doi:10.3969/j.issn.1672-1292.2022.01.011

基于外点检测的加权 k-means 算法

胡豪杰[1],陈辉[2],穆婷婷[3],姚敏立[1],何　芳[1],张峰干[1]

(1.火箭军工程大学,陕西 西安 710025)
(2.中国航天科技集团有限公司第四研究院,陕西 西安 710025)
(3.北京新时代环球进出口有限公司,北京 100027)

[摘要]　为解决 k-means 聚类算法中异常样本点破坏数据分布,致使簇中心发生较大偏差的问题,通过计算样本点与潜在簇中心的距离赋予样本点不同的权重,降低外点对数据分布的影响,并通过对权重向量施加 ℓ_0-norm 范数在聚类模型中自适应移除外点. 采用交替最小化优化算法求解模型,在人工合成数据集和真实数据集上的实验表明,所提模型能有效降低外点对聚类的影响,可得到更有效的聚类效果.

[关键词]　聚类,k-means,外点检测,ℓ_0-norm
[中图分类号]TP391　[文献标志码]A　[文章编号]1672-1292(2022)01-0075-06

Weighted k-means Algorithm Based on Outlier Detection

Hu Haojie[1],Chen Hui[2],Mu Tingting[3],Yao Minli[1],He Fang[1],Zhang Fenggan[1]

(1.Rocket Force Engineering University,Xi'an 710025,China)
(2.The Fourth Academy of China Aerospace Science and Technology Corporation,Xi'an 710025,China)
(3.Beijing New Era Global Import and Export Co.,Ltd.,Beijing 100027,China)

Abstract:In this paper,to solve the problem of that few outliers can easily destroy the cluster structure,leading to a significant deviation for the obtained centroids in k-means clustering algorithm,we assign different weights on the data points based on their distance from the potential cluster center to alleviate the negative impact on the data structure. Moreover,we also incorporate outlier detection in our clustering model by imposing ℓ_0-norm constraint on weight assignments. To optimize the model,we introduce an efficient alternating minimization algorithm. Extensive experiments on both synthetic and real datasets show the effectiveness of the proposed model.

Key words:clustering,k-means,outlier detection,ℓ_0-norm

聚类是将物理或抽象对象的集合划分成多个类簇,使得簇内样本相似度较高,而簇间样本的相似度较低[1-2]. 作为一种基本的数据分析方法,聚类在数据挖掘、模式识别、信息检索等领域得到了广泛应用[3]. 在众多聚类算法中,k-means 算法以其简单高效的优点成为最具代表性的聚类方法之一. 然而众多实验结果表明,k-means 对外点或离群点很敏感,这限制了聚类效果的进一步提升. 针对此问题,通常有两种方法来处理,一是使用诸如 ℓ_1-norm 之类的特定范数来削弱外点的影响[4-5];另一种方法是将外点检测引入聚类算法中,直接消除外点[6-8],这种方法更为直接有效.

近年来,已有许多基于外点检测的聚类算法被提出. Hautamäki 等[9]提出了 ORC(outlier removal clustering)算法,该算法包括两个阶段,第一阶段即原始的 k-means 聚类过程,第二阶段是迭代移除离簇中心较远的样本点. Ahmed 等[10]提出了一种改进的结合外点检测与聚类的方法,该算法首先将每个样本点分配到距其最近的簇,然后计算 SSE(sum of squared error)/SST(total sum of squares)以检测异常点,若检测到任何异常值,则从数据集中删除,重复该过程直至算法收敛. Whang 等[11]提出了一种非穷尽的重叠 k-means 算法,该算法假设数据点可能不属于任一簇,或者属于多个簇. Gan 等[12]提出了 KMOR(k-means with outlier removal)算法,该算法引入了一个额外的簇,该簇包含所有外点. 根据 Holoentropy 在外点检测

收稿日期:2021-08-31.
通讯作者:何芳,博士,讲师,研究方向:机器学习. E-mail:fanghe1107@gmail.com

中的应用,Liu 等[13]提出了 COR(clustering with outlier removal)算法,在实现外点检测的同时达到了良好的聚类效果.

