文章编号:1001-506X(2011)01-0202-06

采用属性相关分析的异常数据检测方法

刘 波,潘久辉
(暨南大学信息科学技术学院计算机科学系,广东 广州 510632)

摘 要:为了发现数据库中的异常数据,提出了两个数据项集之间相关可信度的新概念,并研究了基于该度量的异常数据检测规则的计算算法,产生的规则适合于离散型属性孤立点的检测。在计算检测规则中,最小相关可信度阈值不需由用户指定,而是根据 1-数据项集的频率确定;利用相关可信度的性质,可以减小检测规则计算算法的时间复杂度。实验结果表明,采用该方法计算获得的相关规则进行异常数据检测,不仅效率较高,而且检测的准确率、查全率也较高。

关键词:异常数据;检测;相关分析;度量
中图分类号:TP 302 **文献标志码**:A **DOI**:10.3969/j.issn.1001-506X.2011.01.41

Study of abnormal data detecting method using attribute correlation analysis

LIU Bo, PAN Jiu-hui
(*Department of Computer Science, College of Information Science and Technology, Jinan University, Guangzhou 510632, China*)

Abstract: In order to discover abnormal data in a database, a new concept of correlated confidence between two data itemsets is proposed, and the algorithm of computing the rules for detecting abnormal data based on the metric is studied. The inferred rules are suitable for detecting discrete attribute outliers. In computing the rules, the minimum threshold of correlated confidence is determined by the frequency of 1-itemsets instead of users, and the temporal complexity of the algorithm for computing rules can be reduced by using the properties of correlated confidence. The experiment results show that the correlated rules inferred by the method for detecting abnormal data have not only high efficiency but also high precision and recall.

Keywords: abnormal data; detect; correlation analysis; measure

0 引 言

数据作为信息化建设和应用的主体,其质量保障是首先值得关注的问题。然而,在实际的数据库系统中,由于操作或传输等原因,常常出现数据拼写错误、缺失、不一致等现象。因此,有必要研究异常数据检测、纠正或处理技术。

成熟的数据清洗工具及流行的商用数据库管理系统的数据清洗构件,大多通过人工定义的数据约束、变换规则等检测与消除异常数据,但存在使用复杂、不能发现违背属性之间关联性的异常数据等问题。

近几年来,利用统计分析、机器学习、数据挖掘等技术研究数据异常的检测与处理方法已成为研究热点。例如:文献[1-5]提出的有关方法通过数据实例和数据源模式自动发现属性之间的关联规则、分类规则、相关关系、数据之间的相似性等,用于检测、纠正异常或不一致的数据,为消除多数据源之间的数据冲突、明确属性含义提供依据。

尽管通常在应用数据库中存在一些异常数据,但毕竟是少数,可以将它们看成孤立点。虽然异常或孤立点数据未必一定错误,但检测是消除错误数据的有效手段。孤立点分为类孤立点和属性孤立点[6]。类孤立点一般是在具有类别属性的数据集中存在的稀有类对象(元组或记录等),而属性孤立点是指数据集对象(元组或记录等)中错误的或偏离正常分布的属性值。研究属性孤立点的检测具有重要意义[7],因为许多数据集并非存在类别属性,而且检测及更正属性孤立点能够保留数据集完整信息。稀有的类对象不一定包含异常的数据,所以发现类孤立点的方法并不适合于发现属性孤立点[7]。

存在较多类孤立点的发现方法[8-9],但属性孤立点分析的方法较少。文献[1,10]提出采用分类、聚类、关联分析等方法发现错误的属性值,均存在一定的局限性:检测模型受

收稿日期:2009-11-06; **修回日期**:2010-07-15。
基金项目:广东省科技计划项目(2010B010600026);广东省教育部产学研结合项目(2007B090400095)资助课题
作者简介:刘波(1965-),女,副教授,主要研究方向为数据挖掘、数据库、信息集成。E-mail:ddxllb@163.com

分类、聚类、关联分析方法及参数影响较大;用全局数据得到的规则度量值评估局部数据是否正常,使得误检率高。文献[7]提出了基于子空间属性相关分析的孤立点检测算法(outlier detection from data subspaces, ODDS),对于一个关系数据集通过投影计算产生所有子关系,再根据所有属性的 O-Measure 或 Q-Measure 度量值及其排序序列的变化率发现属性孤立点边界,进而得到属性孤立点,这种方法虽然发现错误数据较全、准确性较高,但时间复杂度高。

