离群点挖掘研究

徐 翔, 刘建伟, 罗雄麟
(中国石油大学 自动化研究所, 北京 102249)

摘 要: 随着人们对欺诈检测、网络入侵、故障诊断等问题的关注,离群点挖掘研究日益受到重视。在充分调研国内外离群点挖掘研究成果的基础上,介绍了数据库领域离群点挖掘的研究进展,并概要地总结和比较了已有的各种离群点挖掘方法,展望了离群点挖掘研究的未来发展方向和面临的挑战。

关键词: 离群点; 数据挖掘; 局部离群点; 高维数据; 数据流
中图分类号: TP311.13; TP391 **文献标志码**: A **文章编号**: 1001-3695(2009)01-0034-07

Research on outlier mining

XU Xiang, LIU Jian-wei, LUO Xiong-lin
(*Research Institute of Automation*, *China University of Petroleum*, *Beijing* 102249, *China*)

Abstract: The problem of outlier mining attracts more and more interests in research when the research fields of fraud detection, intrusion detection, fault diagnosis and so on receive wide attentions. This paper presented a survey for the research results of outlier mining at home and abroad, and based on this survey, introduced the research process of outlier mining in the areas of database. It also presented 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.

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

一直以来,人们都比较重视数据集中的离群数据,通常认为这些数据改变了数据集的原有信息或数据产生机理。因此,发现离群点并减少其对数据分析的影响是一项很有意义的工作。然而,一个人的噪声可能是另一个人的信号[1];在欺诈检测、故障诊断、网络入侵等领域的应用中,罕见事件通常比常规事件更有吸引力。简单地删除离群点的做法可能导致一些重要信息的丢失,因此,在这些领域都十分注重离群点的研究。目前,离群点问题的研究是信息科学中一个较活跃的分支,在数据挖掘研究领域也备受关注。

1 离群点的定义与相关概念

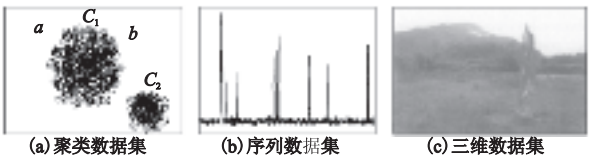
1.1 离群点的定义

在不同的应用领域,离群点有不同的定义。一个经常被引用的定义首先由 D. Hawkins 提出。

定义 1 离群点。一个离群点是这样一个观测值,它与其他观测值偏离特别多,让人们怀疑它是由不同的机理产生[2]。

图 1 给出了几种常见的数据集中离群点的例子。其中,图 1(a)聚类数据集中的点 a 和点 b;(b)序列数据集中的几个最大偏离和(c)三维数据集中突起的岩石都可以被认为是数据集中的离群点。

产生离群点的原因[3]很多,最主要有三种:a)数据来自不同的类。举例来说,进行信用卡欺诈的人(即离群数据),与那些合法使用信用卡的人相比,属于信用卡用户的一个不同的类。b)数据的固有变化。观测数据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,并从侧面反映了数据集的分布特征。c)数据测量和采集错误。人为操作错误、测量仪器的缺陷或故障也可能导致部分数据成为离群点。



(a)聚类数据集 (b)序列数据集 (c)三维数据集

图1 数据集中离群点示例

根据不同的分类标准,可将离群点分成不同的类别[4]。图 2 形象地给出了离群点的一种分类情况。图 2 中,坐标轴表示分类标准;刻度表示相应标准下的分类情况。这样的分类尽管不是很完备,但从一定程度上反映了数据集中离群数据的多样性。因此,离群点的检测是件很复杂的工作,需要采用有效的策略和算法。

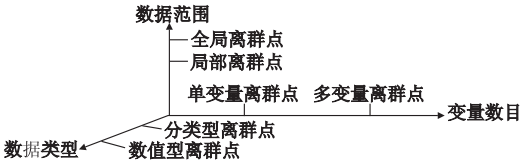


图2 离群点分类

1.2 离群点挖掘的概念

数据挖掘就是从大量的数据中自动/半自动地获得有用信息的过程。随着数据挖掘技术的不断发展,其在金融、证券、通信、零售等行业的安全管理、客户关系管理系统中得到了广泛的应用。

数据挖掘的任务可以被分成四类[1,5,6],即依赖性检测、类型识别、类型描述和异常/离群点检测。前三类任务对应着数

收稿日期: 2008-04-04; **修回日期**: 2008-06-08

作者简介:徐翔(1984-),男,江苏泰州人,硕士研究生,主要研究方向为数据挖掘(mason1200_cn@163.com);刘建伟(1966-),男,新疆石河子人,副教授,博士,主要研究方向为机器学习、非线性控制;罗雄麟(1963-),男,湖南汨罗人,教授,博导,主要研究方向为控制理论与应用、复杂系统检测、控制与优化、模式识别与智能系统、系统工程.

据集中适用于绝大多数对象的模式。数据挖掘中大多数的研究,如关联规则挖掘[7]、分类[8]、数据聚类[9,10]等都属于这三类。相反地,第四类任务关注数据对象中一个非常小的部分,通常被忽略或被当成噪声处理掉了。实际上,从知识发现的角度来看,罕见事件经常比普通事件更有价值。例如在信用卡欺诈检测中,人们不关心那些正常的信用卡使用记录,恰恰是异常的行为如异常的交易金额、交易时间和地点等,这些才是所关注的。所以,研究离群点挖掘方法是很有意义的一项工作。

离群点挖掘可描述如下[11]:给定一个有 n 个数据点或对象的数据集和期望的离群点数目 k,找出与数据集中其余数据显著不同的、异常的或不一致的前 k 个对象。离群点挖掘问题可以被看做两个子问题:a)定义在给定的数据集中,什么样的数据可以被认为是不一致的;b)找到一个有效的方法来挖掘所定义的离群点。

离群点挖掘的目标[12]是找出那些与数据集中大多数对象有非常不同的行为的罕见数据。研究离群点的异常行为可以帮助揭开隐藏的有价值的知识,辅助决策者作出决定,获得利益或改进服务质量。因此,离群点挖掘是数据挖掘的一个重要分支,有着很广泛的应用,包括信用卡欺诈检测、发现电子商务中的犯罪行为、天气预报和市场营销等。

