DBSCAN 聚类算法的参数配置方法研究

宋金玉[1],郭一平[1],王 斌[2]
(1.解放军陆军工程大学 指挥控制工程学院,江苏 南京 210007;
2.解放军陆军工程大学 教学考试中心,江苏 镇江 212000)

摘 要: 随着互联网技术的飞速发展,海量数据涌现。在海量的数据中,存在大量无用甚至错误的“脏数据”,这些低质量的数据难以提供有价值的信息。数据质量低的一个方面就是数据异常。对数据异常检测问题进行了研究,将基于密度的 DBSCAN 聚类算法应用于数据的异常检测,并针对该算法在应用过程中对参数设置敏感的问题,提出了一种邻域阈值(Eps)和点数阈值(Minpts)的配置方法。该方法可根据数据集本身的统计特性以及图表的可视化展示来为算法确定合适的参数。利用 MATLAB 工具,编程实现了 DBSCAN 聚类算法及辅助参数的计算,并在 Iris 数据集上进行了实验验证。实验结果表明,用该方法进行 DBSCAN 聚类算法参数的设置是可行的,弥补了 DBSCAN 聚类算法参数设置的传统做法单靠经验的不足,使得检测结果的准确性和可伸缩性更好。

关键词: 数据异常检测; 聚类算法; DBSCAN; 参数配置
中图分类号: TP311 **文献标识码:** A **文章编号:** 1673-629X(2019)05-0044-05
doi: 10.3969/j.issn.1673-629X.2019.05.009

Research on Parameter Configuration Method of DBSCAN Clustering Algorithm

SONG Jin-yu[1],GUO Yi-ping[1],WANG Bin[2]
(1.School of Command and Control Engineering,Army Engineering University of PLA,Nanjing 210007,China;
2.Center of Teaching and Testing,Army Engineering University of PLA,Zhenjiang 212000,China)

Abstract: With the rapid development of Internet technology,massive data emerge. There are a large number of useless or even wrong “dirty data” in these data,and these low quality data are difficult to provide valuable information. Data exception is one aspect of low data quality. This paper discusses the application of DBSCAN clustering algorithm in the detection of abnormal data. Aiming at the problem that the algorithm is sensitive to parameter setting in the application process,we propose a configuration method of the neighborhood threshold (Eps) and the point threshold (Minpts) by applying the DBSCAN algorithm based on density to the anomaly detection of data. This method can determine the appropriate parameters according to the statistical characteristics of the data set itself and the visual presentation of the graph. Using MATLAB tool,the DBSCAN clustering algorithm and the calculation of auxiliary parameters are programmed,and the experimental verification is carried out on the Iris data set. The experiment shows that the method is feasible to set the parameters of DBSCAN clustering algorithm,which makes up for the lack of experience alone of the traditional method. The accuracy and scalability of the detection result are improved.

Key words: abnormal data detect; clustering algorithm; DBSCAN; parameter configuration

0 引 言

全球知名咨询公司麦肯锡称“数据,已经渗透到当今每一个行业和业务职能领域,成为重要的生产因素。”但现实生活中,人们常常抱怨“数据丰富,信息贫乏”。这是因为在海量的数据中,存在大量无用甚至错误的“脏数据”,根据“垃圾进,垃圾出(garbage in, garbage out)”[1]原理,低质量的数据难以提供有价值的信息,反而会带来负面影响,会因各种数据/信息质量(data/information quality,DQ/IQ)问题给用户带来麻烦甚至损失[2-4]。

数据质量低的一个方面就是数据异常,即数据集中出现明显区别于其他正常数据的数据。由于数据异

收稿日期: 2018-06-27 **修回日期:** 2018-10-30 **网络出版时间:** 2019-03-06
基金项目: 国家自然科学基金(61371196)
作者简介: 宋金玉(1967-),女,硕士,副教授,研究方向为数据工程; 郭一平(1994-),女,硕士,研究方向为数据工程。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20190306.0907.032.html

