

智能数据分析中的异常数据管理

姚卫新¹ 黄丽华²

(复旦大学管理学院, 博士生¹; 教授、博士生导师² 上海 200433)

[摘要] 数据是信息时代的重要资源, 智能数据分析为发掘它自身的价值起到了很大的作用, 但异常数据却给智能数据分析带来了新的挑战。本文提出了异常数据的管理方法, 包括不同类型异常数据的分类方法、蕴知性异常数据的筛选方法、从源头上减少噪音性异常数据的方法, 以及当前对异常数据进行研究必须采取的措施。

[关键词] 异常数据 管理 智能数据分析 [中图分类号] F27 N94 [文章编号] 1000-7857(2004)07-0054-03

OUTLIER MANAGEMENT OF INTELLIGENT DATA ANALYSIS

YAO Wei-xin HUANG Li-hua

(Management School, Fudan University, Shanghai 200433, China)

Abstract: Data is one of critical resources in the information era. Intelligent data analysis is being challenged by outlier though it can mine valuable information from data. This paper proposes a management method of the outliers which includes the distinction of different outliers, identification of the outliers containing knowledge, reduction of noise the outliers from source. It also points out some measures necessary to be taken in the research of outlier.

Key Words: outlier, management, intelligent data analysis

坚实的相关领域的知识, 才能取得较好的效果。

1 智能数据分析概述

数据已成为信息时代的重要资源, 数据被采集后在企业之间或企业内部的信息系统中共享, 数据量的增加导致高效的基于计算机的分析方法的出现, 如智能数据分析, 它是指运用统计学、模式识别、机器学习、数据抽象等数据分析工具从数据中发现知识的分析方法。智能数据分析的目的是直接或间接地提高工作效率, 在实际使用中充当智能化助手的角色, 使工作人员在恰当的时间拥有恰当的信息, 帮助他们在有限的时间内作出正确的决定。

信息系统中积累的大量数据, 其原始数据的价值很小, 只有通过智能化分析方法抽取其中的精华, 才能转变为信息“金矿”, 为人类造福。智能数据分析方法研究已经有数十年的历史, 研究人员将人工神经网络、贝叶斯网络、决策树、遗传算法、基于范例的推理法、归纳逻辑编程法等智能数据分析方法应用到具体工作中, 先后取得了很大的突破, 解决了许多疑难问题。智能数据分析方法主要为两种类型: 一是数据抽象(Data Abstraction), 主要涉及数据的智能化解释, 以及如何将这种解释以可视化或符号化的形式表示出来; 二是数据挖掘(Data Mining), 主要涉及从数据中分析和抽取知识, 目的是为了支持业务管理或预测趋势。

已经开发出的这两种类型的数据分析系统数不胜数, 或多或少在有关领域取得了一些成果, 但都是从分析正常数据入手。目前许多检测仪器和设备开始向智能化方向发展, 尤其是计算机的使用, 提高了检测效率和效果, 但也出现了新的问题, 即检测出的数据中包含大量的异常数据(Outliers), 这些异常数据与通常的数据模型或数据规律不符合^[1], 对正常工作造成一定的影响。从分析异常数据的角度对数据进行智能化分析还刚刚起步。信息系统中的数据类型有多种多样, 如数值、图像、声频、视频等, 对不同类型的异常数据进行分析不仅要选择恰当的方法, 而且要具备

2 异常数据的处理

2.1 异常数据的检测

异常数据的检测在数理统计和计算机领域研究得比较多, 目前检测异常数据的方法大致分为以下几种。

2.1.1 基于统计的异常数据检测方法^[2] (Statistics-Based Outlier Detection) 即假设所给定的数据集存在一个分布或概率模型(如正态分布), 然后根据相应模型并通过不一致性测试来发现异常数据。应用这种测试需要了解数据集参数的有关知识(如数据分布情况)、分布参数知识(如均值和方差), 以及所预期的异常数据个数。

2.1.2 基于距离的异常数据检测方法^[3,4] (Distance-Based Outlier Detection) 即假设一个数据集 S 中的对象至少有 p 部分与对象 o 的距离大于 d , 则对象 o 是一个带参数 p 和 d 的基于距离的异常数据, 记为 $DB(p, d)$ 。换句话说, 将基于距离的异常数据看作是那些没有“足够多”邻居的对象, 这里的邻居是基于距离给定对象的距离来定义的。

