计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(7):200−203 [doi: 10.15888/j.cnki.csa.005936]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

数据整合中异常检测算法研究①

方　正[1,2], 高　岑[2], 田　月[2], 王　嵩[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)

摘　要: 传统的数据整合方案[1]中存在结构上的不严谨性, 在整合期间由于各种原因导致整合后的结果存在很多异常离群点, 而且并没有有效的措施进行检测和避免. 本文提出了基于角度的改进后的三阶段离群点检测算法, 通过对数据整合后的结果进行检测, 有效地消除了存在的大量疑似离群点. 这种改进算法减小了传统算法中对离群点误判的可能性, 考虑到数据动态变化的因素, 二次验证疑似离群点的异常情况的真实性. 本文以生产事故应急救援平台系统项目为背景.

关键词: 数据整合; 异常检测; 离群点检测; 基于角度; 生产事故

引用格式: 方正,高岑,田月,王嵩.数据整合中异常检测算法研究.计算机系统应用,2017,26(7):200–203. http://www.c-s-a.org.cn/1003-3254/5936.html

Research on Anomaly Detection Algorithm in Data Integration

FANG Zheng[1,2], GAO Cen[2], TIAN Yue[2], WANG Song[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: Traditional data integration solutions in the presence of the structure are not precise. During the integration period, the integrated result due to various reasons has many abnormal outliers, which cannot be detected and avoided with effective measures. This paper proposes an improved three stage outlier detection algorithm based on angle, which is mainly to detect the results after data integration, and effectively solve the problem of a large number of suspected outliers. This improved algorithm reduces the possibility of outliers in the traditional algorithm, taking into account the factors of dynamic changes in the data, verifying the abnormal real situation of suspected outliers for two times. This paper is backgrounded on the project of production accident emergency rescue system.

Key words: data integration; anomaly detection; outlier detection; based angle; production accident

引言

作为数据挖掘的一项技术, 离群点检测已经在众多领域得到了广泛的应用, 如计算机入侵检测、灾难异常检测等, 其中, 常用的检测方法主要有: 基于密度的离群点检测[2,4], 基于聚类的离群点检测以及基于邻近性方法的检测[5,6]. 本文研究案例背景是生产事故应急救援平台系统[3], 在平台构建时需要融合来自不同信息系统的数据, 其中遇到的核心问题就是融合后数据出现异常状况. 为了降低数据融合后的异常状况, 基于数据特征众多且复杂, 引入了高维数据下的离群点检测相关算法.

文献[7]中提到了利用基于角度的方法进行高维数据集的离群点检测, 这种方法的优点是避免了对高维数据的降维, 不会使高维数据退化, 适用于各种复杂的高维数据. 但其在计算离群点因子时并没有考虑到数据动态变化的因素. 本文对传统的基于角度的离群点检测算法进行改进, 并引进时间序列推进数据演变的概念, 在已有数据基础上增加数据维度, 提出一种基于

① 收稿时间: 2016-08-19; 收到修改稿时间: 2017-01-23

角度的三阶段离群点检测算法(Angle-Based Three Stage Outliter Detection), 相比较于传统的离群点算法, 此时的离群点总数量下降了9%左右.

1 异常检测算法提出

异常对象一般被称为离群点, 异常检测也称偏差检测和例外挖掘[8]. 常见的异常成因: 数据来源多样、自然变异以及数据测量或收集误差. 异常数据挖掘是一个非常有趣的研究课题, 在国内外受到了广泛的研究, 并且研究成果丰富, 大致可分为三类[7-9]: (1)基于模型的技术, 首先建立一个数据模型, 异常是那些通用模型不能完美拟合的对象, 如果模型是簇的集合, 则异常就是不显著属于任何簇的对象, 如在使用回归模型时, 异常是相对于远离预测值的对象; (2)基于邻近度的技术, 通常可以在对象之间定义邻近性度量, 异常对象是那些远离其他对象的对象; (3)基于密度的技术, 仅当一个点的局部密度显著低于它的大部分近邻时才将其分类为离群点.

