第 38 卷第 4 期　　重庆工商大学学报(自然科学版)　　2021 年 8 月
Vol. 38　No. 4　　J Chongqing Technol & Business Univ(Nat Sci Ed)　　Aug. 2021
doi:10. 16055/j. issn. 1672-058X. 2021. 0004. 005

基于局部密度离群点检测 k-means 算法*

刘　凤, 戴家佳**, 胡　阳
(贵州大学 数学与统计学院,贵阳 550025)

摘　要:针对数据集的聚类过程容易受到离群值的影响这一问题,提出了局部密度离群值检测 k-means 算法,即先对数据集使用局部密度离群值检测方法检测离群值,先把离群值去除,再进行 k-means 聚类,算法的有效性通过 Davies-Bouldin 指标(DB)、Dunn 指标和 Silhouette 指标进行评价,在人工生成的数据集与 UCI 数据集上验证,去除离群值,再使用 k-means 算法得到的聚类结果相比原始数据集进行 k-means 算法聚类结果较好,并且用在疫情数据分析上,对安徽省、北京市、福建省、广东省等 24 个省、市、自治区 2020 年 2 月 18 日新型冠状病毒肺炎确诊人数进行聚类分析,得到的去除离群值在使用 k-means 算法相比原始数据集进行 k-means 算法聚类结果较好,该结果能帮助更好地在实际中怎么去做决策以及更好地降低经济损失。

关键词:k-means;离群点;LOF;评价指标
中图分类号:O212. 4　　**文献标志码**:A　　**文章编号**:1672-058X(2021)04-0030-06

0　引　言

聚类是无监督学习中一个重要的方法,它在一定意义上检测数据集中相似的对象,并形成群集的过程,这在大数据时代是很有必要的。聚类的最主要目的是使得群集与群集之间的距离最大化,群集内的对象距离最小化。同一个簇中的对象与该簇中其他的对象是非常相似的,不同群集之间中对象彼此差异是比较大的。MacQueen[1] 提出 k-means 算法,该算法是一种最著名的划分聚类算法,其算法原理在聚类算法中是比较简单的,并且易于聚类。在实际生活中聚类分析也得到比较广泛的应用[2-4]。虽其原理简单,但是该算法也存在一些缺点,最主要有以下几个方面:对球状形数据集聚类结果较好,其他的数据集聚类结果相对较差;在聚类的时候容易局部收敛;k 是随机的,需要在聚类的时候人为给定;聚类算法中初始聚类中心的选取问题;若数据集中有离群点、数据类型不一致等对聚类结果影响比较大。考虑这些不足之处,有大量学者对其不足进行改进,Arthur 和 Vassilvitskii[5] 对其初始聚类中心随机性做改进,首先在数据集中随机选择一个数据点作为第一个初始聚类中心;其次在数据集中选择与第一个聚类中心相对最远的点作为第二个类簇中心,依据这种原理,直到选出 k 个群集质心。然而聚类中心的选择会影响聚类数目的准确性,Alibuhtto 和 Mahat[6] 提出了一种基于距离的 k-means 算法,来确定聚类中的聚类数目。Masud 等[7] 提出了聚类数目与聚类中心的自动选择方法,Zhou 等[8] 提出了基于遗传算法的初始聚类中心的自动选择。大量学者也在算法中对 k 值进行了改进,Qi 等[9] 提出了层次 k-means 算法来确定聚类中的 k 值。程明畅等[10] 提出了基于类簇之间分位数半径的动态 k-means 算法,改进的 k-means 聚类方法取得了良好的聚类效果。在聚类中,数据类型不一致以及离群点等会对聚类结果起到一定的影响,因此聚类中离

收稿日期:2020-08-26;**修回日期**:2020-10-15.
* **基金项目**:贵州省数据驱动建模学习与优化创新团队(黔科合平台人才〔2020〕5016).
作者简介:刘凤(1992—),男,贵州晴隆人,硕士生,从事概率论与数理统计研究.
** **通讯作者**:戴家佳(1976—),女,贵州贵阳人,教授,博士,从事生物与医学统计、生存分析、应用统计、大数据建模与分析、拟合优度检验等研究. Email: jjdai@ gzu. edu. cn.