2　离群点挖掘研究现状

离群点通常被当做聚类挖掘的副产物,所以许多聚类挖掘算法都将其作为噪声删除[9,10]。近年来,随着人们对离群点挖掘重要性的认识不断加深,以及其越来越广泛的应用,离群点挖掘日益受到重视,成为数据挖掘领域的研究热点之一。

2.1　离群点挖掘方法

先前的离群点挖掘方法大致可以分为以下几种类型:a)基于统计学的方法,包括基于分布的方法、基于深度的方法。b)基于距离的方法,包括基于索引的算法、嵌套—循环算法、基于单元的算法。c)基于密度的方法。d)基于聚类的方法。e)基于偏离的方法,包括序列异常技术、OLAP 数据立方体技术。这些方法之间的相互关系可以形象地用图 3 表示。当然各种方法之间也有一定的重叠,这里不作详细分析。

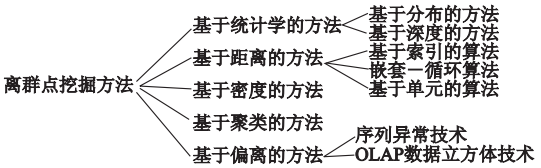


图3　离群点挖掘方法分类

2.1.1　基于统计学的离群点挖掘方法

许多现有的离群点挖掘研究都集中在统计学领域[2,12~15]。Barnett 给出了统计学领域的离群点定义。

定义 2　离群点。一个离群点是这样的数据点,基于某种度量,该数据点与数据集中其他的数据点有着明显的不同[12]。

1)基于分布的离群点挖掘方法

统计学上基于分布的方法首先对给定的数据集假设一个分布或概率模型(如正态或泊松分布),然后使用不一致检测辨识出关于这个模型的离群点。这个检测要求事先知道数据集的参数(如假设的数据分布)、分布的参数(如均值和方差)和期望的离群点数目[12]。

一个统计学的不一致检测需要检查两个假设,即工作假设和替代假设。一个工作假设 H 是这样一个命题:有 n 个对象的整个数据集都来自一个初始的分布模型 F,即

$$H:\ o_i \in F;i=1,2,\cdots,n$$

如果没有统计上显著的证据支持拒绝工作假设,那么工作假设就被保留下来。不一致检测校验一个对象 o_i 关于分布 F 是否显著的大(或小)。已经有很多不同的检验统计量被提出来作为一个不一致检测使用,这取决于数据的可利用的信息。假设一些统计量 T 已经被选来进行不一致检测,且关于对象 o_i 的统计量值为 v_i,然后构造出 T 的分布。估算出显著性概率 $SP(v_i)=\mathrm{Prob}(T>v_i)$。如果 $SP(v_i)$ 是充分小的,那么对象 o_i 是不一致的,工作假设被拒绝。同时,一个声明 o_i 来自另一个分布模型 G 的替代假设 $\overline{H}$ 被采用,它说明数据对象 o_i 来自另一个分布模型,所以该对象是一个离群点。结果在很大程度上取决于选择的模型 F,因为 o_i 可能在一种模型下是离群点,而在另一种模型下是一个完全正确的值[11]。

在确定检测的推翻假设的几率方面,替代分布是非常重要的,也就是当 o_i 真是一个离群点时,工作假设被拒绝的概率。替代分布主要有三种:固有替代分布、混合替代分布和滑动替代分布[11]。

有两种基本类型的方法来检测离群点:a)块(block)方法。在这种情形下,所有可疑的对象或者都被视为离群点,或者都被认为是一致的。b)顺序(sequential)方法。由里及外(inside out)的方法是其中的一个例子。其主要思想是首先检测最不可能是离群点的对象。如果它被发现是一个离群点,那么所有更极端的对象也都被认为是离群点;否则,接着检测下一个最不可能是离群点的对象,并重复上面的过程。这个方法比块方法更有效率[11]。

统计学领域基于分布的离群点挖掘方法很多。Yamanishi 等人[14]使用一个高斯混合模型来描述常规的行为,并且每个数据都被给定一个分数,这个分数基于数据点与模型的相异程度。现在已经有了一些对于这个方法的改进方法,其中有一个方法结合监督学习的方法来获得离群点的一般模式[15]。这些统计学的方法都依赖于数据分布、分布的参数(均值和方差)、期望的离群点数目、期望的离群点类型。可是,这些方法都遭遇到两个严重的问题:a)绝大多数不一致检验都是相对于单变量(也就是单属性)的。这个局限使得这些方法不适合多维数据库。b)它们都是基于分布的。在绝大多数情况下,不知道一个特殊的属性符合一个正态分布、一个泊松分布等,则必须执行大量的测试来找到一个匹配这些属性的分布[1]。为了改进这些不足之处,发展出了基于计算统计学的方法,被称为基于深度的方法。

2)基于深度的离群点挖掘方法

基于某些深度的定义(例如,如果 k 是为了发现点 P 而必须要被移除掉的数据点的最小数目,那么 P 的深度就是 k[16]),数据对象被组织在数据空间的层中。基于深度的方法的理论根据是:相比较于较深的层,较浅的层中更可能含有离群点。剥离(peeling)和等深线(depth contour)是在深度研究方面的两个重要概念[17,18]。这些基于深度的方法避免了分布匹配的问题,并且理论上允许处理多维的数据对象。可是,实际上,k 维层的计算依赖于 k 维凸包(convex hull)的计算。因为计算一个 k 维凸包的复杂性的下限是 $\Omega(N^{\lceil k/2\rceil})$,所以基于深度的方法对于超过四维的大数据集来说是不现实的。事实上,现有的基于深度的方法仅仅在 $k\leqslant 2$ 时能够提供可接受的性能[18]。

2.1.2　基于距离的离群点挖掘方法

基于距离的离群点概念和挖掘方法首先由 Knorr 和 Ng 提出,该方法能够有效地处理五维以上的大数据集。下面介绍基于距离的离群点挖掘的概念及相关算法。

定义 3　基于距离的离群点 $DB(p,d)$。如果数据集中至少有 p(分数,$\in[0,1]$)部分的对象与对象 o 的距离大于 d,那么 o 就是一个基于距离的离群点,即 $DB(p,d)$[1,5,6,19]。

