文章编号：2095-2163(2021)05-0205-04 中图分类号：TP391.41 文献标志码：A

K-means++算法优化及其在地震前兆分析中的应用研究

苏保玉，李 忠，朱 婷，张 伟
（防灾科技学院 应急管理学院，河北 燕郊 065201）

摘 要：K-means++算法是近年来发展起来的一种聚类分析方法，解决了经典 K-means 算法的初始聚类中心随意确定的问题，在一定程度上提高了收敛速度。本文针对 K 值难以确定的问题，采用拐肘法、轮廓系数法和 CH 指标法联合确定 K 值，从而优化了 K-means++算法，并用于地震地磁数据聚类分析中。计算结果表明，优化后的 K-means++聚类算法，能够较好地发现离群点，并与发生的地震对应，明显优于经典 K-means 算法，对于地震监测预报工作具有重要的现实意义。

关键词：K-means 聚类算法；拐肘法；轮廓系数法；CH 指标法

Optimization of K-means + + algorithm and its application in the seismic-geomagnetic analysis

SU Caiyu, LI Zhong, ZHU Ting, ZHANG Wei
(School of Emergency Management, Institute of Disaster Prevention, Yanjiao Hebei 065201, China)

【Abstract】 K-means++ algorithm is a kind of clustering analysis method developed in recent years. It solves the problem that the initial clustering center of classical K-means algorithm is determined at random, and improves the convergence speed. In this paper, the elbow method, contour coefficient method and CH index method are used to determine the K value, which optimizes the K-means++ algorithm and is applied to the clustering analysis of seismic and geomagnetic data. The calculation results show that the optimized K-means++ clustering algorithm can find outliers and correspond to the earthquakes, which is obviously better than the classical k-means algorithm, which has important practical significance for earthquake monitoring and prediction.

【Key words】 K-means cluster algorithm; elbow method; silhouette coefficient method; CH index method

0 引 言

K-means++算法是 2007 年由 David Arthur 和 Sergei Vassilvitskii 提出的，是 K-means 算法的改进版本。传统 K-means 算法的聚类效果以及运行时间在很大程度上受到初始聚类中心选择的影响，K-means++算法对此进行了改进。虽然 K-means++算法在初始化簇中心时，增加了计算量，但在整个聚类过程中，能够显著地提升计算效率和改善聚类结果误差[1]，被广泛应用于分类[2]、聚类[3-4]等领域。

但是，由于 K-means++依旧存在难以确定合适的 K 值问题，可能导致聚类结果会产生局部最优解。基于此，本文针对 K 值难于确定的问题，采用 3 种方法联合确定最佳的聚类中心个数，从而改进 K-means++算法，并将优化算法应用于地震地磁数据聚类分析中。

1 K-means++聚类算法

经典的 K-means 算法是无监督的聚类方法[5]，具有简单、高效、易于实现等特点。局部搜索能力较强，且对于大数据样本空间的聚类有较高的效率[6]，在数据聚类分析领域应用广泛[7]。但是，K-means 也存在 2 个致命缺点：一是 K 值多以使用者的经验来确定，因此存在很大的主观性，在不同应用场景使用时造成很大困扰；二是初始值采用随机选取方式，可能造成算法收敛速度较慢、计算效率较低的问题。

基于上述问题，有学者提出了 K-means++算法。在初始值选取时，采用样本点与中心点的距离作为概率值，距离越远则被选中的概率越大，距离越近则概率越小，从而有效解决了上述第二个问题。

K-means++聚类算法实现步骤如下：

(1)从样本集 X 中随机选择一个初始聚类中心点，加入到聚类中心集 C 中；

基金项目：廊坊市科技支撑计划项目(2017013157)；大学生创业训练项目(202011775154)。

作者简介：苏保玉(2000-)，女，本科生，主要研究方向：数据挖掘；李 忠(1966-)，男，博士，教授，主要研究方向：人工智能、大数据技术。
通讯作者：李 忠 Email：lizhong@ cidp.edu.cn

收稿日期：2021-02-02

(2)遍历样本集,计算每一个样本点 x_i 与已存在聚类中心点 c_k 的距离 $dis(x_i,c_k)$,选取最短距离 $D(x_i)$,如式(1)所示:

$$D(x_i)=\min\{dis(x_i,c_k)\}, \tag{1}$$

依次计算每个样本点的概率,如式(2)所示:

$$P=\frac{D(x_i)^2}{\sum_{x\in X}D(x_i)^2}. \tag{2}$$

其中,P 为一个样本点 x_i 被选中为聚类中心的概率;$D(x_i)$ 代表一个样本点 x_i 到已有聚类中心的最短距离;$\sum_{x\in X}D(x_i)^2$ 代表所有样本点到每个已存在聚类中心的最小距离的平方和。选择概率最大的样本点作为下一个聚类中心点,直至找到 K 个初始聚类中心;

