Hans Journal of Data Mining 数据挖掘, 2022, 12(1), 20-26
Published Online January 2022 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2022.121003

Hans汉斯

基于机器学习的K-Means聚类优化算法研究

李　贞，刘海燕*，刘　策，李庆钰，刘　刚

北华航天工业学院，河北　廊坊

收稿日期：2021年11月28日；录用日期：2021年12月28日；发布日期：2022年1月6日

摘　要

K均值聚类(K-Means)算法是基于划分的聚类算法中的一个典型算法，是机器学习研究算法的基础。通过将相似的样本自动归到一个类别，合理地确定K值和K个初始类簇中心点，使聚类效果更好。经过适当的预处理，可以对数据做初步分析，甚至挖掘出隐含的价值信息。相比于SVM、GBDT等机器学习算法，具有操作简单、采用误差平方和准则函数、对大数据集处理上有较高的伸缩性和可压缩性的优点。但是，这种聚类算法仍然存在随机初始聚类中心导致算法不稳定、K值的选取不好把握、非凸性数据集非常难收敛等问题。为提升数据挖掘中聚类分析的效果，本文在分析数据挖掘、聚类分析、传统K-Means算法的基础上，提出一种改进的K-Means算法，经过实验证明，K-Means的改进算法可以有效地提高簇的质量，以及算法的效率和稳定性，使其提供更加精准有效的服务，并且减少了算法开销。

关键词

改进K-Means算法，Mini Batch K-Means算法，数据挖掘

Research on K-Means Clustering Optimization Algorithm Based on Machine Learning

Zhen Li, Haiyan Liu*, Ce Liu, Qingyu Li, Gang Liu

North China Institute of Aerospace Engineering, Langfang Hebei

Received: Nov. 28th, 2021; accepted: Dec. 28th, 2021; published: Jan. 6th, 2022

Abstract

K-Means Clustering (K-Means) algorithm is a typical algorithm based on the clustering algorithm

*通讯作者 Email: 37206446@qq.com

文章引用: 李贞, 刘海燕, 刘策, 李庆钰, 刘刚. 基于机器学习的K-Means聚类优化算法研究[J]. 数据挖掘, 2022, 12(1): 20-26. DOI: 10.12677/hjdm.2022.121003

of division, which is the basis of the machine learning research algorithm. By automatically categorizing similar samples into one category, the K value and K initial cluster center points can be determined reasonably to make the clustering effect better. After proper pre-processing, the data can be analyzed and even the implied value information can be excavated. Compared with machine learning algorithms such as SVM and GBDT, it has the advantages of simple operation, the use of error square and standard functions, and the high flexibility and compressibility of large data sets. However, this clustering algorithm still has the problems such as random initial clustering center leading to algorithm instability, poor grasp of K value selection and non-convex data set is very difficult to converge. In order to improve the effect of clustering analysis in data mining, this paper puts forward an improved K-Means algorithm on the basis of analyzing data mining, clustering analysis, and the traditional K-Means algorithm. Experiments have proved that the improved K-Means algorithm can effectively improve the quality of clusters as well as the efficiency and stability of the algorithm; and make it provide more accurate and effective service, and reduce the algorithm overhead.

Keywords

Improved K-Means Algorithm, Mini Batch K-Means Algorithm, Data Mining

1. 引言

聚类学习[1]最早被用于模式识别、机器学习、数据分析以及数据挖掘等多类领域之间，而且还被应用于各种大型的数据库中，因此聚类算法的研究分析越来越受到研究学者的关注。聚类是机器学习算法的一种技术，也是一种无监督学习的方法，同时也是许多研究领域常用来统计数据分析的技术，它涉及数据点的分组。我们通过给定一组数据点，按照某个特定标准把一个数据集分割成不同的组或簇，也就是使用聚类算法将每个数据点划分为一个特定的组。在理论逻辑上，同一组的数据点的对象属性和特征的相似性尽可能大，而不同组中的数据点的属性和特征的差异性也尽可能大。在聚类过程中，我们需要将聚类后同一类的数据尽可能聚集到一起，不同数据尽量分离。

