

基于改进距离和的异常点检测算法研究

李春生, 于 澍, 刘小刚

(东北石油大学 计算机与信息技术学院 黑龙江 大庆 163318)

摘要: 为了降低原始数据中的勘误影响,提高数据质量,深入分析了常用的基于距离的异常点检测算法,提出了一种新的基于改进距离的异常点检测算法,舍去了传统算法中对 $DB(d, p)$ 参数的设置。首先,为了解决终端的不确定性选择属性困难的问题,引入了“属性隶属度”的概念,简化了检测属性的选择方式;其次,为了解决由于数据分布不均匀而导致的检测准确率较低的问题,改进了常用的距离度量,并采用改进的加权距离进行计算,得到距离矩阵,通过分析计算距离的总值,给出了一种异常评价方法用来判断异常点的异常程度;最后,以股票交易数据进行实验,与传统基于距离和的检测算法进行比较,结果表明该改进算法在异常点检测的准确度方面具有明显的改善。

关键词: 数据挖掘;改进距离;异常数据检测;距离和

中图分类号: TP301.6

文献标识码: A

文章编号: 1673-629X(2019)03-0097-04

doi: 10.3969/j.issn.1673-629X.2019.03.021

Research on Outlier Detection Algorithm Based on Improved Distance

LI Chun-sheng, YU Shu, LIU Xiao-gang

(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: In order to reduce the influence of errata in the original data and improve the data quality, we deeply analyze the commonly used distance-based outlier detection algorithm and propose a new outlier detection algorithm based on the improved distance, omitting the setting of $DB(d, p)$ parameter in the traditional algorithm. First of all, in order to solve the problem of terminal uncertainty selection attributes, the concept of “attribute membership degree” is introduced to simplify the selection of detection attributes. Secondly, in order to solve the problem of low detection accuracy caused by uneven data distribution, the commonly used distance measurement is improved, and the improved weighted distance is used for calculation to obtain the distance matrix. By analyzing the total value of the calculated distance, an anomaly evaluation method is proposed to judge the anomaly degree of the abnormal points. The experiment is conducted with the stock trading data. Compared with traditional distance-based detection algorithm, it shows that the improved algorithm has a significant improvement in accuracy of abnormal point detection.

Key words: data mining; improved distance; outlier detection; distance and distance

0 引言

随着网络和信息的广泛普及,人们在日常生活中会产生大量的数据,而从这些数据中发现某种未知的规律,挖掘其中隐含的信息与知识,成为了研究者的研究乐趣。在这个过程中,数据预处理占据了较大的一部分,其中很重要的一方面就是对于异常数据的识别,即异常点检测技术,它会影响后续数据挖掘模型的预测效果和实用性。异常点又叫做噪声、孤立点、离群点等。然而发现的异常数据并不能简单地直接剔

除掉,不仅会影响数据的连续性,而在某些领域中,这些被认为是不正常的的数据隐藏着更有价值的信息,例如把异常点检测技术应用到以下领域:保险理赔、银行交易中的欺诈检测及风险分析;医药研究中药品所产生的异常反应^[1];网络安全方面入侵行为的检测。

目前,异常数据的挖掘算法大致有基于统计、基于距离、基于密度、基于聚类等。其中传统的基于距离的异常数据挖掘算法原理较为简单,使用方便,在数据集分布均匀的情况下,检测效果较好,但存在以下缺陷:

收稿日期: 2018-04-09

修回日期: 2018-08-09

网络出版时间: 2018-12-19

基金项目: 国家自然科学基金面上项目(51774090);黑龙江省自然科学基金面上项目(F2015020);黑龙江省教育科研专项引导性创新基金项目(2017YDL-12);黑龙江省教育规划重大课题(GJ20170006)

作者简介: 李春生(1960-),男,博士,教授,博导,研究方向为数据挖掘与智能系统、软件集成技术、图像处理与模式识别、智能仪器与计算机控制系统;于 澍(1992-),女,硕士研究生,研究方向为数据挖掘。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20181219.1510.014.html>

