异常点挖掘研究进展

王宏鼎,童云海,谭少华,唐世渭,杨冬青
(北京大学 视觉与听觉处理国家重点实验室,北京 100871)

摘 要:异常点是数据集中与其他数据显著不同的数据.一个人的噪声对另一个人而言可能是有用的数据,因此,随着人们对数据质量、欺诈检测、网络入侵、故障诊断、自动军事侦察等问题的关注,异常点挖掘在信息科学研究领域日益受到重视.在充分调研国内外异常点挖掘研究文献基础上,系统地综述了数据库研究领域中异常点挖掘的研究现状,对已有各种异常点挖掘方法进行了总结和比较,并结合当前研究热点,展望了异常点挖掘未来的研究方向及其面临的挑战.

关键词:异常点;挖掘方法;局部异常点;数据流;高维数据
中图分类号:TP182;TP311 **文献标识码**:A **文章编号**:1673-4785(2006)01-0067-07

Research progress on outlier mining

WANG Hong-ding, TONG Yun-hai, TAN Shao-hua, TANG Shi-wei, YANG Dong-qing
(National Laboratory on Machine Perception, Peking University, Beijing 100871, China)

Abstract: An outlier is a data point that is significantly different from the others in a data set. One person's noise could be another person's signal, and therefore the problem of outlier mining attracts more and more interests in research of information science when the research fields of data quality, fraud detection, intrusion detection, fault diagnosis, military scout and so on receive wide attentions. In this paper, a survey was presented for the problem of outlier mining from the basic concepts to the principal research problems and the underlying techniques, including origination of outlier, definition of outlier and the comparison of popular outlier mining methods. A summary of the current state of the art of these techniques, a discussion on future research topics, and the challenges of the outlier mining were also presented.

Key words: outlier; outlier mining; local outlier; data stream; high-dimensional data

长期以来,人们十分关注数据集中异常的数据,这些数据通常被认为污染了数据集,即改变数据集的原有信息或数据产生机理.因此,发现异常点并减少异常点对数据分析的影响是一项很有意义的研究.然而,一个人的噪声可能是另一个人需要的信号,简单地剔除异常点的方式可能导致一些重要信息的丢失,因此,在提高数据质量、发现欺诈行为、进行故障诊断以及辅助军事侦察等领域都十分关注异常点的研究.过去的一个多世纪中,异常点问题的研究经历了几次盛衰交替.目前,它再次成为信息科学中一个活跃的分支,在数据库和数据挖掘研究领域受到广泛关注.

1 异常点挖掘

异常点有多种别名,如噪声、新颖点、异常物、偏离点、例外点等,除上述外,国内译名还有孤立点、离群点等,这里通称为异常点[1-12].

1.1 异常点的定义

异常点定义有多种,但具有代表性的是 V. Barnett[1]在统计学研究领域中给出的定义.

定义 1 异常点:一个异常点是这样的数据点,基于某种度量而言,该数据点与数据集中的其他数据有着显著的不同[1].

图 1 给出了数据集中几种异常点示例,其中,图 1(a) 聚类数据集中的点 A 和点 B,图 1(b) 序列数据集中的最大偏离点和图 1(c) 三维数据集中陡然突起的岩石都可被认为是数据集中的异常点.

除定义 1 外,许多研究者根据特定的研究背景,给出了不同的异常点的定义[2-7,13-19],尽管它们不尽相同,但都反映了异常点的特点:首先,异常点看起来是令人惊讶的,它是异常点的关键特征之一;其次,异常点是一个相对的定义,如果初始分布模型的

收稿日期:2006-03-01
基金项目:国家自然科学基金资助项目(60403041).