随着大数据时代的到来，聚类算法技术也迎来了又一次的历史新高，聚类算法在数据挖掘、统计学、机器学习、空间数据库技术、生物学以及市场营销等各个技术领域方面都有渗透。如今，各种聚类方法也被专家学者不断改进，针对各种不同类型的数据，使每种的方法适合它自己的应用场景，因此对各种聚类算法、以及聚类的效果的差别成为我们非常值得研究的一个课题，本文主要研究 K-Means 聚类算法以及 Mini Batch K-Means [2]算法，通过对大量数据集进行测试，最后对数据集聚类的效果进行比较和分析。

2. 国内外发展现状

2.1. 国外发展现状

聚类算法有着深远的历史发展，1963 年层次聚类算法由 Ward 提出，在 1957 年，Lloyd 根据划分思想，首次提出来 K-Means 算法的雏形，直到 1967 年，K 均值算法的理论才真正的由 James MacQueen 提出，进而逐步发展至今，该算法是所有聚类算法中研究最多、改进最成功一种算法。发展至今，聚类算法已经非常成熟，也被大量学者和研究者应用以及改进提升。目前，数据聚类方法[3]主要可以分为四大

类别：划分式聚类方法(Partition-Based Methods)、基于密度的聚类方法(Density-Based Methods)、层次化聚类方法(Hierarchical Methods)等，而我们研究的 K-Means 算法是一种基于划分的经典聚类算法。

2.2. 国内发展现状

目前，国内对数据挖掘算法的研究学习也有十几年的历史。在 2004 年，我国的学者在 K-Means 聚类算法思想的基础上外加核学习的理论，提出了一种核 K-Means 聚类算法，从而加快了大量数据集的运算速度。在 2008 年，徐义峰和徐去青等多名专家同样也是根据 K-Means 聚类算法思想中随机选择初始聚类中心的缺点，提出了一种新型的基于数据样本分布选取初始聚类中心的方法，该方法更加精准细致的提高了 K-Means 聚类算法的精准度。聚类学习是最早被用于模式识别及数据挖掘任务的方法之一，并且被用来研究各种应用中的大数据库，因此用于大数据的聚类算法受到越来越多的关注。K-Means 算法已经被国内外学者研究多年，并且在商业、工业、科学技术、大数据、数据挖掘、人工智能、统计学等很多领域应用广泛，在现实中的应用也十分成熟，如用于金融行业中银行客户信息的细分，账目数据的分类，微博、抖音等娱乐 APP 中热点词汇挖掘，淘宝京东等电商平台市场关注以及规模销量的数据采集，图形分割等。

3. 机器学习 K-Means 聚类算法

3.1. K-Means 和 Mini Batch K-Means 算法原理

K-Means 算法思想[4]：对给定的样本集，事先确定聚类簇数 K，让簇内的样本尽可能紧密分布在一起，使簇间的距离尽可能变大，以 K 为参数，把 N 个数据对象分成 K 个簇，使簇内具有较高的相似度，而簇间的相似度较低，该算法试图使集群数据分为 N 组独立数据样本，使 N 组集群间的方差相等，即最小化惯性或集群内的平方和，利用各簇中对象的均值来进行计算的。

Mini Batch K-Means 算法思想[5]：该算法是 K-Means 算法的变种，采用小批量的数据子集减小计算时间，同时仍试图优化目标函数，而我们所说的小批量是指每次训练算法时所随机抽取的数据子集，采用这些随机产生的子集进行训练算法，它将大大减小了数据的计算时间，同时与其他的聚类算法比较，可以缩短 K 均值的收敛时间，小批量 K 均值产生的结果，通常情况下只是稍微与标准算法差点。

