当代学术文丛

量子聚类算法的研究及其在异常检测中的应用
基于语义属性数据离群聚类的异常检测
基于趋势的时间序列数据析取之异常检测算法
基于语义属性数据核分类方法的异常检测

Anomaly Detection Algorithms in non-Numerical Data

非数值属性数据异常检测算法

李志华　张海涛
孙　雅　耿振民　编著

江西人民出版社
Jiangxi People's Publishing House
全国百佳出版社

非数值属性数据异常检测算法

Anomaly Detection Algorithms in non-Numerical Data

李志华　张海涛
孙　雅　耿振民　编著

江西人民出版社
Jiangxi People's Publishing House
全国百佳出版社

图书在版编目(CIP)数据

非数值属性数据异常检测算法 / 李志华等编著.
—南昌: 江西人民出版社, 2015.12
ISBN 978-7-210-08140-1

Ⅰ. ①非… Ⅱ. ①李… Ⅲ. ①数据检测 Ⅳ.
①TP274

中国版本图书馆 CIP 数据核字(2015)第 305750 号

非数值属性数据异常检测算法
李志华 张海涛 孙雅 耿振民 编著
责任编辑:徐 旻 李月华
出 版:江西人民出版社
发 行:各地新华书店
地 址:江西省南昌市三经路 47 号附 1 号
发行部电话:0791-88629871
编辑部电话:0791-86898815
邮 编:330006
网 址:www.jxpph.com
E-mail:gjzx999@126.com web@jxpph.com
2015 年 12 月第 1 版 2015 年 12 月第 1 次印刷
开 本:787 毫米×1092 毫米 1/16
印 张:14.75
字 数:200 千字
ISBN 978-7-210-08140-1
赣版权登字-01-2015-895

定 价:36.00 元
承 印 厂:南昌三联印务有限公司
赣人版图书凡属印刷、装订错误,请随时向承印厂调换

前　言

计算机技术、通信技术、网络技术的迅猛发展将人类带进了大数据时代。

数据挖掘(Data Mining)是对所观测的数据对象进行分析,以便找到这些数据对象中便于理解且有用的某种模式、新规律、新知识的过程。近年来,在该领域虽然新的和改进的算法不断涌现、应用领域也正在逐渐扩大,但是同样面临许多新问题,其中更加丰富的数据类型如流、时间序列、图、音频、视频和图像等,这些大多数是非数值属性数据,需要改变传统的数据挖掘算法去适应针对这些数据类型数据的挖掘需要。当然,丰富数据挖掘领域中的算法和技术,以便适应更加广泛的应用需求,依然是数据挖掘发展的主旋律。

异常(Anomaly)指的是在数据集中偏离大部分数据,即这些偏离数据与大部分数据存在明显的差异。人们研究发现这些偏离的数据并非全部由随机因素或偶然误差而产生,而是产生于某些完全不同的机制,所以这些异常数据中通常隐藏有有价值的重要信息和可以帮助人们深入认知被检测对象的重要线索。异常检测(Anomaly Detection)是数据挖掘领域研究的最基本问题之一,主要用来发现数据集中与其他数据存在明显不同的异常。

虽然,数据挖掘研究与应用在近年来得到了前所未有的快速发展。但是不论是数据挖掘还是异常检测都需要进一步与数理统计、神经网络、模式识

别、机器学习、知识库系统、高性能计算和可视化处理技术等学科领域相结合，特别是与这些前沿学科中出现的新理论和新技术紧密结合，形成学科交叉，面对大数据的挑战。

为了进一步推动异常检测技术向非数值属性数据等其他领域拓展，以适应不断涌现的新型应用，并反映该领域的最新研究成果，让相关专业的研究生、高年级本科生和工程技术人员比较深入地理解并快速掌握新兴的异检测技术，特编写此书。本书的另外两名作者是本人指导的硕士，他们协助完成本书部分章节的工作。本书撰写过程中融合了作者在攻读博士学位期间所做的主要研究成果和近几年在时间序列异常检测领域的研究工作，同时也吸纳了国内外具有代表性的研究成果。并且在撰写过程中，围绕异常检测这个主题，从不同角度进行展开，力求以点带面。

