

基于最小超球面密度的孤立点检测算法

冯宇¹, 苑易伟²

(1. 长安大学 电子与控制工程学院, 陕西 西安 710064;

2. 西安理工大学 自动化与信息工程学院, 陕西 西安 710048)

摘要:定义了最小超球面密度的概念,提出了一种基于最小超球面密度的孤立点检测算法(minimum hyper sphere density, MHSD)。该算法根据数据的 k 近邻和反 k 近邻获得数据的有效近邻,并使用最小超球面密度和有效近邻计算每个数据的密度背离程度,进而计算每个数据的孤立程度,将孤立程度超过规定阈值的数据视为孤立点。实验数据为一个二维人工数据集和两个高维实际数据集,检测三个数据集的孤立点,对算法性能进行评估,并与经典的局部离群因子算法(local outlier factor, LOF)、离群影响因子算法(influenced outlierness, INFLO)和密度相似邻域离群因子算法(density similarity neighbor based outlier factor, DSNOF)进行比较。实验结果表明,基于最小超球面密度的孤立点检测算法可以准确检测出数据中的孤立点,且性能优于三种经典算法。

关键词:孤立点检测;最小超球面;有效近邻;局部密度差;密度背离程度

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2019)06-0032-05

doi:10.3969/j.issn.1673-629X.2019.06.007

An Outlier Detection Algorithm Based on Minimum Hyper Sphere Density

FENG Yu¹, YUAN Yi-wei²

(1. School of Electronics and Control Engineering, Chang'an University, Xi'an 710064, China;

2. School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: Minimum hyper sphere density (MHSD) is defined and an outlier detection algorithm based on MHSD is proposed. The effective neighbors are obtained according to k -nearest neighbors and reverse k -nearest neighbors. The density deviation degree of each datum is calculated using minimum hyper sphere density and effective neighbors. Then the isolation degrees can be calculated. Data are regarded as outliers when their isolation degrees are higher than the threshold. A two-dimensional artificial data set and two high-dimensional real data sets are used to evaluate the algorithm performance. The mining results are compared with those of three classical algorithms, which are local outlier factor (LOF), influenced outlierness (INFLO) and density similarity neighbor based outlier factor (DSNOF). The experiment shows that MHSD can find outliers accurately and its performance is better than the three classical algorithms.

Key words: outlier detection; minimum hyper sphere; effective neighbor; local density difference; density deviation

0 引言

孤立点也被称作离群点或异常点,是指数据中不符合一般规律或明显偏离其他数据的数据。在数据挖掘研究领域,孤立点检测是一个重要的研究方向,其任务是通过算法寻找与大部分数据有着不同属性或含义的少量数据^[1]。一方面,在数据预处理阶段发现并剔除这些数据,可以降低孤立点对数据挖掘结果的负面影响;另一方面,罕见的数字不一定是无用的数据,可

能蕴含更大的研究价值,通过研究这些数据,会发现一些新的知识^[2]。孤立点检测技术已经广泛应用于工业控制、网络安全、电子商务等各个领域^[3-4]。

1 相关研究

孤立点检测算法大致可以分为四类,分别为基于统计、基于聚类、基于距离和基于密度的算法。基于统计的孤立点检测算法需要已知数据集的统计学分布规

收稿日期:2018-09-06

修回日期:2018-12-27

网络出版时间:2019-03-20

基金项目:中央高校基本科研业务费专项资金(300102328103);陕西省自然科学基金(2017JQ6075)

作者简介:冯宇(1984-),男,博士,讲师,研究方向为数据挖掘、生物信号测量与分析。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20190319.1729.002.html>

律^[5-6],并且应明确孤立点不服从这种分布规律。然而实际中孤立点的分布规律往往并不清楚,与此同时,结构复杂的高维数据集会明显减弱这类算法的效果。基于聚类的孤立点检测算法是通过聚类分析将数据划分为若干个簇,并给定阈值,小于设定阈值的簇即视为孤立点^[7]。这类算法在聚类过程中就能检测孤立点,但其检测的效果与簇形成的质量密切相关^[8]。基于距离的孤立点检测算法需要计算每个对象与其近邻的距离,并且认为孤立点与其近邻的距离比正常数据远。这类算法应用简单,并且其有效性得到了充分的论证^[9],但如果距离参数选取不合适,算法的效果会变得很不理想^[10]。基于密度的孤立点检测算法根据数据的局部孤立程度来寻找孤立点,该思路更符合孤立点定义,因此基于密度的孤立点检测算法近年来发展迅速^[11-12]。

