Hans Journal of Data Mining 数据挖掘, 2024, 14(3), 143-148
Published Online July 2024 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2024.143013

Hans汉斯

K-Means聚类算法中确定k值的改进方法

李自刚，刘叶青，赵致远，吴佳雪，周　达，秦　玥

河南科技大学数学与统计学院，河南　洛阳

收稿日期：2024年6月3日；录用日期：2024年7月3日；发布日期：2024年7月11日

摘　要

针对传统k-means聚类算法过于依赖聚类数k的问题，本文提出了确定最佳聚类数k的一种新方法——双均值法。该算法不依赖于预先设定的k值，而是通过计算簇内平均距离与簇间平均距离的比值来动态确定最优的k值。该方法的创新之处在于，它结合了簇内的紧密度和簇间的分离度，从而更加精确地反映了数据的真实结构。通过在多个公共数据集上求得的k值与数据的真实类别数比较，或手肘法求得的k值相比较，说明新方法有效。

关键词

数据挖掘，聚类分析，K-Means算法，手肘法

Improvement Methods for Determining the Value of k in the K-Means Clustering Algorithm

Zigang Li, Yeqing Liu, Zhiyuan Zhao, Jiaxue Wu, Da Zhou, Yue Qin

School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang Henan

Received: Jun. 3rd, 2024; accepted: Jul. 3rd, 2024; published: Jul. 11th, 2024

Abstract

The issue of traditional k-means clustering algorithm relying too heavily on the number of clusters, k. A new method for determining the optimal number of clusters, k, has been proposed—the double mean method. This algorithm does not rely on a pre-defined k value, but rather calculates the ratio of intra-cluster average distance and inter-cluster average distance to dynamically determine the optimal k value. The innovation of this method lies in the fact that it combines intra-cluster density and inter-cluster separation, thus more accurately reflecting the true structure

文章引用: 李自刚, 刘叶青, 赵致远, 吴佳雪, 周达, 秦玥. K-Means 聚类算法中确定 k 值的改进方法[J]. 数据挖掘, 2024, 14(3): 143-148. DOI: 10.12677/hjdm.2024.143013

of the data. By comparing the k value obtained on multiple public datasets with the true number of classes in the data or with the k value obtained using the elbow method, the effectiveness of the new method is demonstrated.

Keywords

Data Mining, Cluster Analysis, K-Means Algorithm, Elbow Method

1. 引言

聚类分析是一个非监督的研究方式，用以发掘数据中潜在的特征，通过测量数据点之间的相似性并把它们分类为不同的组合。K-means 算法是最经典也是应用最广泛的聚类算法[1]。将数据集中的观测点分为不同的群组或簇。此方法采用欧氏距离函数来评估数据对象间的相似性，并以误差平方和作为评价标准。通过计算得到的结果可以确定聚类数目和最佳划分数以及最优初始中心。K-means 算法具有识别分布相对均衡的球状集群的能力。其优点为简单、高效、易于实现。同时 k-means 聚类算法还具有若干局限性，如对初始聚类中心的选择十分敏感、在应用此算法前需先确定 k 值等[2]。

针对上述的缺点，来自不同领域的学者纷纷提出了相应的改进算法。王森等人在文献[3]中，对当前的改进算法进行了基于密度和距离的分类总结，这些算法主要包括初始聚类中心点的选取、k 值的确定以及对离群点的处理。另外，他们还对各种改进方法的优势以及存在的弊端，作出了深刻的剖析。李波等人在文献[4]中，根据经典的 K-means 方法，在数据分布不均时的特殊情况下，即随意选取初始聚类中心点会造成的结果不平衡，给出了一个新方法。这个算法通过计算数据样本的空间密度来确定初始聚类中心。具体而言，它挑选了簇内误差平方和最小的点作为首个初级聚类中心，同时也剔除了那些以这个样本点为核心、平均距离为半径的区域。接着，对剩余的数据集重复之前的步骤，继续寻找初始的聚类中心，直到找到 K 个聚类中心点。冯波等人在文献[5]中，为解决传统 K-means 算法在初始聚类中心上的敏感性问题，提出了一种优化策略。该方案基于数据样本的分布情况，动态地选取初始聚类中心。具体来说，算法首先通过计算数据点之间的距离，构造出一个最小生成树。接着，通过修剪这棵最小的生成树，从而产出了 k 组初始的数据。黄苏雨[6]提出了确定聚类数的 EE 法，即将簇内距离和与簇间距离两个值的比值作为评价聚类效果的指标。根据定义，这个比值容易出现波动，不易确定聚类数。本文在此基础上进行改进[6]。

