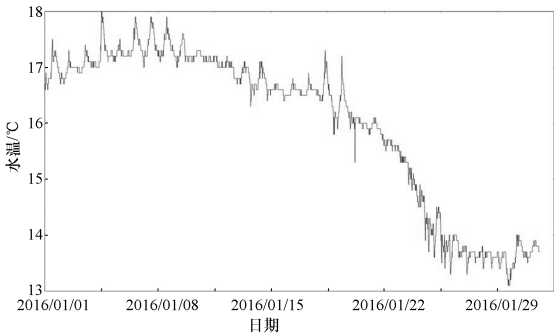


图 5 经方案处理的水温数据序列

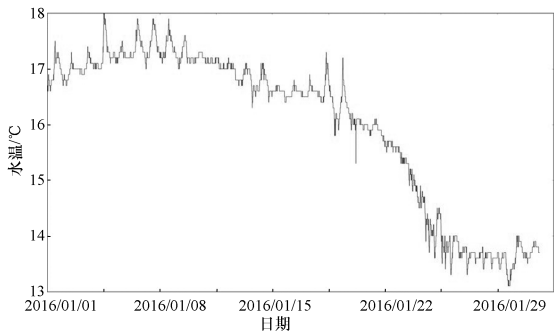


图 8 消除异常值的水温数据序列

样的较大值,为验证方法有效,随机将序列中的一些数据调至较大的值,通过采用离差法进行探测和修正并保留连续 3 次出现的异常值,并消除所有异常值,得到的序列分别如图 6、图 7 和图 8 所示。可见,使用模型可成功消除异常值。

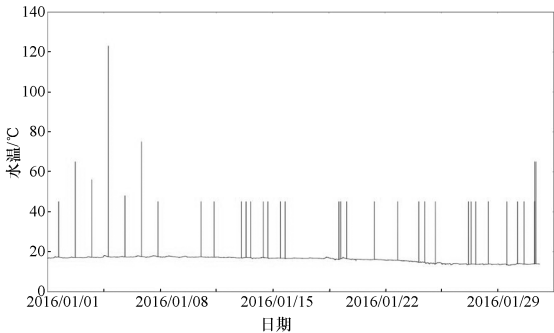


图 6 随机调整数值后的水温数据序列

4 结语

目前的海洋研究往往需要多学科和多数据源间的协作,而海洋领域的数据挖掘技术研究和应用较其他领域更为复杂,其业务化推广还存在许多困难,因此应建立数据仓库,以实现多源异构数据源

的共享和互操作以及多学科的综合分析。数据清洗作为数据挖掘的首要步骤,也是数据仓库在数据准备阶段的重要工作之一,对于把关数据质量十分重要。通过初步建立数据清洗规则并将其应用到样本数据,可帮助不断修改完善已有规则,直到对所得结果满意为止;之后将这些规则应用到整个数据集中,从而建立数据质量更高的数据仓库,为海洋综合研究提供更有力的数据支持。

参考文献

[1] 王欣,王得隽.离散信号的滤波[M].北京:电子工业出版社,2002.

[2] 李明叁.序统计滤波估计检测海洋探测异常数据[J].海洋测绘,2007,27(1):7-11.

[3] 邱越峰,田增平,季文赞,等.一种高效的检测相似重复记录的方法[J].计算机学报,2001,24(1):69-77.

[4] 夏登文,石囊祥,于戈.海洋数据仓库及数据挖掘方法研究[J].海洋通报,2006,24(3):60-64.

[5] 俞荣华,田增平,周傲英.一种检测多语言文本相似重复记录的综合方法[J].计算机科学,2002,29(1):118-121.