2 关键技术

文中提出一种基于最小超球面密度的孤立点检测算法(minimum hyper sphere density, MHSD),在介绍算法之前,需对相关的定义和概念进行阐述。

2.1 相关定义

定义1:给定正整数 n ,一个 n 维球面是 $n+1$ 维空间中距离某个点为 R 的所有点的集合。

特别地,一个0维球面是长度为 R 的线段的两个端点,一个1维球面是半径为 R 的圆,一个2维球面是半径为 R 的球体的面。维数大于2的球面称为超球面。最小球是包含对象 p 的 k 近邻所有超球中半径 R 最小的球面, p 不一定是球心。

定义2: $NNk(p)$ 是对象 p 的 k 近邻序列, $RNNk(p)$ 是 k 近邻中包含 p 的所有对象,即反近邻。

以图1为例介绍 $RNNk(p)$ 的构造方法。假设数据集 $A = \{p, q_1, q_2, q_3, q_4, q_5\}$,当 $k=3$ 时, $NNk(p) = \{q_1, q_2, q_3, q_4\}$, $NNk(q_1) = \{p, q_2, q_4\}$, $NNk(q_2) = \{p, q_1, q_3\}$, $NNk(q_3) = \{q_1, q_2, q_5\}$, $NNk(q_4) = \{p, q_1, q_2, q_5\}$, $NNk(q_5) = \{q_1, q_2, q_3\}$ 。则根据定义2, p 的 $RNNk(p)$ 为 $\{q_1, q_2, q_4\}$,同样也能得到其他对象的 $RNNk$ 。显然, $NNk(p)$ 与 $RNNk(p)$ 不一定相等。

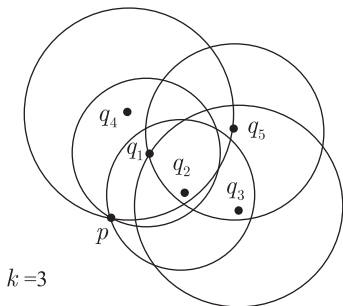


图1 反近邻构造方法示意

定义3:对于给定的正整数 k ,对象 p 的有效近邻定义为 p 的 k 近邻对象的 k 近邻中包含 p 的所有对象,即: $ENP(p) = NNk(p) \cap RNNk(p)$, $ENP(p)$ 可能为空集。

以图2为例介绍 $ENP(p)$ 的构造方法。图中数据集 $D = \{a, b, c, d, e, f, g, h, i, j, k\}$,当 $k=4$ 时, d 的 k 近邻 $NNk(d) = \{e, g, h, j\}$,反近邻 $RNNk(d) = \{e, j\}$; g 的 k 近邻 $NNk(g) = \{e, f, h, i\}$,反近邻 $RNNk(g) = \{e, f, h, i, k\}$ 。则 d 和 g 的有效近邻分别为 $ENPk(d) = \{e, j\}$ 和 $ENPk(g) = \{e, f, h, i\}$ 。

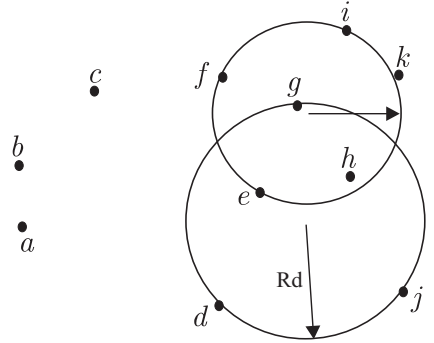


图2 有效近邻构造示意

根据上述定义,文中对最小超球面密度的定义为:

定义4:对象 p 的最小超球面密度是 p 的有效 k 近邻个数与最小超球面半径 R 的比值,即

$$spden(p) = \frac{|ENP(p)|}{R(p)} \quad (1)$$

其中, $|ENP(p)|$ 为有效近邻的个数。

2.2 算法描述

文中算法首先计算所有数据的有效 k 近邻,然后根据 k 近邻中的数据和对象 p 的距离从小到大排列,得到对象的最近近邻序列,通过计算对象 p 与有效 k 近邻中的最小超球面密度差来计算对象 p 的密度背离程度,最后根据有效 k 近邻中的所有数据的密度背离程度计算对象 p 的孤立程度。算法具体描述如下:

步骤1:计算每两个数据间的距离,并找到所有数据的 k 近邻。

步骤2:构建每个数据的近邻序列NNS。

设 $NNS(p)$ 的初始值是 $\{p\}$,每次迭代把 $NNk(p)$ 中距离 p 最小的对象加入到 $NNS(p)$ 中,当 $NNk(p)$ 中所有对象都加入到 $NNS(p)$ 时,迭代结束。