基于角度的离群点检测[10,11], 适用于高维度数据模型, 通常避免邻近性度量, 并且采用新的启发式方法来检测离群点, 避免了在多维数据采用降维手段所造成的数据退化影响. 由于影响生产事故因素复杂, 其表现出来的数据特征差异性较大, 所以不能盲目认为某个点就是离群点, 原有的基于角度的离群点检测算法就显得不太适用, 故本文采用分段模式对原算法进行改进, 第一阶段是传统的高维离群点检测, 第二阶段是推进第一阶段的数据进行必要的演练, 得到在时间序列上相对于第一阶段的未来趋势. 第三阶段主要是分析第二阶段末的时候第一阶段中的疑似离群点是否仍然在检测中表现异常, 如果是则将其判定为离群点, 反之则判定为正常数据.

2 基于角度的三阶段离群点检测算法

2.1 传统的基于角度的离群点检测算法

算法描述[7]:

(1) 图1中的点除了形成了两个小型的簇, 还有疑似离群点O. 在簇1中选取两个数据点, 分别命名为对象X、对象Y, 此处的对象代表具有多属性的事物, 例如煤矿生产事故监控模块中的瓦斯传感器或负压传感器, 对象X与对象Y的区别是对象所处时间段不同, 对应的属性数据也表现不同. 连线OX、OY形成一个角度$\angle XOY$.

(2) 对于点簇中心的点, 构成的角度差别很大, 对于远离簇中心的点, 角度的变化较小. 对于离群点O, 角度变化显著地小. 这样就可以使用点的角度来确定一个点是否是离群点.

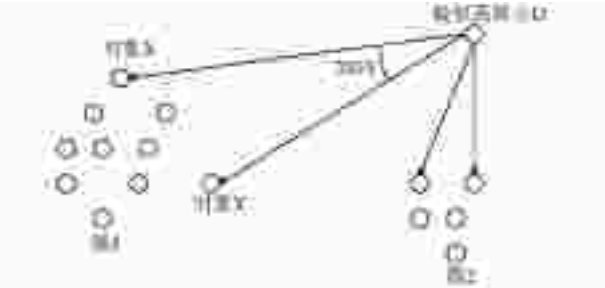

图1 基于角度的离群点

(3) 结合角度和距离来对离群点建模. 对于每个点对象, 使用距离加权的角度方差(distance-weighted angle variance)作为离群点得分. 即给定一个点集D, 对于每个点(属于D)定义基于角度的离群点因子(Angle-Based Outlier Factor, ABOF)为:

$$ABOF(o)=\arccos\left(\frac{<\overrightarrow{ox},\overrightarrow{oy}>}{\left\|\overrightarrow{ox}\right\|\cdot\left\|\overrightarrow{oy}\right\|}\right)*\frac{1}{dist(\overrightarrow{ox})*dist(\overrightarrow{oy})} \quad (1)$$

注: $<\overrightarrow{ox},\overrightarrow{oy}>$是点内积操作, dist(,)是标准距离. 显然, 点离簇越远, 点的角度的方差越小, ABOF越小. 基于角度的离群点检查方法对每个点计算ABOF, 并且按ABOF递增序输出数据集中点的列表.

2.2 基于角度的三阶段离群点检测算法

2.2.1 算法及算法说明

算法: 基于角度的三阶段离群点检测算法

说明: x_i, y_i为点簇中的任意点, O代表疑似离群点, n为点簇数目, m为疑似离群点数目

```
1  WHILE i < m DO
2    WHILE j<n DO
3      计算疑似离群点O与点簇中的点(x_i, y_i)的ABOF
4      按ABOF递增序列记录存储到L
5    END
6  END
7  IF SIZE(L)>n*10%
8    提高维度, 由三维推向多个维度
9    将L中疑似离群点映射到多维空间
10   WHILE i < SIZE(L) DO
11     WHILE j < n DO
12       重新计算ABOF, 并存储到L'
13     END
```

14 END

15 比较L’与L中的点记录, 取出共同存在的点, 即为最终离群点

16 END IF

算法说明: 低维问题向高维问题展开是一个数据推演的过程, 当然随之计算ABOF也有所变化, 如公式(2), 参数α代表升维后其他数据特征的权重.

$$ABOF'(o)=\mathrm{A}BOF(o)+\alpha\left(\frac{<\overrightarrow{ox_1},\overrightarrow{oy_1}>}{dist(o,x_1)^2 dist(o,\mathrm{y}_1)^2}\right) \quad (2)$$

本文案例数据来源于生产事故应急救援平台, 考虑到影响事故破坏程度的因素众多, 且随时间推移, 先前相关性较小的特征数据的影响力可能也越来越大. 比如, 生产车间有毒化学气体泄漏, 后期伴随着天气变化, 在有风与无风条件下其扩散范围与速度是有很大差别的. 所以, 在异常检测过程中, 本文改进后的算法采取分段处理, 共分为三个阶段:

(1) 第一阶段: 即算法描述中第1到第6步, 根据传统的基于角度的离群点检测算法, 计算出点簇中每个点的离群点因子(ABOF), 并按递增序排列输出, 划分出疑似离群点;

(2) 第二阶段: 即算法描述中第8到14第步, 若离群点数量超过总数一定的比例, 则对原数据集D中的点簇进行升维, 并映射到高维空间, 也就是说扩大数据特征, 得到新的点集D’;

(3) 第三阶段, 即算法描述中第15步, 根据得到的新的点集D’计算每个点的离群点因子, 按递增序输出, 并与从(2)中得到的离群点进行比较, 得到新的离群点;

(4) 算法步骤7给出了离群点总数量的阈值, 所以循环迭代执行算法步骤7到步骤16, 达到阈值后算法退出.

算法图示: 在离群点异常检测过程中, 主要有两种可能性, 一是有很大不同, 特别是离群因子前后变化较大, 则说明疑似离群点有被误认的可能性, 如图2所示; 二是前后序列变化并不大, 则说明疑似离群点异常的可能性很大, 就将其划分为离群点, 如图3所示.

3 实验与验证

为验证本文提出的改进算法的有效性, 基于沈阳市应急救援管理平台系统[3]所提供的数据, 设计实验与常见异常检测算法进行比较, 并将离群点占比作为衡量标准. 离群点占比等于离群点与总数据量的比. 离群点比例随着实验次数的增加, 逐渐处于平稳的趋势, 此时离群点比例较低的则说明对应的异常检测算法较好.

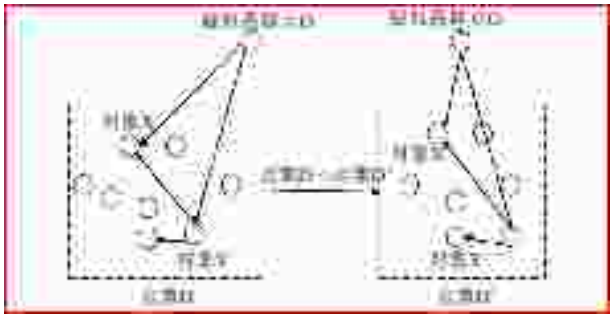

图2 离群点融合到新的点集

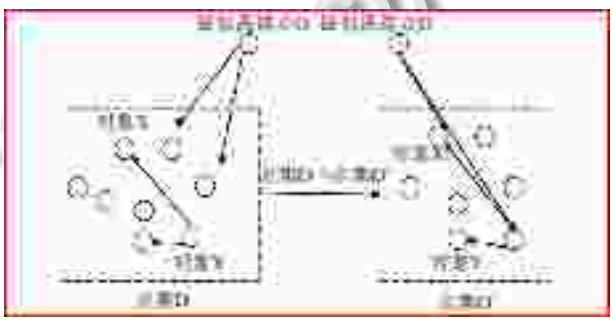

图3 离群点没有融合到新的点集

3.1 实验及数据对象

实验数据来源为实验室在做项目沈阳市应急救援管理平台系统数据, 项目背景之一是在生产事故发生后进行实时评估事故破坏程度, 由此给上层决策系统带来数据支撑. 在数据整合[1]过程中, 伤亡率以及事件严重程度数据整合出错率较高, 故选择数据系统有: 企业内部资源管理系统, 员工排班系统, 地域分布系统, 如图4. 为了使实验结果达到平稳状态, 选择数据量达到10万条以上. 数据点簇的构造中考虑到的数据特征有: 周围地域, 引起爆炸物种类, 事件发生时间等.

图4 采取的相关数据系统

3.2 不同异常检测算法比较

由于异常检测算法众多, 为验证不同算法对于异常

数据检测的准确率, 本文设计如下对比试验. 分别选取基于距离的离群点检测算法、基于角度的离群点检测算法以及本文2.2节提出的改进后的三阶段检测算法.

由表1可知, T1算法下检测出的疑似离群点较多, 说明数据整合期间异常情况可能较多, 但经过T2算法下的数据推演, 二次进行异常检测, 发现近似有30%到40%疑似离群点被判为无异常情况, 归属于正常范围.