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

常往往使数据表现为孤立点[5]，也称之为离群点或异常点。这些数据可能是需要消除的错误数据，也可能是重要的报警点，预示出现问题或者发生了重要变化。

离群点检测(又称为异常检测)是找出其行为不同于预期对象的过程，通过检测并去除数据源中的这些孤立点可达到消除数据异常的目的，从而提高数据源的数据质量。

数据挖掘技术中的聚类分析工具，可以用于离群点(噪声)检测。聚类分析采用某种算法将大的数据集合根据数据相似性把所有数据划分成簇，即采用相似度进行归类，相似度较高的归为一类，明显不属于任何一类的单个(或少量)数据集可认为是异常数据。聚类算法有很多种，其中基于密度的方法(density-based method)考虑数据集中的每个对象，根据一定距离内数据密度来划分簇数，数据比较密集的可以被认为一个簇，而比较稀疏的区域则被认为是噪声。

有代表性的基于密度的全局邻域(density-based spatial clustering of applications with noise，DBSCAN)算法将数据空间中的数据抽象为数据点，通过计算点之间距离和点密度来进行聚类，可将噪声或离群点从簇内分离。在 DBSCAN 算法使用中，需设置邻域阈值(Eps)和点数阈值(Minpts)两个参数，根据参数将有一定密度的区域划分为簇，且聚类结果对参数值敏感。

目前已有许多文献对 Eps 和 Minpts 参数值的设定方法进行了研究。对于密度均匀的数据，文献[6]通过分析数据的统计特性来自适应确定 Eps 和 Minpts。对于不同密度数据的聚类，文献[7]采用自适应的 Eps 参数; 文献[8]根据基于网络与基于密度的聚类算法间的等效规则来计算不同密度的密度阈值; 文献[9]提出基于数据分区的 PDBSCAN 算法; 文献[10]提出基于网格分区来确定 Eps 的方法。

文中根据数据的统计特性，利用图表的可视化结果，提出了一种确定 DBSCAN 算法参数的方法。

1　DBSCAN 聚类算法的分析与实现

由于数据集中相似重复记录的个数是不确定的，因此，要求聚类算法应具有能够发现任意形状簇的能力。DBSCAN 算法[11]可将具有足够高密度的区域划分为一个簇，簇数事先是不确定的，点的邻域的形状取决于两点间的距离函数 $\mathrm{dist}(p,q)$，对象间的距离是根据对象的属性值计算得来的。因此，聚类结果取决于选择哪些属性变量、采用何种距离度量以及如何计算度量的属性。

下面介绍用来描述算法的相关概念[12]。

(1) 距离。

对象间的距离采用求数据相异度的方法。假设 X_1，X_2 代表数据集中的两个数据对象，n 是参与计算的数据对象的属性个数，每个数据对象用 n 维向量(X_{i1}，X_{i2}，…，X_{in})表示，X_{1k}，X_{2k} 分别为两个数据点的第 k 维坐标。d_{12} 是两点间距离，有多种形式的距离度量可采用。如欧几里德函数，则 d_{12} 可由如下公式计算:

$$d_{12}=\sqrt{\sum_{k=1}^{n}(x_{1k}-x_{2k})^2} \tag{1}$$

其中，d_{12} 就表示这两个数据对象的相异度。当两个对象越相似或接近时，d_{12} 值越接近 0，而当两个对象越不相同或相距较远时，d_{12} 值越大。还可以根据每个属性的重要性为其赋一个权重。

(2) 邻域、密度、核心点、边界点。

数据集中任意一点的邻域记为 $N_{\mathrm{Eps}}(p)$，是数据集中与 p 点的距离小于给定 Eps 的点的集合。

$$N_{\mathrm{Eps}}(p)=\{q\in D\parallel \mathrm{dist}(p,q)\leqslant \mathrm{Eps}\} \tag{2}$$