(3)以各个簇内样本点的中心点更新聚类中心,得到 K 个新聚类中心;

(4)计算各样本点到 K 个中心点的距离,按照最短距离归属原则归类,形成 K 个新类;

(5)重复步骤(3)、(4)直到聚类中心点不变或者达到迭代次数结束计算,并记录最终的 K 个聚类中心点和簇内数据。

2 K-means++算法改进

从上述介绍可见,K-means++算法解决了 K-means 算法的初始中心选择问题,但对 K 的选择依然没有改变思路。为解决这个问题,本文采用拐肘法(Elbow Method)、轮廓系数法(Silhouette coefficient)和 Calinski-Harabaz 指标法(CH)联合确定 K 值。

2.1 拐肘法

拐肘法的核心思想,是通过计算各个 K 值时的畸变程度,画出畸变程度与 K 值的关系图,找出拐点。畸变程度就是每个簇的质点与簇内样本点的均方差。一个簇畸变程度的高低,表明簇内成员的松散程度高低。随着 K 值的增大,到达拐点后,畸变程度会得到很大改善,下降幅度趋于平缓,此时这个拐点就可以考虑为聚类性能较好的点,可以确定其对应 K 值作为聚类个数,如式(3)所示:

$$SSE=\sum_{i=1}^{k}\sum_{p\in c_i}|p-m_i|^2. \tag{3}$$

其中,SSE 是畸变程度;c_i 代表第 i 个簇;p 是 c_i 中的样本点;m_i 是 c_i 的质心,即 c_i 中所有样本的均值。

利用拐肘法获得的 K 值记作 K_1。

2.2 轮廓系数法

轮廓系数(Silhouette Coefficient)是由 Kaufman 等人提出的一种用来评价算法聚类质量的有效性指标。该指标结合了凝聚度和分离度,不仅用以评价聚类质量,还可用来获取最佳聚类数[8]。假设样本集 X 包含 n 个样本点 $x_1,x_2,\ldots,x_n$,将样本集分为 K 个簇,每个样本点的轮廓系数如式(4)所示:

$$S_i=\frac{b_i-a_i}{\max\{a_i,b_i\}}. \tag{4}$$

其中,S_i 为第 i 样本点的轮廓系数;a_i 是样本点 i 到其所属簇中所有其它点的平均距离;b_i 为样本点 i 到其它簇中所有点的最小距离。S_i 介于 [-1,1],接近-1 则说明样本点 i 更应该分类到另外的簇,接近 0 则说明样本 i 在 2 个簇的边界附近,越趋近于 1 则聚类效果越优。

所有点的轮廓系数求平均,得到最终的平均轮廓系数。对于现有的分类数,求取轮廓系数的最大值 S_k,与之对应的 K 值就是最佳聚类数[9]。

利用轮廓系数法获得的 K 值记作 K_2。

2.3 CH 指标法

CH 指标由分离度与紧密度的比值得到[10]。假设数据集被划分为 K 个类,CH 指标计算如式(5)所示:

$$CH(K)=\frac{trB(K)/(K-1)}{trW(K)/(N-K)}. \tag{5}$$

其中,N 为数据集总样本数;K 为类别数;tr 表示对矩阵求迹,即矩阵主对角线元素之和;$B(K)$ 为类别之间的协方差矩阵;$W(K)$ 为类别内部数据的协方差矩阵。

CH 越大代表类内越密集,类间越分散,聚类结果性能越好,对应的 K 值也是最优的聚类数。利用 CH 指标法获得的 K 值记作 K_3。

2.4 联合法确定 K 值

从上述 3 种求 K 值的算式和求取过程可以看出:拐肘法实际上计算聚类的最小距离,理论上距离越小越好,但聚类个数太多就失去意义了。因此,K 值是在下降率突然变缓时刻的值,认为此时聚类效果最佳。轮廓系数法强调的是簇内部凝聚度最大、簇间分离度最大时聚类效果最佳,符合客观实际,其值域在[-1, 1],越靠近 1 说明对应的 K 值越佳。CH 指标法实质上要求类别内部数据的协方差越小越好,类别之间的协方差越大越好,类似于轮廓系数法,而且指标值越大,对应 K 值的聚类效果越佳。