以上方法虽然都在一定程度上对算法的聚类结果进行了优化,但都遵循一个理想的假设,即每个样本点严格属于某一簇[14],否则即是外点. 而在但实际数据集中,外点的分布是模糊的. 为使聚类效果更加理想,本文计算数据点与最近簇中心的欧式距离并分配不同的权重,以此计算加权后的簇中心,并采用交替最小化优化算法求解该模型,其中权重为零的样本点被认为是外点. 实验结果表明,本文算法提高了外点检测的精确度,优化了聚类效果.

1　*k*-means 算法

k-means 算法是一种经典的聚类算法. 给定数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}\in\mathbf{R}^{d\times n}$,*k*-means 的基本思路是将这 n 个样本点划分成 k 个簇,使得每一簇内的样本点具有较高的相似度,簇与簇间具有较低的相似度,相似度通过样本间的欧式距离来测算. *k*-means 的目标函数可表示为:

$$\min_{\boldsymbol{Y},\boldsymbol{C}}\sum_{i=1}^{n}\sum_{j=1}^{k}y_{ij}\|\boldsymbol{x}_i-\boldsymbol{c}_j\|_2^2, \tag{1}$$

式中,$\boldsymbol{c}_j=\dfrac{\sum_{i=1}^{n}\boldsymbol{x}_i y_{ij}}{\sum_{i=1}^{n}y_{ij}}$. $\boldsymbol{Y}=\{y_{ij}\}_{n\times k}$为聚类指示矩阵,若 $\boldsymbol{x}_i$ 属于第 j 簇,则 $y_{ij}=1$,反之 $y_{ij}=0$. *k*-means 算法的优化步骤如下:

(1)从数据集中随机选取 k 个样本点作为初始簇中心;

(2)计算每个样本点到各个簇中心的欧式距离,并将其赋给最近的簇;

(3)计算每个簇的平均值,作为新的簇中心;

(4)不断重复步骤 2 与步骤 3,直至聚类结果不再发生变化.

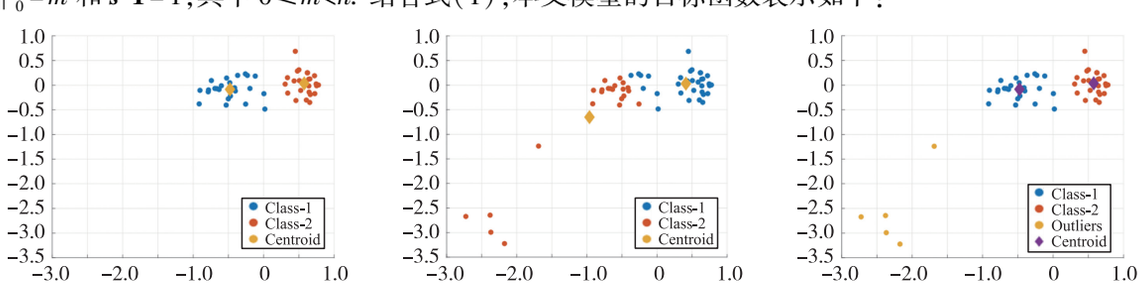
2　基于外点检测的加权 *k*-means 算法

本节提出一种改进的基于外点检测的加权 *k*-means 算法,并采用迭代优化算法来解决该模型.

2.1　模型介绍

外点或离群点是指明显偏离数据集中其他样本对象的样本点. *k*-means 算法中基于欧式距离的损失函数对异常值极为敏感,具体来说,若 $\boldsymbol{x}_i$ 是一个外点,会使簇中心的位置严重偏离其真实位置,导致较差的聚类性能. 图 1(a)为正确的无外点的数据分布. 如图 1(b)所示,*k*-means 算法将 6 个样本点错误归类,无法正确恢复图 1(a)所示的正确数据分布. 如图 1(c)所示,将聚类与外点检测相结合,可以清楚地区分不同的簇.