$DB(p,d)$ 统一了其他关于离群点的定义[19],所以又被称为一致离群点(unified outliers)。例如,有一个正态分布数据集 T,如果假设数据集中一个对象 t 与均值 μ 的偏差大于或等于 3 倍方差 σ,则认为 t 就是一个离群点,那么这类离群点可以用 $DB(0.998,0.13\sigma)$ 来定义[19]。

直观而言,可将基于距离的离群点看做是那些没有足够多邻居的对象。这里的邻居是基于与给定对象间的距离来定义的。目前,研究人员已经提出了许多高效的基于距离的离群点挖掘算法,代表性的是:

a)基于索引的算法。给定一个数据集,基于索引的算法[6]使用多维的索引结构,如 R-树或 k-d 树来搜索每个对象 o 的在半径 d 之内的邻居。假设 M 是一个离群点的 d-邻域内的对象的最大数目。一旦对象 o 的 $M+1$ 个邻居被搜索到,则 o 就不是一个离群点。这种算法最差情况下的复杂性为 $O(n^2k)$。其中:n 是数据集中对象的数目;k 是维数。基于索引的算法随着 k 的增加能够很好地扩展。但即使这样,这个复杂性估计也只考虑了搜索时间,而索引的建立是一个巨大的潜在花费,这通常致使基于索引的算法失去竞争力。

b)嵌套—循环算法。它与基于索引的算法有相同的计算复杂性,但避免了建立索引结构,且尽量地最小化 I/O 数目。它将存储缓冲器空间分为两半,并将数据集分成若干个逻辑块;通过精心地选择块被装载到每一半中的顺序,能够获得较好的 I/O 效率。

Ghoting 等人[20]对嵌套—循环算法进行了扩展,提出了一种适用于高维数据集的二段算法 RBRP。该算法的第一阶段使用一个如同 k-均值的聚类算法将数据集划分成若干个小的聚类,每个聚类都可以看成是一个桶(bin);第二阶段使用扩展的嵌套—循环算法找出已经被划分成若干个桶的数据集中的离群点。Bay 等人[21]提出了一个改进的嵌套—循环算法,结合随机化和一个简单的修剪规则,在大型的真实数据集中有接近线性的时间性能,但是 I/O 花费却是二次的。Tao Yu-fei 等人[22]进一步改进了 Bay 等人的算法,提出了 SNIF 算法,改进了 I/O 效率,通过扫描数据库最多两次就能找出所有的离群点。

c)基于单元的算法[1,5,6,23]。对于内存常驻数据集,为了避免 $O(n^2)$ 的计算复杂性,发展出基于单元(cell)的算法[6,23]。其复杂性是 $O(c^k+n)$。其中:c 是一个取决于单元数目的常量;k 是维数。在这个方法中,数据空间被划分成若干个边长等于 $d/2\sqrt{k}$ 的单元。每个单元有围绕在它周围的两层:第一层有一个单元厚;第二层有 $|2\sqrt{k}-1|$(趋于最近的整数)个单元厚。图 4 为点数据的 cell 结构示意图。这种算法在单元到单元的基础上计算离群点,而不是在对象到对象的基础上。对于一个给定的单元,它存储三个计数:单元中的对象数目、单元中和第一层中的对象数目、单元中和两个层中的对象数目。假设这些计数分别为 cell_count、cell_+_1_layer_count 和 cell_+_2_layers_count,M 是可能出现在一个离群点的 d-维邻域内的对象的最大数目。

(a)在当前单元中的一个对象 o,仅在 cell_+_1_layer_count 小于或等于 M 时,才被认为是一个离群点。如果这个条件不能保证,那么在这个单元中的所有对象就可以被移除掉,无须进一步研究,因为它们不可能是离群点。

(b)如果 cell_+_2_layers_count 小于或等于 M,那么单元中的所有对象都认为是离群点;否则,如果这个数目大于 M,那么可能单元中有一些对象是离群点。为了检测出这些离群点,使用对象到对象的处理,对于单元中的每个对象 o,检查 o 的第二层中的对象。对于单元中的对象,仅仅那些在它们的邻域中有不超过 M 个点的对象是离群点。一个对象的 d-邻域由对象所在的单元、所有的第一层和一部分第二层所组成。

这种算法相对于前面算法的一个变化是关于 n 呈线性的,并保证扫描数据库不超过三次。它适用于较大的内存常驻数据集,但是不能很好地测量高维数据库。

基于距离的离群点挖掘方法对相对高维的数据集有较好的挖掘效果,但也存在两个主要问题:a)距离函数和相关参数的选择。基于距离的离群点挖掘方法需要用户设置参数 p 和 d,寻找这些参数的合适设置可能涉及到许多实验并可能导致试错。b)它只能发现全局(global)离群点而丢失局部(local)离群点。

2.1.3　基于密度的局部离群点挖掘方法

统计学上的和基于距离的离群点挖掘方法都是从全局的角度考虑给定的数据集,但是遇到不同的密度分布时,这些方法就不适用了。这里首先介绍局部离群点的概念。如图 5 所示的二维数据集,C_1 和 C_2 是数据集中两个密度不同的聚类,可以直观地看出 C_2 比 C_1 稠密。依照 Hawkins 的定义,o_1 和 o_2 都应该是离群点,而聚类 C_1 和 C_2 中的对象不是离群点;相反地,依据基于距离的离群点定义,仅有 o_1 可能是一个合理的 $DB(p,d)$ 离群点。因为对于 C_1 中的每个对象 q,q 与它的最近邻之间的距离大于 o_2 与 C_2 之间的距离(即 $d(o_2,C_2)$),实际上可以发现没有合适的 p 和 d 值能使 o_2 是一个 $DB(p,d)$ 离群点而在 C_1 中的对象却不是离群点。从本例可以看出,基于距离的离群点在数据集中聚类密度存在不同时会出现问题。为了解决这个问题,Breunig 等人[24]提出了基于密度的局部离群点挖掘方法,从而保证离群点 o_1 和 o_2 都能被发现。