步骤3:根据定义4计算所有数据的最小超球面密度。

步骤4:计算每两个数据的最小超球面密度差:

$$\Delta spden(x, y) = |spden(x) - spden(y)| \quad (2)$$

且 $\Delta spden(x, y) = \Delta spden(y, x)$ 。

步骤5:根据最小超球面密度差,计算对象的最近近邻序列的密度背离程度(NDD)。

$$\text{NDD}(p)=\frac{\sum_{i=1}^r\Delta\text{spden}(p,c_i)}{i}$$

(3)

其中, $r=|\text{NNk}(p)|$ 。

步骤 6: 计算每个数据的孤立程度。

$$\text{NDDOF}(p)=\frac{|\text{NNk}(p)|\times\text{NDD}(p)}{\sum_{o\in\text{NNk}(p)}\text{NDD}(o)}$$

(4)

该值越小,说明 p 的孤立程度越低;该值越大,说明 p 越有可能是孤立点。根据设定的阈值便可检测出孤立点。

3 实验结果与分析

3.1 检测人工数据集的孤立点

人工二维数据集如图 3 所示,数据集包含 1 700 个正常数据和 7 个孤立点。这 1 700 个正常数据分为一个由 200 个服从高斯分布的数据组成的低密度簇 C_1 和三个由 500 个服从高斯分布的数据组成的高密度簇 C_2 、 C_3 和 C_4 。7 个孤立点的编号分别 o_1,o_2,\cdots,o_7 。

使用三种经典的孤立点检测算法与文中提出的

MHSD 方法进行检测结果比较,三种方法分别为局部离群因子算法 (local outlier factor, LOF)、离群影响因子算法 (influenced outlierness, INFLO) 和密度相似邻域离群因子算法 (density similarity neighbor based outlier factor, DSNOF)^[13-15],各方法检测结果中孤立程度最高的 7 个点如表 1 所示。可以看出,INFLO 无法检测出 o_1 、 o_5 、 o_6 和 o_7 ,DSNOF 无法检测出 o_4 ,LOF 和 MHSD 均可检测出 7 个孤立点,MHSD 每个孤立点的孤立程度比 LOF 更大,说明 MHSD 的检测效果最理想。

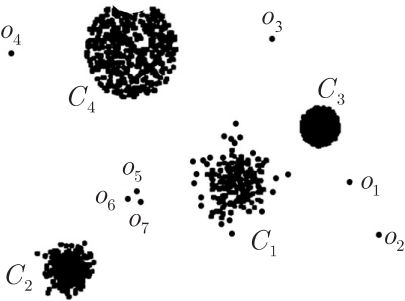


图 3 人工数据集示意

表 1 人工数据集孤立点检测结果

排序	LOF		INFLO		DSNOF		MHSD	
	编号	孤立程度	编号	孤立程度	编号	孤立程度	编号	孤立程度
1	o_2	29.2	o_2	10.0	o_2	8.2	o_1	50.0
2	o_1	15.1	o_3	10.0	o_1	6.5	o_2	50.0
3	o_3	14.1	o_4	10.0	o_3	5.4	o_3	50.0
4	o_5	6.0	1 089	1.4	o_7	2.9	o_7	50.0
5	o_7	4.9	1 054	1.3	o_5	2.8	o_7	49.3
6	o_6	4.7	1 689	1.3	o_6	2.7	o_6	22.6
7	o_4	3.6	1 193	1.3	438	2.3	o_5	19.7

3.2 算法性能评估

引入 precision (Pr), recall (Re) 和 rank power (RP) 三个参数评估算法的性能^[16]。Pr 表示算法检测出的孤立点的孤立程度在孤立程度最大的 m 个数据中的比例,Re 表示算法在检测出的孤立程度最大的 m 个数据中孤立点个数占孤立点总数的比例,RP 表示算法检测出的 n 个孤立点在孤立程度最大的 m 个数据中所占的位置参数。三个参数的值越大,表明算法性能越好。

数据集 1 选用经典的皮马印第安人糖尿病数据

集^[17],该数据集分为两类,第一类包含 500 个数据,第二类包含 268 个数据。将第一类数据随机删除 488 个,得到一个随机分布的簇,作为此实验中的孤立点^[18]。为了比较在不同参数选择下不同算法的效果,将近邻个数 k 选为数据的维数、数据量的 5% 和 10%,即 k 的值选择为 8、14 和 28。各个算法的性能评估结果如表 2 ~ 表 4 所示,其中 Nrc 为算法检测出的孤立点个数。