本文提出了两个数据项集之间相关可信度的新度量,据以推导出强相关性的数据项集组合,产生一些异常数据检测规则,应用到异常数据的检测中。通常,关联或相关规则是在频繁项集基础上获得的,然而非频繁的项集并非一定包含异常数据,因此本研究对异常数据的检测并不以数据项集是否频繁为依据,而是直接通过两个数据项集之间的相关可信度获得相关规则。本研究工作提出的方法适合于离散型属性孤立点的检测,与同类相关工作文献[7]比较,不仅自动获取检测规则,而且人工指定参数数目少、算法时间复杂度较低。

1　有关概念

从一个关系数据集经过关联或相关分析所产生的规则多少、规则是否有趣,取决于相关度量的选择及度量阈值的设定。采用不同度量,对属性之间相关性评估结果可能不同。常用的关联或相关规则度量有:支持度、置信度、提升度、全置信度、余弦等[11]。这里提出一种新的度量,即相关可信度。

定义 1　相关可信度。假定两个数据项集 X_1 和 X_2,X_1 和 X_2 的属性集交集为空集,它们之间的相关可信度为

$$R\_Conf(X_1, X_2) = \frac{sup(X_1, X_2)}{\max(sup(X_1), sup(X_2))} \quad (1)$$

式中,$sup(X_1, X_2)$为同时满足项集 X_1 和 X_2 的支持度计数;$sup(X_1)$,$sup(X_2)$分别是 X_1 和 X_2 的支持度计数;max ($sup(X_1)$, $sup(X_2)$)是 $sup(X_1)$和 $sup(X_2)$中的较大者。

例 1　表 1 为一关系数据集实例,记录了网上书店客户情况,其中包括 3 条记录的异常数据(即用斜体字表示:*Low*、*Australia* 、*GD*)。下面是数据项集之间的相关可信度计算示例。

表 1　关系数据集实例

CustID	Country	Province/State	City	Proffession	Amount
C1	China	Guangdong	Guangzhou	Teacher	High
C2	China	Guangdong	Guangzhou	Teacher	High
C3	China	Hunan	Changsha	Teacher	*Low*
C4	China	Hunan	Changsha	Teacher	High
C5	*Australia*	Hunan	Changsha	Salesman	High
C6	China	Hunan	Changsha	Student	Middle
C7	USA	California	LA	Salesman	Middle
C8	China	*GD*	Guangzhou	Student	Middle
C9	China	Guangdong	Guangzhou	Student	Middle
C10	China	Jiangxi	Nanchang	Student	Middle

$R\_Conf(\{Guangzhou\},\{Guangdong\}) = 3/4 = 75\%$

$R\_Conf(\{USA,California\},\{LA\}) = 1/1 = 100\%$

$R\_Conf(\{China, Jiangxi\},\{Nanchang\}) = 1/1 = 100\%$

$R\_Conf(\{Teacher\},\{High\}) = 3/4 = 75\%$

$R\_Conf(\{China,Guangzhou\},\{GD\}) = 1/4 = 25\%$

$R\_Conf(\{Changsha\},\{Australia,Hunan\}) = 1/4 = 25\%$

$R\_Conf(\{Teacher\},\{Low\}) = 1/4 = 25\%$

从表 1 容易得知,虽然{USA, California, LA}、{China, Jiangxi, Nanchang}均为稀有数据项集,但{USA, California}与{LA}、{China, Jiangxi}与{Nanchang}的相关可信度为 100%;{China, Guangzhou, GD}、{China, Hubei, Changsha}、{Teacher, Low}也为稀有数据项集,但{China, Guangzhou}与{GD}、{ Changsha }与{ Australia, Hunan }、{ Teacher }与{Low}的相关可信度低,所以包含这些数据项集的记录很有可能存在异常数据。

定义 2　相关可信。给定最小相关可信度阈值 $\theta(\theta<1)$,若 $R\_conf(X_1, X_2) \geqslant \theta$,则 X_1 和 X_2 是相关可信的;若 $R\_conf(X_1, X_2) < \theta$,则 X_1 和 X_2 不是相关可信的。