如果一个对象相对于它的局部邻域,特别是对于邻域的密度是离群的,那么它就是一个局部离群点。根据这个观点,图 4 中的 o_2 相对于 C_2 的密度是一个局部离群点。o_1 也是一个离群点,而 C_1 中没有对象会被错认为是离群点。这构成了基于密度的局部离群点挖掘的基础。这个方法的另一个关键思想是:不同于先前的方法,它不确切地指明一个对象是否为离群点,而是评估一个对象可能是离群点的程度。这个离群的程度利用一个对象的局部离群因数(LOF)来计算。这个程度取决于对象相对于周围的邻域是多么的孤立。这个方法既可以检测全局离群点,也可以检测局部离群点。

基于密度的局部离群点的确切定义以及局部离群因数的计算方法比较复杂[24],其基本思想来源于基于密度的聚类算法[9,25~27]。根据定义,如果一个对象 p 不是局部离群点,那么 $LOF(p)$ 就近似等于 1;如果 p 是一个局部离群点的可能性越大,$LOF(p)$ 的值也就越大。因此,决定一个点 p 是否为局部离群点就基于 $LOF(p)$ 的计算。

Tang 等人[28]介绍了一个基于连接的离群点因数(COF)方

案,提高了 LOF 方案的效率。Jin 等人[29]提出了一种 LOF 算法的改进算法,给定一个 n,仅寻找数据集中前 n 个局部离群点(top-n local outliers),避免计算绝大多数对象的 LOF 值。Papadimitriou 等人[30]使用局部相关积分(LOCI)方法,进一步扩展了基于密度的方法。

基于密度的局部离群点挖掘方法即使在数据有不同密度的区域也能适用。在算法的时间复杂性方面,对于低维数据,可通过使用特殊的数据结构减少到 $O(n \log n)$;但在通常情况下,它与基于距离的方法类似,也有 $O(n^2)$ 的时间复杂性(n 为对象的数目)。而且,相关参数的选择也是一个难题。

2.1.4　基于聚类的离群点挖掘方法

聚类分析找到紧密相关的对象的群,而离群点挖掘是找出与其他对象不紧密相关的对象。通常离群点都是聚类挖掘算法的副产物,但是通过对聚类算法进行一定的改进,也可以应用于离群点挖掘。基于聚类的离群点定义[3]如下:

定义 4　如果一个对象不严格地属于任何聚类,那么这个对象就是一个基于聚类的离群点。

一个使用聚类来进行离群点挖掘的方法是将那些远离其他聚类的小聚类作为离群点丢弃。这个方法能够使用任何聚类技术,但是要求知道最小聚类的尺寸和一个小聚类与其他聚类之间距离的阈值。通常,这个过程被简化成丢弃所有小于最小尺寸的聚类。这个方法对选择的聚类数是很敏感的。Jiang 等人[31]提出的两阶段聚类算法就使用这种思想。离群点被定义为远离绝大多数点的小聚类。首先将数据点划分成若干聚类,每个聚类中的点要么都是离群点,要么都不是。换句话说,当包含在同一个聚类中的点不是足够靠近时,聚类就可能被分成两个更小的聚类。在划分完数据点之后,可以很容易地看出寻找离群聚类的时间复杂性减少了。这是因为相似的点被合并到相同的聚类中,并且被处理的是聚类而不是点。在算法的第一阶段中,通过使用一个直观推断:"如果一个新的输入模式足够地远离所有聚类的中心,那么就指定它为一个新的聚类中心",来修改传统的 k-Means 算法;然后在算法的第二阶段中,构建阶段中聚类的一个最小扫描树(MST),并作为森林的一个成员,从森林中移除一个树的最长的边,并用两个新生成的子树代替原来的树。小的聚类有较少数目的节点的树,被选中并认为是离群点。此外,重复地从森林中移除最长的边直到树的数目足够多。He 等人[32]提出了基于聚类的局部离群点的概念。

一些聚类算法有线性的或接近线性的时间和空间复杂性,因此基于这样的算法的一个离群点挖掘技术也是高效的。但是,通过这种算法生成的离群点的质量取决于生成的聚类的质量,而每个聚类算法仅适用于某一种特定类型的数据,因此需要谨慎地选取聚类算法。

2.1.5　基于偏离的离群点挖掘方法

基于偏离的离群点挖掘不使用统计测试或基于距离的度量来辨识异常对象,而是通过检测一组对象的主要特征来辨识离群点,过分偏离给定描述的对象被认为是离群点[11]。基于偏离的离群点挖掘方法主要有两种:

a)序列异常技术[33]。该技术模仿了人类可以从一系列类似的对象中识别出异常对象的行为。它使用隐含的数据冗余。给定一个包含 n 个对象的数据集 D,建立一个子集合序列 $\{D_1, D_2, \cdots, D_m\}$。其中 $2 \leqslant m \leqslant n$,并有 $D_{j-1} \subset D_j$,这里 $D_j \subseteq D$。评估序列中子集之间的差异度。这个技术介绍了如下几个关键术语:

(a)异常集　就是偏离的集或离群点。它被定义为对象的最小子集,这个子集的移除导致数据集中剩余部分的差异度有最大量的减少。

(b)差异度函数　这个函数不要求对象间的一个度量距离。它可以是任何函数,只要满足条件,即给定一个对象集,如果某个对象与另一个对象是相似的,那么就返回一个较低的值。对象之间的差异度越大,函数返回值也越大。

(c)平滑因子　对于序列中每个子集,这个函数都会被计算。它估价了从原始对象集中移除子集后,差异度能减少多少。这个值通过集的基数来测量。平滑因子最大的子集就是序列中的异常集。

这个算法选择一个子集的序列来进行分析。对于每个子集,它确定这个子集与序列中前一个子集的差异度。为了避免输入顺序对结果的任何可能的影响,处理过程可以重复若干次,每一次都有子集的一个不同的随机顺序。在所有的迭代中,有最大平滑因子值的子集就是异常集。

b) OLAP 数据立方体技术[34]。一个进行偏离检测的 OLAP 方法使用数据立方体来辨识高维数据中的异常区域。为了提高效率,偏离检测过程可以与立方体计算过程重叠进行。该方法是一种探索驱动的方法,预先计算的指示数据异常的度量,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。如果数据立方体中的一个单元值与基于一个统计模型的期望值显著不同,那么这个单元值就被认为是一个异常。这个方法采用如背景颜色那样的可视化提示来反映每个单元的异常程度。用户可以选择对标志为异常的单元进行向下钻取。一个单元的度量值可能反映发生在该立方体上的更细节或更低层次上的异常,因为这些异常在当前层次上可能是不可见的。