群点的研究是近年来一个较为热门的课题。Zhang 等[11]提出了一种加权距离测度的高斯函数优化 k-means 聚类算法。Jones 等[12]提出了一种新的 FilterK 算法,通过降低离群值的影响来改善 k-means 聚类的结果,Yu 等[13]将剔掉离群点检测,应用到隐私保护中,Neelima 和 Kumar[14]提出基于离群点检测应用到色调映射,数据集中有离群点对聚类的结果影响都比较大,基于离群点对数据集在聚类中影响比较大。

以上这些学者对 k-means 算法进行改进,使得 k-means 算法在聚类过程中 k 值的随机性得到一定的解决,并提供了类簇中心选择方法,但是对于数据集中的离群值对聚类过程产生较大的影响,笔者对数据集使用局部密度离群点 LOF 检测方法对数据集进行处理,将离群值从数据集中剔除之后再使用 k-means 算法聚类。

1　k-means 算法与局部密度离群点检测方法

1.1　k-means 算法

k-means 算法是一种迭代算法,它是将数据集划分为 k 个预先定义的不重叠的群集。在这种情况下,每个数据点都属于一个群集。将数据集点分配到 k 个群集中,使数据点与群集的质心之间距离的平方之和达到最小。

k-means 算法的 5 个主要步骤如下:

输入:数据集 $D=\{x_1,x_2,\cdots,x_n\}$,类簇数为 k。

输出:k 个类簇数 $C=\{C_1,C_2,\cdots,C_k\}$。

Step 1　从数据集 D 中随机选择 k 个对象作为聚类中的初始中心。

Step 2　计算数据集 D 中的每一个对象到选取的 k 个对象之间的距离。

$$dist(x_i,C_j)=\|x_i-C_j\|^2,i=1,2,\cdots,n;j=1,2,\cdots,k$$

Step 3　根据 Step 2 将数据集 D 中的每一个对象划分到距离最近的类簇中,即满足下式:

$$dist(x_i,C_j)=\min\{dist(x_i,C_j),i=1,2,\cdots,n;j=1,2,\cdots,k\}$$

Step 4　重新更新每个群集的聚类中心,即

$$C_j=\frac{1}{N_j}\sum_{i=1}^{N_j}x_i,j=1,2,\cdots,k$$

其中,N_j 表示为第 j 个类簇中对象个数。

Step 5　反复进行 Step 2、Step 3、Step 4,直到群集质心不再发生变化或者达到预先给定最大迭代次数,算法终止。

1.2　局部密度离群点检测法(Local Outlier Factor,LOF)[15]

先给出几个局部密度离群点检测 LOF 方法的相关定义:$d(X,Y)$ 表示对象 X 与对象 Y 的距离。对象 X 的第 k_距离(k_distance):对给定的正整数 k,在样本空间 Ω 中,它与对象 X 之间的距离 $d(X,\Omega)$。样本空间 Ω 中最远的对象 W 与对象 X 的距离,不包括对象 X,即 $k\_distance(X)=d(X,W)$。对象 X 的第 k_距离邻域($N_k(X)$):

$$N_k(X)=\{Y|d(X,Y)\leqslant k\_distance(X)\}$$

可达距离(Reach Distance):数据集中 X 与 Y 之间的可达距离为

$$reach\_dist(X,Y)=\max\{k\_distance(X),dist(X,W)\}$$

局部可达密度(Local Reach Density,LRD):

$$LRD_k(X)=|N_k(X)|/\sum_{W\in N_k(X)}reach\_dist(X,Y)$$

局部密度离群值因子:

$$LOF_k(X)=\sum_{W\in N_k(X)}\frac{LRD_k(W)}{LRD_k(X)}/|N_k(X)|$$

2　聚类有效性评价指标

评估聚类结果的好坏,聚类有效性指标应能提供一些关于数据集分类质量的评价。最重要的体现在“类簇间相似度比较低,类簇内相似度较高”。聚类有效性评价方法有很多,采用的聚类评价指标有:DB 指标、Dunn 指标和 Silhouette 指标。

