

一种基于数量关联的孤立点检测算法

余宇峰¹, 万定生¹, 熊 强²

(1 河海大学 计算机与信息学院, 江苏 南京 210098; 2 江西省南昌市水文局, 江西 南昌 330018)

摘 要: 孤立点检测是数据挖掘的重要研究方向之一, 目标是发现数据集中不具备数据一般特性的数据对象. 通过挖掘数据属性间内在的数量关联规则, 标记产生的弱关联规则中置信度小于阈值的极小概率事件为孤立点, 提出了一种基于数量关联的离群点检测算法. 实例表明, 算法能够有效检测数据集中的孤立点, 具有应用价值.

关键词: 数据挖掘; 孤立点检测; 数量关联; 离散化

中图分类号: TP302.1

文献标识码: A

文章编号: 1000-7180(2012)09-0164-04

Outlier Detection Algorithm Based on Quantitative Association

YU Yu-feng¹, WAN Ding-sheng¹, XIONG Qiang²

(1 College of Computer & Information, HoHai University, Nanjing 210098, China;

2 Nanchang Hydrology Bureau of Jiangxi Province, Nanchang 330018, China)

Abstract: Outlier detection, aim of which is to discover the abnormal data objects in the data set, is an important research aspect in the data mining. This paper proposes a new method to detect outliers by discovering quantitative association rules between the data attributes. The outliers are defined as the small probability events which confidence less than the threshold in weak association rules. Experimental results show that the algorithm can effectively on outlier detection and has some application prospect.

Key words: data mining; outlier detection; quantitative association; discretization

1 引言

孤立点(或称离群点, 异常点)指数据集 D 中与其他数据点表现不一致或者大大地偏离其他数据点以至于怀疑它是由不同的机制生成的小部分对象^[1]. 虽然孤立点的存在不符合数据的一般规律, 而且通常是测量、记录错误等因素造成的, 但孤立点也可能潜藏中应用领域感兴趣的、有意义的信息. 孤立点产生的原因很多, 大致分为三类: 数据来源于异类, 如欺诈、不寻常的实验结果等; 数据变量固有变化引起, 是自然发生的, 反映了数据集的数据分布特征, 如气候变化、基因突变等; 数据采集误差, 主要是由于人为错误、设备故障等引起, 这类离群不能提供有用信息, 只会降低数据质量和分析的准确性, 因此很多孤立点检测的目标是要消除这类孤立点.

孤立点的存在, 会降低数据质量, 并对数据分析、管理和决策水平产生不良影响. 本文在关联挖掘算法和异常检测算法的基础上, 提出了一种基于数量关联的离群点检测算法. 算法通过挖掘数据属性间内在的数量关联规则, 标记产生的弱关联规则中置信度小于阈值的极小概率事件为孤立点. 算法除了产生强关联规则外, 还产生可能包含了孤立点的弱关联规则.

2 相关研究

2.1 孤立点检测

所谓孤立点检测, 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^[2]. 孤立点检测可以形式化地描述为: 给定一个包含 n 个数据点或对象的集合及预期的孤

收稿日期: 2012-02-23; 修回日期: 2012-03-20

基金项目: 国家自然科学基金项目(51079040); 水利部“九四八”项目(201016)

立点数目 k , 发现与剩余的数据相比是显著异常的、孤立的或不一致的前 k 个对象的过程。孤立点检测需要解决二个问题: 在给定的数据集中定义什么样的数据是不一致的; 找到一个有效的方法来检测这样的不一致数据。

孤立点检测方法大多建立在统计学的基础上, 可分为以下几类:

(1) 基于偏离的方法^[3]: 通过检查一组对象的主要特征来确定孤立点, 该方法中序列异常的概念没有得到普遍认可, 遗漏了较多异常数据。

(2) 基于分布的方法^[4]: 通过假设给定的数据集服从一个随机分布(如正态分布), 把偏离这些模型的对象当作离群点。但大多数分布模型只能直接应用于单变量的特征空间, 难以应用于多维空间。

(3) 基于距离的方法^[5-7]: 根据数据点或对象之间的距离来检测孤立点。文献[5]提出数据集中至少有 p 部分对象与对象 o 的距离超过阈值 d , 则 o 就是离群点。文献[6]根据 k -最近邻距离对离群点进行排序, 计算排在最前面的 n 个离群点。文献[7]针对不确定数据, 提出基于距离的不确定数据离群点检测方法。

