$K-Means$ 改进算法研究综述

周梦豪

西安交通大学管理学院工业工程 71 2174111346

摘要： $K-Means$ 算法作为一种经典的无监督学习算法，因其具有简洁高效，收敛性好等特点，成为应用最广泛的一种聚类算法。但 $K-Means$ 算法也存在固有的一些局限性，例如，算法存在超参数聚类数 k，对初始聚类中心敏感，欧氏距离具有局限性，对鲁棒性差，容易陷入局部最优等。针对 $K-Means$ 算法的这些缺陷国内外学者分别提出了不同的改进策略，并取得了较好的效果。本文对这些改进算法进行了概括比较，并对 $K-Means$ 算法的发展方向和趋势进行了展望。

关键词： $K-Means$ 算法 超参数 聚类中心 欧氏距离 离群点 局部最优

中图分类号： TP301.6

1 引言

随着人类社会进入信息时代其是进入大数据时代以来，与人们生产生活息息相关的诸多领域被数据所覆盖，越来越多的数据被人们获得。人们意识到数据信息已经成为一种重要的资源。而要利用这种资源获得效益，就需要使用数据挖掘从大量非结构化数据中获得知识。聚类作为数据挖掘六大功能之一（数据挖掘具有六大功能[1]：统计、分类、聚类、关联分析、预测和偏差检测），是一种无监督学习，因其简单、高效实用、更贴合现实的特点，在各个领域被广泛应用。聚类分析的核心是将混杂的数据划分合适的类，从而发现隐藏信息。

目前的聚类算法可以分为基于划分的聚类、基于层次的聚类、基于密度的聚类、基于网格的聚类、基于模型的聚类以及基于模糊的聚类[2]。

基于划分的聚类算法需要人为设定聚类数目和初始聚类中心，通过不断迭代，直至收敛，典型的该类算法有：$K-Means$ 算法以及其变体包括 K-medoids[3]、k-models 等。

基于层次的聚类算法有两类：自底而上(Bottom-up)和自上而下(Top-down)，这类聚类的典型算法包括 BIRCH 算法[4]、CURE 算法[5]、Chameleon 算法[6]等。

基于密度的算法是基于邻近区域的密度对数据进行分类[7]，典型算法例如：DBSCAN 算法、OPTICS 算法、DENCLUE 算法等。

基于网格的聚类算法采用了并行处理的思想，将原始空间分割，分别在网格单元中聚类，典型算法有：Sting 算法、Wave-cluster 算法等。

基于模型的聚类算法的基本思想是假定每一类有一最佳的模型，常见的算法有：混合高斯算法(GMM)，SOM 算法等。

基于模糊的聚类算法采用模糊数学的思想，以概率表示数据归属某一类的可能性，典型算法有：FCM 算法等。

$K-Means$ 聚类算法由 James MacQueen 在 1967 年提出[8]，$K-Means$ 算法作为一种经典的无监督学习算法[9]，因其具有简洁高效，收敛性好等特点，成为应用最广泛的一种聚类算法。但 $K-Means$ 算法也存在固有的一些局限性，例如，算法存在超参数聚类数 k，对初始聚类中心敏感，对离群点敏感，容易陷入局部最优等。针对 $K-Means$ 算法的这些缺陷国内外学者分别提出了不同的改进算法，并取得了较好的效果。

本文首先介绍了 $K-Means$ 算法的原理与流程，接着介绍了针对该算法固有缺陷的改进，并对 $K-Means$ 算法的发展方向和趋势进行了展望。

2 传统 $K-Means$ 聚类算法

2.1 算法思想

$K-Means$ 算法的基本思想是：从原始数据中选择 k 个初始聚类中心，计算其余数据与聚类中心的欧氏距离，找到最近的聚类中心，将数据归为该聚类中心对应的类，使用类均值作为新的类中心，进行迭代直至收敛获得达到最大迭代次数。

从机器学习的视角描述 $K-Means$ 算法可以分成 4 部分：

(1) 训练数据：给定样本集 $D=\{x_1,x_2,\cdots,x_m\}$，

(2) 表现度量: $\min E=\sum_{i=1}^{k}\sum_{x\in C_i}\left\|x-\mu_i\right\|_2^2$

其中，利用 $K-Means$ 算法划分类 $C=\{C_1,C_2,\cdots,C_k\}$，$\mu_i=\frac{1}{|C_i|}\sum_{x\in C_i}x$ 是类 C_i 的均值向量。

