

文章编号:1001-9081(2009)11-3077-03

孤立点用户意义分析在质量管理中的应用

王 越^{1,2}, 刘亚辉¹, 徐传运²

(1. 重庆邮电大学 计算机科学与技术学院, 重庆 400065; 2. 重庆理工大学 计算机软件及应用研究所, 重庆 400050)

(wangyue@cqit.edu.cn)

摘 要: 现存有关孤立点分析的研究很少解释识别出的孤立点的用户意义, 而孤立点通常都包含着重要的信息, 在许多应用领域中对于孤立点意义的解释和孤立点本身同等重要。因此, 给出孤立点用户意义的定义, 并提出一种基于距离和的孤立点用户意义分析算法(DSCM), 对每一孤立点给出相应的解释, 以帮助用户更好地理解孤立数据。应用到质量管理中的结果表明, 该算法是有效的和实用的, 且易用性较强。

关键词: 孤立点; 用户意义; 距离和; 质量管理; 数据挖掘

中图分类号: TP311.13 **文献标志码:** A

Application of outlier customer meaning analysis in quality management

WANG Yue^{1,2}, LIU Ya-hui¹, XU Chuan-yun²

(1. College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China;

2. Institute of Software Technology and Application, Chongqing University of Technology, Chongqing 400050, China)

Abstract: The customer meaning for outlier explanation is rarely provided in the current studies. The outliers usually contain important information, and for many applications, the explanations are as important to the user as the outliers. A new definition of outlier customer meaning was given, and a new outlier customer meaning analysis algorithm named DSCM was put forward based on distance sum. The algorithm gave an explanation of every outlier, which improved the user's understanding of the data. Then the algorithm was applied to quality management, and the results show that the algorithm is effective and practical, and more easy to use.

Key words: outlier; customer meaning; distance sum; quality management; data mining

0 引言

孤立点分析是数据挖掘技术中一个非常重要的研究方向, 它是从大量复杂的数据中检测出存在于小部分异常数据中新颖的、与常规数据模式显著不同的数据模式。孤立点也叫作异常数据, 而异常数据往往会带来严重的后果或包含着重要信息。

在质量管理领域, 质量数据出现异常的其中一个原因是由系统性原因(系统因素)或可以避免的原因而造成的产品质量波动。这类原因在生产过程中并非大量存在, 表现为具有方向性或周期性、突然而至的对产品质量产生影响。这类原因虽少, 但对产品质量造成的影响往往较大, 如设备出现故障、操作者违反操作规程、原材料性质变化等。一般情况下, 异常波动在生产过程中不允许存在, 一旦出现, 必须立即查明原因, 消除异常波动^[1]。在质量管理中, 为了提高产品的质量也必须消除这种异常波动, 因此, 针对历史产品数据记录中的那些异常数据, 对其进行分析, 确定产生该异常波动的实际原因(即是由哪一个或哪几个属性值发生异常引起的), 从而采取相应的措施以提高产品的质量, 具有十分重要的现实意义。

孤立点分析过程通常可以看作三个子问题^[2]: 1) 什么样的数据是不一致的, 即孤立点的定义; 2) 有效挖掘孤立点的方法; 3) 孤立点的意义, 即对孤立点的合理解释。目前的研

究大多集中在解决前两个子问题, 即对孤立点的定义和有效挖掘孤立点的方法, 而对挖掘出的孤立点的意义的解释却很少涉及。本文采用一种基于距离和的孤立点用户意义分析算法对检测出的孤立点进行分析, 给出了引起该数据点发生异常的具体原因, 即分析出是由哪一个或哪几个属性引起的(这里称之为原因属性), 并且进一步给出了每个原因属性的孤立程度, 最后将算法应用到质量管理中, 用具体实验证实了该方法的有效性和实用性。

1 基于距离和的孤立点分析方法

1.1 基于距离的孤立点分析

基于距离的孤立点分析是这样定义的: 如果数据集合 S 中对象至少有 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(Distance-based, DB)的孤立点, 即 $DB(P, d)$ ^[3]。换句话说, 可以将基于距离的孤立点看作是那些没有“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。它又分为三个基本类型: 基于索引(index-based)的算法、基于嵌套循环(nested-loop)算法和基于单元(cell-based)的方法。但它们分别存在输入参数很难确定, 并且对于参数相当敏感, 不同参数的结果有很大不稳定性; 不能给定孤立点的孤立程度; 算法的复杂度较高等缺点^[4-5]。