表1 数据整合中的疑似离群点被二次验证后的情况

簇总点数	95	112	115	230	310	…
T1	7	23	25	34	41	…
T2	/	10	8	15	26	…

注: T1, 代表在基于角度的离群点检测算法下的疑似离群点数目; T2, 代表在基于角度的三阶段离群点检测算法下的疑似离群点数目.

由图5可知, 随着数据量的不断增大, 离群数量处于一个平稳增长趋势, 改进后的基于角度的三阶段离群点检测算法相对于其他两种算法来说, 上下波动较小, 误判率降低, 说明效果较好. 由此得出一般性规律:

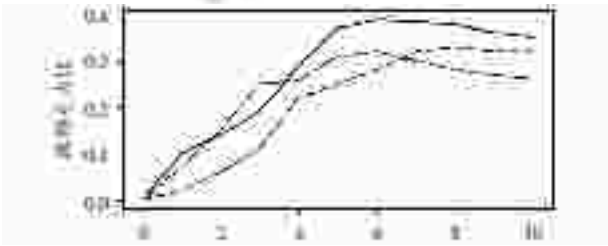

图5 不同异常检测算法下疑似离群点实验记录

注: 图5中实线代表基于距离的离群点检测算法; 短线代表基于角度的离群点检测算法; 短点相间线代表本文2.2节提出的改进后的三阶段检测算法.

(1) 数据动态性变化对于异常情况的出现影响较大, 也就会导致某个对象偏离了正常的点集D, 这就会影响了ABOF计算的准确性;

(2) 在随着时间推进去适当扩大数据后, 再次计算ABOF时发现, 原离群点的离群因子并不显著变小, 那我们就认为发生融合了, 不再假设此离群点异常;

(3) 另外一种情况是在数据推进后, 计算出疑似离群点的离群因子还是显著的小, 那就必然断定疑似离群点异常.

3.3 应用意义

本文是基于沈阳市生产事故应急救援平台系统实例. 在平台系统架构上, 因为数据源的多样性和复杂性特征, 导致数据源在平台融合阶段出现大量的不一致数据, 很多有效数据在融合后出现异常情况, 结合传统的数据异常算法检测后效果甚微, 离群点异常保持在30%左右. 在利用本文提出的基于角度的三阶段异常检测算法后, 数据异常离群点下降了9%左右, 实验数据如图5所示.

4 总结

通过运用基于角度的三阶段离群点检测算法, 对大量的整合后的数据进行离群点检测, 然后对实验中出现的很多疑似离群点进行避免并纠正, 经过二次计算检测部分疑似离群点融合到原点集中, 符合预期的实验假设与算法理论推测, 同时, 该算法也提高了离群点检测准确性, 缺点是增加了异常检测算法的时间复杂度, 需要进一步改进.

参考文献

1 李立博. 面向服务的多源异构数据整合平台的设计. 计算机工程与设计, 2011, 32(1): 141–144, 308.

2 韩超. 场景分类与道路场景异常识别算法研究[硕士学位论文]. 北京: 北京交通大学, 2016.

3 董钊. 智慧工厂构建过程中的有效数据整合问题研究[硕士学位论文]. 天津: 天津财经大学, 2015.

4 胡彩平, 秦小麟. 一种基于密度的局部离群点检测算法DLOF. 计算机研究与发展, 2010, 47(12): 2110–2116.

5 江峰, 杜军威, 眭跃飞, 等. 基于边界和距离的离群点检测. 电子学报, 2010, 38(3): 700–705.

6 王敬华, 赵新想, 张国燕, 等. NLOF: 一种新的基于密度的局部离群点检测算法. 计算机科学, 2013, 40(8): 181–185.

7 胡婷婷. 数据挖掘中的离群点检测算法研究[硕士学位论文]. 厦门: 厦门大学, 2014.

8 吴骞, 吴绍春. 基于离群分析的水位异常识别研究. 硅谷, 2010, (24): 45. [doi: 10.3969/j.issn.1671-7597.2010.24.037]

9 Su XG, Tsai CL. Outlier detection. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011, 1(3): 261–268. [doi: 10.1002/widm.v1.3]

10 Wang Y, Wang XC, Wang XL. A spectral clustering based outlier detection technique. Perner P. Machine Learning and Data Mining in Pattern Recognition. New York, NY, USA. 2016. 15–27.

11 Zheng ZG, Jeong HY, Huang T, *et al*. KDE based outlier detection on distributed data streams in multimedia network. Multimedia Tools and Applications, doi: 10.1007/s11042-016-3681-y.