这个模型考虑了一个单元所属的所有维上的度量值的变化和模式。例如,假设有一个销售数据的数据立方体,并且要查看每个月的销售概况。在可视化提示的帮助下,注意到与其他月相比,十一月的销售量有所增加,在时间维上这像是一个异常。可是,通过下载十一月来观察这个月的每一项的销售,可以注意到十一月中其他项的销售也有一个相似的增加。因此,如果考虑到项目维,十一月的总销售上的增加就不是一个异常。这个模型考虑了隐藏在所有数据立方体集合分组操作后面的异常。对于这样的异常,因为搜索空间非常大,特别是当存在许多包含若干层的概念层次的维时,人工探测就非常困难了。为了改善以上不足,李翠平等人[35]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中,然后再从中找出离群点,有效地改进了挖掘效率。

2.1.6　其他离群点挖掘方法

上面介绍了离群点挖掘的五种常用方法。最近,一些学者将其他研究领域的方法移植到离群点挖掘领域,提出了许多新颖的离群点挖掘方法。

一些研究者提出了基于子空间的离群点挖掘方法。Aggarwal 等人[36]讨论了一个新的离群点探测技术,通过研究数据集的投影密度分布来挖掘离群点。He 等人[37]提出了一个基于频繁模式的离群点检测方法,旨在利用不同子空间中的频繁模式来定义高维空间中的离群点。Li 等人[38]介绍了一个基于超图模型的离群点挖掘方法,并用来检测分类数据集中的离群点。

一些研究者提出了基于支持向量的离群点挖掘方法。Tax 等人[39]首次提出了支持向量新颖探测器(SVND)。Cao 等人[40]提出了改良的 SVND,在性能上有了较大提高。

还有一些研究者提出了基于神经网络的离群点挖掘方法。Harkins 等人[41]利用重复因子神经网络(RNNs)提供了一个对于离群数据的度量;Williams 等人[42]将 RNN 离群点挖掘方法与其他三种方法进行了对比。

以上介绍了检测数据集中离群点的一些方法,这些挖掘方法主要从数据模型的角度出发。根据不同的分类标准,离群点挖掘方法可以分成不同的类型。比如可以分成参数方法和非参数方法,基于统计学的方法是参数方法,而基于距离、基于聚类、基于密度的方法等都属于非参数方法;还可以分成一维的和多维的挖掘方法,这是根据适用的数据集维度来分类的。现在绝大多数的离群点挖掘方法都没有考虑数据对象的类标签,He 等人[43]和 Papadimitriou 等人[44]研究了类离群点挖掘的问题,并提出了相应的算法。He 等人[45]考虑了类离群点检测问题:"给定一个有类标签的对象集,在考虑类标签的情况下,找到那些引起怀疑的对象",发展了类离群点概念,提出的实用的解决方法是通过扩展现有的离群点挖掘算法到这种情况下来实现的。显然,根据类标签(异常或正常)是否可见,可以将离群点挖掘方法分成有监督、无监督和半监督的方法[3]。

2.2　离群点挖掘研究热点

本节重点介绍离群点挖掘领域的研究热点,从调研情况看,当前离群点挖掘的研究热点主要集中在高维数据集、数据流和 Web 数据的离群点挖掘上。

2.2.1　高维数据集离群点挖掘

在大量实际应用中,需要寻找高维数据集中的离群点。这些高维数据集的数据维数通常高达上百维,这对已有的离群点挖掘算法是一个巨大的挑战。目前,许多挖掘方法是利用数据集之间的关系相似度概念来挖掘离群点的,然而在高维数据集中,数据是相当稀疏的,相似度也就失去了意义。事实上,基于相似的定义,稀疏的高维数据集中的每一个点都可以看成是一个离群点。因此,对于高维数据集而言,挖掘出有意义的离群点变得十分复杂[8]。与数据挖掘中大量的其他问题相似,离群点挖掘算法在高维数据集中也将失去其有效性。尽管已有的离群点挖掘算法[5,6,21,33]可以部分或全部满足高维数据集中离群点挖掘的要求,但实际上没有一个算法可对高维数据集中离群点进行有效的挖掘[36]。

文献[36]通过研究投影的密度分布进行离群点检测。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,而该区域表现出了非正常的低密度,就称这样的点为离群点。基于这种思想,文献[36]介绍了一种基于进化算法的高维数据集离群点挖掘方法。使用进化算法的主要目的是找到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[36]还利用该方法对实际数据集进行测试,结果表明,采用进化算法的处理效果很好,甚至对具有 279 个属性的高维数据集也能有效地进行离群点挖掘。另外,借鉴高维数据集的聚类算法,文献[46]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离群点挖掘的方法。

2.2.2　数据流离群点挖掘

目前,在许多如网络监控、通信服务、传感网络和金融服务等应用领域产生一种新型的数据——流数据。这些数据以网络性能参数、电话记录、传感器的读入值和金融数据等体现,它们是实时的、连续的、无限的,是按照数据项到达的顺序控制的,一组有序的点 $x_1, x_2, \cdots, x_n$ 的序列。它们不是存储在磁盘上的有限数据集合,而是无限的、连续的、快速的、实时的、以流的形式出现的数据集合。流数据中每个数据单元是一个元组,元组的个数是无穷的。与传统的数据相比,流数据的特点表现在量大、连续性、实时性,随机存取采用的是单一线性数据扫描,完整地将数据流存储到本地是不可行的。

数据流的诸多特征,以及算法上的有限存储、一次遍历等要求,使其成为近年来数据挖掘技术的研究热点之一。数据流挖掘方面的研究很多[47~54],文献[47]认为挖掘数据流的变化特征是一个关键问题。对数据流变化的研究可分成三类,即变化的模型化和表示、挖掘方法和变化的交互式探查。从目前的研究现状来看,以上问题还没有进行过系统的研究,所以有关数据流变化的研究是数据流挖掘研究的一个十分重要的方向。数据流中离群点的出现是导致数据流变化的一个重要因素,所以研究数据流中离群点挖掘问题是十分必要的。