收稿日期: 2009-05-22; 修回日期: 2009-07-10。

基金项目: 重庆市科技攻关资金资助项目(CSTC, 2009AB2049; CSTC, 2009AC2068)。

作者简介: 王越(1961-), 男, 北京人, 教授, 主要研究方向: 数据挖掘、数据库、嵌入式系统; 刘亚辉(1982-), 男, 河南驻马店人, 硕士研究生, 主要研究方向: 数据库、数据挖掘; 徐传运(1979-), 男, 重庆人, 讲师, 博士研究生, 主要研究方向: 数据挖掘、软件工程。

1.2 基于距离和的孤立点分析方法

针对基于距离的孤立点分析算法存在的缺点,提出了基于距离和 (Distance Sum, DS) 的孤立点分析算法^[6]。与 $DB(P, d)$ 孤立点一样, DS 孤立点分析算法使用同样的距离函数,如绝对距离或者欧氏距离,但并不根据 p 和 d 判定孤立点,而是首先计算数据集中对象两两之间的距离,然后计算每个对象与其他对象的距离之和,设 M 为用户期望的孤立点个数,则距离之和最大的前 M 个对象即为要挖掘的孤立点,这样则可消除需要用户设置参数 p 和 d 的要求。

为检测基于距离和的孤立点,算法需要计算 n^2 次数据对象间的距离,考虑到质量管理中的数据量一般不会很大,算法仍可在较短的时间内返回结果。

2 基于距离和的孤立点用户意义分析算法

2.1 相关概念定义

给定一个原始数据集 S ,该数据集是一个 n 行 m 列的数据表,共有 n 个数据对象, m 个属性,据此,给出以下几个概念的定义。

定义 1 属性空间。由数据集 S 中的 m 个属性构成的集合称为属性空间,记作 $P_m, P_m = (p_1, p_2, \dots, p_m)$ 。

定义 2 属性子空间。属性空间 P_m 的子集称为属性子空间,记作 $P_{m'}$,其中, $1 \leq m' \leq m$, m 和 m' 表示集合中元素的个数。

属性空间 P_m 和属性子空间 $P_{m'}$ 满足关系: $P_{m'} \subseteq P_m$, 特别地,当 $m' = m$ 时, $P_{m'} = P_m$ 。

定义 3 属性子空间的超空间。属性子空间 $P_{m'}$ 的超空间是指属性空间 P_m 的所有子集中包含 $P_{m'}$ 的那些集合,记作 $SP_{m'}$ 。

为了帮助理解以上 3 个定义,现给出一个简单的例子:假设数据集 S 包含 3 个属性分别为 A, B, C ,则其属性空间为集合 $\{A, B, C\}$;其属性子空间有 7 个,分别为集合 $\{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}$;其中属性子空间 $\{A\}$ 的超空间有 3 个,分别为 $\{A, B\}, \{A, C\}, \{A, B, C\}$ 。

定义 4 原因属性集。能够解释某一数据对象成为孤立点的最小属性子空间组成的集合,称为原因属性集,记为 $CP, CP = (P_{m'}^i) (1 \leq i \leq m)$ 。其中,元素 $P_{m'}^i$ 表示该属性子空间 $P_{m'}$ 由 i 个属性组成。

原因属性集可以理解为由若干个集合组成的集合,即集合的集合,其中集合中的每一个元素都表示一个原因属性。

定义 5 孤立点用户意义。由该孤立点的原因属性集及对原因属性集中每一个原因属性的描述组成的知识。

定义中对原因属性集中每一原因属性的描述非常灵活,没有一定的限制,要根据数据集的特征以及用户的需求而定。

2.2 算法基本思想

根据给定的数据集 S 和已知检测出的孤立点集合 O ,首先把数据集的所有属性子空间 $P_{m'}$ 按照所含元素个数升序排列放进一个队列 Q 中,针对孤立点集合中每一个元素 O_i ,从队列 Q 中取出一个属性子空间 $P_{m'}$,然后在该属性子空间中采用基于距离和的算法检测孤立点,并输出孤立点集 O' 。随后,按以下两种情况处理:

1) O' 中含有对象 O_i ,则移除队列 Q 中当前属性子空间 $P_{m'}$ 的超空间 $SP_{m'}$,把 $P_{m'}$ 添加到原因属性集 CP 中,然后再从

队列 Q 中取出下一个元素,循环直至队列为空;

2) O' 中不含有对象 O_i ,则从队列 Q 中取出下一个属性子空间 $P_{m'}$,再按以上方法循环进行,直至队列为空。

经过以上处理后,最终得到一个包含若干元素的原因属性集 CP , CP 中的每一个属性子空间都是能够解释该数据对象成为孤立点的可能原因,需要引起用户的注意。算法最后要结合数据集 S 中数据的特征以及用户的需求给出一个对原因属性集 CP 中的每一个属性子空间的有意义的描述,即给出孤立点的用户意义。