2.1.3 基于偏差的异常数据检测方法^[5] (Deviation-Based Outlier Detection) 即不采用统计检验或基于距离的度量值来确定异常对象, 而是通过检查一组对象的主要特征来确定异常数据, 偏离特征描述的对象就被认为是异常数据。

2.1.4 高维数据异常检测方法^[6] (High-Dimension Outlier Detection) 即把高维数据集映射到低维子空间, 根据子空间映射数据的稀疏程度来确定异常数据是否存在。

这些检测方法各具有不同的优点和缺点, 对于不同类型的异常数据, 需要选择不同的检测方法, 以便取得最佳效果。

2.2 异常数据区分

有些异常数据是由检测仪器和操作者行为造成的测量性异

常数据 (Metrical Outliers), 是一种噪音性数据。这样的异常数据可能产生于检测仪器不完善或操作者行为不稳定, 甚至数据录入过程中也会输入错误的数值。如在一种新颖的青光眼电脑视野测试仪的使用过程中, 受试者对仪器不适应、眼睛疲劳、情绪不集中等原因都可能产生测量性异常数据。这样的异常数据是不受欢迎的, 需要尽量排除。

有些异常数据却是由事件本身的变化引起的, 包含有意义信息, 但用目前的知识是无法解释的, 是一种蕴涵新知识的蕴知性异常数据 (Knowledge-Containing Outliers)。同样以青光眼电脑测试为例, 由于目前对青光眼的认识并不完整, 而且该疾病本身具有长期进行性的特点, 所以, 有时采集到的数据即使是有意义的蕴知性异常数据, 但用目前的青光眼知识却无法进行判断。这样的异常数据要尽量保留以便作进一步分析。例如, 在 1880 年, 英国物理学家 Rayleigh 用不同方法测量氮时, 发现密度值有细微的差异, 即从空气中测得的氮密度大, 因而他认为空气中一定包含少量的其它气体。这个猜测最终导致氦气的成功分离, 他因此于 1904 年获得诺贝尔奖^[7]。

无论是测量性异常数据还是蕴知性异常数据, 它们总是混杂在一起, 因此需要找到合适的方法对它们进行区分, 以便提高分析的质量或发现新的知识。目前还没有比较成熟的方法可以解决这一问题, 但可以从以下两个思路来进行探索。

一是建立一个“测量性异常数据模型”, 凡符合此模型的异常数据被认为是由于检测仪器不完善或操作者行为不稳定造成的, 可以被排除掉, 而保留那些不符合此模型的异常数据, 以便作进一步分析。

二是建立一个“蕴知性异常数据”模型, 保留符合该模型的异常数据, 而排除那些不符合此模型的异常数据。

通过这两个步骤, 就保留了大部分的“蕴知性异常数据”, 再运用已知的知识, 通过合理的假设, 在实践中进行检验, 很可能开启发现新知识的大门。

2.3 从源头上减少测量性异常数据

如果数据质量提高了, 那么分析结果会更好, 也更有可能发现新知识。因此需要尽可能减少测量性异常数据, 而减少的最好方法是从源头上, 即在数据采集阶段采取有力措施避免它出现。

首先对测试仪器的各个部件进行完善, 保证产生的数据具有极高的可信度, 同时在测试过程中按照操作规程细致地完成每一个步骤。

由于操作者个人的行为波动是产生测量性异常数据的主要原因之一, 因此要尽量保证操作者的心绪平静。友好的测试环境、测试界面和测试方法有助于操作者集中精力, 减少分心的机会。

智能化的测试界面也有助于减少行为因素引起的噪音数据。可以根据不同的情况, 选择表示不同情绪的图形、符号或音乐来调节操作者的注意力和情绪, 让操作者完全融入测试过程中, 以便采集到最真实的数据。

选择合理的重复测试次数也能提高数据的质量, 减少测量性异常数据。可以根据操作者在测试中的表现增加或减少测试次数, 当然要做到这一点不仅需要丰富的主观经验, 也需要制定一个客观的衡量标准。