在数据流研究中,数据的变化一般对任何数据处理算法都有深远的影响。例如在构建一个数据流挖掘模型时[51,54],变化到达之前的数据可能偏离那些与不再保持的特征所相关的模型,如果想处理对于数据流的查询,想再用到前期某一时间段内的数据就比较困难了。绝大多数现有的研究集中在通过丢弃旧数据或给其较小的权重来匹配变化分布的算法[51],但没有考虑分布的变化是何时发生的以及怎样发生的。因此文献[55]提出了检测数据流什么时候发生变化的方法,并对变化进行量化和描述。它采用两个数据窗口,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时,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。该算法的基本思想是,假设两个数据集 S_1 和 S_2,它们分别由两个不同的分布 P_1 和 P_2 产生,根据数据集 S_1 和 S_2 判断是 $P_1 = P_2$ 还是 $P_1 \neq P_2$,从而利用两个数据集分布的关系来判断数据流的变化,并开发出一整套检测数据流变化的算法。文献[56]介绍了一个从数据流中挖掘报警事件(即离群点)的系统 MAIDS。该系统的主要功能组件包括流查询引擎、流数据分类器、流模式发现器、流聚类分析器和流挖掘观测仪五部分;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的离群点挖掘问题。文献[57]提出了一种基于数据空间动态网格划分的快速数据流离群点检测算法。该算法利用动态网格对空间中的稠密和稀疏区域进行划分,过滤处于稠密区域的大量主题数据,有效地减少算法需要考察的数据对象的规模;而对于稀疏区域中的候选离群点,采用近似方法计算其离群度,具有较高的离群度的数据就作为离群点输出。文献[58]在频繁模式的离群因子概念基础上,提出了类别属性数据流离群度量——加权频繁模式离群因子(WFPOF),并针对数据流特点提出了基于 WFPOF 的高维数据流离群点检测算法 FODFP-Stream。该算法通过动态发现和维护频繁模式来计算离群度,能有效处理高维类别属性数据流,并进一步扩展到数值属性和混合属性数据流;同时通过数据衰减系数的设定,可以有效地处理数据流数据中的概念转移问题。

2.2.3　Web 离群点挖掘

Web 挖掘就是从 Web 数据中分析有趣的和有用的模式,但是,现有的 Web 挖掘算法处理的问题大多是发现 Web 中的频繁模式,通常忽略被称为噪声或离群点的非频繁模式。Web

数据中的离群点被定义为这样的观测点:它偏离其他的观测值太远,以至于人们怀疑它源自不同的机理或该对象与数据集中其余的数据明显不一致[59,60]。文献[59]把 Web 离群点分为不同类别,如图 6 所示。

2	2	2	2	2	2	2
2	2	2	2	2	2	2
2	2	1	1	1	2	2
2	2	1	c	1	2	2
2	2	1	1	1	2	2
2	2	2	2	2	2	2
2	2	2	2	2	2	2

图4　Cell结构示意图

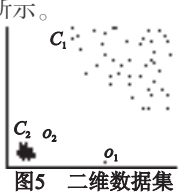


图5　二维数据集

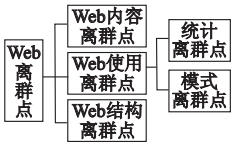


图6　Web离群点分类

Web 数据主要由半结构化或非结构化类型的数据所构成,这对基于 Web 数据的信息发现提出了挑战,将已有的离群点挖掘算法直接应用到 Web 数据中是一种不明智的做法。对 Web 离群点的挖掘首先需要对 Web 数据进行预处理,然后再进行离群点挖掘。文献[59,60]分别给出了 Web 内容离群点挖掘的一个框架和比较初级的算法。文献[61]提出了一种基于距离的 Web 使用离群点挖掘方法,利用 Web 使用数据,分析用户的上网行为模式。

3　结束语

离群点挖掘是一个非常有应用价值的研究领域,近年来受到越来越多的关注,但由于离群点定义的主观性和相对性,挖掘数据集中的离群点仍是相当复杂的,至今没有通用的、高效的方法来解决这个问题。本文通过对该领域的深入研究,重点介绍了离群点挖掘方法中基于统计学的方法、基于距离的方法、基于密度的方法、基于聚类的方法和基于偏离的方法,还简要介绍了其他一些新的离群点挖掘方法,并总结了它们各自的优缺点。同时,根据目前离群点挖掘方向的研究热点,本文对高维数据集离群点挖掘、数据流离群点挖掘和 Web 离群点挖掘进行了简要介绍。当然,这并不能代表目前离群点挖掘的所有研究趋势,但它们确实是离群点挖掘研究的难点。

根据调研,认为以下几个方面是未来离群点挖掘研究的发展方向:

a)采用接近人类思维的、智能化的挖掘算法来进行离群点挖掘,这将有效地减少离群点定义的主观性和相对性带来的影响。

b)数据可视化技术的应用将会大大地简化离群点的挖掘工作。

c)研究高维数据集、数据流和 Web 数据中离群点挖掘的高效算法仍将是热点,尤其是在欺诈检测、网络入侵等具体应用领域上。

d)重视对离群点挖掘的后期分析和处理。目前,在几乎所有的离群点挖掘研究中都没有关注这方面的问题,一般仅对发现的离群点进行简要的说明,如果能够结合实际的背景意义,对发现的离群点进行详细的分析处理,将会带来更大的实际应用价值。

参考文献:

[1] KNORR E M, NG R T. Algorithms for mining distance-based outliers in large datasets[C]//Proc of VLDB'98. San Francisco, CA: Morgan Kaufmann Publishers, 1998:392-403.

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

[3] TAN Pang-ning, STEINBACH M, KUMAR V. Introduction to data mining[M]. Boston: Pearson Addison-Wesley Education Inc, 2006.

[4] 王宏鼎, 童云海, 谭少华, 等. 异常点挖掘研究进展[J]. 智能系统学报, 2006,1(1):67-73.