相关可信度具有如下几个重要性质:

性质 1　相关可信度具有零不变性,即相关可信度度量的值不受零事务的影响。

证明　根据文献[11]给出的零事务定义,零事务是指不包含任何考查项集的事务。定义的 X_1 和 X_2 两个数据项集的相关可信度计算式,仅与包含 X_1、X_2 的事务计数有关,与 $\overline{X_1 X_2}$(既不包含 X_1 也不包含 X_2)的事务计数无关,即不受零事务个数的影响。　　证毕

由于大型数据集常常具有许多零事务,因此在为相关分析选择合适的兴趣度量时,考虑零不变性是重要的[11]。

性质 2　假定两个数据项集 X_1 和 X_2,且 $sup(X_1) \geqslant sup(X_2)$,则 $R\_Conf(X_1, X_2) \geqslant R\_Conf(X_1, X_2')$,$X_2'$ 是 X_2 的超集。

证明　因为添加任何项到项集 X_2 中都不可能增加 $sup(X_1, X_2)$,而且 $sup(X_1) \geqslant sup(X_2)$,也不可能减少 max ($sup(X_1)$, $sup(X_2)$),因而 $R_Conf(X_1, X_2')$不会增加。

证毕

用实例还容易证明,相关可信度量并不具有向下封闭性[11],即如果两个数据项集是相关可信的,则它们的子集之间未必是相关可信的。但是,如果两个数据项集不是相关可信的,则支持度较小的数据项集进一步增长而支持度较大的数据项集不变,也使得它们仍然不是相关可信的。这可用性质 3 描述。

性质 3　两个数据项集 X_1 和 X_2,如果它们是非相关可信的,且 $sup(X_1) \geqslant sup(X_2)$,则 X_1 和 X_2 的任何超集也不是相关可信的。

证明 对于两个项集 X_1 和 X_2，由于它们是非相关可信的，则 $R\_conf(X_1, X_2) < \theta$；根据性质 2，在 X_2 中添加任何数据项，不会增加 $R\_conf(X_1, X_2)$，也就仍然不会满足最小相关可信度阈值。 证毕

定义 3 相关规则。给定两个相关可信的数据项集 X_1 和 X_2，相关规则记为：$X_1 \leftrightarrow X_2$。

这里定义的相关规则与通常以支持度-置信度框架或其他度量定义的关联规则或相关规则的判别条件不同，本文定义的相关规则与数据项的频繁度无关，且具有对称性质的相关要求，即同时满足 $X_1 \rightarrow X_2$ 和 $X_2 \rightarrow X_1$ 两条规则，更适合于准确检测异常数据项。

定义 4 相关规则的长度。相关规则 $X_1 \leftrightarrow X_2$ 的长度为数据项集 X_1 和 X_2 的长度之和，即 $|X_1| + |X_2|$，X_1 和 X_2 中包含数据项的总数。长度为 $k(k \geqslant 2)$ 的相关规则称为 k-相关规则。

2 相关规则的计算算法

2.1 基本思路

如何高效计算相关规则，以及如何利用相关规则检测异常数据，是本文着重研究的两个问题。本节首先研究第一个问题。

对于关系数据集 D，假设 n 为记录的个数，m 为每一记录具有属性的个数。如果要计算所有数据项集的相关性，复杂度非常高，因为有 2^m 个可能的属性子集，各属性取值还存在多种可能。用穷举法计算相关规则是不现实的，本研究在所定义的子空间中计算，并利用一些启发式策略，使得计算过程有效且得到的相关规则有意义。主要思路说明如下：

(1) 一些属性之间显然不存在相关性，因此可以通过选择属性子集以减少相关规则的计算量。

一般地，对于取值个数等于记录个数或为 1 的属性一定不会与其他属性存在相关性，如：表 1 中 CustID(客户编号)显然与其他任一属性不存在相关性，在相关规则的计算中不考虑该类无用属性。