3 异常数据分析的实际运用

对异常数据的分析还处在起步阶段, 对其进行管理和分析的方法需要不断摸索, 但可以借鉴已经具备的业务知识和其它学科的知识, 尽快制定有关标准, 聚集各方力量开展工作。例如在基于

电脑的青光眼检测仪器的实验中, 利用获得的异常数据, 大大提高了早期诊断青光眼的能力, 同时发现以前对导致眼睛失明的盘尾丝虫病认识存在不足。以前认为盘尾丝虫病和青光眼在视网膜的损伤方面情况相似, 而通过异常数据的分析和实践的检验, 发现情况并非如此, 因此丰富了有关该疾病的医学知识。又如 2003 年上半年流行的 SARS, 起初将其认定为普通肺炎进行治疗, 没有取得明显效果, 在病例大量增加后才引起有关部门和人员的重视。在采集了许多病例并进行分析后, 采取新颖的治疗方法, 才及时控制了该疾病的蔓延。这实际上包含着对异常数据的分析。目前对该疾病还有许多疑问, 如为什么治愈后的 SARS 病人容易股骨头坏死? 为什么华人受感染的群体中, 死亡比例比其他种族高? 为什么儿童不容易受感染? 对这些异常现象常规的医学知识目前还很难解释, 可以尝试从监控体系中获得的异常数据入手, 结合现有的医学知识, 进行合理的推理, 也许能找出答案。

在许多商业领域中, 异常数据具有显著的商业价值, 如客户流失、银行的信用卡欺诈、电信中移动话费拖欠等。通过对异常数据的分析可以迅速准确地甄别这些异常事件。

俄罗斯科学院地磁电离层和电波传播研究所利用等效偶极源方法对 MAGSA 卫星地磁总强度异常数据进行分析, 得到了亚洲地区地壳内部的视磁化强度分布图, 表明磁化强度异常分布与地质构造具有较好的相关性。

地球表层的土壤是由多种不同矿物质的混合物组成的, 其中一小部分 (主要是氧化铁) 因为含铁而具有磁性。人类扰动原生土壤的活动会改变原生土壤的铁含量及其磁性, 所以古代的道路、沟渠以及砖石结构等遗迹的磁性, 都与周围原生土壤的磁性有着一定的差异。这种差异可以通过磁力仪测定和记录下来, 由此获得异常数据。通过对异常数据的判读, 就能认识地下遗迹的状况。

异常数据对人类是一个持续的挑战, 需要我们通力合作、加强研究、运用创新的方法, 寻找合适的途径加深对其的认识。在这一点上, 计算机技术的发展为我们提供了很好的帮助, 它的调整运算能力和大容量存储能力可以加速对异常数据的研究。

4 存在的问题和解决方法

对于异常数据, 目前存在着态度冷漠、缺乏标准、研究力量分散的问题。为推动异常数据的分析研究, 可从以下方面入手。

4.1 改变对异常数据的态度

由于异常数据在实际工作中造成许多不便, 且很多异常数据是不具备任何意义的噪音性数据, 因此人们往往对其采取冷漠和摒弃的态度, 将有意义的异常数据也一起删除, 从而丧失了发现新知识的机会。量变才能导致质变, 很难从单个或少量的异常数据中分析出有意义的知识, 但可以从细微处入手, 不断积累, 将工作过程中出现的异常数据进行收集, 并将出现异常数据的案例进行详细备案, 在积累到一定数量后, 结合已经具备的相关知识, 进行合理的推断, 再进行必要的实践, 一定能获得令人欣喜的收获。

4.2 制定统一标准

目前分散在各个企业的信息系统基本没有统一的标准, 各自采集的数据难以共享, 给大规模数据分析的开展造成了人为的障碍, 因此有必要加快制定统一的标准。这些标准包括数据共享的标准、数据化条款; 术语、词汇和格式的标准; 数据的可视化标准; 不同数据源之间的接口标准; 异种数据类型 (包括图像、信号) 的集成标准。

制定这些标准是一项困难的任务, 但必须拿出高屋建瓴的思路、迎难而上的勇气, 为建设高效实用的信息系统打下基础, 也为异常数据的分析开辟一条捷径。

中国农地金融制度体系的基础制度构建^{*}

李爱喜¹ 罗剑朝²
(浙江财经大学¹ 杭州 310012;西北农林科技大学² 陕西杨凌 712100)