[5] KNORR E, NG R. Finding intentional knowledge of distance-based outliers[C]//Proc of VLDB'99. Edinburgh: [s.n.], 1999:211-222.

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

[7] AGRAWAL R, IMIELINSKI T, SWAMI A. Mining association rules between sets of items in large databases[C]//Proc of SIGMOD'93. New York: ACM Press, 1993:207-216.

[8] BREIMAN L, FRIEDMAN J H, OLSHEN R A, *et al*. Classification and regression trees[M]. New York: Chapman & Hall, 1984.

[9] ESTER M, KRIEGEL H P, SANDER J, *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise [C]//Proc of KDD'96. Portland, Oregon: AAAI Press, 1996:226-231.

[10] NG R T, HAN Jia-wei. Efficient and effective clustering method for spatial data mining[C]//Proc of VLDB'94. Vancoaver, BC: University of British Columbia, 1994:144-155.

[11] HAN Jia-wei, KAMBER M. Data mining: concepts and techniques [M]. New York: Morgan Kaufmann Publishers, 2001.

[12] BARNETT V, LEWIS T. Outliers in statistical data[M]. 2nd ed. New York: Wiley, 1994.

[13] ROUSSEEUW P J, LEROY A M. Robust regression and outlier detection[M]. New York: Wiley, 1987.

[14] YAMANISHI K, TAKEUCHI J, WILLIAMS G, *et al*. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms[C]//Proc of KDD'00. New York: ACM Press, 2000: 320-324.

[15] YAMANISHI K, TAKEUCHI J. Discovering outlier filtering rules from unlabeled data-combining a supervised learner with an unsupervised learner[C]//Proc of KDD'01. New York: ACM Press, 2001: 389-394.

[16] JOHNSON T, KWOK I, NG R. Fast computation of 2-dimensional depth contours[C]//Proc of KDD'98. New York: [s.n.], 1998: 224-228.

[17] PREPARATA F, SHAMOS M. Computational geometry: an introduction[M]. New York: Springer-Verlag, 1988.

[18] RUTS I, ROUSSEEUW P J. Computing depth contours of bivariate point clouds[J]. Computational Statistics & Data Analysis, 1996,23(1): 153-168.

[19] KNORR E, NG R. A unified motion of outliers: properties and computation[C]//Proc of KDD'97. California: AAAI Press, 1997:219-222.

[20] GHOTING A, PARTHASARATHY S, ERIC M. Fast mining of distance-based outliers in high-dimensional datasets, NGS-CNS-0406386 [R]. Ohio: the Ohio State University, 2005.

[21] BAY S D, SCHWABACHER M. Mining distance-based outliers in near linear time with randomization and a simple pruning rule[C]// Proc of KDD'03. New York: ACM Press, 2003:29-38.

[22] TAO Yu-fei, XIAO Xiao-kui, ZHOU Shui-geng. Mining distance-based outliers from large databases in any metric space[C]//Proc of KDD'06. New York: ACM Press, 2006:394-403.

[23] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[C]//Proc of SIGMOD'00. New York: ACM Press, 2000:427-438.

[24] BREUNIG M M, KRIEGEL H P, NG R T, *et al*. LOF: identifying density-based local outliers[C]//Proc of SIGMOD'00. New York: ACM Press, 2000: 93-104.

[25] AGRAWAL R, GEHRKE J, GUNOPULOS D, *et al*. Automatic subspace clustering of high dimensional data for data mining applications [C]//Proc of SIGMOD'98. New York: ACM Press, 1998:94-105.

[26] WANG Wei, YANG Jiong, MUNTZ R. STING: a statistical information grid approach to spatial data mining[C]//Proc of VLDB'97. Athens: [s.n.], 1997:186-195.

[27] HINNEBURG A, KEIM D A. An efficient approach to clustering in large multimedia databases with noise[C]//Proc of KDD'98. New York: AAAI Press, 1998:58-65.