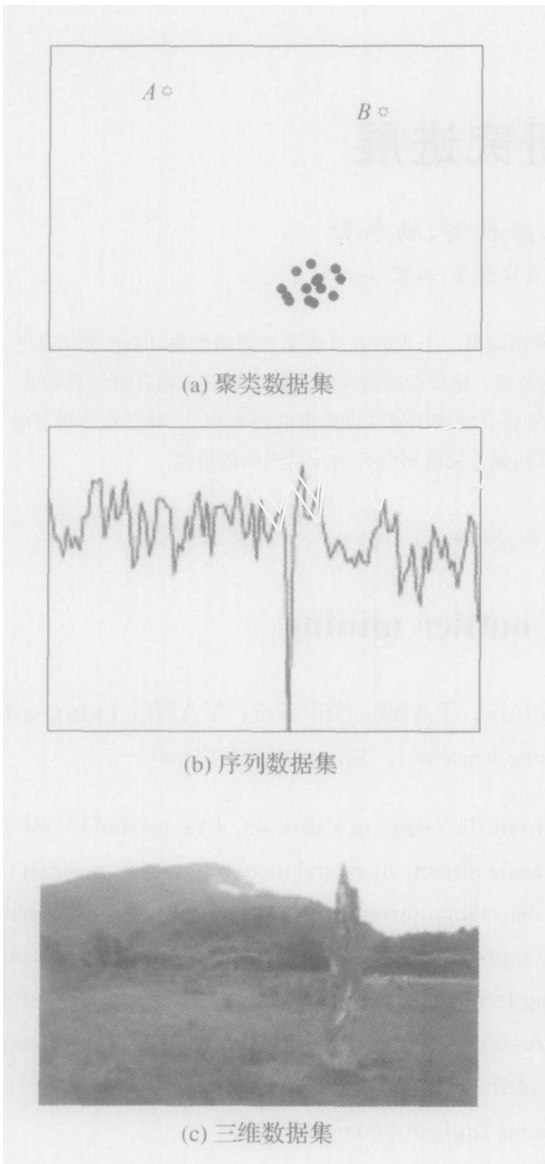


图1 数据集中异常点示例

Fig. 1 Examples of outliers in data sets

假设不同,会产生不同的结论;最后,异常点有较强主观性,几乎所有研究者进行异常点挖掘研究时都定义特有的挖掘背景.

异常点出现的原因很多,但可归为3类:1)数据变量固有变化引起,即观测值在样本总体中发生了变化,这种变化是样本总体自然发生的特征,是不可控的,并且从侧面反映了数据集的数据分布特征;2)测量错误引起,由于测量仪器的一些缺陷导致部分测量值成为异常点;3)执行错误引起,如黑客网络入侵、系统机械故障的出现导致数据集出现异常点.

同时,根据不同分类角度,异常点可分为不同类别,图2给出了异常点的一个分类情况,其中,坐标轴代表不同角度,刻度代表每个角度下的分类情况.尽管该分类不很完备,但侧面反映了数据集中异常点类型的多样性,因此,有效发现数据集中的异常点并不是一件容易的工作,需要采用有效的策略和算法.

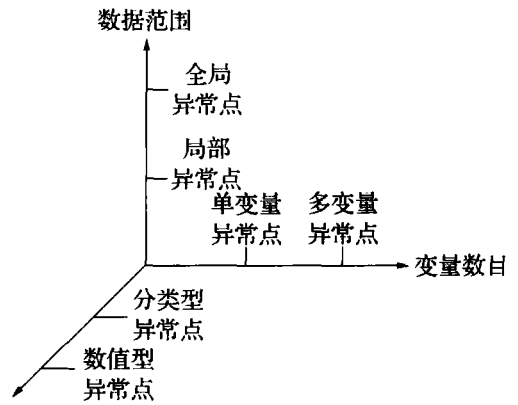


图2 异常点分类

Fig. 2 Classification of outliers

1.2 异常点挖掘

数据挖掘就是从大量数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[3].随着数据挖掘技术的不断成熟,目前已在金融、电信、零售等行业的风险管理、客户关系管理和决策支持系统中逐步获得广泛应用.异常点挖掘,就是从数据集中自动发现异常点,它是数据挖掘研究领域的一个分支.

通常,数据挖掘技术以其功能和发现的模式被分为4类[14,17,20]:依赖性检测、类型识别、类型描述、异常点检测.许多数据挖掘研究——关联规则挖掘[21]、分类[22]和数据聚类[23-24]——都属于前3类,它们研究数据集中的绝大多数对象,而第4类由于只占数据集中的较少部分,通常被看作聚类过程的副产品,当作噪声处理.因此,起初的许多数据挖掘算法通常被设计得比较健壮以包容异常点[23-24].但是,一个人的噪声可能是另一个人需要的信号,相对稀有事件的出现可能比常见事件更有意义,所以研究异常点挖掘方法是很有意义的一项工作.

异常点挖掘可以被形式化的描述[3]:给定一个含有 n 个数据点或对象的集合,预期的异常点数目 K,发现集合中与其余数据相比显著相异的、异常的或不一致的前 K 个对象,所以,异常点挖掘问题可被看作2个子问题:

1)在给定的数据集合中定义什么样的数据被认为是不一致的;

2)找到一个有效的方法来挖掘这样的异常点.

可见,异常点挖掘问题属于 Top-k 问题.