为消除外点的影响,本文引入权重向量 $\boldsymbol{s}\in\mathbf{R}^{n\times 1}$,使得距簇中心越远的样本点具有越小的权重,并满足 $\|\boldsymbol{s}\|_0=m$ 和 $\boldsymbol{s}^{\mathrm{T}}\mathbf{1}=1$,其中 $0\leqslant m<n$. 结合式(1),本文模型的目标函数表示如下:



(a) 无外点的数据分布　(b) 在带外点的数据分布上*k*-means的聚类结果　(c) 基于外点检测的聚类结果

图 1　聚类结果示意图

Fig. 1　Schematic diagram of clustering results

$$\begin{aligned}&\min \sum_{i=1}^{n} s_i \sum_{j=1}^{k} \| \boldsymbol{x}_i-\boldsymbol{c}_j \|_2^2 y_{ij}+\gamma \| \boldsymbol{s} \|_2^2.\\&\text{s.t. } \boldsymbol{Y}\in \boldsymbol{Ind},\ \| \boldsymbol{s} \|_0=m,\boldsymbol{s}^{\mathrm{T}}\boldsymbol{1}=1,\boldsymbol{0}\leqslant \boldsymbol{s}\leqslant \boldsymbol{1},\boldsymbol{C}.\end{aligned} \tag{2}$$

式中,$\boldsymbol{Ind}\in \mathbf{R}^{n\times k}$ 为聚类指示矩阵,表示矩阵每一行只有 1 个元素为 1,其余为 0. 权重向量中 $\boldsymbol{s}$ 中有 $n-m$ 个零元素,对损失没有贡献,而相应的样本点被视为外点. 式(2)中的第二项是正则化项,其中 γ 是正则化参数. 注意到若 γ 取值足够大,则式(2)等价于 k-means 算法.

2.2 优化求解

式(2)中有三个变量($\boldsymbol{Y}$,$\boldsymbol{C}$ 和 $\boldsymbol{s}$)需要优化,本文采用迭代优化算法求解该问题.

固定 $\boldsymbol{C}$ 和 $\boldsymbol{s}$,求解 $\boldsymbol{Y}$,则问题(2)转化为求解

$$\min_{\boldsymbol{Y}\in \boldsymbol{Ind}} \sum_{i=1}^{n} \sum_{j=1}^{k} s_i \| \boldsymbol{x}_i-\boldsymbol{c}_j \|_2^2 y_{ij}. \tag{3}$$

类似于求解式(1),通过将每个样本点分配给距其最近的簇中心来更新 $\boldsymbol{Y}$. 也即,对于每一数据点 $\boldsymbol{x}_i$,找到使 $s_i \| \boldsymbol{x}_i-\boldsymbol{c}_j \|_2^2$ 最小的 j,令 $y_{ij}=1$.

固定 $\boldsymbol{s}$ 和 $\boldsymbol{Y}$,求解 $\boldsymbol{C}$,则计算式(2)对 $\boldsymbol{c}_j$ 的导数并将其设置为零,可得

$$\sum_{i=1}^{n} \sum_{j=1}^{k} s_i(\boldsymbol{x}_i-\boldsymbol{c}_j)y_{ij}=0\Rightarrow \boldsymbol{c}_j=\frac{\sum_{i=1}^{n} s_i\boldsymbol{x}_i y_{ij}}{\sum_{i=1}^{n} s_i y_{ij}}. \tag{4}$$

固定 $\boldsymbol{C}$ 和 $\boldsymbol{Y}$,求解 $\boldsymbol{s}$,则问题(2)等效于求解

$$\begin{aligned}&\min \sum_{i=1}^{n} s_i \sum_{j=1}^{k} \| \boldsymbol{x}_i-\boldsymbol{c}_j \|_2^2 y_{ij}+\gamma \boldsymbol{s}^{\mathrm{T}}\boldsymbol{s}.\\&\text{s.t. } \boldsymbol{s}^{\mathrm{T}}\boldsymbol{1}=1,\boldsymbol{0}\leqslant \boldsymbol{s}\leqslant \boldsymbol{1},\ \| \boldsymbol{s} \|_0=m.\end{aligned} \tag{5}$$