邻域中点的个数称为该点的密度，若其大于或等于给定的最小值 MinPts，则称点 p 为核心点，否则称为边界点。

(3) 直接密度可达、密度可达、密度相连。

数据集中任意两点 p，q，如果 $q\in N_{\mathrm{Eps}}(p)$，且 $|N_{\mathrm{Eps}}(p)|\geqslant \mathrm{MinPts}$，则称点 q 是从点 p 关于 Eps 和 MinPts 直接密度可达的。

如果 p，q 两点间存在一个点的序列 p_1，p_2，…，p_n，且 $p_1=p$，$p_n=q$，p_{i+1} 是从 p_i 直接密度可达的，则称点 q 是从点 p 关于 Eps 和 MinPts 密度可达的。

如果存在一个点 o，q 和 p 都是从点 o 关于 Eps 和 MinPts 密度可达的，则称点 q 是从点 p 关于 Eps 和 MinPts 密度相连的。

(4) 簇。

数据集中基于密度的簇是基于密度可达的最大密度相连的点的集合。簇中的任意两点是关于 Eps 和 MinPts 密度相连的。

给定参数 Eps 和 MinPts，DBSCAN 算法的实现就是生成相应的簇。DBSCAN 算法从任意点 p 开始，检索所有从点 p 关于 Eps 和 MinPts 密度可达的点。如果 p 是核心点，就生成一个关于 Eps 和 MinPts 的簇; 如果 p 是边界点，且没有从 p 密度可达的点，算法就去处理数据集中的下一个点。算法实现的流程参见图 1。

相比其他聚类算法，例如基于层次的算法等，DBSCAN 算法的优点是可以发现数据集中任意形状的簇，它的聚类速度比较快，聚类能力也很强。但必须为每个簇指定恰当的 Eps 和 MinPts，及每个簇中的至少一个点。由于很难事先获得数据集中所有簇的相关信息，DBSCAN 算法实现时对所有簇采用相同的全局参数值 Eps 和 MinPts，但把确定参数的任务留给用户，而且算法生成的结果对参数是敏感的。如若根据数据集

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

中存在的比较密集的区域，选取了一个较大的 Minpts 值，那么数据集中其他区域会因为密度不够大而不能被划分成簇，会造成噪声点过多现象；若根据数据集中存在的比较稀疏的区域，选取了一个较小的 Minpts 值，那么整个数据集很容易直接被划成一个大簇，参数值的微小变化往往会导致差异很大的聚类结果。