由于数据集中数据的动态性、多维性和多样性,

发现数据集中的异常点问题通常比较困难.

2　异常点挖掘研究进展

如前所述,异常点通常被作为聚类挖掘的副产品,并且,许多聚类挖掘算法把异常点作为干扰数据剔除[23-24].因此,在数据挖掘研究领域,异常点挖掘当初并不是研究主流,但随着人们对其重要性认识的加深,异常点挖掘日益受到重视.

2.1　异常点挖掘方法

由于异常点是数据集中与其余数据有显著不同的数据点,因此,比较直观的方法是建立数据集中绝大部分数据的数据模型,从而把不满足该数据模型的那一部分数据认为是异常点.这样,出现了基于不同数据模型的异常点挖掘方法,如图 3 所示,后面逐一介绍.

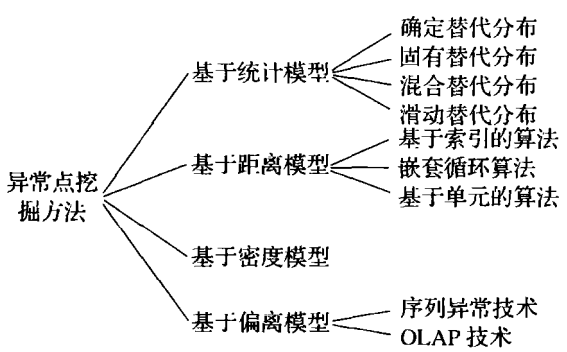


图 3　异常点挖掘算法分类

Fig. 3　Mining algorithms of outliers

2.1.1　基于统计模型的异常点挖掘方法

基于统计模型的异常点挖掘方法的思想来自于统计学方法,因为统计中常用的方法是先对给定的数据集合假设一个分布或概率模型(例如一个正态分布),然后根据该模型,采用不一致检测确定异常点.该方法需要事先知道数据集数据模型(例如假设的数据分布)、分布参数(例如平均值和方差)和假设的异常点的数目[1].

任何统计检测不可避免地要检测数据集的工作假设和替代假设,如果没有在统计上的显著证据支持拒绝工作假设,工作假设被保留.不一致检测验证一个对象关于工作假设分布是否显著的大(或者小),如果估算显著性概率是足够的小,那么对象是不一致的,工作假设被拒绝,同时,替代假设被接受,它说明了数据对象来自于另一个分布模型,所以该数据对象是一个异常点.基于统计模型方法的替代模型主要有 4 种:确定性替代分布、固定替代分布、混合替代分布和滑动替代分布[1].

尽管已有的异常点检测方法大多来自于统计学领域[1-2,4,25],但由于没有一个统一的、大家公认的异常点定义,导致了不同环境下发现异常点的上百种方法,它们依赖于:数据的分布,分布的参数(例如均值和方差),期望的异常点的数目,以及期望的异常点类型(例如在一个有序的采样集中上界异常点或下界异常点)[1].除此之外,该方法还遇到了 2 个关键问题:一是绝大多数一致性检验是针对单个属性的,而许多数据挖掘问题要求在多维空间中发现异常点;二是统计学方法要求知道关于数据集合参数的知识,例如数据分布,但许多情况下,数据分布是未知的,尤其当没有特定检验时,统计学方法不能保证所有的异常点都被发现,或者观察到的分布不能恰当地被标准分布来建模描述[14].

2.1.2　基于距离模型的异常点挖掘方法

为改进基于统计模型异常点挖掘方法的缺点,研究人员对部分基于统计模型的异常点挖掘方法进行完善[24-25],从而克服关于数据分布和维数限制的困难,但事实上,这些方法仅对二维以下的数据集挖掘可达到令人满意的效果[25-26].此外,著名数据挖掘原型系统如 DBSCAN[23]、CLARANS[24]、BIRCH[27]等聚类方法通常把异常点作为聚类的副产品处理,这些算法针对数据集中的聚类挖掘进行优化,并且数据集中的异常点是通过聚类间接获得的,所以这些方法对异常点挖掘并没有进行过多的研究探讨.

为解决统计学方法带来的缺陷(大数据集和多维问题),数据挖掘研究领域中引入了基于距离的异常点的概念和挖掘方法[14,17,20,28],它们可以有效地处理五维以上的大数据集.下面介绍基于距离模型的异常点概念及相关挖掘算法.