(3) 优化算法：迭代优化算法

最小化 $\min E=\sum_{i=1}^{k}\sum_{x\in C_i}\left\|x-\mu_i\right\|_2^2$ 很困难，这是一个 NP 难问题[10]，因此，$K-Means$ 算法采用了贪心策略，通过迭代优化来求近似解。

(4) 超参数调整：聚类数 k

2.2 算法流程

$K-Means$ 算法的流程图如图 1 所示。

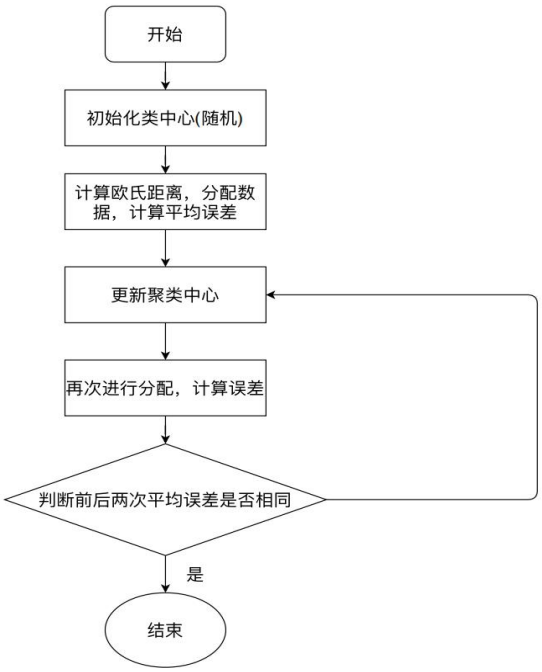

图 1 $K-Means$ 聚类算法流程图

2.3 $K-Means$ 算法的性能

$K-Means$ 算法的时间复杂度为 $O(nlkm)$，空间复杂度为 $O(mn)$。其中，m 为每个元素字段个数，n 为数据量，l 为迭代次数。一般 l,k,m 均可认为是常量，因此，$K-Means$ 聚类算法的时间和空间复杂度可以简化为：$O(n)$，因此，$K-Means$ 算法的性能很好.

2.4 算法优缺点

该算法具有简洁高效，收敛性好等优点，但是也具有一些局限性：

1）聚类数 k 需要认为设定，具有很强的主观性，不同的 k 对聚类效果影响很大

2）聚类中心点的选取会影响聚类结果；

3）采用欧式距离作为度量，存在局限性；

4）对噪声敏感，离群点对最终聚类结果的影响很大；

5）易陷入局部最优解[10]，以获得全局最优解

3 针对存在超参数 k 的改进算法

在 $K-Means$ 算法中，需要人为主观设定聚类数 k，这种设定对使用者提出了很高的要求，严重依赖使用者对数据的了解程度，因此，在实际应用时稳定性较差。针对这种缺陷，不同学者提出了不同解决方案。

3.1 手肘法

手肘法是一种基于误差平方和 (SSE) 的算法[12]。

$$SSE=\sum_{i=1}^{k}\sum_{x\in C_i}\left\|x-\mu_i\right\|_2^2$$

如图 2 随着聚类数 k 的增加，SSE 会逐渐下降，当 k 接近真实聚类数时，SSE 会大幅下降，之后会趋缓，因此会出现一个“手肘”形状，“肘点”就是理想聚类数 k。