2.1　Silhouette 指标($f_{\text{silhouette}}$)[16]

计算数据集中每一个数据点的轮廓系数如:

$$f_{\text{silhouette}}=\frac{d_b(i)-d_w(i)}{\max(d_b(i),d_w(i))}$$

其中,$d_b(i)$ 表示为数据集中点 i 与其他各类簇中点距离平均值中的最小值,$d_w(i)$ 表示为数据集中点 i 与该类簇中其他点距离的平均值。$f_{\text{silhouette}}$ 值在 -1 和 1 之间,若为 1,则数据集中点 i 分配在相应的类簇是恰当的;若为 -1,则数据集中点 i 应该分配在其他的类簇;若为 0,则数据集中点 i 可以分配在该类簇中,也可以分配到其他类簇中。

2.2　Davies-Bouldin 指标(f_{DB})[17]

数据集中 f_{DB} 指标计算如下:

$$f_{\text{DB}}=\frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}$$

其中,σ_i,σ_j 表示对应的类簇内距离的平均值,$d(c_i,c_j)$ 表示质心 c_i,c_j 之间的距离,指标值越小,聚类结果中类簇内部越紧密,类簇间越分离。

2.3 Dunn 指标(f_{Dunn})[18]

数据集中的 Dunn 指标如下:

$$f_{\text{Dunn}}=\frac{\min\limits_{1\leqslant i\leqslant j\leqslant n} d(i,j)}{\max\limits_{1\leqslant k\leqslant n} d(k)}$$

其中,$d(i,j)$表示数据集中任意两个类簇之间的距离,$d(k)$表示数据集中任意类簇内距离。f_{Dunn}值越大,说明数据集的聚类结果较好。

3 实验结果及其分析

3.1 人工随机产生的数据集

人工生成的数据集,第一组为 $mean_{x_1}=3$,$mean_{y_1}=2$,方差为0.25 的正太分布随机数;第二组为 $mean_{x_2}=3$,$mean_{y_2}=8$,方差为 0.25 的正太分布随机数;第三组为 $mean_{x_3}=9$,$mean_{y_3}=5$,方差 0.25 的正太分布随机数;第四组为 $mean_{x_4}=15$,$mean_{y_4}=2$,方差为 0.25 的正太分布随机数;第五组为 $mean_{x_5}=13$,$mean_{y_5}=7$,方差为 0.25 的正太分布随机数,每一组产生 100 个正太分布随机数;用 runif 函数生成均匀分布随机数,最小值为 9,最大值为 20。人工生成的数据集用来验证去除离群点之后的数据集与原始数据集做 k-means 聚类比较,其聚类图以及评价指标如图 1—图 3。