令 $d_i=\sum_{j=1}^{k} \| \boldsymbol{x}_i-\boldsymbol{c}_j \|_2^2 y_{ij}$,组成向量 $\boldsymbol{d}\in \mathbf{R}^{n\times 1}$,则式(5)可以向量形式简化表示为

$$\min_{\boldsymbol{s}^{\mathrm{T}}\boldsymbol{1}=1,\boldsymbol{0}\leqslant \boldsymbol{s}\leqslant \boldsymbol{1},\ \| \boldsymbol{s} \|_0=m} \left\| \boldsymbol{s}+\frac{1}{2\gamma}\boldsymbol{d} \right\|_2^2. \tag{6}$$

首先仅考虑权重向量 $\boldsymbol{s}$ 的前两个约束条件,上述问题的拉格朗日函数为

$$\boldsymbol{L}(\boldsymbol{s},\theta,\boldsymbol{\zeta})=\frac{1}{2}\left\| \boldsymbol{s}+\frac{\boldsymbol{d}}{2\gamma} \right\|^2-\theta\left(\sum_{i=1}^{n} s_i-1\right)-\boldsymbol{\zeta}\cdot \boldsymbol{s}, \tag{7}$$

式中,θ 与 $\boldsymbol{\zeta}\geqslant 0$ 是拉格朗日乘数. 根据 KKT 条件,有

$$s_i=\left(-\frac{\boldsymbol{d}}{2\gamma}+\theta\right)_+. \tag{8}$$

不失一般性,假设 $d_1,d_2,\cdots,d_n$ 从大到小排序. 根据约束 $\| \boldsymbol{s} \|_0=m$,则 $s_m>0,s_{m+1}=0$. 由此可得

$$\begin{cases}-\dfrac{d_m}{2\gamma}+\theta>0,\\-\dfrac{d_{m+1}}{2\gamma}+\theta\leqslant 0.\end{cases} \tag{9}$$

根据约束 $\boldsymbol{s}^{\mathrm{T}}\boldsymbol{1}=1$,有

$$\sum_{i=1}^{m}\left(-\frac{d_i}{2\gamma}+\theta\right)=1\Rightarrow \theta=\frac{1}{2m\gamma}\sum_{i=1}^{m} d_i+\frac{1}{m}. \tag{10}$$

结合式(9)和式(10),参数 γ 满足以下不等式

$$\frac{m}{2}d_m-\frac{1}{2}\sum_{j=1}^{m} d_j<\gamma\leqslant \frac{m}{2}d_{m+1}-\frac{1}{2}\sum_{j=1}^{m} d_j. \tag{11}$$

因此,为了获得满足所有约束条件的权重向量 $\boldsymbol{s}$,可将 γ 设置为

$$\gamma=\frac{m}{2}d_{m+1}-\frac{1}{2}\sum_{j=1}^{m} d_j. \tag{12}$$

根据式(8)、(10)和(12),$\boldsymbol{s}$ 可计算如下

$$\hat{s}_i=\begin{cases}\dfrac{d_{m+1}-d_i}{md_{m+1}-\sum\limits_{j=1}^{m}d_j}, & i\leqslant m;\\ 0, & i>m.\end{cases}\tag{13}$$

权重向量 $\boldsymbol{s}$ 中零元素对应的样本点被视为外点,而向量 $\boldsymbol{s}$ 是不断更新迭代的,外点亦会不断更新直至算法收敛,于是算法自适应地移除了离群点. 本文算法的具体流程如下所示:

算法 1　基于外点检测的加权 k-means 算法.

输入:数据集 $\boldsymbol{X}\in\mathbf{R}^{d\times n}$,类簇数 k,有效样本点数 m;

输出:聚类指示矩阵 $\boldsymbol{Y}$.