本书共分八章。第一章主要介绍异常检测的相关概念、检测方法的研究现状、检测方法的评估指标和经典的异常检测方法等；第二章主要介绍了常见非数值属性数据如语义数据、异构数据、时间序列数据等的概念、特点、表达形式、度量方法等；第三章主要介绍了量子理论、量子理论中的粒子分布机制与数据样本分布的同质性分析，以及基于量子理论的聚类、分类等数据挖掘算法和基于这些数据挖掘算法的异常检测方法；第四章主要通过分析离群数据与异常的内在关系，介绍了一种基于度量的离群数据聚类算法和基于该算法的异常检测方法；第五章介绍了语义属性数据在核空间中的分布情况和内积计算方法、语义属性数据的核分类方法，以及基于该方法的异常检测方法；第六章介绍了信息论相关的知识和用熵表达数据样本集中的结构特征，通过结构的挖掘实现数据样本的聚类，以及基于该算法的异常检测方法；第七章介绍了时间序列的表示方法、析取方法，以及基于这些数据析取方法的异常检测方法；第八章在介绍了时间序列的聚类和分类方法基础上，结合时间序列异常检测的特殊性，进一步介绍了时间序列的可视化异常检测技术。

感谢实验室的研究生张海涛、孙雅、刘庭绪、胡振宇、陈超群等同学，他们

有的参加了部分章节的撰写，有的做了大量的编辑排版工作。

由于作者水平有限，书中难免存在不妥之处，恳请广大读者、同行和专家批评指正。

李志华

2015 年 10 于江南大学

目 录

第一章

绪　论

1.1　引　言

随着社会信息化的不断发展，信息技术应用领域的不断拓展，各个应用领域包括经济、医疗、建筑、环境等均积累了越来越多的数据。自 20 世纪 80 年代开始，世界各地的数据总量飞速增长，每年甚至几个月便会增长一倍，然而，面对这些海量级别的复杂数据，如何有效地对这些历史数据进行组织和整合，然后通过数据挖掘[1]（Data Mining）来发现其中隐藏的存在潜在价值的知识和信息成为当前最为迫切的问题，是数据挖掘学科面临的一个巨大的挑战。

所谓数据挖掘是对所观测的数据对象进行分析，拟从大规模的复杂类型数据中挖掘出潜在的、未知的、有价值的知识和信息的过程，其主要目的是发现那些出现频率较高的模式。当今，数据挖掘技术在多种行业、各个领域中均得到了广泛、深入的应用，其中异常检测就是其成功应用的范例之一。基于数据挖掘中各种算法的异常检测方法或检测技术在信息安全领域、网络安全领域、金融领域、医疗领域、气象灾害领域、地质灾害监测领域等都得到了

成功应用,取得了良好的效果,创造了可观的社会效益。如今,异常检测技术已在工业设备诊断、网络安全、自然灾害检测、疾病诊断等国民生产的各行各业中显示出强大的生命力和进一步发展的潜力。

本章首先对异常检测的相关定义、所面临问题、发展现状、评价指标等进行一个比较系统的介绍。

1.2 异常检测相关概念

通常把那些符合期望(正常类或目标类)行为的数据或数据模式称为正常,而那些不符合期望(反类或异常类)行为的数据模式称为异常[2][3][4]。数据挖掘本质上是期望寻找出事物的某种发展规律,所以常常会丢掉那些出现频率很小的模式。而在很多领域中,那些极少出现的模式或者说异常通常比频繁出现的模式有着更重要的科研及实用价值。异常检测能够发现一些真实存在的而又出乎意料的信息,"异常"中蕴含显著的、至关重要的行为信息[2]。但是在传统数据挖掘领域,由于针对异常的采样代价高昂或者由于采样非常困难,使得对异常数据或异常行为知道的不多,甚至一无所知。然而,异常检测技术的价值潜力不可忽视。截至目前,异常检测在环境监测、网络入侵、信用卡(或保险)欺诈检测、疾病诊断等方面都得到了广泛的应用[3][4]。

1.3 异常检测方法

作为数据挖掘的主要任务之一,异常检测在近年来得到的关注和研究越来越多,研究人员也从不同角度提出了很多方法,从异常检测原理的角度概括如下:

(1)基于分布的方法

基于分布的方法[5]是最早被提出来的异常检测方法,主要发展于统计学领域。该类方法假设数据集的分布为已知分布(如正态分布、泊松分布等),