[摘要] 根据我国目前农地金融制度体系存在的有关制度缺陷,提出了建立农地金融制度的基础制度:(1)改革农地经营方式,明确农村土地所有权,稳定土地承包权,促进农地使用权合理流转,健全土地产权法规;(2)采用避免风险、控制损失与转嫁风险等措施,防范农地金融风险;(3)建立科学的农地评估体系,提高评估从业人员素质,正确处理政府、评估师协会、评估机构和评估师的关系,促进评估机构开展公平竞争,允许国外土地评估机构进入中国评估市场。
[关键词] 农地金融 金融风险 农地经营方式 土地评估 [中图分类号] F3 F8
[文章编号] 1000-7857(2004)07-0056-03

THE BASIC CONSTRUCTION OF RURAL LAND FINANCE SYSTEM IN CHINA

LI Ai-xi¹ LUO Jian-chao²
(1. Zhejiang University of Finance & Economics, Hangzhou 310012, China)
2. Northwest Sci-Tech University of Agriculture and Forestry, Yangling, 712100 China)

Abstract: The essay puts forward the idea that the basic construction of rural land finance system should be established in China according to the limitation in current rural land finance system: to innovate the rural management method, to make sure the ownership of the land and stabilize the land contract right, to accelerate the land use reasonably and perfect the land property regulations; to take some measures such as avoiding risk, controlling loss and imputing to others, to keep away the rural land finance risk; to build up a scientific evaluating system, to improve the employees' quality, to deal exactly with the government, evaluation association, institution and appraisers so as to inspire the fair competition and allow foreign institutions to enter China.
Key Words: rural land finance, financial risk, rural land management method, land evaluation

1 农地金融制度概述
在世界农业发展的历史进程中,农地金融制度作为农村金融体系不可分割的一个部分,在促进农地资源合理配置、增加农业
* 本文为国家自然科学基金资助项目(项目编号 70141024)。

信贷投入、提高农业效率与竞争力等方面发挥了十分重要的作用。随着世界经济一体化趋势的加强,中国农业不可避免地要卷入到国际农业发展的合作与竞争当中。因此,借鉴国外农地金融制度的成功经验,构建适合中国国情和农情的农地金融制度,对推动中国农村的制度创新和管理创新,具有现实和深远的意义。

4.3 联合攻关、政府推动
企业的员工虽然拥有丰富的业务知识,但一般缺乏分析异常数据的知识和工具。他们即使积累了大量的异常数据,却往往不知道如何分析这些异常数据。因此,除了员工的积极参加外,还需要计算机专家的参与,寻找分析异常数据的最佳方法。
从异常数据中分析出的知识最初只有理论上的意义,要使其能被真正应用到实际工作中,还必须经过实践的检验。因此还需要依靠社会的力量,在有关单位进行大量的实验,在实际中不断改进异常数据的分析方法,提高所获知识的质量。
由于社会力量一般比较分散,也比较薄弱,仅仅靠他们的力量很难取得大的作为。因此,依靠政府的强势力量,制定统一的规划,进行必要的投入,协调各方的力量,才能早出、多出成果。

参考文献
[1]D.Hawkins.Identification of Outliers.Chapman and Hall,London, 1980
[2]V.Barnett,T.Lewis.Outliers in Statistical Data.John Wiley and

Sons,NY 1994
[3]E.Knorr and R.Ng. Finding Intensional Knowledge of Distance-based Outliers.VLDB Conference Proceedings,1999
[4]E.Knorr and R.Ng. Algorithms for Mining Distance-based Outliers in Large Data Sets.VLDB Conference Proceedings,1998
[5]Rakesh Agrawal,Prabhakar Raganan.A Linear Method for Deviation Detection in Large Databases. KDD Conference Proceedings,1995
[6]J.C.Aggarwal and P.Yu.Outlier Detection for High Dimensional Data.In Proc.of ACM SIGMOD'2001
[7]Brachman,R.J.and Anand, T.,The process of knowledge discovery in databases:a human-centered approach,in Advances in Knowledge Discovery and Data Mining (eds.U.M.Fayyad,G. Piatetsky-Shapiro,P.Smyth and R.Uthurusamy),1996:37~57, AAAI Press/MIT Press

(责任编辑 肖庆山)