(4) 基于密度的方法^[8-10]: 关注对象周围邻居的密度, 强调对象的局部性。文献[8]提出了局部离群因子 LOF, 通过数据空间的所有维度来计算对象的距离和对象的可达密度, 并通过局部的离群度来判断离群点。在此算法基础上产生了基于局部信息熵的加权子空间离群点检测算法^[9]、局部空间离群测度^[10]等方法。

(5) 基于关联的方法^[11-12]: 通过发现频繁项集来检测孤立点。He 提出频繁模式孤立点因子 FPOF, 认为关联规则算法发现的频繁模式反映了数据集中的普遍模式, 而不包含频繁模式或只包含极少频繁模式的数据其实就是孤立点。

2.2 关联规则挖掘

给定项集 $I = \{i_1, i_2, \dots, i_m\}$, I 的每个元素是数据库中的数据项, 关联规则^[12]是如下形式的蕴涵, $A \Rightarrow B$, 其中: $A \subseteq I, B \subseteq I$, 且 $A \cap B = \emptyset$ 。关联规则设定 2 项约束, 最小支持度(min_sup)和最小可信度(min_conf)。规则的支持度是指数据库中支持 $A \cup B$ 的元组数占所有元组的比例; 可信度是指支持 A 的元组中, 支持 B 的元组所占比例。最小支持度和最小可信度是用户说明的 2 个阈值, 关联规则的挖掘就是发现数据库中支持度和可信度大于最小支持度和最小可信度的全部规则。关联规则挖掘的关键在于

发现支持度大于最小支持度的全部项集(强项集)。

数量关联规则是指包含 Boolean 属性和连续属性的关联规则, 如 $(Balance \in [v_1, v_2]) \Rightarrow (CardLoan = yes)$ 。数量关联规则同样要满足支持度和可信度的约束条件。如果规则的可信度大于用户给定的最小可信度, 则称该规则是可信的。如果规则的支持度大于用户给定的最小支持度, 则称该规则是强规则; 否则称该规则为弱关联规则。当弱关联规则可信度低于某一阈值, 通常认为该规则对应的数据为异常点。

3 基于数量关联规则的孤立点检测算法

3.1 算法基本思想

关系数据对象各指标之间往往存在着较强的内在的相关性, 如股市行情和股票价格、水位与流量等。因此, 可利用数据挖掘技术对数据集进行关联规则分析, 挖掘数据序列中的高置信规则和孤立点, 并结合业务知识分析孤立点的成因, 发现数据序列中的异常模式和噪声数据。算法的基本思想如下:

(1) 建立数量关联规则挖掘参照表

对待检测数据集的连续属性 R, S 采用数量关联规则分析方法得到形如 $C_R \in [y_1, y_2] \Rightarrow C_S \in [x_1, x_2] (\text{conf} = a\%)$ 的关联规则参照表。

(2) 数据审核

将待检测数据与步骤(1)得到关联规则参照表进行比对, 标记其中出现的极小概率事件; 如果该事件出现的概率小于某一数值, 则认为该事件为孤立点。

(3) 数据分析

分析步骤(2)得到的孤立点, 确定孤立点的成因, 去除其中的噪声数据, 从而达到提高水质数据质量的目标。

3.2 算法的基本步骤

3.2.1 连续属性离散化

数值型属性离散化处理是将连续数据分割为多个有限的区间, 过程如下:

假设一个包含有 M 个样本的数据集 D , 其中每一个样本包含 K 个属性, F 描述为样本中的一个数值型属性。离散算法就是在 F 上找到一个最佳方案 L , 将 F 分割成包含有多个数据的离散区间:

$$L: \{(d_0, d_1], (d_1, d_2], \dots, (d_{n-1}, d_n]\} \quad (1)$$

其中 d_0 是 F 的最小值, d_n 是 F 的最大值。方案 L 就是决策树生成多分枝的依据, 如图 1 所示。

属性 F 的每一个值都可以划分到由公式(1)中

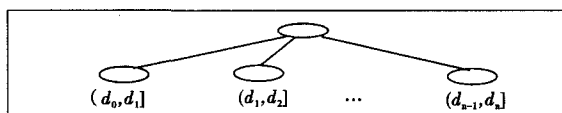


图1 连续数值型属性离散化方案

定义的 n 个区间中的一个。连续属性离散化将得到一个表示属性与划分方案间对应关系的一维数组，其形式如表 1 所示。其中 Q_i 表示类别 T_i ($F \in (d_{i-1}, d_i]$) 上的样本数。