参数设置复杂,检测效果对参数要求较高;异常数据并不是存在于所有属性上,只存在于某些影响数据稳定性的属性上;当数据集分布不均匀时,可能会出现偏差,不能有效地检测出异常点。

为了解决上述问题,提出了一种基于改进距离和的异常数据检测算法;对于终端的不确定性选择属性困难的问题,引入了“属性隶属度”的概念;用改进的距离进行异常点的检测,给出了异常点的异常程度。

1 基于距离的异常点检测

1.1 异常点检测方法

最先提出基于距离的异常点检测方法的是 Knorr 和 Ng^[2],他们认为的异常点是这样的:在数据集中至少有 k 部分数据点之间的距离大于某一阈值 d 。其实质是将异常点看作是在 d 范围内邻居较少的点。基于距离的异常点检测算法一般分为以下三种:

(1) 基于索引的算法(index-based)^[3]。

该算法首先建立了一个多维索引结构,然后根据该索引结构查找出各个数据点在半径为 d 的范围内的邻居。需事先设定临界个数 k ,在查找过程中当某点的邻居个数一旦大于 k ,则说明该点不是异常点,立即停止查找,并标记该点为正常的数据点。直至将所有点都做好是否为异常点的标记。在最不理想的情况下该算法的复杂度为 $O(s * n^2)$,其中 s 表示数据的维数, n 表示数据点的个数。由于构建索引结构需要的时间较长,所以该算法是非常耗时的,因此应用较少。

(2) 基于循环嵌套的算法(nested-loop)^[4]。

为了避免建立索引结构,提出了一种基于循环嵌套的算法。首先将数据集均匀分成若干数据块,同时将内存缓冲区均分为 A 和 B 两部分。然后将未在该区存储过的数据块读入 A 部分,其他数据块依次由 B 部分读取(每次只读取一个)。最后计算 A 和 B 两部分的数据块中每对数据对象之间的距离,并记录 A 部分的每个数据对象的邻居数。一旦大于 k 个,则停止计算, A 部分继续读取下一个数据块;若小于 k 个,则 B 部分继续读入下一数据块。直至读取完所有的数据块,并累加邻居数。该算法的时间复杂度与第一种算法的复杂度相同。

(3) 基于单元的算法(cell-based)^[4]。

基于单元的算法是想要避免时间复杂度与前两种方法相同,因此它将数据集划分为一个个单元,对每个单元的异常点的个数进行计算而不是对每个数据对象进行计算。该算法的时间复杂度为 $O(C^k + n)$ 。其中 $C^k = m(1 + 2k + 1)^k$, m 为单元数, k 为数据的维数。而早些年前,Knorr 和 Ng 通过实验证明了只有当 $k \leq 4$ 时该算法的运行时间才会优于 NL 算法^[2]。

1.2 距离的度量及异常点的定义

常用的距离度量方法有曼氏距离、欧氏距离、切比雪夫距离等。

对于两个 m 维空间中的数据点 x_i 与 x_j ,它们之间的欧几里得距离和曼哈顿距离可以由明考斯基距离来概括,即:

$$d_{ij} = \left[\sum_{k=1}^m |x_{ik} - x_{jk}|^q \right]^{1/q} \quad (1)$$

其中,当 $q=1$ 时为曼哈顿距离,表示为:

$$d_{ij} = \sum_{k=1}^m |x_{ik} - x_{jk}| \quad (2)$$

当 $q=2$ 时,表示为最常用的欧几里得距离:

$$d_{ij} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2} \quad (3)$$

其中, x_{ik} 表示在第 k 维的第 i 个分量; x_{jk} 表示第 k 维的第 j 个分量。

切比雪夫距离表示为各坐标数值差的最大值,即:

$$D(p, q) = \max_i (p_i - q_i) \quad (4)$$

常用距离度量种类比较多,使用哪种距离度量要看算法应用的具体领域。用不同的距离度量进行计算,得到的结果也可能是不同的。

常见的基于距离的异常点定义有以下几种:

(1) 在数据集 S 中,若点 O 是异常点,那么至少有 p 部分对象到点 O 的距离大于 d ,也就是说,如果点 O 在以 d 为半径的范围内,有不超过 M 个点的邻居,那么点 O 就是一个带有参数 p 和 d 的异常点,表示为 $DB(p, d)$,这里 $M = n(1-p)$, n 是数据集 S 中的数据对象个数。

(2) 异常点是指数据集 S 中的 n 个点,这 n 个点到其第 k 个最近邻点的距离是其他所有点中最大的,其中 n 是异常点个数的估计值。

(3) 在数据集 S 中,取前 n 个与其 k 个最近邻点的距离之和最大的点,则这 n 个点就是异常点。其中 n 是异常点个数的估计值。

尽管以上几种定义都各不相同,但都是基于距离的异常点的定义。文中将使用第三种定义来判断数据集集中的某个点是否为异常点。

2 基于改进距离和的异常点检测

公式中的参数及其意义如表 1 所示。

表 1 参数及意义

参数	意义
μ_k	第 k 个属性的属性隶属度
ω_{ik}	第 i 条数据的第 k 个属性
ξ_k	第 k 个属性的平均值
$D_{i(k)}$	x_i 到其 k 个最近邻居的距离和
$D_{j(k)}$	x_j 到其 k 个最近邻居的距离和
ϕ_i	第 i 个点的异常程度

2.1 检测属性的选取

在数据集中,异常点并不是存在于每个属性上,它们只是在其中某一部分属性上出现异常状态。一般情况下,都是由该领域的专家来选取这些有研究价值的属性,但是当终端操作人员不具备相关的专业知识时,要从众多数据中选取能够影响数据稳定性且具有研究价值的属性比较困难,为此引入了“属性隶属度”的概念。它能够反映每个属性的检测价值,即使无领域专家在场的情况下,终端操作人员也能够通过计算每个属性的“属性隶属度”,选取到最适合的检测属性。

属性隶属度 μ :对于数据集中任意一条数据 ω 的任意属性,都有一个数 $\mu(\omega)$ 与之对应, μ 为该属性的“属性隶属度”, $\mu(\omega)$ 为该属性的隶属度函数,表示为:

$$\mu_k(\omega) = \frac{\sum_{i=1}^n |\omega_{ik} - \xi_k|}{\sum_{i=1}^n |\omega_{ik}|} \quad (5)$$

属性的 $\mu(\omega)$ 值越大,说明该属性值的波动越大,检测价值也越高,就越有可能成为被检测的对象; $\mu(\omega)$ 值越小,该属性值的波动也越小,检测价值也越低,就越有可能被忽略。

2.2 改进距离度量

为了解决由数据分布不均匀而导致的检测准确率低的问题,文中改进了距离度量,以明考斯基距离为例,表示为:

$$d_{ij}^M = \lambda_k \left[\sum_{i=1}^m |x_{ik} - x_{jk}|^q \right]^{\frac{1}{q}} \quad (6)$$

其中 λ_k 定义为:

$$\lambda_k = \frac{k^2}{D_{i(k)} D_{j(k)}} \quad (7)$$

对于分布不均匀的数据集来说,基于常用距离的异常点检测算法的检测效果较差^[5-7],当数据点分别分布在密集区域和稀疏区域时,由 k 个最近邻点所组成的局部区域范围是不同的,按照传统基于距离的算法进行异常点检测时,就有可能将原本正常的的数据点标记为异常点^[8]。