(2) 用于检测异常数据的规则并非需要涉及较多属性，也不需要计算出所有相关规则。

首先，计算出有关属性 1-数据项集之间的相关可信度，得到 2-相关规则集 $Rule = \{(X_1, X_2, A_1, A_2) \mid R\_Conf(X_1, X_2) \geqslant \theta, sup(X_1) \geqslant sup(X_2), |X_1| = 1, |X_2| = 1, A_1$ 和 A_2 分别对应 X_1 和 X_2 的属性名字符串$\}$，非相关规则集 $Prune = \{(X_1, X_2, A_1, A_2) \mid R\_Conf(X_1, X_2) < \theta, sup(X_1) \geqslant sup(X_2), |X_1| = 1, |X_2| = 1, A_1$ 和 A_2 分别对应 X_1 和 X_2 的属性名字符串$\}$。接着，再通过扩展 *Prune* 中的 X_1 项集计算 k-相关规则($k > 2$)。

值得注意的是，对于两个数据项集 X_1 和 X_2，在 *Rule* 或 *Prune* 中的保存顺序代表支持度计数的大小顺序，对于规则本身而言无关紧要，但对进一步选择项集扩展以及避免产生重复规则具有启发作用。

利用相关可信度性质 2，对于 *Rule* 中的 2-相关规则，可以通过抑制 X_2 项集的扩展，减少较弱相关规则的产生；虽然，X_1 的扩展还可能产生更多相关规则，但也没有必要进一步扩展，因为，规则越简单越易于检测异常数据，而且由 *Rule* 集中数据项扩展得到的相关规则也可能由 *Prune* 集中的数据项扩展得到。例如，在表 1 中，利用相关规则 {Guangzhou}↔{Guangdong} 可以检测到客户 C8 的“GD”属性值异常，没有必要用 {China, Guangzhou}↔{Guangdong} 检测，而且 {China, Guangzhou}↔{Guangdong} 也可由 *Prune* 中的 ({China}, {Guangdong}, Country, Province/State) 通过扩展 {China} 数据项集得到。

利用相关可信度的性质 3，也可以控制相关可信的数据项集的进一步扩展，减少一些非相关可信的项集对判断。对于 *Prune* 中的长度为 2 的非相关规则，只需在 X_1 中添加 1-数据项，将产生的 3-相关规则存入 *Rule* 集合，而非相关的、长度之和为 3 的 X_1 和 X_2 数据项集仍保留在 *Prune* 中，且满足 $sup(X_1) \geqslant sup(X_2)$。

同样，对 *Prune* 中长度之和为 $k-1$ 的 X_1 和 X_2 数据项集，扩展 X_1，将产生的 k-相关规则并存入 *Rule* 集合，而非相关的、长度之和为 k 的 X_1 和 X_2 数据项集仍保留在 *Prune* 中，且满足 $sup(X_1) \geqslant sup(X_2)$。

(3) 通过限制相关规则的最长长度，减少多余或复杂的检测规则产生。

需要的检测规则满足如下限定：$X_1 \leftrightarrow X_2$，$sup(X_1) \geqslant sup(X_2)$，$2 \leqslant |X_1| + |X_2| \leqslant k$，$k(2 \leqslant k \leqslant m)$ 为规则的最大长度。因为，这种限定不仅使得计算复杂度降低，而且产生的规则简单实用。

(4) θ 对 *Rule* 集合的结果影响较大，鉴于数据项集 X_1 和 X_2 的相关可信度与数据集 D 的局部属性有关，设 X_1 和 X_2 的对应属性包括：$A_1, \cdots, A_k$，各属性所有的 1-数据项在数据集 D 中的最大频率为 $m_1, \cdots, m_k$，数据集 D 中的记录数为 n，则关于属性 $A_1, \cdots, A_k$ 的 θ 值为

