文章编号: 1673- 9469(2011) 02- 0079- 03

孤立点检测及在煤矿安全预警系统中的应用

朱秀莉, 顾洪博, 杨冬黎
(东北石油大学 计算机与信息技术学院, 黑龙江 大庆 163318)

摘要: 针对孤立点检测算法的一些弊端和煤矿安全预警系统的实际情况, 选择一个基于距离和孤立点检测算法对瓦斯浓度的异常数据进行处理, 分析出设备异常数据、噪声数据和瓦斯突出孤立点数据, 通过挖掘孤立点来发现真实的潜在的信息, 保证安全预警的准确性。

关键词: 孤立点检测; 距离和; 煤矿安全预警; 瓦斯浓度
中图分类号: TP301. 6 文献标识码: A

Outliers detection algorithm in the early warning of mine safety system and it's application

ZHU Xie-li, GU Hong-bo, YANG Dong-li
(School of Computer& Information Technology, Northeast Petroleum University, Heilongjiang Daqing 163318, China)

Abstract: Aimed to the shortages of algorithm and the complicated actuality of the warning system, the outliers detection of distance sum- based algorithms were used in the system of early warning of mine safety. The research analyzed the noisy data, the outliers based in the abnormal equipment and the abnormity data of the concentration of gas . It is showed that the model can find out the actual latent information by excavating outliers and the model could make sure the veracity of safety warning.

Key words: outliers detection; distance sum- based; early warning of mine safety; concentration of gas

孤立点检测是数据挖掘领域中一个重要的研究方向。其任务是发现数据集中明显不同于其他数据的对象。孤立点的应用主要有信用卡与保险欺诈、入侵检测、气象预报、病例分析与诊断等[1]。孤立点检测一般包括: 孤立点的定义、寻找孤立点。现有的孤立点算法根据孤立点定义的角度不同, 分为基于统计的、基于距离的、基于密度的、基于深度的和基于偏离的算法[2]。煤矿安全预警系统需要从不完整的、模糊的、随机的大量数据中, 挖掘有用的信息和知识, 直观表征采煤区域的总体状况, 为煤矿安全生产提供可靠依据[3]。本文采用孤立点检测的方法对瓦斯浓度的异常数据进行处理, 保证预警的准确性。

1 孤立点检测算法

1.1 基于统计

基于统计的孤立点检测算法主要思想是假定数据集服从某种分布或概率模型, 通过不一致检验把那些严重偏离分布曲线的数据视为孤立点[4]。从已知值的分布找出统计参数, 如均值和方差, 再根据统计参数和孤立点期望数目建立阈值。阈值之外的可能是孤立点, 如: 阈值= 均值 ± 2× 标准差。

若已知数据集的概率分布 (如正态分布, 均值) 时, 用基于统计的方法。此方法主要局限在科研计算领域, 它适用于数值型数据, 而不适用于高维、分类数据的挖掘。

1.2 基于距离

1998 年, Konr 提出基于距离的孤立点检测算法[5], 2000 年 Ng 进行了更新[6], 因此孤立点的定义是: 数据集 D 中, 至少有 p 部分对象与对象 O 的距离大于 d, 那么对象 O 就是一个带参数 p 和 d 的基于距离的孤立点, 记为 $DB(p, d)$。

此方法使用全局阈值故不能处理具有不同密

收稿日期: 2011- 03- 06
基金项目: 黑龙江省自然科学基金(F200603)
作者简介: 朱秀莉(1962-), 女, 福建省厦门市人, 工程师, 从事人工智能及数据挖掘的教学和科研工作。

度的数据集。此外, 算法需要事先确定参数 pct 和 $d_{\min}$, 这是比较困难的。对于给定的不同参数 $d_{\min}$, 检测结果通常具有很大的波动性。后来又提出基于距离和[7]、基于属性距离和[8] 的孤立点检测算法。

1.3 基于密度

密度常用邻近度是指任意一点和 p 点距离小于给定半径 r 的邻域空间内的数据点的个数。一般定义是点到其 k 近邻的平均距离, 平均距离小则密度小。基于密度的孤立点检测, 就是探测局部密度, 认为孤立点是在低密度区域中的对象。经典的基于密度的孤立点检测算法是LOF 算法[9]