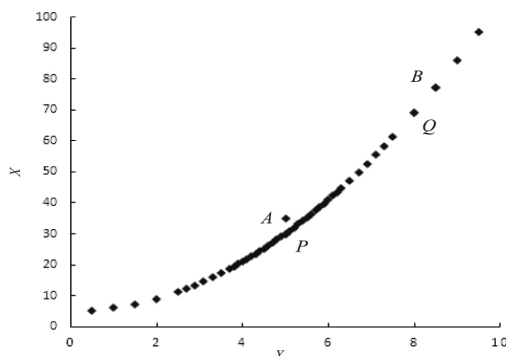


图1 不均匀分布的散点图

图1所示为非均匀分布的抛物线形数据集 S ,已知点 A 为数据集 S 中的一个异常点,点 B 为数据集 S 中正常的数据点,若点 B 到其 k 个最近邻点的距离之和大于点 A 到其 k 个最近邻点的距离之和,传统基于距离的算法就可能会把 B 点当作异常点来处理,而把点 A 认为是正常的数据点^[9-10]。

改进后的距离由式5可知。设 $d_{A(k)} < d_{B(k)}$,如果点 A 到其最近邻点(设为点 P)的距离与点 B 到其最近邻点(设为点 Q)的距离相等,即 $d_{AP} = d_{BQ}$,那么由于 $d_{P(k)} < d_{Q(k)}$ (假设 $k=4$),所以 $d_{AP}^M > d_{BQ}^M$ 。由此可以看出,改进后的距离“放大”了异常点与其临近点之间的距离,而“缩短”了正常点与其临近点之间的距离。换句话说,改进后的距离突出了异常点,使之看起来更加异常,更方便判断^[11-12]。

2.3 基于改进距离和的异常点检测算法

传统基于距离的算法中,对 $DB(p, d)$ 的参数设置比较复杂^[13],需要不断进行测试,找出符合用户需求的参数,而所得结果对参数是敏感的。为了避免参数的设置,提出了一种基于改进距离和的异常点检测算法,并且给出了异常点的评价方法。步骤描述如下:

Step1: 假设数据已经过标准化,计算数据集中第一条数据与其他数据之间的距离 d_{ij} 。

Step2: 由Step1得到该数据点与 k 个最近邻点的距离之和 $D_{i(k)}$ 。

Step3: 计算 λ_k ,求出数据点与其他数据点之间的改进距离 $d_{ij}^M = \lambda_k d_{ij}$ 。

Step4: 循环Step1到Step3,直至将数据集中所有数据点的 d_{ij}^M 都计算出来,形成一个主对角线元素为0的对称矩阵 P :

$$P = \begin{bmatrix} 0 & d_{12}^M & \cdots & d_{1n}^M \\ d_{21}^M & 0 & \cdots & d_{2n}^M \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1}^M & d_{n2}^M & \cdots & 0 \end{bmatrix}$$

Step5: 计算 $\phi_i = \sum_{j=1}^{n-1} d_{ij}^M$,则 ϕ 值最大的 M 个数据即为异常数据(M 为期望的异常点个数)。

对此算法的说明如下:

(1) 矩阵 P 中每个元素代表两个点之间的距离,例如 d_{12}^M 表示第一个点与第二个点之间的距离。

(2) ϕ_i 为评价异常点异常程度的关键,即矩阵 P 中第 i 行元素之和, ϕ_i 的值越大,就说明数据点 i 离其他点越远。即该点比其他点异常。

3 实验设计

实验选取了100组股票交易数据进行异常点检

测,每条记录包含 6 个属性。通过计算每个属性的 μ 值,最终选择了 μ 值最大的两个属性即交易量与交易金额。经筛选后的数据集如表 2 所示。

表 2 筛选后的部分数据集

序号	交易量	交易金额
1	58	2 900
2	100	5 000
3	122	6 100
4	200	10 000
5	200	4 792
6	1 072	25 685.1
...
99	131.4	1 333.67
100	154	18 072.6

实验中所用的距离度量是选取 $q=2$ 时的明考斯基距离进行计算,当 $k=30$ 时,距离和矩阵 P 为:

$$P = \begin{bmatrix} 0 & 3.12 & 5.38 & \cdots & 29.33 \\ 3.12 & 0 & 0.29 & \cdots & 10.08 \\ 5.38 & 0.29 & 0 & \cdots & 6.28 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 29.33 & 10.08 & 6.28 & \cdots & 0 \end{bmatrix}$$