本文基于 k-means 聚类算法，提出确定聚类数 k 值的改进方法[7]。为了验证新方法的效果，和常使用的手肘法在同一数据集上来求最优的 k 值，发现求出的 k 值相同，说明改进方法是有效的。

2. 双均值法确定最优聚类数

2.1. 确定最优聚类数 k 值

文献[6]中，第 i 个簇内的平均距离

$$E_i = \sum_{x \in D_i} \frac{1}{n_i} \sum_{i=1}^{k} \left\| x - M_i \right\|^2 ,$$

簇间的距离

$$E_0 = \frac{1}{k}\sum_{i=1}^{k}\left\|M_i - M\right\|^2$$

其中 D_i 为第 i 个簇，n_i 为第 i 个簇中样本个数，k 表示簇的个数，M_i 为第 i 个簇的质心，M 为所有数据的中心。计算 E_i 与 E_0 的比值 $\frac{E_i}{E_0}$，当 $\frac{E_i}{E_0}$ 取局部最小即下降最快时，求出最优聚类数。在应用时发现，$\frac{E_i}{E_0}$ 的值有时不单调，容易出现反复。为了处理这种情况将两个公式分别改进。

首先，计算簇内平均距离 d。簇内平均距离定义为每个簇中数据与该簇质心之间的距离的总和的平均值，其计算公式为：

$$d = \frac{1}{n}\sum_{x\in D_i}\sum_{i=1}^{k}\left\|x - M_i\right\|^2$$

其中 n 为样本的总个数。d 越小，说明簇内的数据越紧凑，聚类效果越好。

接下来，计算簇间平均距离 d_0。簇间平均距离定义为每个簇的质心之间距离的平均值，d_0 的计算公式为：

$$d_0 = \frac{2}{k(k-1)}\sum_{i=1}^{k}\sum_{j=1}^{k}\left\|M_i - M_j\right\|^2 (i < j)$$

簇之间的距离 d_0 越大，类间的相似度就越小，聚类效果越好[8]。

最后，计算簇内平均距离与簇间平均距离的比值，并将此值命名为 dd 值。其表达式为：

$$dd = \frac{d}{d_0}$$

这个值可以作为评估聚类效果的标准。当簇内样本紧密而簇间距离较远时，聚类效果通常较好，此时 dd 值较小。反之，如果簇内样本分散或簇间距离过近，聚类效果较差，此时 dd 值会增大。dd 值越大，聚类效果越差。显然当 dd 的值取局部最小值时，聚类效果最好，此时的 k 值即为最优聚类数，这种确定 k 值的方法称为双均值法。

2.2. 改进的算法描述

基于优化初始聚类中心的 K-means 算法可以描述如下：

1) 初始化过程：设置 k 的初始值为 1，并计算此时的 EE 值(误差平方和或能量函数值)。

2) 逐步增加 k 值：将 k 增加 1，并重新计算 EE 值，将结果记录下来。

3) 确定最优 k 值：随着 k 的逐渐增大，并在每个 k 值下运行传统的 K-means 算法，并记录每次运行得到的 EE 值。选择使得 EE 值局部最小的 k 值作为最优聚类数，并确定最终的 k 值。