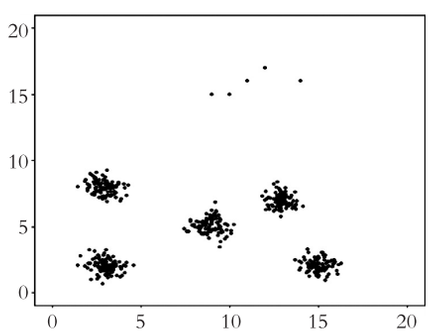

图 1 原始数据集

Fig. 1 The initial data set

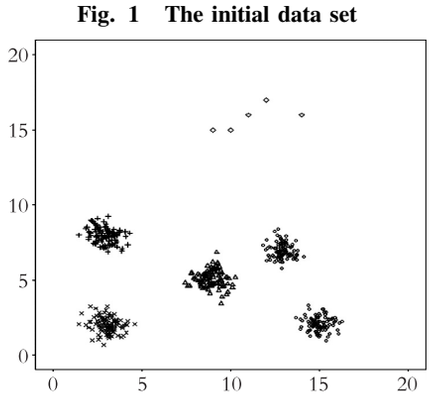

图 2 原始数据集聚类图

Fig. 2 The initial data set clustering diagram

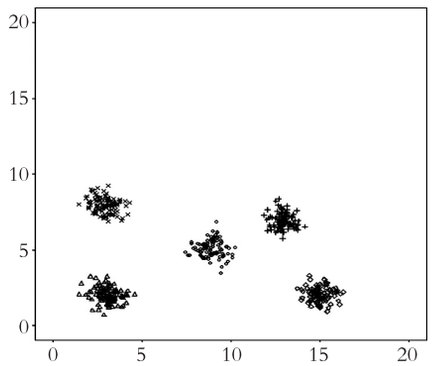

图 3 去除离群值聚类图

Fig. 3 The cluster diagram of removal of outliers

图 2、图 3 具体体现由表 1 给出,表 1 给出了 f_{DB} 指标 f_{Dunn} 指标和 $f_{\text{Silhouette}}$ 指标。原始数据集 k-means 算法的 f_{DB} 指标为 0.442 5,去除离群值 k-means 算法的 f_{DB} 值为 0.281 5,去除离群值 k-means 算法的 f_{DB} 值比原始数据集 k-means 算法的 f_{DB} 值小,f_{DB} 值越小表明类簇内对象之间的距离越小,同时类簇间对象之间的距离越大,说明 f_{DB} 指标的值越小越好,去除离群值 k-means 算法相比原始数据集 k-means 算法较好;f_{Dunn} 指标和 $f_{\text{Silhouette}}$ 指标越大表明类簇对象之间的距离越大,且类簇内对象之间的距离越小,说明这两种指标的值越大越好,原始数据集 k-means 算法的 f_{Duun} 值为 0.414 1,$f_{\text{Silhouette}}$ 值为 0.718 8,去除离群值 k-means 算法的 f_{Duun} 值为 0.606 9,$f_{\text{Silhouette}}$ 值为 0.820 8,去除离群值 k-means 算法相比原始数据集 k-means 算法较好。表 1 证明了人工生成的数据集去除离群值在再用 k-means 得到的聚类结果比直接使用 k-means 聚类算法还要好。

表 1 人工产生的数据集聚类评价指标

Table 1 Evaluation index of manually generated data cluster

评价指标	f_{DB}	$f_{\text{Silhouette}}$	f_{Dunn}
原始数据集 k-means	0.442 5	0.718 8	0.414 1
去除离群点 k-means	0.281 5	0.820 8	0.606 9

3.2 University of California Irvine(UCI)数据集

为了证明所提出的算法有效性,选了 UCI 数据集中的 Wine 和 Seeds 两个数据集来进行证明。表 2 给出了各个数据集的量,表 3 给出了 Wine 数据集使用 k-means 算法以及先使用局部密度离群点 LOF 检测方法剔掉离群点,再使用 k-means 聚类结果的 3 种聚类评价指标,表 4 为 Seeds 数据集使用 k-means 算法以及先使用局部密度离群点 LOF 检测方法剔掉离群点,再使用 k-means 聚类结果的三种聚类评价指标。

表 2　数据集各个量

Table 2　Each quantity of the data set

数据集	样本总数/个	维　数	类簇个数/个
Wine	178	13	3
Seeds	210	7	3

表 3　Wine 数据聚类结果的 3 种评价指标

Table 3　Three evaluation indexes of Wine data clustering results

评价指标	f_{DB}	$f_{Silhouette}$	f_{Dunn}
原始数据集 k-means	1.468 1	0.284 9	0.232 3
去除离群值 k-means	1.196 8	0.335 2	0.336 7

表 4　Seeds 数据聚类结果的 3 种评价指标

Table 4　Three evaluation indicators for the clustering result of Seeds data

评价指标	f_{DB}	$f_{Silhouette}$	f_{Dunn}
原始数据 k-means	0.843 3	0.471 9	0.085 5
去除离群值 k-means	0.702 0	0.527 4	0.077 3