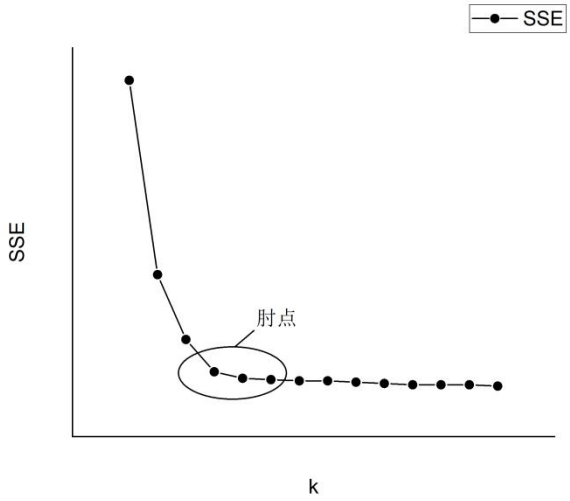

图 2 SSE 与 k 的关系

该方法具有简单高效的优点，但是无法保证一定出现明显的“肘点”，此时 k 值的选择会对最终结果产生很大影响，聚类效果很不稳定。

3.2 内部有效性指标

内部有效性指标是指对同一聚类算法在不同的超参数下聚类效果的表示，依据的思想是好的聚类应该是“类内紧凑，类间松散”。应用较为广泛的有：SC 指标、CH 指标[13]和 DB 指标[14]。

1） SC (Silhouette Coefficient)指标基于凝聚度和分离度，定义为：

$$S=\frac{(b-a)}{\max(a,b)}$$

其中，a 表示样本 i 到同一类内其他点不相似程度的平均值，b 表示样本 i 到其他类内平均不相似程度的最小值。

2） CH (Calinski-Harabasz)指标基于类内和类间离差矩阵，定义：

$$CH(k)=\frac{trB(k)/k-1}{trW(k)/n-k}$$

其中，$trB(k)$ 表示类间离差矩阵，$trW(k)$ 表示类内离差矩阵。

3） DB (Davies-Bouldin)指标基于类内散度和类中心距离，定义：

$$DBI=\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}(\frac{\overline{S_i}+\overline{S_j}}{\|w_i-w_j\|_2})$$

其中，$\overline{S_i}$ 为第 i 类数据到类中心距离的平均值，$\|w_i-w_j\|_2$ 是第 i,j 类中心距离。

4 针对初始聚类中心选取的改进

$K-Means$ 算法对初始聚类中心的选取较为敏感，目前对初始聚类中心的选取经常使用随机方法，但是这对聚类效果的稳定性产生了影响，为了改进该缺陷，学者们希望通过提出一些准则对不同聚类中心进行评价，并提出了不同的模型。

Rodriguez A 等[14]提出了类中心处在局部密度比较大的位置。在该思想的基础上，Xiong 等[16]将大于密度平均值的数据对象作为高密度点集，从高密度点集中选取密度值最大的点作为第一个初始聚类中心，以此类推，得到所有的聚类中心，但此方法存在聚类中心集中的缺陷。Du 等[17]在此基础上，加入点之间的距离一定程度克服了该缺陷。贾瑞玉等[18]在此基础上，重新定义了局部密度，但是对稀疏数据的聚类效果不好。

Lei Gu[19]等采用减法聚类确定初始聚类中心。减法聚类是一种密度聚类算法，根据下列公式计算每个点的密度指标，选取其中最大的作为聚类中心，之后去除该点，重新选择，不断迭代，直至结束。该方法存在两个缺陷：鲁棒性较差，对离群点敏感；存在的超参数较多。

$$H(x_i)=\sum_{j=1}^{n}e^{-\alpha\times d(x_i,x_j)}$$

其中，$d(x_i,x_j)=\|x_i-x_j\|_2$，α 为超参数。

5 针对距离度量的改进

$K-Means$ 算法中对相似性和距离的度量采用了欧氏距离[20]。欧氏距离存在的缺陷是将数据的所有特征看成同等重要，这和现实中的绝大多数情况并不相符。因此，为了克服该缺陷，可以从两方面改进：1. 提出合适的权重生成策略；2. 采用其他更合适的度量指标。

关于第一种方式，Xu 等[21]采用信息论中的信息熵概念和特征选取算法，计算特征权重，得到加权距离进行聚类，取得了改进效果；此外还有，基于密度，自适应加权等。

关于第二种方式，W.Xue 等[22]采用空间密度相似性度量，发现对非线性流形聚类的效果很好，J. P. Singh 和 N. Bouguila[23]针对比例数据，提出用 Aitchison 距离度量来对比例数据进行聚类，陈磊磊[24]发现余弦距离和谷本距离更适合对文本的聚类。