鉴于 3 种方法从不同的角度求取 K 值,因此本

文将 3 种方法获得的 K 值综合求取平均值，即：

$$K = (K_1 + K_2 + K_3)/3. \quad (6)$$

依据四舍五入原则确定最终的 K 值，这样求取的 K 值既兼顾了簇内集中簇间分散的要求，又兼顾了距离最小原则，达到最优聚类结果。

3　地震地磁前兆数据分析

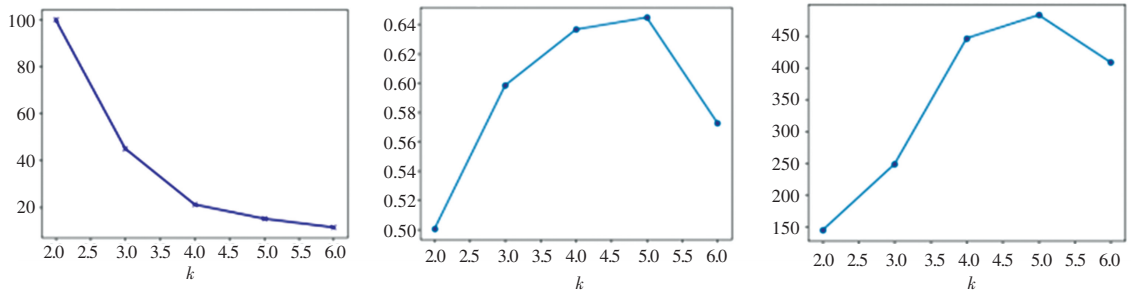
聚类分析方法在地震监测、地震裂缝分布预测[11]等地震数据分析领域应用较多，能够从大量的地震数据中获取规律性的知识，避免了不同学者的主观性问题。地震地磁前兆数据，是地震台站通过地磁仪长期观测到的时间序列数据，单位为奥斯特。根据地震学家的研究结果表明，在地震前 1 个月内，震中一定范围内会出现地磁异常变化。因此可以通过长期的观测并通过数据挖掘方法搜索离群点，从而发现地震发生前期地磁变化，为地震预测带来可能。

3.1　数据来源及预处理

本次实验数据选取 2008 年 1～6 月份四川省成都市的地震地磁前兆数据。由于地磁前兆数据存在数据缺失、非标准化等问题，需要进行插值、整理、规范化处理等预处理操作，以保证数据的可用性。缺失数据采用均值法补齐，规范化采用 z-score 标准化方法进行完成，并去掉量纲。经过处理的数据符合标准正态分布，按照时间排列构成标准化的数据集。

3.2　计算结果与分析

首先以拐肘法、轮廓系数法和 CH 指标法分别计算 K_1、K_2 和 K_3 值，如图 1 所示。



（a）拐肘法曲线　（b）轮廓系数曲线　（c）CH 指标曲线

（a）Elbow Method curve　（b）Silhouette coefficient curve　（c）Calinski-Harabaz curve

图 1　不同算法求得 K 值折线图

Fig. 1　Line graph of K-value by different algorithms

根据聚类性能标准评估原理，从图 1(a)的拐肘法得到 K_1 值为 4，从图 1(b)的轮廓系数法得到 K_2 值为 5，从图 1(c)的 CH 指标法得到 K_3 值为 5。

其次求取平均值作为最优 K 值，$K = (4 + 5 + 5)/3 = 4.7 \approx 5$。因此 $K = 5$ 是最佳的聚类类别个数，此时可以达到最好的聚类效果，计算结果如图 2 所示。

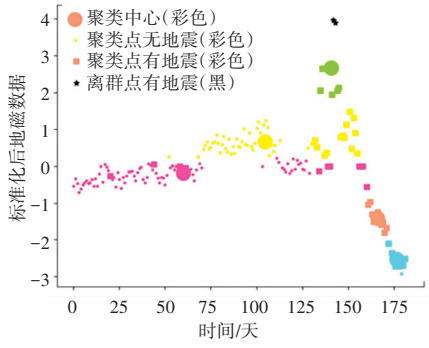


图 2　K-means++的聚类结果

Fig. 2　Clustering results of K-means + + algorithm

由图 2 可见，存在 5 个聚类中心。即图中大圆点，检测到 2 个异常点，即图中五角星，正好与 2 次地震对应。详尽信息见表 1。

作为对比，同样以 $K = 5$ 作为聚类数，利用传统 K-means 算法对成都地震地磁前兆数据进行聚类分析，结果如图 3 所示。

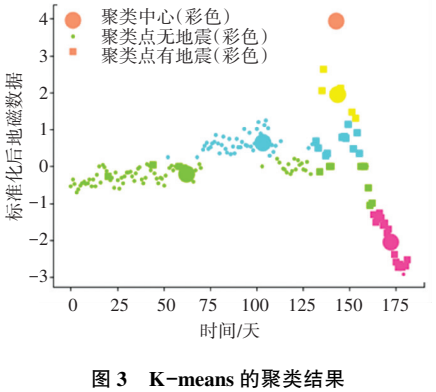


图 3　K-means 的聚类结果

Fig. 3　Clustering results of K-means algorithm

表 1 两次地震情况

Tab. 1 Details of two earthquakes

日期	时间	纬度	经度	震源深度	里氏震级	位置
20080520	00:57:37.6	31.62	104.15	15km	4.4	四川省绵阳市安州区高川乡 619 乡道
20080521	13:59:7.5	31.36	104.01	13km	4.5	四川省德阳市什邡市红白镇园家坪