3.2. K-Means 聚类算法流程

聚类过程：

采集数据：选择样本集，将簇的数目划分为 K，最大迭代数为 N；

清理数据：对采集好的数据进行清理，为每个聚类选择一个初始聚类中心；

(1)选取数据：将样本集按照最小距离原则分配到最邻近聚类；

(2)数据转换：使用每个聚类的样本均值更新聚类中心；

(3)重复步骤二、三，直到聚类中心不再发生变化；

(4)输出数据：对处理好的数据进行聚类运算，并实现雷达图，输出聚类中心和 K 个簇划分；

4. 算法实现

聚类算法是一种非常有效的非监督的机器学习算法，研究如何在没有训练的条件下把样本划分为若干类，即把相似度较大的对象自动归类到一个类别中，相异度大的则划分为不同的类型，聚类算法属于无监督学习，即事先不会给出标记信息，通过对无标记样本的学习来解释数据的内在性质及规律，为进一步的数据分析提供基础，因此，聚类算法在数据挖掘中起着至关重要的作用。

4.1. K-Means 算法分析

输入：样本集 D={x1,x2,x3,x4,...,xn};聚类簇数 K。

算法过程：

(1)从 D 中随机选择 K 个样本作为初始均值向量{a1,a2,a3,a4,...,ak}；

(2)重复此操作；

(3)令 Fi = ∅(1≤i≤K)；

(4)For j = 1,2,...,m do；

(5)计算样本 xj 与各均值向量 ai(1≤i≤K)的距离：dji = ||xj−μi||2；

(6)根据距离最近的均值向量确定 xj 的簇标记：bj = argmini∈{1,2,3,...,k}dji;

(7)将样本 xj 划入相应的簇：Cbj = Cbj⋃{xj}；

(8)end for

(9)for i = 1,2,3,...,k do

(10)计算新均值向量 ai’

(11)if ai’ ≠ ai then

将当前均值向量 ai 更新为 ai’

(12)else

保持当前均值向量不变

(13)end if

(14)end for

(15)Until 当前均值向量均为更新

输出：簇划分 C = {C1,C2,C3,...,Ck}

4.2. Mini Batch K-Means 算法分析

Mini Batch K-Means 算法流程与 K-Means 类似，Mini Batch K-Means 使用了一个种叫做分批处理的方法对数据点之间的距离进行计算。Mini Batch K-Means 在计算过程中不必使用所有的数据样本[6]，则是通过从不同类型的样本数据中抽取到一部分样本数据来分别代表各自类型数据进行计算。但是由于计算样本数据量十分少，所以会相应的减少运行时间，同时抽样时也必然会造成数据的准确度减低。

算法过程：

第一，Mini batch 是原始数据集中的子集，这个子集是在每次训练迭代时抽取的样本，在这里我们默认为 100 个，通过 batch_size 进行设置。接下来，我们先从不同类别的样本中抽取一部分样本数据集,通过使用 K-Means 算法来构建一个拥有 K 个聚簇点的模型[7]。

第二，我们继续抽取训练数据集中的一部分数据集的样本数据,并将它们添加到上述的模型中,并且将其分配给距离最近的一个聚簇中心点。

第三，更新聚簇的中心点值。

第四，我们继续循环迭代上述步骤中的第二步和第三步操作,直到中心点稳定或者达到迭代的次数时,再停止计算操作。

4.3. 实验结果及分析

4.3.1. 聚类结果评价

在实验中，使用包含 1000 个样本的数据集，并且每个样本具有两个属性的二维数据集，将数据集分

为 3 簇，即 K 值等于 3。首先，计算多个点的中心，检查两个点是否有差别，针对每个点寻找距离其最近的中心点，重新计算中心点。需要注意的是，在计算中心点的时候，需要将原来的中心点算进去，判断中心点是否发生变化，即判断聚类前后样本的类别是否发生变化，最终返回聚类的结果。

首先随机获取一个样本集，用于测试 K-Means 算法，在各个区域内随机产生一些点，自定义中心点，本次的目标聚类个数为 5，因此选定 5 个中心点，获取聚类结果图。图 1 是聚类结果图。