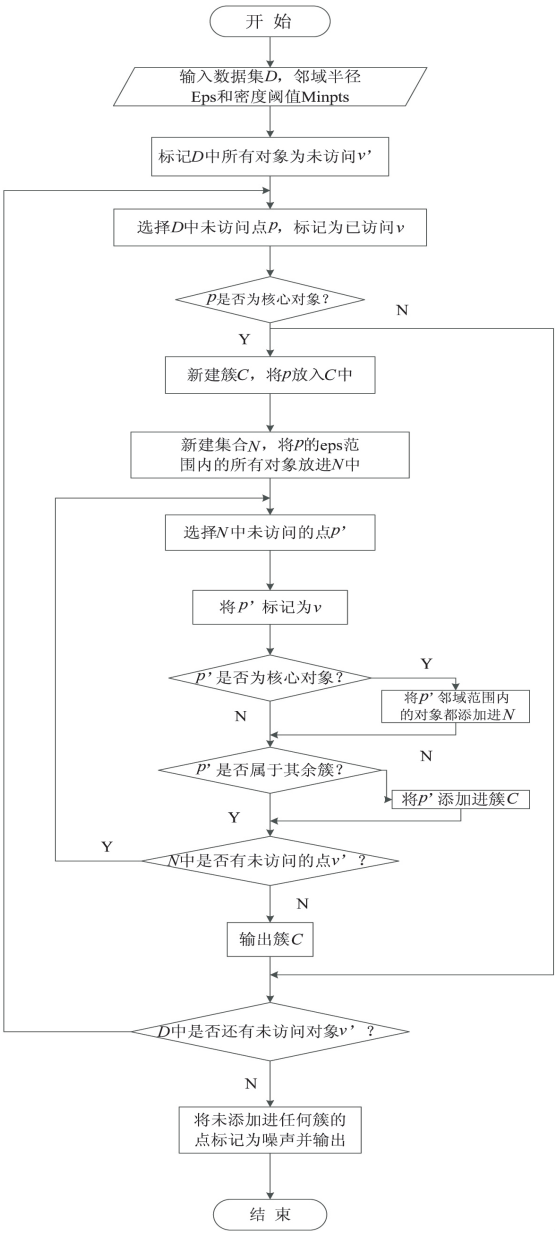

图 1 DBSCAN 算法流程

2 DBSCAN 聚类算法的参数配置

传统 DBSCAN 算法中 Eps 和 Minpts 两个参数是根据经验设置的，并根据聚类结果进行调整。这样做显然盲目性大，工作量也大，而且效果也不一定好。因此，文中提出了一种参数的判断方法，该方法的主要思想是根据数据集本身的统计特性以及图表的可视化结果由人工来选择参数。

2.1 Eps 参数的确定方法

首先按式 1 计算数据对象间的距离，得到距离矩阵 $\mathbf{Dist}_{n\times n}$。

$$\mathbf{Dist}_{n\times n}=\{\mathrm{dist}(i,j),1\leqslant i\leqslant n,1\leqslant j\leqslant n\} \quad (3)$$

其中，n 是数据集 D 中的数据对象个数，每个元素表示对象 i 到对象 j 的距离。

求出矩阵后，将行向量按升序排序。这样，每行就是相应数据点到其他所有点距离的一个排序。则矩阵 $\mathbf{Dist}_{n\times i}$ 中第 i 列的数据的意义是距每个数据点最近的第 i 个距离值的集合。为观察取不同 i 值(如 1,2,…,7)时数值集合的统计特点，绘制图形，其中图 2 是距离值的概率密度分布曲线，图 3 是对第 i 列数据进行升序排序后的曲线。数据采用的是随机生成的含有 150 个数据点的二维数据集 Dataset1。

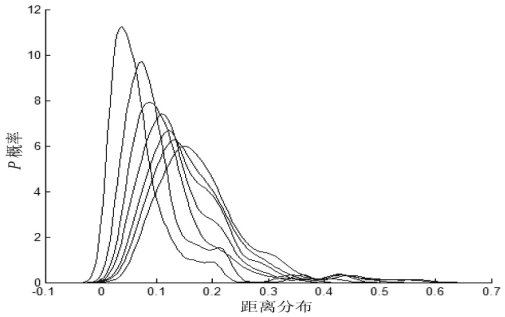

图 2 距离值的概率密度分布曲线

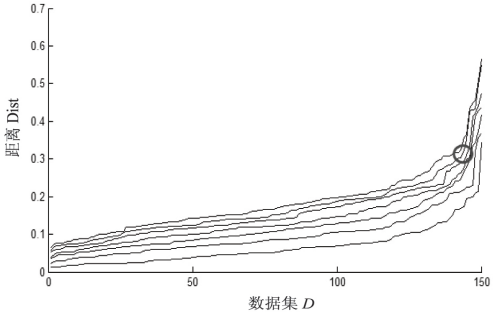

图 3 距每个数据点最近的第 i 个距离值的升序曲线

图 2 中，曲线均值越大的是对应 i 值越大的曲线。可以看到，无论 i 取多少，曲线的分布都大概成泊松分布。曲线右侧，距离比较大的地方(图中接近 0.4)密度已变得非常小，这些比例很小却和其他点相比距离明显偏大很多的点，噪声的可能性较大。因此，可以考虑以此来确定 Eps 参数。

图 3 中，曲线由下至上依次是 i 值增大对应的曲线。曲线的趋势大致相同，前期和中间都比较平缓，末端陡峭。可以发现，当 i 大于 4 以后，曲线的陡峭点都大概集中在一个区域，即图中圆形标注内。在 Martin Ester 等的研究[13]中，对此问题也有描述。若取陡峭点对应的距离作为邻域阈值，可以估计，当 i 大于 4 以后，对噪声的划分情况是近似一样的，也就是聚类和噪声检测结果趋于稳定。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