$$\theta(A_1, \cdots, A_k) = (m_1/n + \cdots + m_k/n)/k \tag{2}$$

如：在表 1 中有

$$\theta(\text{Country, Province/State, City}) = (0.8 + 0.4 + 0.4)/3 = 53.3\%$$

2.2 算法描述

根据以上思路，提出了基于相关可信度的相关规则挖掘算法 RCMine，得到的规则适用于异常属性值的检测，如图 1 所示。

```
算法:RCMine //基于相关可信度的相关规则挖掘
输入:
    D:关系数据集(去掉了无用属性);
    m:属性个数;
    k:规则的最大长度(2≤k≤m);
输出:相关规则子集 Rule。
方法:
    (1) for i=1 to m
    (2)         计算 A_i 所有 1-数据项集的频率及其最大值 m_i;
    (3) for i=1 to m
    (4)         产生包括 A_i 和 A_{i+1}属性的 D 数据集投影 S;
    (5)         对 S 中的每一对满足 sup(X1,X2)≠0 的 X1
                和 X2
    (6)             If R_Conf(X1,X2)≥θ(X1,X2) then
    (7)                 将(X1,X2,A1,A2)保存到 Rule 集合,满
                        足 sup(X1)≥sup(X2);
    (8)             Else
    (9)                 将(X1,X2,A1,A2)保存到 Prune 集合,
                        满足 sup(X1)≥sup(X2);
    (10) Endfor;
    (11) For Prune 中的每一元素
    (12)    If |X1|+|X2|<k then
    (13)        从 Prune 中移出该元素 e;
    (14)        在 e 的 X1 和 A1 中增加 1-数据项及属性,对每
                一对满足 sup(X1,X2)≠0 的 X1 和 X2
    (15)        If R_Conf(X1,X2)≥θ(X1,X2) then
    (16)            将(X1,X2,A1,A2)保存到 Rule 集合,满足
                    sup(X1)≥sup(X2);
    (17)        Else
    (18)            将(X1,X2,A1,A2)保存到 Prune 集合,满足
                    sup(X1)≥sup(X2);
    (19) Endfor
```

图 1　RCMine 算法

图 1 中的 RCMine 算法分两个部分,分别说明如下。

(1) 第一部分,句(1)～句(10)

句(1)和句(2)计算所有 1-数据项集的频率,用于不同属性数据项组合的最小相关可信度阈值计算。句(3)～句(10),计算 2-相关规则。由 D 数据集的两个属性产生投影子集 S,计算 S 中所有满足 $R_conf(X_1,X_2)\geqslant\theta(X_1,X_2)$的数据项 X_1和 X_2,将(X_1,X_2,A_1,A_2)保存到 *Rule* 集合中,$|X_1|=1$、$|X_2|=1$,并且 $sup(X_1)\geqslant sup(X_2)$。同时,产生所有满足 $R_Conf(X_1,X_2)<\theta(X_1,X_2)$的数据项 X_1和 X_2,将(X_1,X_2,A_1,A_2)保存到 *Prune* 集合中,$|X_1|=1$、$|X_2|=1$,并且 $sup(X_1)\geqslant sup(X_2)$。

(2) 第二部分,句(11)～句(19)

计算 k-相关规则($k>2$)。对于 *Prune* 集合中的每一元素,若$|X_1|+|X_2|<k$,添加一个非 A_1、非 A_2属性的 1-数据项到 X_1项集,检测新的 X_1和 X_2是否相关可信;若是,则 A_1中也增加新数据项的属性名,并将(X_1,X_2,A_1,A_2)添加到 *Rule* 集合中,同时调整顺序,满足 $sup(X_1)\geqslant sup(X_2)$;否则,将$(X_1,X_2,A_1,A_2)$仍保存在 *Prune* 中,用于添加另一个非 A_1、非 A_2属性的 1-数据项到 X_1项集中,继续判断与计算。

2.3　算法分析

给定一个包括 m 个属性的关系数据集 D,共有 n 个元组,k 为规则的最大长度($2\leqslant k\leqslant m$),$n_1,n_2,\cdots,n_m$为各属性取值个数。

(1) 第一部分,句(1)～句(10)

计算所有属性 1-数据项集的频率需扫描 n 个元组,复杂度为 $O(n)$。

含有 2 个属性的投影数据集个数为

$$C_m^2=\frac{m!}{(m-2)!\times 2}=\frac{1}{2}m(m-1)\tag{3}$$

式中,m 为属性关系中包括的属性个数。

每个 2-投影数据集计算两个项集之间的相关可信度时,须扫描 n 个元组,因此,第一部分的时间复杂度为 $O(nm^2)$。

(2) 第二部分,句(11)～句(19)

算法第一部分产生的 *Rule* 和 *Prune* 数据集中元素总数为

$$n_1(n_2+\cdots+n_m)+n_2(n_3+\cdots+n_m)+\cdots+n_{m-1}n_m=$$
$$n_1n_2+\cdots+n_1n_m+n_2n_3+\cdots+n_2n_m+\cdots+n_{m-1}n_m\ //$$
$$\text{共}\frac{1}{2}m(m-1)\text{ 个项式之和}\leqslant\frac{1}{2}m(m-1)r^2=O(m^2r^2)\tag{4}$$

