第 34 卷 第 21 期 计 算 机 工 程 2008 年 11 月
Vol.34 No.21 Computer Engineering November 2008
·博士论文· 文章编号：1000—3428(2008)21—012—03 文献标识码：A 中图分类号：TP311

基于置信区间的偏离群数据检测方法

夏秀峰，谢光宇，石祥滨，徐 蕾
(沈阳航空工业学院计算机学院，沈阳 110136)

摘 要：异常数据检测与处理是数据仓库系统中数据清洗领域的研究热点。该文提出一种基于置信区间的偏离群数据检测方法，从总体中筛选出有效样本，利用遗传算法从中找到可信样本，利用可信样本确定置信区间，基于置信区间对总体进行检测及处理。该方法所处理的数据不需要与时间相关，且可以快速地识别、检测出大数据量中的“脏数据”。实验结果表明，该方法能有效地解决无规则状态下的偏离群数据的检测，并在实际应用中取得了良好效果。

关键词：脏数据；置信区间；偏离群数据；遗传算法

Detection Method of Deviated Group Data Based on Confident Interval

XIA Xiu-feng, XIE Guang-yu, SHI Xiang-bin, XU Lei
(School of Computer, Shenyang Institute of Aeronautical Engineering, Shenyang 110136)

【Abstract】 It is a hot topic to detect and dispose the exceptional data in the field of data-cleansing operation of data warehouse system. After analyzing the current detection technology, a detection method of the deviated group data based on confident interval is proposed, in which an effective stylebook is screened out from the group data, a credible stylebook is found from the effective stylebook using a genetic arithmetic, a confident interval is obtained based on credible stylebook, then the group data will be detected and disposed using the confident interval. The data disposed by this method can be irrelative to the time scales, and the detection and identification speed of the “dirty data” in a large volume of data is fast. Experimental results indicate that this method can effectively implement the detection of the deviated group data in the random data, and good effects are obtained in practical applications.

【Key words】 dirty data; confident interval; deviated group data; genetic arithmetic

1 概述

数据仓库包含从各种数据源中导入的大量数据，其质量是制约数据仓库应用的“瓶颈”之一[1]。由于组织机构中已经存在的遗留系统(legacy system)[2]建设于不同历史时期，且在实施过程中缺乏全局统一的规划，这些数据在导入到数据仓库之后，不可避免地存在一定量的冗余、不一致及缺失等现象。因此，数据仓库系统需要功能强大的数据清洗工具来检测并消除上述“脏数据”，最终为分析决策支持提供正确一致的信息。

“脏数据”中的异常数据的处理一直以来都是数据仓库领域的研究热点。为消除异常数据，文献[3]提出了一种混沌的异常数据动态检测方法，可以较好地解决异常数据分析中的“屏蔽效应”及异常数据识别问题。但该方法所分析的数据必须与时间相关(如跟踪电厂的用电量，用于检测某段时间内的异常点)。文献[4]引入了数据支持量和有效支持量的概念，针对不同分类数给出了滤波阈值，该算法可以消除异常数据的影响。但该方法同格拉布斯检验法(1959 年提出)和狄克逊检测方法(1953 年提出)一样，一般用于检测数据异常点。

本文提出了一种基于置信区间的检测数值型“脏数据”的方法，能够较好地实现数据量巨大且偏差并不明显的数值型“脏数据”检测，源数据中含有非数值型数据和缺失数据(空值)不影响算法的检测：(1)该方法所检测的数据与时间因素线性无关。(2)不必先判断各个异常点后再进行检测，而是采取一次处理批量检测的方法。本文的例子来自于某市特种设备监督检验数据仓库系统。限于篇幅，仅对某一种机电类特种设备——电梯的层、站、门数进行数据清洗。实际应用证明了本文算法的可行性和有效性。

2 脏数据检测

2.1 检测原理

通常，对数据量巨大且偏差不明显的数值型“脏数据”采用制定相应规则[3]的方法予以解决。但当没有相关的规则去检测大量数值型数据中含有部分“脏数据”时，用户在处理过程中会遇到无法判断有效取值范围的困难，进而影响数据抽取、转换、装载及清洗(extract transform load and cleansing)[5]过程中抽取数据的质量。