所以,取 i 等于 4 时,曲线的陡峭点对应的距离值作为 DBSCAN 算法中的 Eps 参数,即图 3 中圆圈标注的纵坐标,大概在 0.3~0.4 之间,与图 2 的分析一致。

2.2　Minpts 参数的确定方法

前面 Eps 的取值是取距每个数据点最近的第 4 个距离值集合升序排序后曲线的陡峭点对应的距离值,即假定 Minpts 为 4。但由于第一步需人工参与判断,很可能出现误差,而且固定的 Minpts 值设定不能保证对任意的数据集检测都有比较好的效果。为了能够更匹配已经确定的 Eps 值,可根据人们实际中对噪声判断的标准,再重新确定 Minpts 值。该思想融合了 Alex Rodriguez 等[14] 提出的新型聚类算法的思想,即对于一般数据集,簇中心被局部密度较高的邻居点所包围,而高局部密度的点之间距离比较大。首先,根据 2.1 中确定 Eps 的方法得到 Eps 的值,然后计算每个数据点 i 的局部密度值 ρ_i,即数据点邻域半径(Eps)内包含的邻居点数,再利用式 4 计算每个数据点 i 距更高密度点的距离 δ_i,对于具有最高密度的点,δ_i 的取值为其到数据集中最远点的距离。

$$\delta_i = \min_{j:\rho_j>\rho_i}(\mathrm{dist}(i,j)) \tag{4}$$

数据集仍采用随机生成的含有 150 个数据点的二维数据集 Dataset1,对每个数据点计算上述两个值,并以点图(如图 4)的形式表现,图中的点就是数据集的每个对象。

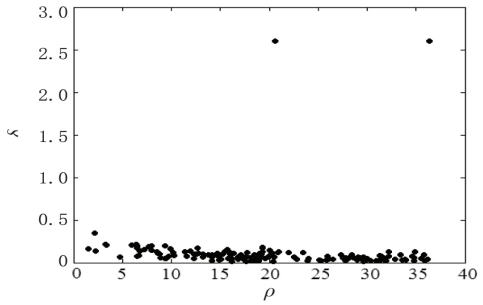

图 4　Dataset1 中每个点 δ_i 与 ρ_i 的函数关系

为了进一步说明该图的意义,又采用随机生成的含有 51 个数据点的二维数据集 Dataset2,计算得到每个点 δ_i 与 ρ_i 的函数关系,如图 5 所示。

在图 5 中,右上角的点 δ_i 比较大并且 ρ_i 也比较高,应该是一个簇的中心,而图中左边的点 ρ_i 非常小、δ_i 又相对比较大的点更多是噪声。根据 δ_i 与 ρ_i 的函数关系点图,在聚类前,就可以得到数据聚类后的一个大概情况。对于数据集 Dataset2,由图 5 可知数据集大概集中在一个区域,并且有少量噪声点,其中,数据对象点 1(图中箭头标注),其 ρ 为 0,并且 δ 非常大,可以肯定是一个噪声点;数据对象点 2 和 3 虽然 ρ 也比较小,但 δ 相对不是很大,所以只是有可能是噪声点,因为 DBSCAN 聚类的簇是密度相连接的点集。若选取 Minpts 为 7(ρ =7),可以预测聚类检测结果,横轴坐标为 7 右侧的点肯定是核心点,不会是噪声,而左边那些比较稀疏的点则有可能存在要寻找的噪声点。在图 4 中,可以判断数据集大概集中在两个区域,也有可能存在噪声。若选取 Minpts 为 2(ρ =2),横坐标为 2 左侧的三个点有可能是要检测的异常噪声点;若选取 Minpts 为 5,横坐标为 5 左侧的六个点都有可能是要检测的异常噪声点。