表1 连续数值属性离散化结果

区间范围	$(d_0, d_1]$	$(d_1, d_2]$	...	$(d_{n-1}, d_n]$
类别	T_1	T_2	...	T_n
数量	Q_1	Q_2	...	Q_n

3.2.2 局部孤立点检测

合并样本数据集 D 中经过离散化后的 K 个连续数值型属性，得到完整的数据集。并根据每个属性离散化结果，生成 K 个属性区间联合分布密度的 K 维空间矩阵（文中主要讨论二维空间）。如表 2。

表2 属性关联二维空间矩阵

指标	R_1 $x \in (x_0, x_1]$	R_2 $x \in (x_1, x_2]$	...	R_n $x \in (x_{n-1}, x_n]$
$S_1 y \in (y_0, y_1]$	T_{11}	T_{12}	...	T_{1n}
$S_2 y \in (y_1, y_2]$	T_{21}	T_{22}	...	T_{2n}
...	...	...	...	...
$S_m y \in (y_{m-1}, y_m]$	T_{m1}	T_{m2}	...	T_{mn}

分别统计属性指标 R 和 S 在各离散区间上的样本记录值并计算 $R_i \in (x_{i-1}, x_i]$ 时 $S_j \in (y_{j-1}, y_j]$ 的分布概率 $P(S_j \wedge R_i | R_i)$ 及 $S_j \in (y_{j-1}, y_j]$ 时 $R_i \in (x_{i-1}, x_i]$ 的分布概率 $P(R_i \wedge S_j | S_j)$ 。其中 $P(S_j \wedge R_i | R_i)$ 和 $P(R_i \wedge S_j | S_j)$ 可以由下式计算得到：

$$P(S_j \wedge R_i | R_i) = T_{ji} / \sum_{i=1}^m T_{ji}$$

$$P(R_i \wedge S_j | S_j) = T_{ij} / \sum_{i=1}^n T_{ij} \quad (2)$$

若 $P(R_i \wedge S_j | S_j) < \min_conf$ ，则标记该区间内分布的数据为局部孤立点。

3.2.3 区间合并

对属性关联二维矩阵进行分析，若属性 R 相邻区间 R_i, R_{i+1} 的二维联合分布密度 $P(R_i \wedge S_j | S_j)$ 、 $P(R_{i+1} \wedge S_j | S_j)$ 均大于某一预设阈值 τ ，则合并 R_i, R_{i+1} ，调整值域为 $(x_i, x_{i+2}]$ 并重新生成属性关联二

维矩阵。同理可以按照上述规则对属性 S 进行相同操作直至所有区间都不在发生变化。若 $P(R_i \wedge S_j | S_j)$ 、 $P(R_{i+1} \wedge S_j | S_j)$ 均大于 $\min_conf$ ，则将该规则加入关联规则参照表；否则将标记该规则对应的数据区间为全局孤立点。

3.3 算法实现

输入：数据集 D ，最小置信度 $\min_conf$ 和阈值 τ 。

输出：规则参照表 Rt ，孤立点集 O 。

算法：

(1) 按照属性 R, S 将数据集 D 分解成两个数据集 DR 和 DS 。

(2) 采用 K-means 算法分别对数据集 DR 和 DS 进行聚类，将连续数值属性离散化成有限个区间，以类标号 R_i 和 S_j ($i, j = 1, 2, \dots, n$) 表示，并计算每个类标号 R_i 对应的值域区间 $x \in (x_i, x_{i+1}]$ 及 S_j 对应的值域区间 $y \in (y_j, y_{j+1}]$ 。

(3) 合并步骤(2)中经过聚类分析后的数据集 DR 和 DS ，得到新的数据集 D' 。计算 D' 中每个类标号 R_i 和类标号 S_j 间的数量对应关系。

(4) 将步骤(3)得到的数量对应关系代入式(2)，计算 R_i 与 S_j 的联合分布概率 $P(S_j \wedge R_i | R_i)$ 及 $P(R_i \wedge S_j | S_j)$ 。若 $P(S_j \wedge R_i | R_i)$ 或 $P(R_i \wedge S_j | S_j) < \tau$ ，则标记该区间内分布的数据为局部孤立点 LO_i 。

(5) Repeat:

合并 $P(S_j \wedge R_i | R_i)$ 、 $P(R_i \wedge S_j | S_j)$ 均大于 $\min_conf$ 的相邻聚类区间 R_i, R_{i+1} ，得到属性 R 新的聚类区间 R'_i ，调整 R'_i 的值域为 $(x_i, x_{i+2}]$ ，重新计算 $P(R'_i \wedge S_j | S_j)$ 并根据结果重新调整 LO_i 。