表3 给出了 Wine 数据集的 f_{DB} 指标、f_{Dunn} 指标和 $f_{Silhouette}$ 指标。原始数据集 k-means 算法的 f_{DB} 指标为 1.468 1,去除离群值 k-means 算法的 f_{DB} 值为 1.196 8,去除离群值 k-means 算法的 f_{DB} 值比原始数据集 k-means 算法的 f_{DB} 值小,f_{DB} 值越小表明类簇内对象之间的距离越小,同时类簇间对象之间的距离越大,说明 f_{DB} 指标的值越小越好,去除离群值 k-means 算法相比原始数据集 k-means 算法较好;f_{Dunn} 指标和 $f_{Silhouette}$ 指标越大表明类簇对象之间的距离越大,且类簇内对象之间的距离越小,说明这两种指标的值越大越好,原始数据集 k-means 算法的 f_{Dunn} 值为 0.232 3,$f_{Silhouette}$ 值为 0.284 9,去除离群值 k-means 算法的 f_{Dunn} 值为 0.336 7,$f_{Silhouette}$ 值为 0.335 2,去除离群值 k-means 算法相比原始数据集 k-means 算法较好。;表 4 为 Seeds 数据集聚类结果的 f_{DB} 指标、f_{Dunn} 指标和 $f_{Silhouette}$ 指标。原始数据集 k-means 算法的 f_{DB} 指标为 0.884 3,去除离群值 k-means 算法的 f_{DB} 值为 0.702 0,去除离群值 k-means 算法的 f_{DB} 值比原始数据集 k-means 算法的 f_{DB} 值小,f_{DB} 值越小表明类簇内对象之间的距离越小,同时类簇间对象之间的距离越大,说明 f_{DB} 指标的值越小越好,去除离群值 k-means 算法相比原始数据集 k-means 算法较好;f_{Dunn} 指标和 $f_{Silhouette}$ 指标越大表明类簇对象之间的距离越大,且类簇内对象之间的距离越小,说明这两种指标的值越大越好,原始数据集 k-means 算法的 $f_{Silhouette}$ 值为 0.471 9,去除离群值 k-means 算法的 $f_{Silhouette}$ 值为 0.527 4,去除离群值 k-means 算法相比原始数据集 k-means 算法较好。在 UCI 数据集上验证了去除离群值再使用 k-means 算法得到的聚类结果都有明显的改善,得到的聚类较好。

4　局部密度离群值 k-means 算法在实际中的应用

新型冠状病毒疫情对经济以及人们的生活带来了极大的影响。选取中国部分省、市、自治区数据集(安徽省、北京市、福建省、广东省、广西壮族自治区、江西省、贵州省、河北省、海南省、河南省、吉林省、黑龙江省、湖南省、江苏省、辽宁省、内蒙古自治区、山东省、山西省、陕西省、四川省、天津市、云南省、浙江省、重庆市),这 24 个省、市、自治区新型冠状病毒肺炎疫情数据,数据来源于各个省、市、自治区的卫生健康委员会官网,表 5 为 2020 年 2 月 18 日中国 24 个省、市、自治区新型冠状病毒肺炎的确诊人数。

表 5　2020 年 2 月 18 日中国 24 个省、市、自治区新型冠状病毒肺炎确诊人数

Table 5　The number of confirmed COVID-19 cases in 24 provinces, municipalities and autonomous regions in China on 18 February 2020

简　称	新型冠状病毒肺炎确诊人数								
	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
粤	416	339	98	91	84	66	62	25	23
浙	168	156	504	10	44	42	55	21	10
豫	268	156	155	139	91	76	58	57	52
湘	241	48	78	35	102	155	79	5	59
皖	173	159	155	108	83	69	41	37	33
赣	230	129	123	118	106	76	72	33	22
苏	93	55	79	51	87	40	48	66	27
渝	110	2	3	20	7	28	8	20	15
鲁	47	59	29	24	47	44	52	32	38
川	141	9	16	20	17	22	40	17	22
京	60	58	47	40	12	14	39	18	16
黑	192	52	47	46	43	23	17	15	14
闽	71	35	20	46	14	55	20	6	26
陕	118	13	17	8	15	8	3	26	25
冀	55	48	34	32	31	30	28	23	10
吉	45	5	15	7	5	6	3	2	2
滇	53	25	15	14	13	13	9	8	7
琼	34	54	15	3	6	13	3	9	6
晋	19	12	4	8	10	8	36	7	19
辽	28	19	4	3	7	12	1	8	3
黔	32	36	10	23	10	10	4	17	0
津	53	15	12	6	6	6	4	4	4
桂	54	24	31	5	44	18	8	8	11
蒙	11	11	9	9	8	7	7	7	3