定义 2　基于距离的异常点 $DB(p,d)$:如果数据集合中至少有 p 部分对象与对象 o 的距离大于 d,则对象 o 是一个带参数 p、d 的基于距离的(Distance-based)异常点,即 $DB(p,d)$[14,17,20,28].

从这个定义可看出,$DB(p,d)$ 统一了异常点的概念,所以被称作一致异常点(unified outliers)[28].例如,设存在一个正态分布数据集,如果数据集中存在与均值之间的距离大于或等于 3 倍偏差的数据对象,则被认为是异常点,那么这类异常点可被 $DB(0.998\,8, 0.13\sigma)$ 的异常点定义所概括;对一个泊松分布,如果定义当参数 $\mu=3.0$,当且仅当 $t\geq 8$ 时,t 是异常点,那么这类异常点可被 $DB(0.989\,2, 1)$ 的异常点定义所概括[28].

直观而言,如果不依赖于统计检验,可将基于距离的异常点看作是那些没有足够多邻居的对象,此处邻居是基于距给定对象之间的距离定义的.目前,

该领域研究人员提供了若干高效的基于距离的异常点挖掘算法,比较有代表性的是基于索引的算法、嵌套—循环算法和基于单元的算法[14-15,17,20],这些算法的主要特点是以对象间的距离作为相似性度量.

基于距离模型的异常点挖掘方法[14-15,17,20]概括了基于统计模型的异常点的含义,并且对相对高维数据集有较好的挖掘效果,但也存在2个主要问题:一是距离函数和参数的选择,二是仅能发现全局异常点(global outliers)而丢失了局部异常点(local outliers)[8].

2.1.3 基于密度模型的异常点挖掘方法

在介绍基于密度的异常点挖掘之前,首先介绍一下Hawkwins对异常点的定义.

定义3 Hawkwins异常点:一个异常点是这样一个测量值,它过分地偏离其他测量值,从而使人们对它产生怀疑,怀疑它由不同的机理产生[2].

为了更好地理解该定义,先看一个2-D数据集的例子,如图4所示,该数据集是一个2维数据集,包含502个对象,在聚类C_1中有400个对象,在聚类C_2中有100个对象,此外还有2个特殊的对象o_1和o_2,该例中,可以看出C_2形成的聚类要比C_1稠密.

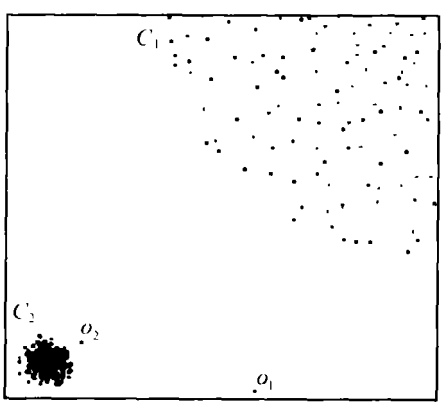


图4 一个二维数据集

Fig. 4 A 2-dimension data set

对于基于Hawkwins的定义,对象o_1和o_2都应被看作是C_2数据集中的异常点,且聚类C_1和聚类C_2中的对象都是聚类中的数据,而不是异常点.既然基于距离的异常点定义能够统一异常点的概念,那么能否找到合适p和d,使得其满足$DB(p,d)$异常点的定义.事实上,利用$DB(p,d)$的定义,只能发现o_1是一个异常点,这是因为对C_1中的每一个对象q,找不到一个合适的参数p和d来满足使q最近的邻域中的对象间的距离大于o_2和C_2间的距离,从而保证o_2是异常点的同时又能保证C_1中的对象不是异常点.从该例可看出,$DB(p,d)$在某些特定的情况下是准确和充分的,但聚类密度如果存在不同就会出现问题,为了解决这个问题,基于密度模型的局部异常点挖掘算法被提出,从而保证o_1和o_2在数据集中都是异常点[8].

基于密度模型的局部异常点定义比较复杂[8,29],基本思想来自于密度聚类方法[9,23,30-31],最后计算局部异常点因素LOF(local outlier factor).在同一个聚类中的任何对象q,可以证明,其LOF近似等于1[22].因此,就图4而言,因为聚类C_1中对象的LOF基本上都近似等于1,所以它们不是异常点.

根据局部异常点的定义及其特征,可通过对数据集中LOF的计算来确定异常点,只要一个对象的LOF远大于1,它可能就是一个异常点,需要引起数据使用者注意.文献[29]给出了一个LOF算法的改进算法,给定一个n,仅找到数据集中前n个局部异常点(top-n local outliers),虽然避开了计算绝大多数对象的LOF,但n的选择仍是一个问题.