2.3 DSCM 主要算法实现

```
//input:数据集 S,孤立点集 O
//output:所有孤立点的原因属性集 CP
DSCM (S, O)
//把 S 中的所有属性子空间  $P_{m'}$  按照所含元素个数
//升序排列放进一个队列 Q 中
FillInQueue (S,  $P_{m'}$ , Q);
FOR i FROM 1 TO O.Count
    WHILE (Q.front != Q.rear)
         $Q_i = Q.front \rightarrow data$ 
        //采用基于距离和的孤立点检测算法,  $M$  是期望的孤立点数
         $O' = FindOutliers(Q_i, M)$ ;
        IF ( $O'$ .Contains ( $O[i]$ ))
             $CP[i].Add(Q_i)$ ;
             $Q.Remove(SP_{m'})$ ; // 移除  $P_{m'}$  的超空间
        ELSE
             $Q.front++$ ;
        END IF;
    END WHILE;
END FOR;
END;
```

3 DSCM 算法在质量管理中的应用实现

3.1 基本步骤

1) 输入原始数据集。实验的数据采用质量管理中的“产品硬度”数据表作为数据源,该数据表共有 50 条记录,4 个数值型属性:“产品硬度”、“材料硬度”、“加工温度”和“添加量”,另外数据表还有两个字符型属性,字符型属性值在质量管理中是作为分层变量来使用的,这里并不参与孤立点用户意义分析算法 (DSCM) 的运算。

为了便于表示和计算,用 n 表示数据表的记录个数, m 表示数据表的数值型属性的个数,则 n 行 m 列的数据表 X 用矩阵表示为:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

2) 标准化预处理。由于质量数据通常由多个属性值组成,而各属性值之间往往由于采用的度量单位不同而有很大的差别,这不利于计算数据对象间的距离,因此,这里要对数据进行标准化处理。标准化的方法是将原来的度量值转换为无单位的值,给定一个变量 f 的度量值,可以进行如下变换:

计算平均绝对偏差 s_f :

$$s_f = \frac{1}{n} (|x_{1f} - m_f| + |x_{2f} - m_f| + \cdots + |x_{nf} - m_f|)$$

这里的 x_{1f}, \dots, x_{nf} 是 f 的 n 个度量值, m_f 是 f 的平均值,即:

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf})$$

计算标准化的度量值(Z-Score):

$$z_{if} = \frac{x_{if} - m_f}{s_f}; 1 \leq i \leq n$$

这个平均的绝对偏差 s_f 比标准差 δ_f 对于孤立点具有更好的鲁棒性。在计算平均绝对偏差时,度量值与平均值的偏差(即 $|x_{if} - m_f|$) 没有被平方,因此孤立点的影响在一定程度上被减小了。采用平均绝对偏差的优点在于孤立点的Z-Score值不会太小,因此孤立点仍可以被发现^[5]。

经标准化处理后的数据集用矩阵 S 表示为:

$$S = \begin{bmatrix} x_{11}' & x_{12}' & \cdots & x_{1m}' \\ x_{21}' & x_{22}' & \cdots & x_{2m}' \\ \vdots & \vdots & & \vdots \\ x_{n1}' & x_{n2}' & \cdots & x_{nm}' \end{bmatrix}$$

3) 输入孤立点集 O 。根据给定的数据集,使用基于距离和的孤立点挖掘算法得到的孤立点集如表 1 所示,其中数据对象列中的数字表示数据对象在数据集中的索引。

表 1 孤立点集 O

数据对象	产品硬度	材料硬度	加工温度	添加量
38	78.13	66.88	860.20	36.80
22	77.95	67.76	850.50	36.93
12	77.81	67.66	810.20	36.90
1	78.21	66.36	825.40	36.33
7	77.98	66.74	780.50	35.89

4) 应用 DSCM 算法。把经过标准化处理后的数据集 S 和孤立点集 O 作为算法的输入,执行 DSCM 算法,得到孤立点集 O 中所有孤立点的原因属性集 CP , 如表 2 所示。其中,DSCM 算法中调用的基于距离和的孤立点检测算法 FindOutliers(Q_i, M) 中的用户设置的期望孤立点个数 M 取值为 5。