Step 1　通过从数据集 $\boldsymbol{X}$ 中随机选择 k 个样本点作为初始化的簇中心 $\{\boldsymbol{c}_j\}_{j=1}^{k}$,并令 $\boldsymbol{s}=\mathbf{1}$;

Step 2　Repeat;

Step 3　通过求解公式(3)更新聚类指示矩阵;

Step 4　通过公式公式(4)更新簇中心 $\{\boldsymbol{c}_j\}_{j=1}^{k}$;

Step 5　通过公式公式(13)更新权重向量 $\boldsymbol{s}$;

Step 6　Until convergence.

3　实验与结果

为检验算法的有效性,将本文算法与 k-mean,NEO-k-means 及 COR 进行比较. 实验环境为:Intel(R) Core(TM)i7-10700K CPU,内存 32GB,操作系统 Microsoft Windows 10,算法编写环境为 MatlabR2020b.

3.1　评价指标

本文采用 4 个评估指标来衡量所提算法在聚类有效性和外点检测两方面的性能,其中 NMI 与 R_n 是两种常用的聚类评价指标,分别定义为

$$\mathrm{NMI}=\frac{\sum_{i,j}n_{ij}\log\dfrac{n\cdot n_{ij}}{n_{i+}\cdot n_{+j}}}{\sqrt{\left(\sum_i n_{i+}\log\dfrac{n_{i+}}{n}\right)\left(\sum_j n_{j+}\log\dfrac{n_{+j}}{n}\right)}},\tag{14}$$

$$R_n=\frac{\sum_{i,j}\binom{n_{ij}}{2}-\sum_i\binom{n_{i+}}{2}\cdot\sum_j\dfrac{\binom{n_{+j}}{2}}{\binom{n}{2}}}{\sum_i\dfrac{\binom{n_{i+}}{2}}{2}+\sum_j\dfrac{\binom{n_{+j}}{2}}{2}-\sum_i\binom{n_{i+}}{2}\cdot\sum_j\dfrac{\binom{n_{+j}}{2}}{\binom{n}{2}}},\tag{15}$$

式中,n_{ij} 表示既属于聚类算法结果的第 i 簇也属于真实类别的第 j 簇的样本数;n_{i+} 和 n_{+j} 分别是聚类算法得到的第 i 簇的样本数以及真实类别第 j 簇的样本数.

为了衡量外点检测的性能,采用 Jaccard 指数和 F-measure 两种评价方法,分别定义为:

$$\mathrm{Jaccard}=\frac{|O\cap O^*|}{|O\cup O^*|},\tag{16}$$

$$F\text{-measure}=2*\frac{precision\cdot recall}{precision+recall},\tag{17}$$

式中,O 和 O^* 分别是算法得到的外点集合以及真实的外点集合;F-measure 是外点准确率和召回率的调和平均值.

3.2 数据集

为测试本文算法的有效性,选择在一组合成数据及多组取自 UCI 中的数据集进行实验. 合成数据集是由高斯分布函数在二维空间中随机生成的包含 90 个样本点的 9 个主要类簇,每个类簇包含 10 个样本点,数据集添加了 6 个具有较大偏差的外点,分布在(1.5,1.5),(2.5,2.5),(3,2.5),(1.5,2.5),(1,2.5),(2.5,1.5)的位置. 4 组真实数据集为 UCI 中的 ecoli 数据集、glass 数据集、yeast 数据集和 zoo 数据集,将每个数据集中最小类的样本点视为外点. 表 1 描述了每个真实数据集的具体信息.