式中,m 为属性关系中包括的属性个数;$r=\max(n_1,n_2,\cdots,n_m)$。

Prune 数据集中产生的元素总数不会超过式(4)分析的个数,对 *Prune* 中的每一元素,扩展次数不超过如下数据:

(1) 计算 3-相关规则:$C_{m-2}^1r=(m-2)r$

(2) 计算 4-相关规则:$C_{m-3}^1r=(m-3)r$

(3) 计算 k-相关规则:$C_{m-k+1}^1r=(m-k+1)r$

即 *Prune* 集中的一个元素,扩展项集总次数不超过

$$(m-2)r+(m-3)r+\cdots+(m-k+1)r\leqslant O(kmr)\tag{5}$$

式中,k 为规则的最大长度;m 为属性关系中包括的属性个数;$r=\max(n_1,n_2,\cdots,n_m)$。

Prune 中的各元素每次扩展完 X_1项集后,计算相关可信度,需扫描 n 个元组,因此,算法第二部分的时间复杂度为 $O(nm^2r^2kmr)$,即 $O(nkm^3r^3)$。

所以,算法 RCMine 总的时间复杂度为:$O(nm^2+nkm^3r^3)$,即 $O(nkm^3r^3)$。一般地,除了个别属性(如客户编码)的取值个数能够达到 n,其他用于相关分析的属性取值个数远远小于 n,即 $r\ll n$,当 k 取最大值 m 时,算法的时间复杂度为 $O(nr^3m^4)$。

3　异常数据的检测方法

3.1　检测准则

由 RCMine 算法计算得到的 *Rule* 规则集,可用于检测已存在或新增元组的异常数据。对于任一$(X_1,X_2,A_1,A_2)\in$

Rule,可以用于检测从源数据集选择 A_1 和 A_2 属性产生的投影子集(下面用 D'表示)中的异常数据。根据相关规则 $X_1 \leftrightarrow X_2$,需要关注 D'中满足 $X_1 \cap \overline{X_2}$ 及 $\overline{X_1} \cap X_2$ 的记录,$X_1 \cap \overline{X_2}$ 表示满足 X_1 中的属性值但不满足 X_2 中的属性值,$\overline{X_1} \cap X_2$ 表示满足 X_2 中的属性值但不满足 X_1 中的属性值。按照所产生的 Rule 规则集的特点,可以归纳以下异常数据检测准则:

准则 1 如果 $X_1 \leftrightarrow X_2$,$|X_1| \geqslant 1$,$|X_2| = 1$,且不存在 $Y \leftrightarrow \overline{X_2}$($X_1 \subset Y$),则在 D'数据集中,满足 $X_1 \cap \overline{X_2}$ 的记录中 A_2 属性异常。

准则 2 如果 $X_1 \leftrightarrow X_2$,$|X_2| \geqslant 1$,$|X_1| = 1$,且不存在 $Y \leftrightarrow \overline{X_1}$($X_2 \subset Y$),则在 D'数据集中,满足 $\overline{X_1} \cap X_2$ 的记录中 A_1 属性异常。

准则 3 如果 $X_1 \leftrightarrow X_2$,$|X_1| > 1$,$|X_2| \geqslant 1$,$X_1 = Y \cup Z$,则在 D' 数据集中,满足 $Y\bar{Z} \cap X_2$ 的记录中 A_z 属性存在异常,其中,A_z 为 Z 的属性,$A_z \subseteq A_1$。

准则 4 如果 $X_1 \leftrightarrow X_2$,$|X_2| > 1$,$|X_1| \geqslant 1$,$X_2 = Y \cup Z$,则在 D'数据集中,满足 $X_1 \cap Y\bar{Z}$ 的记录中 A_z 属性存在异常,其中,A_z 为 Z 的属性,$A_z \subseteq A_2$。

例 2 在表 1 中,根据 RCMine 算法得到相关规则集,可以按照上述检测准则发现所有异常属性值。

(1) 由{Guangzhou}↔{Guangdong},根据检测准则 1,可以发现 C8 客户的 Province/State 属性值"GD"异常;