表 2 $k=8$ 时的算法性能评估结果

m	LOF				INFLO				DSNOF				MHSD			
	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP
10	0	0	0	0	0	0	0	0	0	0	0	0	1	0.1	0.1	0.3
20	0	0	0	0	1	0.1	0.1	0.1	0	0	0	0	6	0.3	0.5	0.3
30	1	0.1	0.1	0.1	3	0.1	0.1	0.1	0	0	0	0	6	0.2	0.5	0.3

续表 2

<i>m</i>	LOF				INFLO				DSNOF				MHSD			
	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP
40	1	0.1	0.1	0.1	3	0.1	0.3	0.1	2	0.1	0.2	0.1	7	0.2	0.6	0.2
50	2	0.1	0.2	0.1	3	0.1	0.3	0.1	3	0.1	0.3	0.1	7	0.1	0.6	0.2
60	3	0.1	0.3	0.1	3	0.1	0.3	0.1	3	0.1	0.3	0.1	9	0.2	0.8	0.2
70	4	0.1	0.3	0.1	3	0.1	0.3	0.1	3	0.1	0.3	0.1	10	0.1	0.8	0.2
80	4	0.1	0.3	0.1	4	0.1	0.3	0.1	3	0.1	0.3	0.1	11	0.1	0.9	0.2
90	4	0.1	0.3	0.1	4	0.1	0.3	0.1	3	0.1	0.3	0.1	12	0.1	1	0.2

表 3 $k=14$ 时的算法性能评估结果

<i>m</i>	LOF				INFLO				DSNOF				MHSD			
	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP
10	0	0	0	0	0	0	0	0	0	0	0	0	4	0.4	0.3	0.3
20	1	0.1	0.1	0.1	0	0	0	0	0	0	0	0	8	0.4	0.7	0.4
30	1	0.1	0.1	0.1	1	0.1	0.1	0.1	0	0	0	0	9	0.3	0.8	0.4
40	3	0.1	0.3	0.1	2	0.1	0.2	0.1	0	0	0	0	10	0.3	0.8	0.4
50	3	0.1	0.3	0.1	2	0.1	0.2	0.1	0	0	0	0	10	0.2	0.8	0.4
60	3	0.1	0.3	0.1	2	0.1	0.2	0.1	1	0.1	0.1	0.1	10	0.2	0.8	0.4
70	3	0.1	0.3	0.1	2	0.1	0.2	0.1	1	0.1	0.1	0.1	11	0.2	0.9	0.3
80	3	0.1	0.3	0.1	3	0.1	0.3	0.1	1	0.1	0.1	0.1	12	0.2	1	0.3

表 4 $k=28$ 时的算法性能评估结果

<i>m</i>	LOF				INFLO				DSNOF				MHSD			
	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP
10	1	0.1	0.1	0.1	2	0.2	0.2	0.3	2	0.2	0.2	0.3	3	0.3	0.3	0.4
20	3	0.2	0.3	0.2	3	0.2	0.3	0.2	3	0.2	0.3	0.2	6	0.3	0.5	0.3
30	3	0.1	0.3	0.2	4	0.1	0.3	0.2	5	0.2	0.3	0.2	8	0.3	0.8	0.3
40	4	0.1	0.3	0.1	4	0.1	0.3	0.2	6	0.2	0.5	0.2	10	0.3	0.8	0.3
50	4	0.1	0.3	0.1	5	0.1	0.4	0.2	7	0.1	0.6	0.2	10	0.2	0.8	0.3
60	5	0.1	0.4	0.1	7	0.1	0.6	0.1	8	0.1	0.7	0.2	12	0.2	1	0.3

当 $k=8$ 时,在前 10 个孤立程度最大的数据中,只有 MHSD 能发现一个孤立点。随着 m 的增大,MHSD 能够检测出的孤立点个数迅速增多,当 $m=90$ 时,所有孤立点全被检测出来。此时 LOF、INFLO 和 DSNOF 分别检测出 4、4、3 个孤立点,MHSD 的性能最好。当 $k=14$ 时和 $k=28$ 时,有类似的结果。真实数据集算法性能评估结果表明,MHSD 算法随着 k 值的增大效

果变好,并且优于 LOF、INFLO 和 DSNOF。

数据集 2 选用本课题组在生物信息实验中实际测量的小鼠心脏窦房结场电位信号数据集,将在常温环境下($22\pm2^{\circ}\text{C}$)测量到的 230 个数据作为正常数据,从低温环境下($5\pm2^{\circ}\text{C}$)测量到的数据中取出 10 个作为孤立点,数据维度为 30,取 $k=30$ 。各个算法的性能评估结果如表 5 所示,MHSD 的性能最优。