Until 所有区间均不发生变化。

(6) 若 $P(R_i \wedge S_j | S_j) \geq \min_conf$ ，则将规则 $x \in (x_i, x_{i+1}] \Rightarrow y \in (y_j, y_{j+1}]$ 加入关联规则参照表；将 LO_i 对应区间内 $P(S_j \wedge R_i | R_i)$ 或 $P(R_i \wedge S_j | S_j) < \tau$ 的所有数据加入全局孤立点 O 。

4 基于数量关联规则孤立点检测应用

由于水质数据各指标之间存在较强的数量关联关系，如水体中的氨氮(NH_3N)是总氮(TN)的组成部分，则 $C_{\text{NH}_3\text{N}} < C_{\text{TN}}$ (C 代表浓度指标)。因此，可以挖掘水质数据序列各指标间的数量关联规则。标记可信度低于阈值的弱关联规则对应的数据为孤立点，结合业务知识分析孤立点的成因，发现水质序列中的异常模式和噪声数据。

选择太湖流域某水质测站 2001~2010 年共

2568 条水质监测数据,利用本文数量关联规则孤立点检测算法进行了多次实验,标记数量关联规则中对应的孤立点。

采用 K-means 分别对数据集内的指标 TN 和 NH₃N 进行连续属性离散化,得到的离散区间如表 3 所示。

表 3 连续属性离散化结果

TN 离散结果		NH ₃ N 离散结果	
TN 类号	TN 值域	NH ₃ N 类号	NH ₃ N 值域
TC1	[0.24,1.63]	NC1	[0.05,0.46]
TC2	[1.64,2.68]	NC2	[0.47,1.19]
TC3	[2.69,3.8]	NC3	[1.2,2.45]
TC4	[3.81,5.24]	NC4	[2.46,4.33]
TC5	[5.27,7.04]	NC5	[4.41,9.11]
TC6	[7.06,11.5]		

合并 TN 和 NH₃N 离散化后的数据集,根据 TN 和 NH₃N 属性离散化结果,得到数量关联规则如表 4 所示。

表 4 数量关联关系分析结果

TN	总数量	NH ₃ N	总数量	可信度
TC1	228	NC1	200	87.72%
		NC2	28	12.28%
TC2	903	NC1	516	57.14%
		NC2	373	41.31%
		NC3	14	1.55%
TC3	896	NC1	92	10.27%
		NC2	581	64.84%
		NC3	221	24.67%
		NC4	2	0.22%
TC4	382	NC2	46	12.04%
		NC3	263	68.85%
		NC4	72	18.85%
		NC1	1	0.26%
TC5	122	NC3	21	17.21%
		NC4	89	72.95%
		NC5	11	9.02%
		NC2	1	0.82%
TC6	37	NC5	29	78.38%
		NC4	8	21.62%

NH₃N∈[1.2,4.33]的置信度为 87.7%,是可信度很高的数量关联规则,可用于检测 TN∈[2.69,3.8]时 NH₃N 取值是否合理;而 TN∈[2.69,3.8]时 NH₃N∈[0.47,1.19]的置信度 0.26%,显然属于异常数据点,应对原始记录中数据进行分析以确定异常数据产生的原因。

5 结束语

本文在关联挖掘算法和异常检测算法的基础上,提出了基于数量关联挖掘的异常检测算法,并将算法应用于太湖流域某水质监测站 10 年水质数据的异常检测。实验结果表明:该方法能有效的挖掘水质数据中高可信度的数量关联区间及的异常点,从而达到提高水质数据质量的目标,具有广泛的应用前景。

参考文献:

[1] Hawkin D. Identification of Outliers[M]. London: Chapman and Hall, 1980.

[2] Han JW, Kamber M. Data mining: concepts and techniques[M]. New York: Morgan Kaufmann Publishers, 2001.

[3] Arning A, Agrawal R, Raghavan P. A linear method for deviation detection in large database[C]//Proc of the Int’l Conf on Knowledge Discovery in Databases and Data Mining. Portland, Oregon, USA: AAAI Press, 1996.

[4] Barnett V, Lewis T. Outliers in statistical data [M]. New York: John Wiley and Sons, 1994.