从图 3 可知,传统 K-means 聚类方法没有找到异常离群点。实际上,在实验过程中,K-means 聚类方法因初始聚类中心选择的随机性问题,该方法难以发现离群点,并且每次计算结果差异较大。这也说明,同样的 K 值下,K-means++聚类方法比 K-means 方法效果更好。

4 结束语

文中在介绍了 K-means++算法基础上,通过拐肘法、轮廓系数法和 CH 指标法联合求取 K 值,优化了 K 值的确定方法,从而很好地解决了 K-means++聚类算法中 K 值难以确定的问题。利用改进后的 K-means++聚类算法对 2008 年 1~6 月份四川省成都市的地震地磁前兆数据进行聚类分析,发现了 2 个离群点,正好与发生的 2 次地震相对应,效果明显优于传统 K-means 结果。得到结论如下:

优化后的 K-means++聚类算法优于传统 K-means 算法,K-means++聚类算法可以较好地解决地震地磁的聚类问题。通过离群点检索可以发现可能发生的地震,对地震监测预报工作具有重要的现实意义。

参考文献

[1] 王振杰,刘杨范,赵爽,等. K-Means++的声速剖面精简方法[J]. 哈尔滨工程大学学报,2020,41(7):985-990.

[2] 杨纲,寇健,严思唯,等. 基于改进 kmeans++算法的用户分类与电价政策影响分析[J]. 电力需求侧管理,2020,22(3):57-62.

[3] 尹林子,关羽吟,蒋朝辉,等. 基于 k-means++的高炉铁水硅含量数据优选方法[J]. 化工学报,2020,71(8):3661-3670.

[4] 卞永明,高飞,李梦如,等. 结合 Kmeans++聚类和颜色几何特征的火焰检测方法[J]. 中国工程机械学报,2020,18(1):1-6.

[5] 张一帆,胡佳浩,李依桥. 基于 k-means 算法实现商品的聚类研究[J]. 数字技术与应用,2020,38(4):108,110.

[6] 段勇强,廖红华,郑才,等. 基于改进 Kmeans 算法的富硒绿茶嫩芽识别[J]. 湖北民族学院学报(自然科学版),2019,37(4):445-448,462.

[7] 柏宇轩. Kmeans 应用与特征选择[J]. 电子技术与软件工程,2018,1:186-187.

[8] 王治和,王淑艳,杜辉. 基于密度敏感距离的改进模糊 C 均值聚类算法[J/OL]. 计算机工程:1-12[2020-07-24]. https://doi.org/10.19678/j.issn.1000-3428.0057901.

[9] ESTIRI H, OMRAN B A, MURPHY S N. kluster: An Efficient Scalable Procedure for Approximating the Number of Clusters in Unsupervised Learning [J]. Big Data Research, 2018,13: 38-51.

[10] 赵谦益. K-means 算法中文文献聚类的 Python 实现[J]. 软件,2019,40(8):89-94.

[11] 龚屹,桂志先,王鹏,等. 基于地震纹理属性聚类分析的裂缝分布预测[J]. 科学技术与工程,2017,17(30): 167-174.

(上接第 204 页)

图 10 地图导航界面

Fig. 10 Map navigation interface

4 结束语

针对城市交通堵塞问题,本文利用大数据、无线通信、定位导航等技术设计出一套客户端软件,其功能包括登录注册、车辆管理、数据管理块与地图导航。实验证明这款软件对缓解交通堵塞问题能够起到一定的作用,但是就目前而言,本文的设计也存在着不足之处,需要后续研究人员进一步开发。

参考文献

[1] 朱惠. 无线通信技术在车联网中的应用实践探微[J]. 数字通信世界,2018, 162(6):202.

[2] 崔志斌. 面向智能网联汽车的云数据平台的设计与实现[D]. 西安:电子科技大学,2020.

[3] 郭振. 基于车联网的车辆信息采集系统的设计与研究[D]. 西安:长安大学, 2015.

[4] 陈植钦,杨云海,陈婵燕,等. 基于 ThinkPHP 的租车商城系统的设计与实现[J]. 现代信息科技, 2019, 3(1): 1-4,10.