表5数据集比较大,先用scale对数据进行标准化再聚类,使用k-means聚类以及先使用局部密度离群值LOF检测方法去除离群值,再使用k-means聚类的3个评价指标如表6所示。

表6 24个省、市、自治区新型冠状病毒肺炎数据集聚类评价指标

Table 6 COVID-19 data cluster evaluation indicators in 24 provinces, mumicipalities and autonomous regions

评价指标	f_{DB}	$f_{Silhouette}$	f_{Dunn}
原始数据k-means	0.7815	0.5537	0.4321
去除离群值k-means	0.5833	0.5814	0.6652

表6给出了24个省市以及自治区新型冠状病毒肺炎数据集聚类的f_{DB}指标,f_{Dunn}指标和$f_{Silhouette}$指标。原始数据集k-means算法的f_{DB}指标为0.7815,去除离群值k-means算法的f_{DB}值为0.5833,去除离群值k-means算法的f_{DB}值比原始数据集k-means算法的f_{DB}值小,f_{DB}值越小表明类簇内对象之间的距离越小,同时类簇间对象之间的距离越大,说明f_{DB}指标的值越小越好,去除离群值k-means算法相比原始数据集k-means算法较好;f_{Dunn}指标和$f_{Silhouette}$指标越大表明类簇对象之间的距离越大,且类簇内对象之间的距离越小,说明这两种指标的值越大越好,原始数据集k-means算法的f_{Dunn}值为0.4321,$f_{Silhouette}$值为0.5537,去除离群值k-means算法的f_{Dunn}值为0.6652,$f_{Silhouette}$值为0.5814,去除离群值k-means算法相比原始数据集k-means算法较好。证明了使用局部密度离群值检测LOF方法对数据集剔除离群值之后再使用k-means算法得到的聚类结果较好,并且使用局部密度离群值检测LOF方法能更好地帮助分析每个地区的新型冠状病毒肺炎疫情情况,能让每个地区政府更好地处理医用物资的调配以及疫情防疫问题,更好降低经济的损失。

5 结束语

基于局部密度离群点检测k-means算法中,通过f_{DB}指标,f_{Dunn}指标和$f_{Silhouette}$3种评价指标进行聚类结果评价:

(1)在人工产生的数据集中,通过3种评价指标得到去除离群值之后得到的聚类结果较好。

(2)在UCI数据集中的Wine与Seeds数据集进行验证,使用本文选用的3种评价指标进行评价,得到的聚类结果较好。

(3)将离群值检测k-means算法应用到新型冠状病毒肺炎疫情数据分析中,得到较好的聚类结果,更好帮助决策者做决策。

(4)未来将对任意数据集的聚类以及聚类的类簇中心的选择进行研究。

参考文献(References):

[1] MACQUEEN J. Some Methods for Classification and Analysis of Multivariate Observations[C]// Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, 1(14): 281—297

[2] ABADI S, SHUKRI K, THE M, et al. Application Model of k-means Clustering: Insights into Promotion Strategy of Vocational High School[J]. International Journal of Engineering & Technology, 2018, 7(27): 182—187

[3] 王菲菲. k-means聚类算法的改进研究及应用[D]. 兰州:兰州交通大学,2017
WANG F F. Research and Application of the Improved k-means Clustering Algorithm[D]. Lanzhou: Lanzhou Jiaotong University,2017(in Chinese)

[4] NALLAMREDDY S, BEHERA S, KARADADI S, et al. Application of Multiple Random Centroid (MRC) Based k-means Clustering Algorithm in Insurancea Review Article[J]. Operations Research and Applications: An International Journal, 2014, 1(1), 15—21

[5] ARTHUR D, VASSILVITSKII S. K-Means++: The Advantages of Carefull Seeding[J]. The Eighteenth Annual ACM-SIAM Symposiumon Discrete Algorithms, Society for Industrial and Applied Mathematics, 2007, 11(6):1027—1035

[6] ALIBUHTTO M C, MAHAT N I. New Approach for Finding Number of Clusters Using Distance Based k-means Algorithm[J]. International Journal of Engineering, Science and Mathematics, 2019, 8(4): 111—122