2.1.4 基于偏离模型的异常点挖掘方法

基于偏离模型的异常点检测不采用统计检验或对象间的距离度量值来确定异常对象[3],而是通过检查一组对象的主要特征来确定异常点,如果一个对象的特征与给定的描述过分"偏离",则该对象被认为是异常点.基于偏离模型的异常点挖掘方法主要有序列异常技术[5]和OLAP数据立方体技术[32]2种.

序列异常技术模仿了人类从一系列类似对象中识别异常对象的方式,利用隐含的数据冗余,给定n个对象的集合S,建立一个子集合的序列,序列中子集合之间利用相异度函数逐个计算相异度,定义平滑因子度量——计算序列中的每个子集造成的从原始对象集合中移走子集合而带来的相异度的降低程度的函数,最终平滑因子值最大的序列子集就是该序列中的异常集.

偏离探测的OLAP方法是在大规模多维数据中采用数据立方体技术来确定反常区域.为了提高效率,偏离的探测过程与立方体的计算是重叠的,该方法采用发现驱动探索的方式,预先计算出的指示数据异常值被用来在数据集合计算的所有层次上指导用户进行数据分析.如果一个数据立方体的单元值显著不同于统计模型得到的期望值,那么该单元的数值被认为是一个异常,并采用可视化的提示进行表示,例如用背景颜色反映每个单元的异常程度.用户可以选择对那些标有异常的单元进行钻取,一个单元的度量值可能反映了发生在该立方体更低层次上的异常,因为这些异常在当前层次上可能是不可见的.该模型考虑了涉及一个单元所属的所有维

的度量值中的变化和模式以及隐藏在数据立方体集合分组操作后面的异常情况.对这种异常,由于搜索空间很大,特别是当存在许多涉及多层概念层次的维时,人工探测变得非常困难.为了改善上述不足,文献[10]通过在挖掘过程中引入约束条件,将数据立方体限制到一个小的多维空间,可以有效地从中发现异常点.

上面介绍了发现数据集中的异常点的几种常用方法,这些异常点挖掘方法主要从数据模型角度出发,具体实现时完全可考虑采用基于启发式规则的方法或机器学习的方法来发现数据集中的异常点.

2.2　异常点挖掘当前研究热点

本节重点介绍数据库研究领域中异常点挖掘的研究热点,根据调研情况来看,当前异常点挖掘的研究热点主要体现在数据流、高维数据集和 Web 数据中的异常点挖掘,下面对这些内容分别进行介绍.

2.2.1　数据流异常点挖掘

数据流的海量、无限、随时间变化等特征,以及算法上的有限存储、一次遍历等要求,使其成为近 2 年数据挖掘技术的研究热点之一.有关数据流挖掘的研究有很多[33-40],文献[33]认为挖掘数据流的变化特征是其中一个关键问题.数据流变化研究可分为 3 类:变化的模型化和表示、挖掘方法以及变化的交互式探查.就目前研究现状而言,以上问题还没有人进行系统地研究,所以有关数据流变化的研究将是数据流挖掘研究的一个十分重要方向.数据流中异常点的出现是导致数据流变化的一个重要因素,所以研究数据流中异常点挖掘问题是十分必要的.

在数据流研究中,数据的变化一般对算法影响的时间距离比较远,例如,在构建一个数据流挖掘模型时[37,40],在变化到达之前,前期数据建立模型的偏差就不再保持,如果再想计算前期每一时间段内的数据就比较困难,现有的方法一般是丢弃原有数据或者给其较小的权重[37],但没考虑何时发生了分布的变化和怎样发生了分布的变化.因此文献[6]给出了检测数据流什么时候发生了变化,并对变化进行量化和描述,它采用 2 个数据窗口,分别称为参照窗口(reference window)和滑动窗口(sliding window),每当一个新数据点出现,滑动窗口向前滑动一次,而参照窗口当且仅当检测到数据流中出现变化时才进行更新.该算法的基本思想基于这样的原理,假设有 2 个数据集 S_1 和 S_2,它们来自于 2 个分布,分别对应于 P_1 和 P_2,根据数据集 S_1 和 S_2 判断 $P_1 = P_2$ 或 $P_1 \neq P_2$ 的方法.从而利用 2 个数据集分布的关系来判断数据流的变化,文献[6]开发了检测数据流变化的一整套算法.此外,文献[7]介绍了一个从数据流中挖掘报警事件(即异常点)的一个系统 MAIDS,该系统主要功能组件包括查询引擎、流数据分类器、流模式发现器、流聚类分析器和流挖掘可视化 5 部分,原理是利用已有的数据挖掘算法,通过组合优化来解决数据流中异常点的挖掘问题.