(2) 由{Teacher}↔{High},根据检测准则 1,可以发现 C3 客户的 Amount 属性值"Low"异常;

(3) 由{Changsha}↔{China, Hunan},根据检测准则 4,可以发现 C5 客户的 Country 属性值"Australia"异常。

检测是一个交互式过程,异常数据由用户确定是否错误,如:上述 C3 客户的 Amount 属性值"Low"异常,但并非一定是错误的。

可能存在这样一种情况:某些异常数据项集是相关可信的,使得相关规则中包括异常数据。如表 1 中,若 A8 客户的 Country 属性值为 Japan ,则 $R\_Conf(\{\text{Japan}\},\{\text{GD}\}) = 100\%$,产生{Japan}↔{GD}。因此,在使用 RCMine 算法得出相关规则集后,应通过与用户交互,删除其中明显不合理或包含拼写错误的规则,才能应用于异常数据的检测。如:删除规则{Japan}↔{GD},才能根据{Guangzhou}↔{Guangdong}检测出"GD"异常,并确认为错误数据后,再由{China, Guangzhou}↔{Guangdong}检测出"Japan"异常。

3.2 检测算法及其复杂度分析

Check 算法如图 2 所示,其对 *Rule* 集合中的每一规则,按照 4 个检测准则对数据集中每一元组进行检测,给出异常数据警告,由用户确认或进行纠正等。

假定一个关系数据集 D,共有 n 个元组,*Rule* 集合包括 n_r 条规则,因为在使用某一检测规则时,还要与其他规则比较,所以检测算法复杂度为 $O(nn_r^2)$。

根据式(4)的分析,n_r 随着 m(元组属性数目)、r(最大的属性取值数目)呈多项式级增长,所以检测算法复杂度也是多项式级的。

Rule 数据集也可用于检测新增数据元组。显然,随着数据集的变化,数据项集之间的相关可信度会改变,需要重新计算或增量维护 *Rule* 集,如何增量维护 *Rule* 集是另一复杂问题,在此不作进一步研究。

```
算法:Check //异常数据检测
输入:
    D:关系数据集(去掉了无用属性);
    m:属性个数;
    k:规则的最大长度(2≤k≤m);
    Rule:相关规则集 Rule。
输出:异常数据
方法:
    (1) For Rule 中的每一元素(规则)
    (2)  产生包括 A1 和 A2 属性的 D 数据集投影 D';
    (3)  对 D'中的每一元组 e
    (4)      If  符合检测准则 1 或检测准则 2 或检测准则
             3 或检测准则 4
    (5)      then  显示异常数据警告,等待用户确认或处
             理异常数据;
    (6) Endfor
```

图 2 Check 算法

4 模拟实验

以第 1 节中的实例数据集为模拟实验背景,分别产生 50 000、100 000、200 000、300 000 条网上书店客户记录,其中 1‰的记录包含 1~3 个属性异常数据(包括拼写错误、缩写等),在 Window XP、CPU 为 intel Pentium Dual Core 的微机环境下,采用 Oracle9.2 数据库管理系统及 PL/SQL 编程,进行如下模拟实验。

在 RCMine 算法中,参数 k 设置为 3,针对不同大小的数据集,相关规则计算及异常数据检测的平均时间、查准确率及查全率如表 2 所示。

表 2 RCMine 算法实验结果

数据集记录数目	50 000	100 000	200 000	300 000
RCMine 的时间/s	0.531	0.953	2.156	5.853
查准率/(%)	100	100	100	100
查全率/(%)	100	100	100	100

此外,按照同类研究文献[7]提出的属性孤立点检测算法 ODDS,采用与 RCMine 算法同样的度量阈值确定方法及关系表投影元组长度阈值 k,分别选用该算法中定义的 Q-Measure和 O-Measure 作为属性度量,计算属性度量及检测异常数据的平均时间、查准率及查全率,结果如表 3 所示。

表 3　ODDS 算法实验结果

数据集记录数目		50 000	100 000	200 000	300 000
Q-Measure 度量	ODDS 的时间/s	10.437	12.047	14.656	17.022
	查准率/(%)	100	100	100	100
	查全率/(%)	100	100	100	100
O-Measure 度量	ODDS 的时间/s	7.234	9.75	12.390	15.025
	查准率/(%)	100	100	100	100
	查全率/(%)	66	67	66.5	66.7