基于密度的方法, 给出了对象是孤立点程度的定量度量, 并且即使数据具有不同密度的区域也能很好地处理。但这些方法必然具有 $O(n^2)$ 的时间复杂度, 其参数选择也是困难的。最近又提出基于平均密度[10] 的和基于 $K-$ 距离[11] 的孤立点检测算法。

总之, 现有的孤立点检测算法能够检测出孤立点, 但也存在局限性。主要是: 高维数据孤立点检测的算法效率的下降、最近邻概念的失效、对背景知识的依赖较多。

2 基于DS 的孤立点检测算法

针对孤立点检测算法的一些弊端, 采用一个基于距离和(Distance Sum, DS) 的孤立点检测算法。首先计算数据集中对象两两间的距离, 然后计算每个对象与其它对象的距离之和, 计算所有对象的距离和均值 H, 则距离之和大于 H 的对象即为部分孤立点。

2.1 基本概念

定义 1 对象 x_i 的距离和 对象 x_i 到数据集中各个对象间的距离和为 $s_i=\sum_{j=1}^{n}\sqrt{\sum_{h=1}^{d}(x_{ih}-x_{jh})^2}$, 其中, 维数为 d。

从定义 1 可以看出, 所有对象的距离和组成一个主对角线为 0 的对称矩阵。

定义 2 对象 x_i 距其它对象的偏离度 $D_i=\left|s_i-\sum_{i=1}^{n}\sqrt{\frac{s_i}{n}}\right|$。

定义 3 孤立点数据集 D 中, 若对象 x_i(维数为 d) 的偏离度 $D_{(i)}>0$, 则称对象 x_i 就是一个带参数 s_i 和 d 的基于距离的孤立点, 记为 $DB(s_i, d)$。

2.2 基于DS 的孤立点检测算法

1) 对原始数据集进行标准化后, 计算 n 个对象两两之间的距离, 并计算出每个对象的偏离度。

2) 若对象 x_i 满足定义 3, 则称对象 x_i 是孤立点。

3) 对检测出的孤立点进行分析。

3 瓦斯浓度异常检测

瓦斯检测对精度要求不高, 但对其可靠性要求非常高。能够连续检测, 并在设定的百分比浓度点准确报警。由于井下环境复杂, 存在着各种干扰源, 传感器输出的信号极易受到污染。因此, 经常会出现瓦斯异常数, 但实际值并不高。系统计算机根本无办法识别这些干扰信号, 系统频频发生误报警。一旦报警, 井下电源就自动切断, 生产停止。据统计, 误报警高达 70%, 正常生产因为频繁的误报警而受到很大影响, 给企业造成很大的损失[12]。

难以克服的“大数污染”问题普遍存在于目前使用的各种系统中。采用孤立点检测的方法对瓦斯浓度的数据进行处理, 防止误报警, 保证报警的准确性。

3.1 数据选取及标准化

本实验所用数据为 8 个矿井某天瓦斯浓度数据, 其中包含瓦斯涌出以及异常数据。每个传感器采集到 400 多个数据, 共搜集了 3 000 多个数据。瓦斯正常浓度在 0~ 1% 之间。为了便于实验, 将获得的数据进行了标准化处理(各乘 100) 得到一个样本集。其中部分数据如表 1。

依照定义 1, 可以计算出每个矿井的各个距离和数据, 如表 2。

表 1 标准化后的部分数据

Tab. 1 Partial database after standardization

井号	1 时	2 时	3 时	4 时	5 时
A	1.36	0.37	1.19	0.35	1.21
B	0.37	0.37	0.34	0.34	1.01
C	0.37	0.34	0.35	0.37	0.5
D	2.37	0.34	0.33	0.33	0.34
E	0.36	0.37	0.34	0.35	0.35
F	0.27	0.32	0.36	0.33	0.35
G	0.34	0.37	0.35	0.35	0.38
H	0.35	0.59	1.69	0.58	1.07

表 2 各个矿井的距离和

Tab. 2 The distance sum of each mines

井号	A	B	C	D	E	F	G	H
A	0	2.05	2.59	2.79	2.71	2.85	2.69	2.1
B	2.05	0	0.58	2.72	0.68	0.84	0.68	1.89
C	2.59	0.58	0	2.22	0.22	0.32	0.2	2.39
D	2.79	2.72	2.22	0	2.08	2.16	2.14	4.61
E	2.71	0.68	0.22	2.08	0	0.18	0.06	2.53
F	2.85	0.84	0.32	2.16	0.18	0	0.18	2.65
G	2.69	0.68	0.2	2.14	0.06	0.18	0	2.49
H	2.1	1.89	2.39	4.61	2.53	2.65	2.49	0