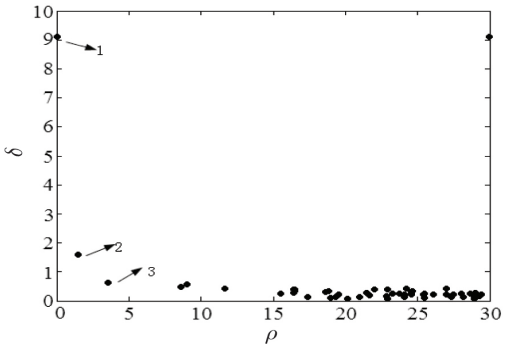

图 5　Dataset2 中每个点 δ_i 与 ρ_i 的函数关系

因此,利用 δ_i 与 ρ_i 的函数关系点图,Minpts 值的设置就转换成了 ρ 阈值的选取问题,可选取图中左边稀疏点和密集点分界处的横坐标(ρ 值)。在实际应用中,若对噪声的要求不是很严格,即偏差度不是很大就可以认为是噪声的话,可选取比较大的 ρ 阈值(Minpts 值);否则,可选取比较小的 ρ 阈值(Minpts 值)。

3　数据异常检测分析

下面在美国加州大学信息与计算机科学系的 Iris(鸢尾花)数据集上,应用 DBSCAN 聚类算法对数据进行异常检测分析,检验所提出的参数配置方法的有效性。

首先对数据集进行改造,添加了两个异常点,即补充了第 151 个数据点(第 152 行),该数据点因花瓣的长度(12)录入有误,造成异常;修改第 70 个数据点(第 71 行),使其 class 属性值由原数据集中的 Iris-versicolor 改为 Iris-virginica。处理后的 Iris(鸢尾花)数据集的部分数据如表 1 所示,数据集有 4 个数值型属性,1 个字符型属性。

DBSCAN 算法通过计算各个数据点的欧氏距离来聚类,要求参与计算的是数值型的属性。但是,在实际中,通常会有字符型属性,若将其舍弃,必然会丧失数据信息的完整性。文中将这类属性分为两类。一类是序数类型属性,比如军衔(少尉,中尉,上尉……)。这类属性的特点是不同属性之间有联系,而且是等距离的。所以,可将其替换为数值 1,2,3……。另一类是标称分类属性的,比如国别(中国,美国,俄罗斯……)。若数据中有多个此类型的属性,通常只选

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

取一个有代表性的参与运算。文中是在数据处理中按标称属性产生概念分层的方法。在聚类前,自动按标称属性分区聚类,不同的分区选取不同的参数设定。这样的处理虽然增加了工作量,但是不仅保证了信息的完整性,而且解决了数据密度不均匀却只有一个全局参数设定的缺陷。

表 1 待检测的 Iris 数据集的部分数据

	Sepal length (萼片长)	Sepal width (萼片宽)	Petal length (花瓣长)	Petal width (花瓣宽)	Class(分析)
2	5.1	8.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
70	6.2	2.2	4.5	1.5	Iris-versicolor
71	5.6	2.5	3.9	1.1	Iris-virginica
72	5.9	3.2	4.8	1.8	Iris-versicolor
150	6.2	3.4	5.4	2.3	Iris-virginica
151	5.9	3	5.1	1.8	Iris-virginica
152	12	3.4	3.5	1.2	Iris-setosa

对 Iris(鸢尾花)数据集的检测按 class 属性值 Iris-setosa,Iris-versicolor,Iris-virginica 进行分区,即分为 3 个区,对每个分区分别进行算法的参数设置。采用前述的参数配置方法,通过可视化的判定,设置 Eps 分别取值 1.5、0.9、1,Minpts 分别取值 12、2、2,从程序输出检测结果可以看到,除了检测出之前添加的两个异常点外,还检测出两个异常点,分别是第 42 个数据点和第 61 个数据点。在原数据集中找到这两个点并进行人工检查,可以发现第 42 个数据点的 sepal width 属性值和其他 Iris-setosa 类鸢尾花相比是最小的,而且偏差比较大,而第 61 个数据点的四个数值属性值和其他 Iris-versicolor 类鸢尾花相比都是比较小的,这也就解释了这两个点被认为是离群噪声点的原因。因此,从数据集的检测结果来看,该方法的参数设置是比较准确的,异常检测的结果也是非常有效的。