本文方法利用数理统计中置信区间的思想，解决总体服从一定区间分布且没有相应规则来检验“脏数据”问题。为准确描述检测策略，下面给出偏离群数据的基本概念。

定义 1 (偏离群数据) 设总体 ε 中的每个元素均为数值型数据，$\xi_1,\xi_2,\cdots,\xi_n$ 为来自其中的相互独立、服从同一分布的

基金项目：辽宁省自然科学基金资助项目(20052007)
作者简介：夏秀峰(1964－)，男，教授、博士，主研方向：数据仓库理论与技术；谢光宇，硕士研究生；石祥滨，教授、博士；徐 蕾，教授
收稿日期：2008-02-12 **E-mail：**xiaxiufeng@syiae.edu.cn

随机变量，ε_i 为 $\xi_i(i=1,2,\cdots,n)$ 的观测值，对于给定的特定数 α_1,α_2 且 $\alpha_1<\alpha_2$，若 $\varepsilon_i\notin[\alpha_1,\alpha_2]$,那么该数据称为总体中的偏离群数据。

一般地，在基于 RDB 的数据仓库系统中，总体 ε 对应一个 RDB 文件中的一个数值型属性的所有值，如：乘客电梯档案表中“电梯层数”属性的所有值。假定某地区的乘客电梯层数都应该在[6,26]区间内，则不在该区间内的其他数据就是偏离群数据。

2.2 检测策略

根据检测处理的数据类型和实际的业务需求进行分析，对偏离群数据的检测及检测后的处理可分为如下 3 个步骤：

(1)样本的初步抽选

样本数量对预测精度的影响，目前尚无成熟的理论或结论，但所抽取的数据量必须大于数理统计中的可信取样数 45[6]。根据切比雪夫定理[6]可知，当样本数据量足够大时，样本数据的算术平均值与总体平均值近似相等，即

$$\lim_{n\to\infty}P(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i-\mu\right|<\varepsilon)=1$$

从总体中抽取样本后，需要判断所抽取的样本数据的有效性，即首先判断样本数据是否存在缺失、非数值型和超过初步筛选区间的无效数据。不同类型的无效数据有不同处理方式，对于缺失型和非数值型数据直接作为不放回抽样的“脏数据”。其他数值型数据需要对其进行初步筛选，有利于遗传算法选取优良基因。由于在初步筛选样本期间不可能准确获得样本的取值范围，因此可以采取设定阈值和手工选取两种方法实现样本的初步筛选。

用于初步筛选的最大阈值和最小阈值由用户设定，当数据在最大阈值和最小阈值之间时，可以认为所选取的数据满足筛选条件。但简单的设定阈值可能会造成将过多数据置为有效数据或无效数据，在某些情况下，可以将最大阈值和最小阈值的精度提高。例如，对于统计某地区 20 世纪 90 年代初期乘客电梯高度时，设定最高层数为 20，最低层数为 4。但资料表明该地区同期乘客电梯层数取值范围为(6, 15)，可将最大阈值置为 15，提升了最大阈值精度；同理，将最小阈值设为 6，提升了最小阈值精度。实验证明，将阈值的精度提升后可明显改善遗传算法中选取优良基因的速度，最终达到了加快检测数据质量的速度。

当样本数量不大时，可以选择手工挑选可信样本数据，但当样本数过大时，手工选择费时费力，并不是一种好的解决方法。为使选择可信样本数据的速度加快、精度加强，可以利用遗传算法从有效样本中选择优良的基因作为可信样本中的数据。

(2)利用遗传算法选取可信样本

遗传算法(Genetic Algorithms, GA)是模拟生物进化过程中“适者生存，优胜劣汰”规律而无需函数梯度信息的自适应全局搜索算法[7]。遗传算法中的种群基因构成染色体，通过选择、交叉等操作反复操作，不断找到适应度好的基因，最终找到最优的染色体。设定经过初步选择后的有效样本为群体，群体中的各个数据为基因，经过算法选定后的一组数据为染色体。为更加清晰地描述算法，下面引入两个定义。

定义 2(衡量重心) 设总体 ε 中的每个元素 $\xi_1,\xi_2,\cdots,\xi_N$ 均为数值型变量，$\varepsilon_i(i=1,2,\cdots,N)$ 为 ξ_i 的对应观测值，从这些观测值中随机抽取 n 个数据作为样本 G，依次记为 $x_i(i=1,2,\cdots,n)$，则 $x=\frac{\sum_{i=1}^{n}x_i}{n}$ 称为样本 G 的衡量重心。

一般地，在基于 RDB 的数据仓库系统中，总体 ξ 对应一个 RDB 文件中的一个数值型属性的所有值，随机地从总体中抽取 n 个数据记为样本 G，计算 G 的样本平均值 x，则 x 为 G 的衡量重心。