[28] TANG Jian, CHEN Zhi-xiang, FU A W, *et al*. Enhancing effectiveness of outlier detections for low density patterns[C]//Proc of PAKDD'02. London, UK: Springer-Verlag, 2002: 535-548.
[29] JIN Wen, TUNG A K H, HAN Jia-wei. Mining top-*n* local outliers in large databases[C]//Proc of SIGKDD'01. New York: ACM Press, 2001:293-298.
[30] PAPADIMITRIOU S, KITAGAWA H, GIBBONS P B, *et al*. IOCI: fast outlier detection using the local correlation integral[C]//Proc of ICDE'03. [S. l.]: IEEE Computer Society, 2003:315-326.
[31] JIANG M F, TSENG S S, SU C M. Two-phase clustering process for outlier detection[J]. Pattern Recognition Letters, 2001, 22(6-7):691-700.
[32] HE Zeng-you, XU Xiao-fei, DENG Sheng-chun. Discovering cluster based local outliers[J]. Pattern Recognition Letters, 2003, 24(9-10):1641-1650.
[33] ARNING A, AGRAWAL R, RAGHAVAN P. A linear method for deviation detection in large databases[C]//Proc of KDD'96. Portland: AAAI Press, 1996:164-169.
[34] SARAWAGI S, AGRAWAL R, MEGIDDO N. Discovery-driven exploration of OLAP data cubes[C]//Proc of EDBT'98. [S. l.]: Springer-Verlag, 1998:168-182.
[35] 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.
[36] AGGARWAL C C, YU P S. Outlier detection for high dimensional data[C]//Proc of SIGMOD'01. New York: ACM Press, 2001:37-46.
[37] HE Zeng-you, XU Xiao-fei, HUANG J Z, *et al*. A frequent pattern discovery based method for outlier detection[C]//Proc of WAIM'04. Berlin: Springer, 2004: 726-732.
[38] LI Wei, QIAN Wei-ning, ZHOU Ao-ying, *et al*. HOT: hypergraph-based outlier test for categorical data[C]//Proc of PAKDD'03. Berlin: Springer, 2003: 399-410.
[39] TAX D M J, DUIN R P W. Support vector data description[J]. Pattern Recognition Letters, 1999, 20(11-13):1191-1199.
[40] CAO Li-juan, LEE H P, CHONG W K. Modified support vector novelty detector using training data with outliers[J]. Pattern Recognition Letters, 2003, 24(14):2478-2479.
[41] HARKINS S, HE Hong-xing, WILLIAMS G J, *et al*. Outlier detection using replicator neural networks[C]//Proc of DWKD. London, UK: Springer-Verlag, 2002:170-180.
[42] WILLIAMS G J, BASTER R A, HE Hong-xing, *et al*. A comparative study of RNN for outlier detection in data mining[C]//Proc of ICDM'02. Washington DC: IEEE Computer Society, 2002:709-712.
[43] HE Zeng-you, DENG Sheng-chun, XU Xiao-fei. Outlier detection integrating semantic knowledge[C]//Proc of WAIM'02. London, UK: Springer-Verlag, 2002:126-131.
[44] PAPADIMITRIOU S, FALOUTSOS C. Cross-outlier detection[C]//Proc of SSTD'03. Berlin: Springer, 2003: 199-213.
[45] HE Zeng-you, XU Xiao-fei, HUANG J Z, *et al*. Mining class outlier: concepts, algorithms and application in CRM[J]. Expert System with Applications, 2004, 27(4): 681-697.
[46] 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.
[47] CHEN Zhi-yuan, 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]. Journal of Computer Science and Technology, 2003, 18(5):538-552.
[48] GUHA S, MISHRA N, MOTWANI R, *et al*. Clustering data streams[C]//Proc of FOCS'00. Redondo Beach: [s. n.], 2000:359-366.
[49] O'CALLAGHAM L, MISHRA N, MEYESON A, *et al*. Streaming-data algorithms for high-quality clustering[C]//Proc of FOCS'01. Washington DC: IEEE Computer Society, 2002:685-704.
[50] DOMINGOS P, HULTEN G. Mining high-speed data streams[C]//Proc of SIGKDD'00. New York: ACM Press, 2000:71-80.
[51] HULTEN G, SPENCER L, DOMINGOS P. Mining time-changing data streams[C]//Proc of SIGKDD'01. New York: ACM Press, 2001: 97-106.
[52] MANKU G S, MOTWANI R. Approximate frequency counts over data streams[C]//Proc of VLDB'02. Hong Kong: VLDB Endowment, 2002:346-357.
[53] CHARIKAR M, CHEN K, COLTON M F. Finding frequent items in data streams[C]//Proc of ICALP'02. London, UK: Springer-Verlag, 2002:693-703.
[54] AGGARWAL C C, HAN Jia-wei, WANG Jian-yong, *et al*. A framework for clustering evolving data streams[C]//Proc of VLDB'03. Berlin: VLDB Endowment, 2003:81-92.
[55] KIFER D, BEN-DAVID S, GEHRKE J. Detecting change in data streams[C]//Proc of VLDB'04. Toronto: VLDB Endowment, 2004: 180-191.
[56] CAI Y D, CLUTTER D, PAPE G, *et al*. MAIDS: mining alarming incidents from data streams[C]//Proc of ACM SIGMOD' International Conference on Management of Data. New York: ACM Press, 2004:919-920.
[57] 杨宜东, 孙志挥, 朱玉全, 等. 基于动态网格的数据流离群点快速检测算法[J]. 软件学报, 2006, 17(8):1796-1803.
[58] 周晓云, 孙志挥, 张柏礼, 等. 高维类别属性数据流离群点快速挖掘算法[J]. 软件学报, 2007, 18(4):933-942.
[59] AGYEMANG M, BARKER K, ALHAJJ R. Framework for mining Web content outliers[C]//Proc of ACM SAC'04. New York: ACM Press, 2004:590-594.
[60] AGYEMANG M, BARKER K, ALHAJJ R. Mining web content outliers using structure oriented weighting techniques and N-grams[C]//Proc of ACM SAC'05. New York: ACM Press, 2005:482-487.
[61] 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.
[62] YU Dan-tong, SHEIKHOLESLAMI G, ZHANG Ai-dong. FindOut: finding out outliers invery large datasets[J]. Knowledge and Information Systems, 2002, 4(4):387-412.

(上接第 8 页)

[33] WEATHERSPOON H, CHUN B G, SO C W, *et al*. Long-term data maintenance in wide-area storage systems: a quantitative approach[C]//Proc of the 1st International Warkshop on Peer-to-Peer Systems. 2005.
[34] SIT E, HAEBERLEN A, DABEK F, *et al*. Proactive replication for data durability[C]//Proc of the 5th International Workshop on Peer-to-Peer Systems. Santa Barbara: [s. n.], 2006.
[35] WEATHERSPOON, KUBIATOWICZ J D. Erasure coding vs replication: a quantitative comparison[C]//Proc of IPTPS. Berlin: Springer, 2002.
[36] RODRIGUES R, LISKOV B. High availability in DHTs: erasure coding vs replication[C]//Proc of the 4th International Workshop on Peer-to-Peer Systems. Berlin: springer, 2005:226-239.
[37] CHUN B G, DABEK F, HAEBERLEN A, *et al*. Efficient replica maintenance for distributed storage systems[C]//Proc of the 3rd Symposium on Networked Systems Design and Implementation. Berkeley, CA: USENIX Association, 2006.
[38] HENDRICKS J, GANGER G R, REITER M K. Low-overhead Byzantine fault Tolerant storage[C]//Proc of SOSP'07. New York: ACM Press, 2007:73-86.
[39] PREGUICA N, LOURENCO J, RODRIGUES R. Byzantium: eficient Byzantine fault-tolerant database replication[EB/OL]. (2006). http://citi. di. fct. unl. pt/project/project. php? id = 60.
[40] LISKOV B, RODRIGUES R. Tolerating Byzantine faulty clients in a Quorum system[C]//Proc of the 26th International Conference on Distributed Computing Systems. Washington DC: IEEE Computer Society, 2006:34-43.
[41] DOUCEUR J R, HOWELL J. Byzantine fault isolation in the farsite distributed file system[C]//Proc of the 5th IPTPS. 2006.
[42] 刘钢, 周敬利, 秦磊华, 等. 纠错码拜占庭容错 Quorum 中错误检测机制[J]. 计算机科学, 2007, 34(5):75-78.