6 针对离群点的改进

$K-Means$ 算法对于离群点敏感，单个离群点会对聚类的最终结果产生很大的影响，因此，如何识别并消除离群点对聚类结果的影响成了一个重要改进方向。

Zhang 等[25]构建了一个基于距离的指标 LDOF 来度量一个数据集的离散程度，以便检测出离群值；Ting Zhang 等[26]提出可以构建一个距离阈值，将和聚类中心距离达到距离阈值的点识别为离群点，并去除该点，以便聚类中心迭代时，离群点对聚类中心产生影响；Breuning 等[27]基于数据点邻近区域的局部密度构建了度量数据点的离群程度的指标，并根据计算结果将明显小于平均水平的异常值识别为离群点。

7 针对局部最优的改进

由于 $K-Means$ 算法的表现度量 $\min E=\sum_{i=1}^{k}\sum_{x\in C_i}\|x-\mu_i\|_2^2$ 是个 NP 难问题，因此求解时

极易陷入局部最优，针对该缺陷的改进思路是将 $K-Means$ 算法和具有一定全局优化能力的启发式算法进行结合，以求提高 $K-Means$ 算法的全局搜索能力。

陈小雪等[28]从改进欧氏距离的权重角度出发，使用萤火虫优化的方法获得加权距离，充分利用了该算法的全局搜索的优势；Kapil[29]等从选择最优聚类中心的角度出发，利用传统的遗传算法对 $K-Means$ 算法进行改进，获得了较好的效果；Shi 等[30]在使用遗传算法之前，使用 SNM 算法对原始数据进行了去重归一化，取得了更好的效果。

8 对未来改进方向的展望

$K-Means$ 聚类算法作为一种经典的无监督学习算法，因其具有简单高效，易收敛的优势在过去几十年被广泛应用于各个领域。虽然传统的 $K-Means$ 算法存在着各种缺陷，但是国内外各个领域的学者一直在各个方向不断地提出改进措施，并取得了不错的效果。虽然我们进入了一个大数据时代，新的使用场景对传统的经典算法提出了各种挑战，但是 $K-Means$ 算法的思想依然具有强大的生命力，针对新的各种挑战对 $K-Means$ 算法进行改进的尝试依然是值得的，这对我们很有启发意义。

$K-Means$ 算法面临诸多新的挑战，这些挑战为我们改进 $K-Means$ 算法提供了方向，例如：

1） 快速增长的数据量对 $K-Means$ 改进算法的复杂度提出了挑战。一方面，大数据时代产生数据的速度越来越快，海量的数据要求算法越简单越好；另一方面，为了克服传统 $K-Means$ 算法的缺陷而提出的各种改进不可避免地提高了算法的复杂度。因此，海量的数据对 $K-Means$ 改进算法的复杂度提出了更高要求。

2） 数据的稀疏性对 $K-Means$ 改进算法提出了挑战。目前产生的大量数据具有稀疏的特点，而传统的 $K-Means$ 算法并不是一种专门针对稀疏数据开发的算法，因此，大量的稀疏数据对包括 $K-Means$ 算法在内的传统经典算法均提出了挑战。

3） 大量的高维数据对 $K-Means$ 改进算法的泛化能力提出了挑战。目前各个行业，尤其是互联网领域产生了越来越多的高维数据，文本、图像、音频、视频等数据类型取代了传统的二维数据，这对算法的泛化能力提出了更高要求。

参考文献：

参考文献

[1] 朱幸燕. 基于消费行为认知的电信企业客户细分方法研究[D]. 华南理工大学, 2011.

[2] 唐东明. 聚类分析及其应用研究[D]. 电子科技大学.

[3] Hae-Sang Park, Jun Chi-Hyuck. A Simple And Fast Algorithm For K-medoids Clustering[J]. Expert Systems with Applications, 2009, 36(2p2): 3336-3341.

[4] Tian Zhang, Ramakrishnan Raghu, Livny Miron. BIRCH: An Efficient Data Clustering Method for Very Large[J]. Acm Sigmod Record, 1996, 25(2): 103-114.