定义 3(衡量距离) 设总体 ε 中的每个元素 $\xi_1,\xi_2,\cdots,\xi_n$ 均为数值型变量，$\varepsilon_i(i=1,2,\cdots,N)$ 为 ξ_i 的对应观测值，从这些观测值中随机抽取 n 个数据作为样本 G，依次记为 $x_i(i=1,2,\cdots,n)$，μ 为样本 G 的衡量重心，定义 $\lambda_i=x_i-\mu$，则称 λ_i 为 x_i 的衡量距离。

衡量距离是决定基因选取的重要标准，当进行基因替换时，一般选择衡量距离偏大的基因进行替换，确保保留基因遗传的良好效果。

在对有效样本进行选择时，利用遗传算法找到“最好”的基因，即找到可信样本。具体遗传算法描述如下：

1)设定基因个数(记为 m)和替换基因个数(记为 n)，且满足 $m>n$。

2)从有效样本中随机抽取 m 个数据，计算该组数据的衡量重心，将结果返回给用户。

3)由用户判断衡量重心是否在认可的区间内，若是则转 5)，否则转 4)。

4)选取该组数据中衡量距离偏大的 n 个数据，将这些数据进行基因替换，转 3)。

5)该组数据最优的染色体，即为可信样本，算法结束。

(3)置信区间的确定及检测

根据列维定理[6]可知，当独立的随机变量均服从相同的分布且有相同的数学期望和方差时，则随机变量之和的极限服从正态分布。由此可推导出，可信样本的算术平均值 $\bar{x}$ 同样服从正态分布，即

$$\bar{x}\sim N(\mu,\sigma^2/n)$$

同时，可以利用所选取可信样本的观测值计算得到样本均值及样本均方差。

根据用户给定的置信水平，结合可信样本可计算得到总体的置信区间。然后逐条抽取待验数据，判断待验数据是否在置信区间内，对于不在置信区间内的偏离群数据直接转入“脏数据”的处理流程；对于在置信区间之内的数据，流程处理结束，可以将该数据放入数据仓库层。在用户的参与下，对所有“脏数据”选择合适的处理方式，如数据填充、区间提示等。

2.3 检测算法描述

偏离群数据的检测的流程具体描述如下：

(1)从总体 M 中随机抽取数据作为样本，记为 P, P 的数量为 x，且 $x>45$，将 P 中的非数值型数据和空值数据筛选掉。

(2)设定初步筛选阈值 y_1 和 y_2，且 $y_1>y_2$，利用阈值对样本 P 进行再次筛选，得到有效样本，记为 Q, $Q>45$。

(3)设定遗传算法中的基因数 m 和替换基因数的个数 n。

(4)利用遗传算法从有效样本 Q 中找到可信样本 $Q_1,Q_1>45$。

(5)利用所选取的可信样本 Q_1 计算得到样本均值：

$$x=\frac{\sum_{i=1}^{N}x_i}{N}$$

样本均方差：

$$S=\sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}}$$

(6)设定置信水平 μ 的值(一般情况下取 μ 为 0.95)。

(7)利用样本均值、样本均方差和给定的置信水平计算置信区间，即

$$(\bar{x}-\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1),\bar{x}+\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1)),\ (\alpha=1\text{-}\mu)$$

(8)通过查表求得 $t_{\alpha/2}(n-1)$ 的值，进而得到置信区间。

(9)利用置信区间检测总体数据。

(10)将偏离群数据隔离，准备进行下一步处理。

需要说明：μ 取值越大，异常数据的错误判断几率越小，把正确数据混入异常数据的也同样增大，取 0.95 为宜。

3 处理“脏数据”

对于非数值型数据和空值数据，经过初步筛选后该类数据不进入数据仓库层。而对于偏离群数据则根据不同情况采取不同方式。

3.1 数据填充

当“脏数据”属于缺失数据或偏离群数据时，可以通过该记录的其他重要信息字段与含有可信数值的记录内容进行比较，找到相似或相同记录进行数据填充。例如：对于电梯档案表中电梯使用年份、安装年份、使用类型等字段是该表的重要字段，找出与该记录重要字段一致的记录信息进行填充。同理，可以将这些相似、相同记录的数据进行累加求平均值的方法进行数据填充。

3.2 区间提示

对于缺失型“脏数据”可以利用 3σ 原则将数据的区间给予用户提示，在用户的参与下对缺失型数据进行填充。由 3σ 的定义可知

$$P(|\xi-\mu|<3\sigma)=0.997\ \ 3$$

随机变量 ξ 在范围 $(\mu-3\sigma,\mu+3\sigma)$ 的概率小于 0.003，可认为所有可能的可信数据均出现在该范围内。其中，μ 和 σ 分别为可信样本的样本均值和样本均方差。