2.2.2　高维数据集异常点挖掘

在许多实际应用中,发现高维数据集中的异常点比较普遍,这些数据集的数据维数甚至高达上百维,它对已有异常点挖掘算法是一个挑战.目前,许多挖掘方法是基于数据集之间的关系利用相似度的概念来发现异常点,然而,在高维情况下,数据十分稀疏,相似度失去了其意义.事实上,基于相似的定义,稀疏的高维数据隐含每一个点几乎都可能是很好的异常点.因此,对高维数据而言,发现有意义的异常点也变得十分复杂和不明显[22].异常点挖掘问题类似于数据挖掘中大量的其他问题,在高维数据集中算法将失去有效性.尽管已有的异常点挖掘算法[14-17]可以部分或全部满足维数据集异常点挖掘算法的要求,但没有一个算法可对高维数据集异常点进行有效地挖掘[13].

文献[13]通过发现投影的密度分布对异常点进行检测,直观而言,如果在某些低维的投影中,一个数据点出现在一个局部的区域,使该区域表现出了非正常的低密度,就称该区域的点为异常点.基于这种思想,文献[13]介绍了一种基于进化算法的高维数据集异常点挖掘方法,利用进化算法的主要目的是发现维的最优组合,即降维,以及计算这些维组合单元的数据密度.文献[13]还利用该方法对实际数据集进行了测试,结果显示,采用进化算法对许多高维数据集的工作效果很好,其中甚至包括对具有 279 个属性的高维数据集进行了异常点挖掘.

此外,借鉴高维数据集中聚类算法,文献[11]提出了一种通过闭频繁项集及其产生的关联规则来进行高维数据集异常点挖掘的方法.

2.2.3　Web 异常点挖掘

Web 挖掘被描述为从 Web 数据中分析有趣和有用的模式,然而,现有的 Web 挖掘算法处理的问题通常是发现 Web 中的频繁模式,忽略了通常被称为噪声或异常点的非频繁模式.Web 中异常点被定义为这样的观察点——偏离其他观测值太远从而让人产生怀疑,怀疑它来自于不同的机理或该对象与数据集中其余的数据明显不一致[18-19].文献[18]把 Web 异常点分为不同类别,如图 5 所示.

由于 Web 数据由不同数据类型构成,这些数据类型主要是半结构化或非结构化的数据,这为基于 Web 数据的信息自动发现提供了挑战,已有的异常

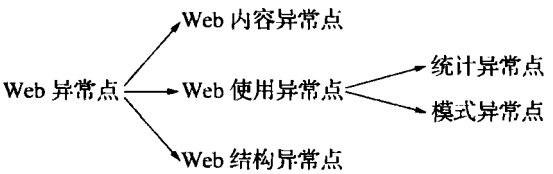


图5 Web 异常点分类

Fig. 5 Classification of Web outliers

点挖掘算法如果直接利用到 Web 数据中不是一种明智的做法,因此,对 Web 异常点的挖掘首先需要对 Web 数据进行预处理,然后再进行异常点挖掘. 文献[18 - 19]分别给出了 Web 内容异常点挖掘的一个框架和比较初级的算法. 并且,为了分析用户的上网行为模式,利用 Web 使用数据,文献[12]给出了一种基于距离的 Web 使用异常点挖掘方法.

3 结束语

异常点挖掘研究是一个非常有应用价值的问题,近年来已引起越来越多的关注,但由于异常点含义的主观性和相对性,发现海量数据集中的异常点仍是比较复杂的问题,至今没有通用、有效的方法来发现数据集中的异常点. 文章通过对该领域的深入系统调研,重点介绍了异常点挖掘方法中的基于统计模型的方法、基于距离模型的方法、基于密度模型的方法和基于偏离模型的方法,并总结了它们各自的优劣.

同时,就目前异常点挖掘问题的热点和难点,文中对数据流异常点挖掘、高维数据集异常点挖掘和 Web 数据异常点挖掘进行了概要介绍. 当然,这还远远不够涵盖目前异常点挖掘的研究趋势,但它们是异常点挖掘的难点所在,仍存在许多值得研究的地方. 同时,根据调研情况来看,异常点挖掘未来的研究方向有以下几方面:

1) 由于异常点的主观性和相对性,采用接近人类思维的、智能化的挖掘算法对异常的挖掘将会更有效.