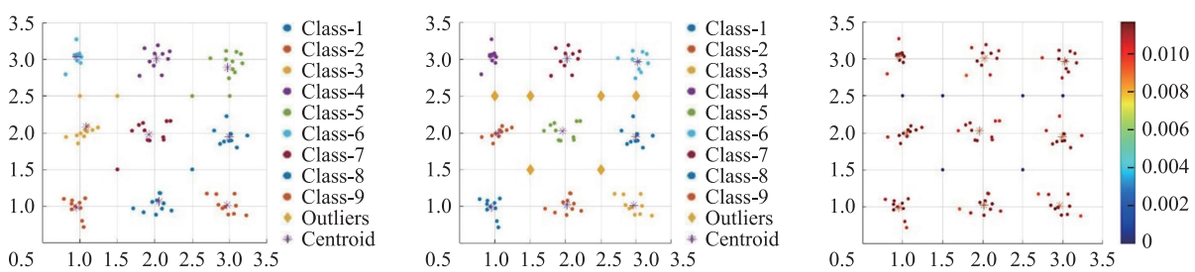
表 1 数据集描述

Table 1 Characteristics of datasets

数据集	样本数	维度	类簇数	外点数
ecoli	336	343	7	2
glass	214	9	4	9
yeast	1 484	1 470	8	5
zoo	101	16	5	4

3.3 实验和结果

实验首先对本文方法与 k-means 算法在合成数据集上的输出进行分析,然后在实际数据集上定量地比较所提出的方法和几种竞争方法. 在合成数据集上得到的聚类结果如图 2 所示. 图 2(a)和图 2(b)分别显示了 k-means 算法与本文算法得到的聚类效果,其中每种颜色表示算法得到的不同类簇,外点由黄色菱形表示,簇中心用星号显示. 可以看出,本文算法可完全区分 9 个类簇和 6 个外点,得到聚类中心也更接近真实的聚类中心. 图 2(c)显示利用本文算法得到的所有数据点的权重分配,可明显看出真实聚类中心区域附近的数据点分配了较大的权重,而离实聚类中心较远的数据点分配了较小的权重,因此可以学习到更精确的簇中心,从而实现了更好的聚类性能.



(a) k-mean聚类结果　(b) 本文算法聚类结果　(c) 本文算法权重分布

图 2 合成数据上的聚类效果

Fig. 2 Performance on the toy dataset

表 2 和表 3 所示为在 4 组数据集上不同算法的实验结果. 由于所比较的算法易受初始化的影响,因此将每个算法随机初始化运行 100 次,并记录平均结果.

表 2 不同算法在 4 组真实数据集上 NMI 与 R_n 取值

Table 2 NMI and R_n of different approaches on four real-world datasets

数据集	NMI				R_n			
	k-means	NEO-k-means	COR	Proposed	k-means	NEO-k-means	COR	Proposed
ecoli	37.1±5.7	42.2±2.8	40.6±5.5	42.4±3.1	17.4±9.8	21.1±6.9	21.2±8.3	22.4±7.8
glass	22.9±5.3	21.6±4.4	26.3±7.8	30.3±4.3	11.3±4.5	9.2±3.2	16.3±7.2	11.1±4.2
yeast	17.0±4.6	15.9±4.9	8.9±1.5	18.0±4.5	8.2±4.4	5.1±4.1	1.8±1.9	8.3±5.3
zoo	68.1±8.2	64.3±4.4	64.6±9.1	72.3±7.1	54.9±14.8	54.5±9.0	59.1±11.8	65.1±13.0

表 3 不同算法在 4 组真实数据集上 Jaccard 与 F-measure 取值

Table 3 Jaccard and F-measure of different approaches on four real-world datasets

数据集	NMI				R_n			
	k-means	NEO-k-means	COR	Proposed	k-means	NEO-k-means	COR	Proposed
ecoli	2.1±5.5	26.3±13.6	4.6±11.6	29.0±11.2	3.7±9.4	39.5±20.4	7.0±17.4	43.5±16.9
glass	0.0±0.0	8.6±2.1	0.0±0.0	32.0±46.8	0.0±0.0	15.7±3.7	0.0±0.0	32.0±46.8
yeast	12.6±9.1	1.2±0.3	1.8±9.4	23.9±40.1	3.9±13.5	2.3±0.6	2.4±12.3	25.6±40.8
zoo	1.0±3.1	1.3±2.3	0.0±0.0	5.0±21.9	0.0±0.0	2.6±4.3	0.0±0.0	5.0±21.9