[7] MASUD M A, HUANG J Z, WEI C, et al. I-nice: A New Approach for Identifying the Number of Clusters and Initial Cluster Centers[J]. Information Sciences, 2018, 466: 129—151

[8] ZHOU X, MIAO F, MA H. Genetic Algorithm with an Improved Initial Population Technique for Automatic Clustering of Low-dimensional Data[J]. Information, 2018, 9(4): 101

[9] QI J, YU Y, WANG L, et al. An Effective and Efficient Hierarchical k-means Clustering Algorithm[J]. International Journal of Distributed Sensor Networks, 2017, 13(8): 1550147717728627

[10] 程明畅,刘友波,张程嘉,等. 基于分位数半径的动态k-means算法[J]. 南京大学学报(自然科学),2018,54

(1):48—55
CHENG M C, LIU Y B, ZHANG C J, et al. Dynamic k-means Algorithm Based on Quantile Radius[J]. Journal of Nanjing University(Natural Science), 2018,54(1):48—55(in Chinese)

[11] ZHANG T, MA F. Improved Rough K-means Clustering Algorithm Based on Weighted Distance Measure with Gaussian Function[J]. International Journal of Computer Mathematics, 2017, 94(1—4):663—675

[12] JONES P J, JAMES M K, DAVIES M J, et al. Filter K: A New Outlier Detection Method for K-means Clustering of Physical Activity[J]. Journal of Biomedical Informatics, 2020(11): 103397

[13] YU Q, LUO Y, CHEN C, et al. Outlier-eliminated K-means Clustering Algorithm Based on Differential Privacy Preservation[J]. Applied Intelligence, 2016, 45(4): 1179—1191

[14] NEELIMA N, KUMAR Y R. Optimal Clustering Based Outlier Detection and Cluster Center Initialization Algorithm for Effective Tone Mapping[J]. Multimedia Tools and Applications, 2019, 78(22): 31057—31075

[15] SANDER J. LOF: Identifying Density - Based Local Outliers[J]. Acm Sigmod Record, 2000, 29(2):93—104

[16] ROUSSEEUW P J. Silhouettes : A Graphical Aid to the Interpretation and Validation of Cluster Analysis [J]. Journal Computational Applied Mathematics, 1987(20): 53—65

[17] DAVIES D L, BOULDIN D W. A Cluster Separation Measure[J]. IEEE Trans Pattern Anal Mach Intell, 1979, PAMI-1(2):224—227

[18] DUNN J C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well - Separated Clusters [J]. Journal of Cybernetics, 1973, 3(3):32—57

The k-means Algorithm Based on Local Density Outlier Detection

LIU Feng, DAI Jia-jia, HU Yang

(School of Mathematics and Statistics, Guizhou University, Guiyang 550025, China)

Abstract: In view of that the clustering process of data set is easily affected by outliers, the local density outlier detection k-means algorithm is proposed. The proposed method firstly detects the outliers of the data set by using local density outlier detection method, removes the outliers at first and then conducts k-means clustering. The validity of the algorithm is evaluated by Davies-Bouldin index, Dunn index and Silhouette index and is verified by artificial data set and UCI data set, and the outliers are removed. The obtained clustering results by using k-means algorithm are better than original data set k-means algorithm clustering results, this method is used for COVID-19 epidemic data analysis and the clustering analysis of the method is conducted on the confirmed infected number of COVID-19 in 24 provinces, municipalities and autonomous regions such as Anhui, Beijing, Fujian, Guangdong and so on on February 18, 2020. The clustering results using k-means algorithm by removing outliers are better than the clustering results of original data set using k-means algorithm, and the results can be conducive to how to make decision in practical work and better reduce economic cost.

Key words: k-means; outliers; LOF; evaluation index

责任编辑:罗姗姗

引用本文/Cite this paper:

刘凤, 戴家佳, 胡阳. 基于局部密度离群点检测 k-means 算法[J]. 重庆工商大学学报(自然科学版),2021,38(4):30—35

LIU F, DAI J J, HU Y. The k-means Algorithm Based on Local Density Outlier Detection[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2021,38(4):30—35