3.3 查找填充

对于缺失型或偏离群数据型“脏数据”，可以通过查找该条记录的有效联系方式等信息进行确认，重新准确填充该值。

4 实验

利用置信区间分析法可以对数值型数据的数据质量进行评估，在衡量数据质量方面具有较好的效果，下面以实验的形式证明方法的正确性和可行性。

4.1 实验验证

实验中，采用的实验总体数据是某市特种设备监督检验系统中某一年度的乘客电梯层数和电梯高度，样本数据从总体中随机抽取。其中总体数量为 50 000 个，给定置信水平 μ=0.95；使用的操作系统是 Windows XP，后台数据库为 Oracle9*i*，编程语言为 Java。

4.1.1 试验数据

在实验中，通过电梯高度与电梯层数相除来计算电梯的每层高度，判断每层高度是否符合标准。通过对各个地区不同年度的电梯高度变化的统计，可以使决策用户观察到电梯每层高度和楼房建筑楼层的变化情况，以便对未来制订、修改新的乘客电梯高度给予指导和帮助。设定最小阈值为 6，最大阈值为 12 作为楼房层数的初步筛选阈值。表 1 为所选数据样本实例(限于篇幅，未列出全部属性项)。

表 1 样例数据

注册编号	电梯层数	电梯高度	安装年份	使用类型
30302101001991120018	10	31.5	1991	乘客电梯
30302101001991120019	3	40.1	1991	乘客电梯
30302101001991120020	15	46.2	1991	乘客电梯
30302101001991120021	11	30.2	1991	乘客电梯
30302101001991120022	10	30.0	1991	乘客电梯
30302101001991120023	10	29.4	1991	乘客电梯
30302101001991120024	8	25.1	1991	乘客电梯
30302101001991120025	7	22.0	1991	乘客电梯
30302101001991120026	66	18.9	1991	乘客电梯
30302101001991120027	13	41.1	1991	乘客电梯

4.1.2 实验结果

利用本文方法，结合表 1 数据得到的置信区间范围为 2.98~3.21，检测出的偏离群数据的注册编号(后 2 位)分别为 19,22,23,26。

对于不同偏离群数据，本文可以采用不同方法进行处理。编号为 19 的电梯层数可以利用数据填充中的重要信息相似字段对比的方法进行提示，如可填充编号为 27 的电梯层数 13；对于编号 22 的电梯高度同样可以采用此方法进行提示，如可填写编号为 23 的电梯高度 29.4；对于编号 23 的信息，通过人工观察发现该编号的数据并不是偏离群数据(属于错检)，不予以处理；利用区间提示的方法，编号为 26 的电梯层数提示结果为 6~7。

编号为 23 的设备信息出现了错检现象，这是由于所选取的样本数据不合适造成的。通过实验验证，即使通过对样本数据量的调整、初步筛选区间的重选、置信水平的更改等一系列措施均不能完全消除错检现象，因此，只能由人工判别。

4.2 有效性和可行性验证

为验证本文方法的有效性，从总体中抽取 5 组不同数量的样本，为总数的 2%(1 000 个)、4%(2 000 个)、6%(3 000 个)、8%(4 000 个)和 10%(5 000 个)。按照置信水平 μ 和 5 组不同数量的样本，经过检测算法处理后得到不同置信区间，利用这些区间所检测出的“脏数据”数量与利用规则检测出的数量相除，得到的百分比与数量的关系以图表形式展示。

图 1 代表检验“脏数据”的效果比较。其中，纵坐标代表不同置信区间检测的“脏数据”数量与利用规则库检测的“脏数据”数量之比。

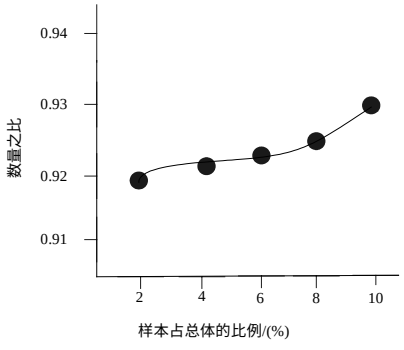

图 1 检测效果

本文的检测算法已经应用到了某市特种设备监督检验系统的数据清洗过程中，通过该方法检测并处理后的数据质量得到了明显地提高，处理后的数据可以直接应用到数据仓库中。综上所述，本文提出的对数值型“脏数据”进行检测的方法能有效地解决无规则状态下，数据量巨大且偏差并不明

（下转第 17 页）