表 5 使用生物信息数据集的算法性能评估结果

<i>m</i>	LOF				INFLO				DSNOF				MHSD			
	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP
5	3	0.7	0.4	0.8	3	0.7	0.3	0.8	2	0.4	0.3	1	4	0.8	0.4	0.8
10	5	0.7	0.6	0.8	4	0.7	0.5	0.6	4	0.4	0.5	0.7	7	0.8	0.7	0.8
15	6	0.6	0.6	0.7	6	0.6	0.6	0.6	5	0.4	0.6	0.6	8	0.6	0.8	0.8
20	6	0.6	0.7	0.7	6	0.4	0.6	0.6	7	0.4	0.6	0.5	10	0.5	1	0.8

4 结束语

文中提出了一种基于最小超球面密度的孤立点检测算法,该算法通过找出对象数据的近邻序列,利用最小超球面密度差计算近邻序列密度背离程度,进而计算每个数据的孤立程度。使用人工数据集和真实数据集的检测结果表明,MHSD 可以检测孤立点,且与经典的 LOF、INFLO 和 DSNOF 三种算法相比,MHSD 的性能最优。

参考文献:

- [1] HAWKINS D. Identification of outliers[M]. London: Chapman and Hall, 1980.
- [2] 任建华,高立明. 基于聚类的两段式孤立点检测算法[J]. 计算机工程与应用, 2016, 52(20): 98–102.
- [3] 乔俊飞,孙玉庆,韩红桂. 改进 K-MEANS 算法优化 RBF 神经网络的出水氨氮预测[J]. 控制工程, 2018, 25(3): 375–379.
- [4] 邓必年. 基于剔除孤立点的物流成本预测模型[J]. 现代电子技术, 2017, 40(13): 114–117.
- [5] HIDO S, TSUBOI Y, KASHIMA H, et al. Inlier-based outlier detection via direct density ratio estimation[C]//Eighth Ieee International Conference On Data Mining. PISA, ITALY: IEEE, 2008: 223–232.
- [6] 牛立尚. 一种基于统计特征的孤立点和边缘点检测算法[J]. 信息技术, 2015(6): 112–114.
- [7] CASSISI C, FERRO A, GIUGNO R, et al. Enhancing density-based clustering: parameter reduction and outlier detection[J]. Information Systems, 2013, 38(3): 317–330.
- [8] WANG C H. Outlier identification and market segmentation using kernel-based clustering techniques[J]. Expert Systems with Applications, 2009, 36(2): 3744–3750.
- [9] GHOTING A, PARTHASARATHY S, OTEY M E. Fast mining of distance-based outliers in high-dimensional datasets[J]. Data Mining and Knowledge Discovery, 2008, 16(3): 349–364.
- [10] ANGIULLI F, BASTA S, LODI S, et al. Distributed strategies for mining outliers in large data sets[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(7): 1520–1532.
- [11] ZHANG J F, ZHANG S L, JIANG Y Y. The local outliers mining system of celestial body spectrum based on constrained concept lattice[J]. Spectroscopy and Spectral Analysis, 2009, 29(2): 551–555.
- [12] ZHAO Haifeng, JIANG Bo, TANG Jin, et al. Image matching using a local distribution based outlier detection technique[J]. Neurocomputing, 2015, 148: 611–618.
- [13] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density-based local outliers[J]. ACM SIGMOD Record, 2000, 29(2): 93–104.
- [14] JIN Wen, TUNG A K H, HAN Jiawei, et al. Ranking outliers using symmetric neighborhood relationship[C]//Pacific-Asia conference on knowledge discovery & data mining. Berlin: Springer, 2006: 577–593.
- [15] CAO Hui, SI Gangquan, ZHANG Yanbin, et al. Enhancing effectiveness of density-based outlier mining scheme with density-similarity-neighbor-based outlier factor[J]. Expert Systems with Applications, 2010, 37(12): 8090–8101.
- [16] YUAN Yiwei, CAO Hui, ZHANG Yanbin, et al. Outlier mining based on neighbor-density-deviation with minimum hyper-sphere[J]. Information Technology and Control, 2016, 45(3): 267–277.
- [17] 何禹德. 基于数据挖掘技术的糖尿病临床数据分析[D]. 长春: 长春工业大学, 2016.
- [18] YE Mao, LI Xue, ORLOWSKA M E. Projected outlier detection in high-dimensional mixed-attributes data set[J]. Expert Systems with Applications, 2009, 36(3): 7104–7113.