[5] Knorr E M, Ng R T. Algorithms for mining distance-based outliers in large datasets [C]//Proc of the 24th Int Conf on Very Large Data Bases. New York: ACM, 1998: 392-403.

[6] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets [C]//Proc of the 2000 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2000: 93-104.

[7] 于浩,王斌,肖刚,等. 基于距离的不确定离群点检测[J]. 计算机研究与发展, 2010, 47(3): 474-484.

[8] Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying density-based local outliers [C] // Proc of ACM SIGMOD Conf. New York: ACM, 2000: 427-438.

[9] 倪巍伟,陈耿,陆介平,等. 基于局部信息熵的加权子空间离群点检测算法[J]. 计算机研究与发展, 2008, 45(7):1189-1194.

就不仅不能保证这段量程的精确度,且也不利于误差的修正;其次,USF 的原理决定了其在小流量上的准确性没大流量的准确.为了确保流量计的精度,增加检定点 $(Q_{\min}+Q_t)/2$.另外由图 1 可知流体在管道中随着流速的变化有三个状态,因此,在检定的时候必须检定到每个状态.因此,针对 DN100,实际检定的流量点是:0.40 m³/s、0.52 m³/s、0.64 m³/s、30 m³/s、70 m³/s、100 m³/s.依据表 2,层流中的检定点:0.4 m³/h,0.52 m³/h;过渡流中的检定点:0.52 m³/h;湍流中检定点:30 m³/h,70 m³/h,100 m³/h,125 m³/h.流体的各个状态都检定到,且检定的流量点基本上均匀分布.实验结果如表 3 所示.每个流量点检定了三次,给出了每次的误差和三次实验的重复性.从表 3 可知,所有七个检定点的误差和重复性都在规定的范围内.因此,实验结果说明本文的检定方法是可行的.

4 结束语

USF 的检定是 USF 投入市场前的一个必须环节,它在 USF 的研发和生产过程中十分的重要.本文针对国标的检定方法进行了研究,并根据实际的情况提出了自己的检定方法,经实验证明方法正确可靠,可以推广使用.

参考文献:

[1] 何群.超声波流量计测量系统的关键技术研究[D]. 哈尔滨哈尔滨工程大学,2004.

[2] 霍殿中,翟秀贞.大口径多声道超声流量计的干校验法检定[J].计量学报,1994,15(4):269-275.

[3] JJG 103-2007.《超声流量计》[S].中国计量出版社,2007.

表 3 测量数据的误差和重复性

流量点	标准值	测量值	误差	重复性
0.4 m ³ /s	202.4	202.05	-0.17%	0.22%
	201.2	200.05	-0.57%	
	203.1	203.06	-0.02%	
0.52 m ³ /s	202.1	201.06	-0.51%	0.17%
	203.6	203.07	-0.26%	
	200.4	200.07	-0.16%	
0.64 m ³ /s	203.8.	203.07	-0.36%	0.14%
	206.2	206.07	-0.06%	
	201.2	201.09	-0.05%	
30 m ³ /s	2010.2	2005.81	-0.22%	0.10%
	2014.5	2006.24	-0.41%	
	2020.5	2012.01	-0.42%	
70 m ³ /s	2008.6	2012.64	0.20%	0.12%
	2010	2013.22	0.16%	
	2005.4	2014.33	0.44%	
100 m ³ /s	2020.5	2025.23	0.23%	0.13%
	2010.3	2019.56	0.46%	
	2009	2018.95	0.50%	
125 m ³ /s	2020	2024.03	0.20%	0.07%
	2018.8	2024.67	0.29%	
	2020.3	2023.22	0.14%	

作者简介:

李 浩 男,(1987—),硕士研究生.研究方向为智能仪表与检测.嵌入式系统.

王让定 男,(1962—),教授,博士生导师.研究方向为智能仪表、嵌入式系统的开发与应用.

(上接第 167 页)

[10] 薛安荣,鞠时光,何伟华,等.局部离群点挖掘算法研究[J].计算机学报,2007,30(8):1455-1463.

[11] He Z Y, Xu X F, Huang J Z, et al. A frequent pattern discovery method for outlier detection [J]. WAIM 2004. 2004,132(19):726-732.

[12] Robert Groth. Data mining: building competitive ad-

vantage[M]. New Jersey: Prentice Hall PTR, 2000. 28-30.

作者简介:

余宇峰 男,(1979—),讲师,硕士.研究方向为数据管理与数据挖掘.

万定生 男,(1963—),教授.研究方向为数据管理与数据挖掘.