表 2 孤立点集的原因属性集 CP

数据对象	原因属性集 CP	原因属性的孤立度
38	加工温度(过大), 产品硬度(过大), 添加量(过大)	209.03, 122.04, 120.14
22	加工温度(过大), 添加量(过大), 材料硬度(过大)	162.98, 142.97, 134.73
12	产品硬度(过小), 添加量(过大), 材料硬度(过大)	162.55, 137.53, 120.96
1	产品硬度(过大)	189.67
7	加工温度(过小)	185.13

3.2 实验结果分析

从算法的输出结果可以看出,针对数据集的孤立点集中的每一个孤立点,算法都给出了其原因属性集,该属性集由原始数据集的属性空间的若干个属性子空间组成。原则上各原因属性之间是并列的关系,即都是能够解释该数据对象成为孤立点的可能原因,但算法又进一步给出了各原因属性的孤立度,即各原因属性对于引起该数据对象成为孤立点的影响程度,这里给出的孤立度,单独看是没有什么意义的,只有放在一起才可以看出各个原因属性对数据对象的影响差异。因此,这里用户就可以根据各原因属性的孤立度去优先考虑那

些孤立度最大的原因属性。例如:第 38 个数据对象成为孤立点的最可能原因是由于“加工温度”异常引起的,其次是“产品硬度”,再次是“添加量”;而对于第 1 个数据对象,由于其原因属性只有一个“产品硬度”,因此,这里给出的孤立度则没有什么意义。

为了能够给用户提供更详细的对于孤立点意义的解释,也即给出一个孤立点原因属性集中每一原因属性的有意义的描述,结合质量数据一般都具有“中间多,两边少”分布规律的特点,可以采用把孤立点的原因属性值 x_{ik} 与原始数据集中该属性下所有数据点的均值 \bar{x}_k 进行比较,进一步给出如下判断:1) 若 $x_{ik} > \bar{x}_k$, 说明是由于该属性值过大引起的;2) 若 $x_{ik} < \bar{x}_k$, 说明是由于该属性值过小引起的。当然也可以采用更加精确的方法来判断,这里不再给出;3) 若 $x_{ik} = \bar{x}_k$, 该情况的出现是没有意义的,可进一步使用更详尽的方法继续判断。

根据以上方法给出的孤立点更加具体化的用户意义如表 2 所示。例如:第 22 个数据对象成为孤立点,最可能是由于“加工温度”“过大”引起的,其次可能是由于“添加量”“过大”引起的,这样用户就可以通过降低加工温度和减少添加量来消除异常的产生。

4 结语

孤立点用户意义分析是一个目前在孤立点分析领域中研究较少的方面,孤立点通常包含着重要的信息,因此给出孤立点的用户意义,以帮助用户更好地理解孤立点具有十分重要的意义。本文给出了一个孤立点用户意义的定义,提出了一个能够给出孤立点用户意义的 DSCM 算法,并将其应用于质量管理中,实验结果表明该算法是有效的,且给出的结果是合理的。该算法只需要用户给出在某一属性子空间下期望的孤立点个数,就能输出想要的结果,减少了用户的参与,易用性较强。如何减小算法的时间复杂度和空间复杂度,如何将算法应用到其他领域,提高算法的通用性,将是下一步研究的重点。

参考文献:

- [1] 李卫红, 杨练根. 质量统计技术[M]. 北京: 中国计量出版社, 2006: 31-33.
- [2] 黄洪宇, 林甲祥, 陈崇成, 等. 离群数据挖掘综述[J]. 计算机应用研究, 2006, 23(8): 8-13.
- [3] KNORR E M, NG R T. Algorithms for mining distance-based outliers in large datasets [C]// Proceedings of the 24rd International Conference on Very Large Data Bases. San Francisco, CA, USA: Morgan Kaufmann, 1998: 392-403.
- [4] JIANG SHENG-YI, LI QINGH-HA, LI KEN-LI, et al. GLOF: A new approach for mining local outlier [C]// 2003 International Conference on Machine Learning and Cybernetics. Washington, DC: IEEE Press, 2003, 1: 157-162.
- [5] KNORR E M, NG R T, TUCAKOV V. Distance-based outliers: Algorithms and applications [J]. The International Journal on Very Large Data Bases, 2000, 8(3/4): 237-253.
- [6] 陆声铨, 林士敏. 基于距离的孤立点检测及其应用[J]. 计算机与数字工程, 2004, 32(5): 94-97.
- [7] HAN J, KAMBER M. Data mining: Concepts and techniques [M]. New York: Academic Press, 2001.
- [8] KNORR E, NG R. Finding intensional knowledge of distance-based outliers [C]// Proceedings of the 25th International Conference on Very Large Data Bases. San Francisco, CA, USA: Morgan Kaufmann, 1999: 211-222.