4) 选择初始聚类中心：从数据集中随机选择第一个数据点作为第一个聚类中心。

5) 重复选择聚类中心：继续从数据集中随机选择数据点，直到选择出 k 个初始聚类中心。

6) 计算距离：计算数据集中每个样本点到最近聚类中心的距离，记为 D(x)。

7) 更新聚类中心：根据每个点 D(x)的值，我们为各个类别挑选出新的聚类中心。选择的依据是概率，这个概率与 D(x)的值成正比，也就是说，距离越近的点被选为新聚类中心的可能性越大。

8) 重复更新过程：不断重复步骤(7)，直到所有的 k 个聚类中心都被选中并更新。

9) 运行 K-means 算法：使用确定的最优 k 值和更新后的聚类中心，运行传统的 K-means 算法进行聚类。

通过上述步骤，我们可以实现基于优化初始聚类中心的 K-means 算法，以改进传统算法的效率和准确性。

3. 实验结果及分析

本文的实验环境为 Intel CPU、16 GB 内存、500 GB 硬盘、Windows11 操作系统和 R 语言软件。为了检测我们的方法相对于先前的方法所呈现的改进，我们用双均值法、文献[4]中的 EE 法以及常用的手肘法进行比较。

手肘法是一种简单但有效的技术[9]，用于确定 k-means 聚类中的最佳簇数。它在实际应用中被广泛使用，尤其是在没有明确先验知识的情况下。因此通过与手肘法的实验结果进行对比来验证双均值法是有效性。

手肘法核心思想：随着聚类数 k 的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么簇内平方和自然会逐渐变小[10]。当 k 值小于真实聚类数时，由于 k 的增大会大幅增加每个簇的聚合程度，故簇内平方和的下降程度会很大；而当 k 值到达真实聚类数时，再增加 k 所得到的聚合程度回报会迅速变小，所以簇内平方和的下降幅度会骤减，然后随着 k 值的继续增大而趋于平缓，也就是说簇内距离之和与 k 的关系图是一个手肘的形状，而这个肘部对应的 k 值就是数据的真实聚类数。

本文采用的 4 个数据集均为公共数据集，分别为 swiss、USArrests、iris、wine.公共数据集 Swiss 是一个经典的数据集，常用于数据挖掘和机器学习算法的教学和研究。它包含了关于瑞士的一些统计数据，其中每条数据包括了 7 个变量：Fertility (生育率)、Agriculture (农业)、Examination (考试)、Education (教育)、Catholic (天主教徒比例)、Infant Mortality (婴儿死亡率)、和 Region (地区)。其优点数据集的变量较少，有助于分析结果的解释合可视化；公共数据集 USArrests 是 R 语言中的一个经典数据集，包含了美国 48 个州和 Washington D.C.的犯罪率数据，其中包含了四个变量：Assault (袭击罪)、Murder (谋杀罪)、UrbanPop (城市人口比例)和 Rape (强奸罪)；公共数据集 iris 是机器学习和统计学中经典的数据集之一，它包含了三种不同品种的鸢尾花(setosa、versicolor 和 virginica)各 50 个样本的数据，其中包括了花萼长度(sepal length)、花萼宽度(sepal width)、花瓣长度(petal length)和花瓣宽度(petal width)四个特征；公共数据集 wine 是另一个经典的机器学习数据集，它包含了三个不同来源的葡萄酒(类别分别为 1、2、3)，每个类别有 59 个样本，共 178 个样本，每个样本有 13 个特征，包括了葡萄酒的化学特征。以上公共数据集均有着数据结构简单，同时包含了多个特征合多个类型，并且被广泛使用合研究，有利于结果的真实性和可信度，故我们采取四个公共数据集同时进行研究，形成对照，其中 iris 数据集和 wine 数据集已知是三类的，swiss 数据集和 USArrests 数据集没有明确的类别划分。