[5] Cheng-Fa Tsai, Chen Zhi-Cheng, Tsai Chun-Wei. MSGKA: an efficient clustering algorithm for large databases[A]//2002.

[6] F Murtagh. A Survey of Recent Advances in Hierarchical Clustering Algorithms[J]. Computer Journal, 1983, (4): 4.

[7] J-Rg Sander, Ester Martin, Kriegel Hans-Peter, et al. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications[J]. Data Mining & Knowledge Discovery, 1998, 2(2): 169-194.

[8] J Macqueen. Some Methods for Classification and Analysis of MultiVariate Observations[A]//1965.

[9] Saroj Kavita. Review: Study on Simple k Mean and Modified K Mean Clustering Technique. 2016.

[10] Daniel Aloise, Deshpande Amit, Hansen Pierre, et al. NP-hardness of Euclidean sum-of-squares clustering[J]. Machine Learning, 2009, 75(2): 245-248.

[11] Cheng-Huang Hung, Chiou Hua-Min, Yang Wei-Ning. Candidate groups search for K-harmonic means data clustering[J]. Applied Mathematical Modelling, 2013, 37(24): 10123-10128.

[12] 成卫青，卢艳红，CHENGWeiqing，等. 一种基于最大最小距离和SSE的自适应聚类算法[J]. 南京邮电大学学报（自然科学版）, 2015, 35(2): 102-107.

[13] T Calinski, Harabasz J. A dendrite method for cluster analysis[J]. Communications in Statistics, 1974, 3(1): 1-27.

[14] David-L Davies, Bouldin Donald-W. A Cluster

Separation Measure[J]. IEEE Trans Pattern Anal Mach Intell, 1979, PAMI-1(2): 224-227.

[15] A Rodriguez, Laio A. Clustering by fast search-and-find of density peaks[J]. Science, 344.

[16] Caiquan Xiong, Zhen Hua, Ke Lv, et al. An Improved K-means Text Clustering Algorithm by Optimizing Initial Cluster Centers[A]//2016.

[17] Xin Du, Xu Ning, Zhou Cailan, et al. A density-based method for selection of the initial clustering centers of K-means algorithm[A]//2017.

[18] 贾瑞玉，李玉功. K-means algorithm of clustering number and centers self-determination%类簇数目和初始中心点自确定的K-means算法[J]. 计算机工程与应用, 2018, 054(007): 152-158.

[19] Lei Gu. A novel locality sensitive k-means clustering algorithm based on subtractive clustering[A]//2016.

[20] Meirong Zhang, Zhao Kai. Mulitilinear Singular Integral on Herz-Hardy Spaces with Variable Exponent[J]. Mathematica Applicata, 2017.

[21] Yan Xu, Fu Xueliang, Li Honghui, et al. A K-means Algorithm Based On Feature Weighting[J]. Matec Web of Conferences, 2018, 232.

[22] Wei Xue, Yang Rong-Li, Hong Xiao-Yu, et al. A novel k-Means based on spatial density similarity measurement[A]//2017.

[23] Jai-Puneet Singh, Bouguila Nizar. Proportional data clustering using K-means algorithm: A comparison of different distances[A]//2017.

[24] 陈磊磊. 不同距离测度的K-Means文本聚类研究[J]. 软件, 2015, (1).

[25] Ke Zhang, Hutter Marcus, Jin Huidong. A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data[J]. 2009.

[26] Ting Zhang, Fang Yuan, Liu Yang. Capped Robust K-means Algorithm[A]//2017.

[27] Jörg Sander. LOF: Identifying Density-Based Local Outliers.[J]. Acm Sigmod Record, 2000, 29(2): 93-104.

[28] 陈小雪，尉永清，任敏，等. 基于萤火虫优化的加权K-means算法%Weighted K-means clustering algorithm based on firefly algorithm[J]. 计算机应用研究, 2018, 035(002): 466-470.

[29] Shruti Kapil, Chawla Meenu, Ansari Mohd-Dilshad. On K-means data clustering algorithm with genetic algorithm[A]//2016.

[30] Haobin Shi, Xu Meng. A Data Classification Method Using Genetic Algorithm and K-Means Algorithm with Optimizing Initial Cluster Center[A]//2018.