由矩阵 P 可计算出 100 个 ϕ 值,对 ϕ 值进行降序排列,设用户期望的异常值为 4,则可得到四个异常点,如表 3 所示。

表 3 异常点检测结果

序号	交易量	交易金额	ϕ 值
6	1 072	25 685.12	912.76
35	2 165	18 012.8	699.55
100	154	18 072.6	698.54
61	1 390	11 564.8	415.09

从输出的结果来看,求得的数据点的距离之和按降序排列,则可取前四条记录,即序号为 6,35,100,61 的数据与其他点距离之和最大,可判定为异常数据。

如图 2 所示,实验将数据量增加至 200,500。并分别对两种算法检测出的异常点个数进行了对比。传统

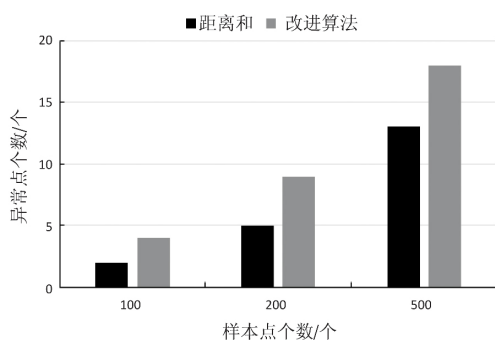


图 2 算法检测结果对比

基于距离和算法的检测效果不理想,准确率约为 70%;而基于改进距离和的异常点检测算法的准确率可以达到 90% 以上,取得了较好的检测效果。

4 结束语

介绍了几种基于距离的异常点检测算法和几种常用的距离度量,给出了一种选择检测属性的方法,提出了一种改进的基于距离和的异常点检测算法。该算法舍去了传统算法中对 $DB(p, d)$ 参数的设置,避免了因参数不合适而产生参差不齐的结果,并给出了数据的异常程度,能够更直观地呈现出来。与传统基于距离和的检测算法进行了比较,证明了该算法在分布不均匀的数据集上有较好的检测效果。现阶段异常点检测技术的难点在于对高维数据的处理方法,因此基于改进距离和的异常点检测算法在高维数据上的应用还有待于进一步研究。

参考文献:

- [1] 张宏翔. 使用 RNN 的基于距离的孤立点检测[J]. 信息与电脑, 2017(8): 81-82.
- [2] KNORR E M, NG R T, TUCAKOV V. Distance-based outliers: algorithms and applications[J]. VLDB Journal, 2000, 8(3): 237-253.
- [3] HAN Jiawei, KAMBER M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2002: 223-259.
- [4] 孟静. 异常数据挖掘算法研究与应用[D]. 无锡: 江南大学, 2013.
- [5] 韦佳, 彭宏, 林毅申. 基于改进距离的孤立点检测方法[J]. 华南理工大学学报: 自然科学版, 2008, 36(9): 25-30.
- [6] 陆声链, 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] 王宏鼎, 童云海, 谭少华, 等. 异常点挖掘研究进展[J]. 智能系统学报, 2006, 1(1): 67-73.
- [8] 邵峰晶. 数据挖掘原理与算法[M]. 北京: 中国水利水电出版社, 2003.
- [9] 赵泽茂, 何坤金, 陈鹏, 等. Web 日志文件的异常数据挖掘算法及其应用[J]. 计算机工程, 2003, 29(17): 195-197.
- [10] 王晓燕. 几种常用的异常数据挖掘方法[J]. 甘肃联合大学学报: 自然科学版, 2010, 24(4): 68-71.
- [11] 杨臻, 张明慧. 基于双倍距离的孤立点检测算法研究[J]. 制造业自动化, 2013, 35(8): 40-42.
- [12] 侯晓晶, 王会青, 陈俊杰, 等. 基于最近邻距离差的改进孤立点检测算法[J]. 计算机工程与设计, 2013, 34(4): 1265-1269.
- [13] 李强, 李振东. 数据挖掘中孤立点的分析研究在实践中的应用[J]. 微计算机应用, 2006, 27(3): 323-327.