从表 2 可以看出,本文所提出的方法在 NMI 指标中均达到了最好效果,在 glass 和 zoo 数据集中,比第

二好的方法实现了超过4%的增益. 在另一指标 R_n 中,虽然本文方法在 glass 数据上并未达到最优,但在其他 3 个数据上实现了最好效果,这是由于本文算法根据数据点与最近簇中心的欧式距离分配不同的权重获得了更为精确的簇中心,因此实现了更好的聚类效果. 表 3 通过 Jaccard 和 F-measure 两个指标定量体现各个算法在不同数据集上外点的检测性能. 本文所提出的算法在所有数据集上的检测性能均明显优于 k-means,NEO-k-means 和 COR. 在数据集 glass 与 yeast 上,指标 Jaccard 和 F-measure 均超过对比方法的 20%. 结果显示,本文算法不仅优化了聚类效果,还提高了外点检测的精确度.

4　结论

本文提出了一种改进的 k 均值模型,该模型可同时实现聚类和外点检测. 为了提高模型的鲁棒性,根据数据点与潜在聚类中心的距离,对数据点分配不同的权重. 为此,通过对权重向量施加 ℓ_0 范数在聚类模型中自适应移除外点. 实验结果证明了该模型在聚类有效性和外点检测方面的优越性.

[参考文献](References)

[1] JAIN A K,DUBES R C. Algorithms for clustering data[M]. New Jersey,USA:Prentice-Hall,1988:227-229.

[2] 吉珊珊. 基于神经网络树和人工蜂群优化的数据聚类[J].南京师大学报(自然科学版),2021,44(1):119-127.

[3] CAMPOS R,DIAS G,JORGE A M,et al. Survey of temporal information retrieval and related applications[J]. ACM Computing Surveys,2014,47(2):1-41.

[4] CAI X,NIE F P,HUANG H. Multi-view k-means clustering on big data[C]//Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Beijing:IJCAI,2013:2598-2604.

[5] NIE F P,HUANG H,CAI X,et al. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization[C]//Proceedings of the 23rd International Conference on Neural Information Processing Systems. Vancouver,Canada,2010.

[6] HUANG S,REN Y,XU Z. Robust multi-view data clustering with multi-view capped-norm K-means[J]. Neurocomputing,2018,311:197-208.

[7] 袁小翠,刘宝玲,马永力. 基于空间邻域连通区域标记法的点云离群点检测[J].计算机应用研究,2020,37(增刊 2):380-382,385.

[8] BEER A,LAUTERBACH J,SEIDL T. MORe++:k-means based outlier removal on high-dimensional data[C]//Proceedings of the 12th International Conference on Similarity Search and Applications. Newark,USA:Springer,2019:188-202.

[9] HAUTAMÄKI V,CHEREDNICHE-NKO S,KÄRKKÄINEN I,et al. Improving K-means by outlier removal[C]//Proceedings of the 14th Scandinavian Conference on Image Analysis. Joensuu,Finland:Springer-Verlag,2005:978-987.

[10] AHMED M,NASER A. A novel approach for outlier detection and clustering improvement[C]//Proceedings of the 2013 IEEE 8th Conference on Industrial Electronics & Applications. Melbourne,Australia:IEEE,2013.

[11] WHANG J J,DHILLON I S,GLEICH D F. Non-exhaustive,overlapping k-means[M]//Proceedings of the 2015 SIAM International Conference on Data Mining. Vancouver,Canada:SIAM,2015.

[12] GAN G J,NG M K P. k-means clustering with outlier removal[J]. Pattern Recognition Letters,2017,90:8-14.

[13] LIU H F,LI J,WU Y,et al. Clustering with outlier removal[J]. IEEE Transactions on Knowledge and Data Engineering,2019. DOI:10.1109/TKDE.2019.2954317.

[14] 许振,吉根林,唐梦梦.基于聚类的兴趣区域间异常轨迹并行检测算法[J].南京师大学报(自然科学版),2019,42(1):59-64.

[责任编辑:严海琳]