2) 数据流、高维数据集以及 Web 数据中异常点挖掘的高效方法研究将仍是异常点挖掘问题的研究热点,尤其对欺诈检测、反洗钱、网络入侵等具体应用领域的研究将具有示范意义.

3) 加强对异常点后期分析和处理问题研究的关注. 就调研情况而言,目前这方面在所有的异常点挖掘研究中没有得到充分重视,一般仅对发现的异常点进行简单的说明,如果能够结合实际背景意义,对发现的异常点进行智能化处理的研究将对实际应用带来更大的价值.

参考文献:

[1] BARNETT V, LEWIS T. Outliers in statistical data: 2nd [M]. New York: John Wiley & Sons, 1994.

[2] HAWKINS D. Identification of outliers[M]. London: Chapman and Hall, 1980.

[3] HAN Jiawei, KAMBER M. Data mining: concepts and techniques[M]. New York: Morgan Kaufmann Publishers, 2001.

[4] QI Hongwei, WANG Jue. A model for mining outliers from complex data sets[A]. In Proc of ACM SAC '04[C]. Cyprus, 2004.

[5] ARNING A, AGRAWAL R, RAGHAVAN P. A linear method for deviation detection in large databases[A]. In Proc of KDD '96[C]. Oregon: Portland, 1996.

[6] KIFER D, BEN-DAVID S, GEHRKE J. Detecting change in data streams[A]. In Proc of VLDB '04[C]. Toronto, 2004.

[7] CAI Y D, CLUTTER D, PAPE G, et al. MAIDS: mining alarming incidents from data streams[A]. In Proceedings of SIGMOD '04[C]. Paris, 2004.

[8] BREUNING M M, KRIEGEL H P, NG R T, et al. LOF: Identifying density-based local outliers[A]. In Proc of SIGMOD '00[C]. Texas, 2000.

[9] HINNEBURG A, KEIM D A. An Efficient approach to clustering in large multimedia databases with noise[A]. In Proc of KDD '98[C]. NY, 1998.

[10] 李翠平,李盛恩,王 珊,等. 一种基于约束的多维数据异常点挖掘方法[J]. 软件学报, 2003, 14(9): 1571 - 1577.
LI Cui-ping, LI Shengen, WANG Shan, et al. A constraint-based multi-dimensional data exception mining approach[J]. Journal of Software, 2003, 14(9): 1571 - 1577.

[11] 陆介平,倪巍伟,孙志辉. 基于关联分析的高维空间异常点发现[J]. 应用科学学报, 2006, 24(1): 60 - 63.
LU Jieping, NI Weiwei, SUN Zhihui. Discovery of high dimensional outliers based on association analysis[J]. Journal of Applied Science, 2006, 24(01): 60 - 63.

[12] 赵泽茂,何坤金,陈 鹏,等. Web 日志文件的异常数据挖掘算法及其应用[J]. 计算机工程, 2003, 29(17): 195 - 197.
ZHAO Zemao, HE Kunjin, CHEN Peng, et al. Algorithms for mining outlier data on web log and its application[J]. Computer Engineering, 2003, 29(17): 195 - 197.

[13] AGGARWAL C C, YU P S. Outlier detection for high dimensional data[A]. In Proceedings of the SIGMOD '01[C]. Santa Barbara: CA, 2001.

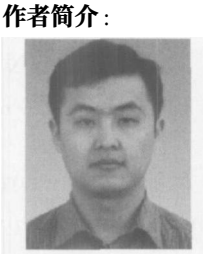
[14] KNORR E M, NG R T, TUCAKOV V. Distance-based outliers: algorithms and applications[J]. The VLDB Journal, 2000, 8(3 - 4): 237 - 253.

[15] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[A]. In Proc of SIGMOD '00[C]. Texas, 2000.

[16] ARNING A, AGRAWAL R, RAGHAVAN P. A linear method for deviation detection in large databases[A]. In Proc