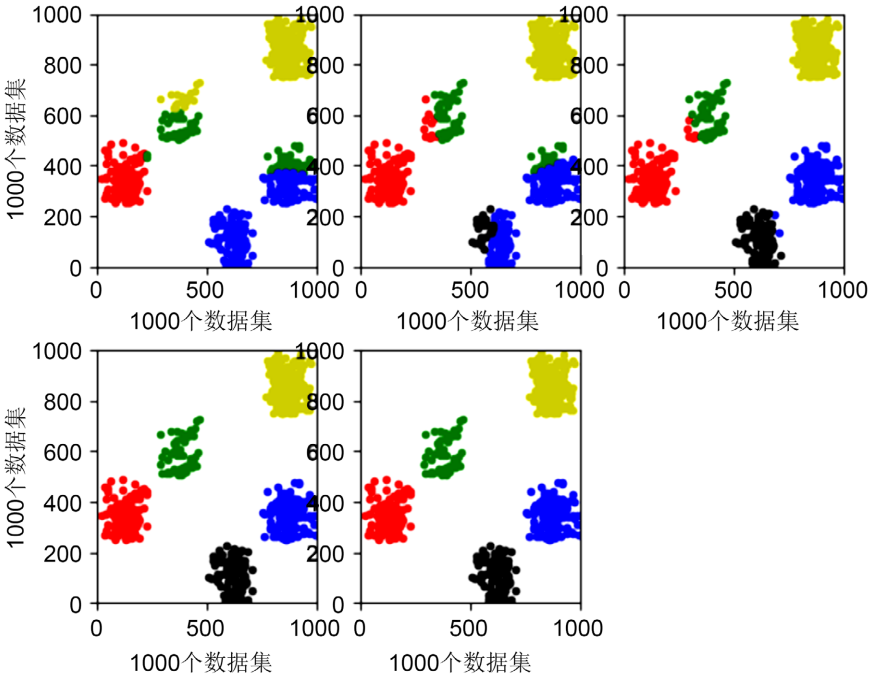

Figure 1. Clustering result graph
图 1. 聚类结果图

上述图例的解析：

第一张图表示：初始的数据集，假设 k = 3。

第二张图表示：我们随机选择了两个 k 类所对应的类别质心，即图中的红色质心和蓝色质心，然后分别求样本中所有点到这两个质心的距离，并标记每个样本的类别为和该样本距离最小的质心的类别。

第三张图表示：经过计算样本和红色质心和蓝色质心的距离，我们得到了所有样本点的第一轮迭代后的类别，此时我们对当前标记为红色和蓝色的点分别求其新的质心。

第四张图表示：新的红色质心和蓝色质心的位置已经发生了变动。

第五张图表示：重复了我们在图二和图三的过程，即将所有点的类别标记为距离最近的质心的类别并求新的质心，最终我们得到的两个类别，即第五张图。

当然在实际 K-Means 算法中，我们一般会多次运行图二和图三，才能达到最终的比较优的类别。

4.3.2. 算法比较分析

给定较多数据，比较两种算法的聚类速度，使用聚类评估算法对 Mini Batch K-Means 算法和 K-Means 算法进行评估。图 2 是 Mini Batch K-Means 算法和 K-Means 算法比较图。

上图我们列出实验结果的原始数据分布图，K-Means 算法聚类结果图与 Mini Batch K-Means 算法聚类结果图，以及 Mini Batch K-Means 算法和 K-Means 算法预测结果的不同点，最终根据训练时间以及两

种算法的不同节点数为 7 个，获取两种聚类算法的速度，从而选择一种最优的算法。从上图，我们看出 Mini Batch K-Means 算法比 K-Means 算法速度快了不止一倍，而效果还是不错的。我们对最基本的 K-Means 算法进行了改进，其中 Mini Batch K-Means 算法可以适用于数据量比较多的情况，且效果也不错。