实验结果如图 1 所示。三种方法的 k 值都是在 y 轴的值下降最快处取得。由 EE 法、双均值法、手肘法得出这四个数据集的最优聚类数均为 3。

本文又将双均值法与 EE 法在不同数据集上求 k 值的时间进行了比较，我们定义了时间差的计算公式如下：

$$\frac{EE-dd}{EE}\times 100\%$$

两种方法在四个数据集上的运行时间差如图 2 所示。

由图 2 可知，*dd-EE* 运行时间差在 5%~25%之间，因此双均值法的时间性能要优于 EE 法，说明双均值算法是更高效的。

4. 结论

本文基于 k-means 算法，通过计算簇内平均距离与簇间平均距离的比值来确定最佳聚类数 k，与原方

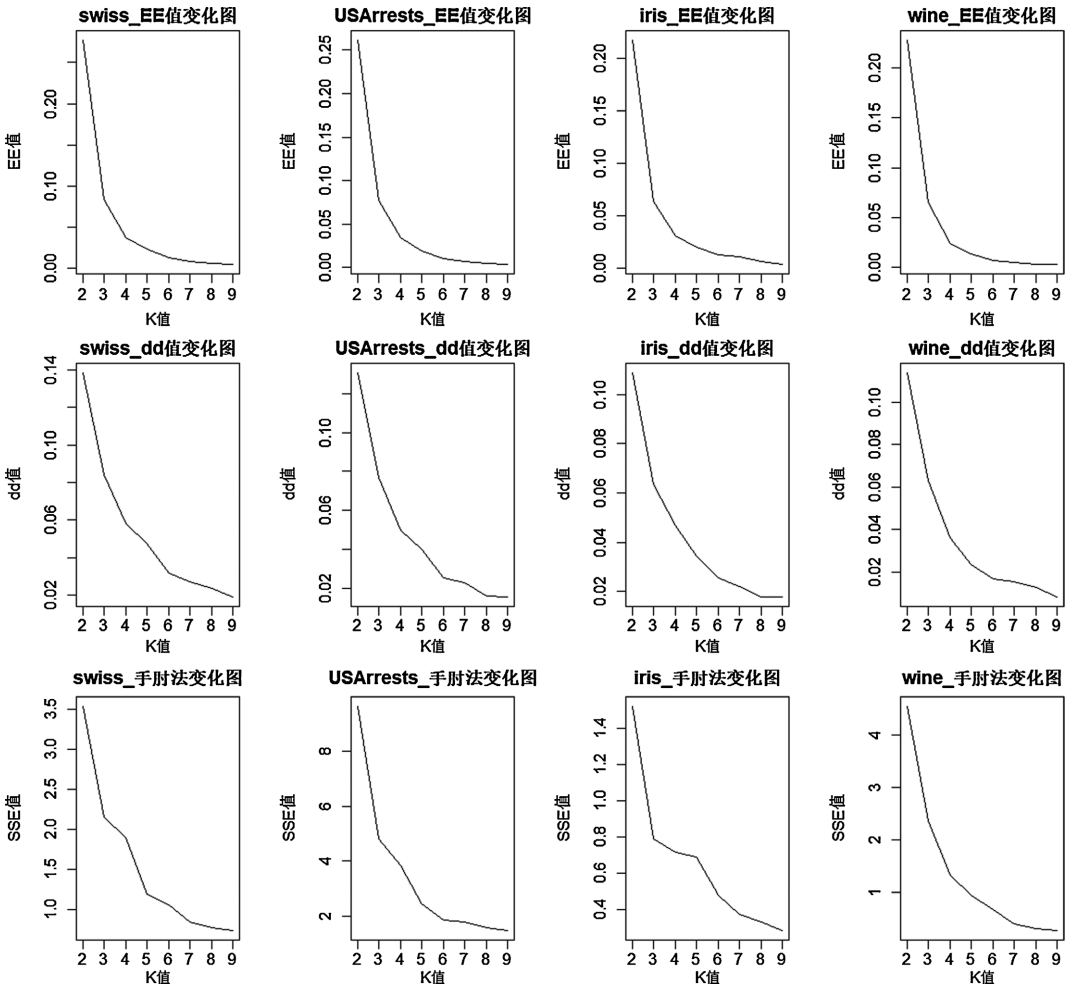

Figure 1. Determination of k value by three methods on four datasets
图 1. 三种方法在 4 个数据集上求 k 值