of KDD'95[C]. Montreals, 1995.
[17] KNORR E, NG R. Finding intensional knowledge of distance-based outliers[A]. In Proc of VLDB'99[C]. Edinburgh, 1999.
[18] AGYEMANG M, BARKER K, AL HAJJ R. Framework for mining web content outliers[A]. In Proc of ACM SAC'04 [C]. Cyprus, 2004.
[19] AGYEMANG M, BARKER K, AL HAJJ R. Mining web content outliers using structure oriented weighting techniques and N-grams[A]. In Proc of ACM SAC'05[C]. NM, 2005.
[20] KNORR E, NG R. Algorithms for mining distance-based outliers in large datasets[A]. In Proc of VLDB'98[C]. NY, 1998.
[21] AGRAWAL R, IMIELINSKI T, SWAMI A. Mining association rules between sets of items in large databases[A]. In Proc of SIGMOD'93[C]. Washington DC, 1993.
[22] BREIMAN L, FRIEDMAN J H, OLSHEN R A, et al. Classification and regression trees[M]. New York: Chapman & Hall, 1984.
[23] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases[A]. In Proc of KDD'96[C]. Oregon, Portland, 1996.
[24] NG R. HAN J. Efficient and effective clustering method for spatial data mining[A]. In Proc of VLDB'94[C]. Santiago, 1994.
[25] KAYA A. Outlier effects on databases[A]. In Proc of ADVIS 2004[C]. Izmir: Turkey, 2004.
[26] JOHNSON T, KWOK I, Ng R. Fast computation of 2-dimensional depth contours[A] In Proc KDD'98[C]. NY, 1998.
[27] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: an efficient data clustering method for very large databases[A]. In Proc. of SIGMOD'96[C]. Montreal, 1996.
[28] KNORR E, NG R. A unified motion of outliers: properties and computation[A]. In Proc of KDD'97[C]. California, 1997.
[29] JIN Wen, TUNG A K H, HAN Jiawei. Mining top-n local outliers in large databases[A]. In Proc of SIGKDD'01[C]. California, 2001.
[30] AGRAWAL R, GEHRKE J, GUNOPULOS D, et al. Automatic subspace clustering of high dimensional data for data mining applications[A]. In Proc of SIGMOD'98[C]. WA, 1998.
[31] WANG W, YANG J, MUNTZ R. STING: A statistical information grid approach to spatial data mining[A]. In Proc of VLDB'97[C]. Athens, 1997.
[32] SARAWAGI S, AGRAWAL R, MEGIDDO N. Discovery-driven exploration of OLAP data cubes[A]. In Proc. of EDBT'98[C]. Valencia, 1998.
[33] CHEN Zhiyuan, LI Chen, PEI Jian, et al. Recent progress on selected topics in database research: a report from nine young chinese researchers working in united states[J]. JSCT, 2003, 18(5):538 - 552.
[34] GUHA S, MISHRA N, MOTWANI R, O'CALLAGHAN L. Clustering data streams[A]. In Proc of FOCS'00[C]. Redondo Beach, 2000.
[35] O'CALLAGHAM L, MISHRA N, MEYESON A, et al. Streaming-data algorithms for high-quality clustering[A]. In Proc of FOCS'01[C]. Las Vegas, 2001.
[36] DOMINGOS P, HULTEN G. Mining high-speed data streams[A]. In Proc of SIGKDD'00[C]. MA, 2000.
[37] DOMINGOS P, HULTEN G, SPENCER L. Mining time-changing data streams[A]. In Proc of SIGKDD'01[C]. California, 2001.
[38] MANKU G S, MOTWANI R. Approximate frequency counts over data streams[A]. In Proc of VLDB'02[C]. Hongkong, 2002.
[39] CHARIKAR M, CHEN K, COLTON M F. Finding frequent items in data streams[A]. In Proc of ICALP 2002[C]. Malaga, 2002.
[40] AGGARWAL C, HAN J, WANG J, et al. A framework for clustering evolving data streams[A]. In Proc of VLDB'03 [C]. Berlin, 2003.

作者简介:



王宏鼎,男,1976 年生,北京大学信息科学技术学院博士研究生,研究方向为智能控制与智能信息处理.
E-mail: hdwang@pku.edu.cn

童云海,男,1971 年生,博士,研究方向为数据仓库系统、联机分析处理和知识发现.发表论文 10 余篇,参与多项国家级和省部级项目并获得多项省部级奖励.

唐世渭,男,1939 年生,教授,博士生导师,中国计算机学会数据库专业委员会副主任.主要研究方向为数据库与信息系统.先后主持多项国家重大科技攻关课题和"973"课题,曾获国家科技进步二等奖等多项奖励,在国内外重要期刊和学术会议发表论文多篇.

杨冬青,女,1945 年生,教授,博士生导师,网络与信息系统研究所副所长,数据库与信息系统研究室主任,中国计算机学会数据库专业委员会委员.主要研究方向为数据库与信息系统.曾获国家科技进步二等奖等多项奖励,在国内外重要期刊和学术会议发表论文多篇.