为了进一步说明参数设置的准确性,将参数 Eps 统一取值为 1,Minpts 统一取值为 4,得到的异常检测结果输出了 7 个异常点,其中一些异常点并不符合对异常点的判断标准。这说明 DBSCAN 算法对参数是敏感的,当参数设置不合适时,其检测结果也将不准确,也验证了文中算法是有效的。

4 结束语

对数据异常检测问题进行了研究,将基于密度的 DBSCAN 聚类算法用于数据的异常检测,并针对该算法在应用过程中对参数设置敏感的问题,提出了一种配置算法邻域阈值(Eps)和点数阈值(Minpts)的方法。该方法可根据数据集本身的统计特性以及图表的可视化展示,为算法确定合适的参数。编程实现了 DBSCAN 聚类算法及辅助参数确定的计算,并利用 MATLAB 工具进行可视化展现。并在 Iris 数据集上进行检测,通过对比测试,验证了用该方法进行 DBSCAN 聚类算法参数的设置是可行的,弥补了 DBSCAN 聚类算法参数设置单靠经验的传统做法,使得检测结果的准确性和可伸缩性更好。

目前,该 DBSCAN 聚类算法参数配置方法需要人工参与判定,仍存在一定的人为因素,同时参数判定的过程还比较麻烦耗时,这些都有待进一步的改进提高。

参考文献:

[1] LEE M L,LU Hongjun,LING T W,et al.Cleansing data for mining and warehousing [C] //Proceedings of the 10th international conference on database and expert systems applications.Florence,Italy: [s.n.],1999: 751-760.

[2] MCGILVRAY D.数据质量工程实践 [M] .刁兴春,曹建军,张健美,等,译.北京: 电子工业出版社,2010: 245-246.

[3] MADNICK S E,WANG R Y,LEE Y W,et al.Overview and framework for data and information quality research [J] .Journal of Data and Information Quality,2009,1(1): 1-22.

[4] 韩京宇,徐立臻,董逸生.数据质量研究综述 [J] .计算机科学,2008,35(2): 1-5.

[5] HAWKINS D M.Identification of outliers [M] .London: Chapman and Hall,1980.

[6] 夏鲁宁,荆继武.Sa-DBSCAN: 一种自适应基于密度聚类算法 [J] .中国科学院研究生院学报,2009,26(4): 530-538.

[7] 赵 文,夏桂书,苟智坚,等.一种改进的 DBSCAN 算法 [J] .四川师范大学学报: 自然科学版,2013,36(2): 312-316.

[8] 谭 颖,胡瑞飞,殷国富.多密度阈值的 DBSCAN 改进算法 [J] .计算机应用,2008,28(3): 745-748.

[9] 周水庚,周傲英,曹 晶.基于数据分区的 DBSCAN 算法 [J] .计算机研究与发展,2000,37(11): 1153-1159.

[10] 庞 洋,徐巧凤.基于网格分区确定 DBSCAN 参数的方法 [J] .计算机与现代化,2010(5): 16-18.

[11] 刘世平.数据挖掘技术及应用 [M] .北京: 高等教育出版社,2010: 33-40.

[12] 李雄飞,董元方,李 军,等.数据挖掘与知识发现 [M] .北京: 高等教育出版社,2010: 172-189.

[13] ESTER M,KRIEGEL H P,SANDER J,et al.A density-based algorithm for discovering clusters in large spatial databases with noise [C] //Proceedings of 2nd international conference on knowledge discovery and data mining.Portland,Oregon: AAAI Press,1996: 226-231.

[14] LAIO A,RODRIGUEZ A.Clustering by fast search and find of density peaks [J] .Science,2014,344(6191): 1492-1496.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net