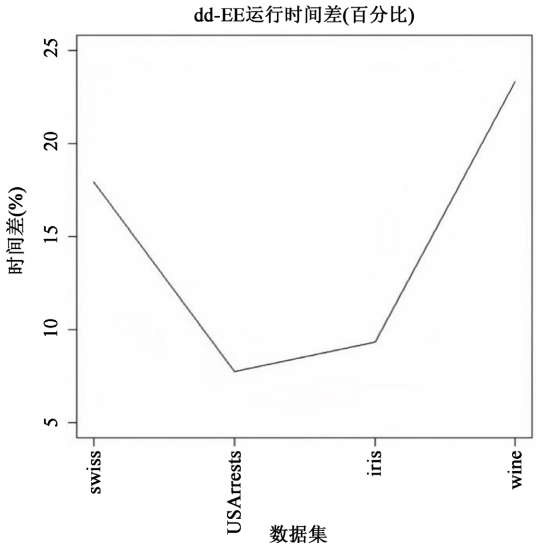

Figure 2. Time comparison of dual mean method and EE method
图 2. 双均值法与 EE 法的时间比较

法和手肘法对比可以得到相同的 k 值，同时为了验证该方法的有效性，又比较了该方法的时间性能，双均值法在 4 个数据集上求出 k 值的时间均少于 EE 法。实验证明，改进之后的双均值法有效。

该方法的提出不仅解决了 k 值选择问题，还为聚类算法的研究提供了新的视角，并在图像分割、社交网络分析、市场细分等领域展现了广阔的应用前景。由于其原理和实现的简单性，双均值法对于非专业数据分析师来说是一个有吸引力的选择。未来工作将探讨该方法在不同数据集上的表现，并研究如何将其与其他聚类算法结合，以进一步提升聚类任务的整体性能。

基金项目

河南科技大学 2023 年大学生创新创业训练计划项目(2023231)。

参考文献

[1] 王实, 高文, 李锦涛. Web 数据挖掘[J]. 计算机科学, 2000, 27(4): 28-3141.

[2] 孙秀娟, 刘希玉. 基于初始中心优化的遗传 K-means 聚类新算法[J]. 计算机工程与应用, 2008, 44(23): 166-168, 182.

[3] 王森, 刘琛, 邢帅杰. K-means 聚类算法研究综述[J]. 华东交通大学报, 2022, 39(5): 119-126.

[4] 李波, 管彦允, 龚唯印, 等. 基于密度的 K-means 初始聚类中心点选取算法[J]. 绥化学院学报, 2022, 42(6): 148-151.

[5] 冯波, 郝文宁, 陈刚, 占栋辉. K-means 算法初始聚类中心选择的优化[J]. 计算机工程与应用, 2013, 49(14): 182-185, 192.

[6] Huang, S.Y. (2022) K-Means Clustering Algorithm Based on Optimization of Initial Clustering Center. *CIBDA*2022, 25-27 March 2022, Wuhan, 297-300.

[7] 李飞, 薛彬, 黄亚楼. 初始中心优化的 K-Means 聚类算法[J]. 计算机科学, 2002, 29(7): 94-96.

[8] 孙红岩, 孙晓鹏, 李华. 基于 K-means 聚类方法的三维点云模型分割[J]. 计算机工程与应用, 2006, 42(10): 42-45.

[9] 王建仁, 马鑫, 段刚龙. 改进的 K-means 聚类 k 值选择算法[J]. 计算机工程与应用, 2019, 55(8): 27-33.

[10] 方姣丽, 左克, 黄春, 刘杰, 李胜国, 卢凯. FD-LSTM: 基于大规模系统日志的故障分析模型[J]. 计算机工程与科学, 2021, 43(1): 33-41.