查准率及查全率的计算公式为

$$\text{查准率} = \frac{\text{被检测到的实际错误记录数目}}{\text{检测到的可疑记录数目}} \quad (6)$$

$$\text{查全率} = \frac{\text{被检测到的实际错误记录数目}}{\text{实际存在的错误记录数目}} \quad (7)$$

对比表 2 和表 3 的结果,可以观察到,采用本文方法计算检测时间较快,且异常数据查准率和查全率高。采用 ODDS 算法,异常数据查准率也高,但时间消耗比本文算法多;且采用 O-Measure 属性度量时,异常数据查全率仅为 66%～67%。

5　结　论

本文采用数据集属性之间的相关可信度,得到相关规则,用于检测数据库离散型属性异常数据,与其他有关方法相比,有如下三方面的优势:

(1) 采用新的度量(即相关可信度)评估属性之间的相关性。这种度量具有三条性质,利用其中两条性质能够减小计算相关规则算法的复杂度,且基于该度量产生的规则具有对称性;

(2) 最小相关可信度阈值不是用户指定的唯一固定值,而根据两个数据项集所包括属性的 1 -数据项集频率自动计算,减少了该参数对结果的影响;

(3) 相关规则计算和异常数据检测算法的复杂度均为多项式级,与其他属性异常检测方法相比,复杂度降低。

文献[7]提出的属性孤立点检测算法 ODDS,采用 Q-Measure作为属性度量时,算法时间复杂度为 $O(nm2^{m-1})$;采用 O-Measure 作为属性度量时,算法时间复杂度为 $O(nm(m+1)2^{m-2})$。其中,n 为数据集包含记录的个数,m 为每一记录具有属性的个数。由此可见,基于这两种度量的算法时间复杂度均为指数级,比本文提出的算法时间复杂度高。实验结果还证明了本文方法对异常数据的查准率和查全率较高。

参考文献:

[1] Chiang F, Miller R J. Discovering data quality rules[C] // *Proc. of the International Conference on Very Large Data Bases*, 2008:1166 - 1177.

[2] Choi O H, Lim J E, Na H S, et al. An efficient method of data quality using quality evaluation ontology[C] // *Proc. of 3rd International Conference on Convergence and Hybrid Information Technology*, 2008:1058 - 1061.

[3] 沈睿芳, 郭立甫, 时希杰. 数据挖掘中的数据预处理模型与算法研究[J]. 计算机应用系统, 2005(7):44 - 46. (Shen R F, Guo L P, Shi X J. Study on the model and algorithms of data preprocessing in data mining[J]. *Computer Application System*, 2005(7):44 - 46.)

[4] Kalashnikov D V, Mehrotra S. A domain-independent data cleaning via analysis of entity-relationship graph [J]. *ACM Trans. on Database Systems*, 2006, 31(2):716 - 767.

[5] Fan W F, Geerts F, Jia X B, et al. Conditional functional dependencies for capturing data inconsistencies[J]. *ACM Trans. on Database Systems*, 2008, 33(2):444 - 491.

[6] Zhu X, Wu X. Class noise vs. attribute noise: a quantitiative study of their impacts[J]. *Artificial Intelligence Review*, 2004, 22(3):177 - 210.

[7] Koh J L Y, Lee M L, Hsu W, et al. Correlation-based detection of attribute outliers[C] // *Database Systems for Advanced Applications*, 2007:164 - 175.

[8] Ghoting A, Otey M E, Parthasarathy S. LOADED: link-based outlier and anormaly detection in evolving data sets[C]// *Proc. of IEEE 4th International Conference on Data Mining*, 2004: 387 - 390.

[9] He Z Y, Deng S C, Xu X F, et al. A fast greedy algorithm for outlier mining[C]// *Proc. of The 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2006:567 - 576.

[10] Ciszak L. Application of clustering and association methods in data cleaning[C] // *Proc. of the International Multiconference on Computer Science and Information Technology*, 2008:7 - 103.

[11] Han J W, Kamber M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 2 版. 北京: 机械工业出版社, 2006:168 - 172. (Han J W, Kamber M. *Data mining concepts and techniques*[M]. Fan M, Meng X F, trans. 2nd ed. Beijing: China Machine Press, 2006:168 - 172.)