依据定义 2, 得到每个对象的偏离度, 如表 3。

表 3 各个矿井的偏离度

Tab. 3 The distance deviation of each mines

井号	A	B	C	D	E	F	G	H	均值
偏离度	17.78	9.44	8.52	18.72	8.46	9.18	8.44	18.66	10.18

依据定义 3, 从表 3 中可以看出, 井号 A、D、H 为孤立点。

3.2 实验分析

根据瓦斯检测数据, 采用基于聚类的孤立点分析, 大致得到 3 类孤立点。

瓦斯浓度远远大于其它时段的浓度。这其实是一种噪声数据。一般由于井下机电设备启停时发出的电磁干扰造成的, 或者井下监控设备所接的电网的强烈电磁干扰。强干扰脉冲能在瞬间完全淹没传感器信号, 结果就造成了"大数"异常现象。频繁的电磁启动脉冲与信号叠加后更会造成严重的"大数"干扰。对这类数据可以不予考虑。

表 4 噪声数据

Tab. 4 The data of noisy

D	2.37	0.34	0.33	0.33	0.34

瓦斯浓度有较少次高于其它时段的浓度。且这类数据变化幅度不大。这种数据的来源是因为监控系统传感器信号抗干扰能力很差, 遇有线路接触不良或电磁干扰就会造成假象信号。如传感器插头氧化、电缆接线盒松动、信号接触不良等造成随机出现的异常现象。

表 5 设备异常孤立点数据

Tab. 5 The outliers based in the abnormal equipment

A	1.36	0.37	1.19	0.35	1.21

瓦斯浓度有较多次高于其它时段的浓度。其高的次数越来越多, 可以考虑这是瓦斯突出, 应报警断电。

表 6 瓦斯突出孤立点数据

Tab. 6 The abnormity data of gas concentration

H	0.35	0.59	1.69	0.58	1.07

瓦斯突出是指随着煤矿开采深度的增加、瓦斯含量的增加, 在煤层中形成了在地应力作用下, 瓦斯释放的引力作用下, 使软弱煤层突破抵抗线, 瞬间释放大量瓦斯和煤而造成的一种地质灾害。

4 结语

本文中采用的孤立点检测方法对煤矿井瓦斯监测数据处理, 滤除大数干扰, 保证正确的报警。将孤立点技术用在煤矿安全监测中, 提高了数据分析的效率, 以有效地挖掘出事故的隐患, 这在保证煤矿的安全生产上有一定的价值。

参考文献:

[1] 牛 琨. 聚类分析中若干关键技术及其在电信领域的应用研究[D]. 北京: 北京邮电大学, 2007.

[2] 杨兰仓. 数据挖掘中聚类和孤立点检测算法的研究[D]. 济南: 山东大学, 2008.

[3] 蔡晓明. 基于地理信息系统的煤矿瓦斯突出预测研究[D]. 昆明: 昆明理工大学, 2006.

[4] 杨永铭, 王 喆. 孤立点挖掘算法研究[J]. 计算机与数字工程, 2008(1): 11- 15.

[5] KNORR E, NG R. Algorithms for mining distance- based outliers in large datasets[C]. Proc of the VLDB Conf, 1998: 392- 403.

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

[7] 陆声链, 林士敏. 基于距离的孤立点检测研究[J]. 计算机工程与应用, 2004, 40(33): 73- 75.

[8] 张忠平, 宋少英, 宋晓辉. ISAD: 一种新的基于属性距离和的孤立点检测算法[J]. 计算机工程与科学, 2009, 31(3): 83- 85.

[9] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density- based local outliers[C]. Proceedings of SIGMOD'00, Dallas, Texas, 2000: 427- 438

[10] 施化吉, 周书勇, 李星毅. 基于平均密度的孤立点检测研究[J]. 电子科技大学学报, 2007, 36(6): 1286- 1288.

[11] 贾晨科. 基于 K- 距离的孤立点和聚类算法研究[D]. 郑州: 郑州大学, 2006.

[12] 肖仁鑫. 煤矿安全预测的研究与集成[D]. 昆明: 昆明理工大学, 2006.

(责任编辑　刘存英)