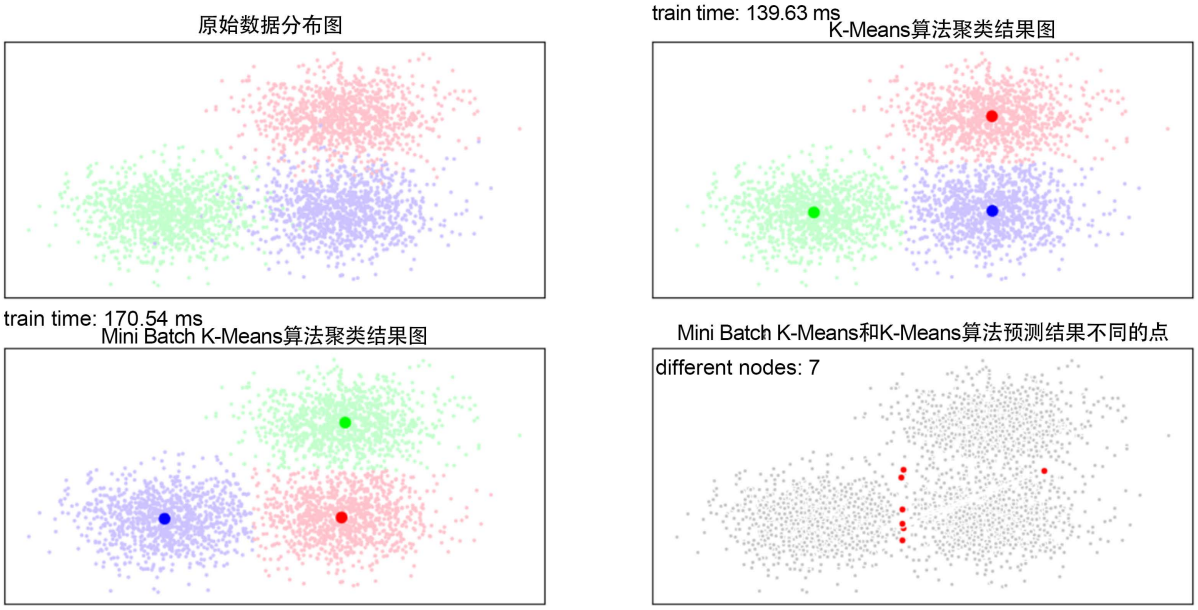

Figure 2. Comparison between Mini Batch K-Means algorithm and K-Means algorithm
图 2. Mini Batch K-Means 算法和 K-Means 算法比较图

4.3.3. 实验结论

K-Means 算法是常用的聚类算法[8]，但其算法本身存在一定的问题，例如在大数据量下的计算时间过长。为此，Mini Batch K-Means 算法便基于 K-Means 的变种聚类算法应运而生。因此，Mini Batch K-Means 是一种在保持聚类准确性下的同时，大幅度缩短计算时间的聚类模型[9]。

以上结果表明，相较于其他优化初始聚类中心的 K-Means 算法[10]，Mini Batch K-Means 算法在计算过程中不必使用所有的数据集样本，而是从不同种类的数据集样本中随机抽取一些样本，从而代表各种不同类型数据来进行聚类计算。正是因为选取的是部分样本数据集[11]，所以数据量比较少，自然而然数据的运行时间也缩短了，但这种方法同样也出现了抽样数据的精准度不高的弊端。但是利大于弊，该聚类算法还是在一定程度上，获取了较好的聚类效果，同时也能缩短聚类时间。在实验中，我们为了增加聚类算法的准确性，可以训练 Mini Batch K-Means 算法的数据多次，通过不同的随机数据采样集获取聚类组，从而选择其中最优的聚类组。

4.4. 结语

本文通过对 K-Means 算法的进行改进，原理比较简单，实现容易，收敛速度快，实验结果证明，该算法能大大提高聚类的准确性。但是也存在一些问题需要我们继续研究。首先，K 值的选取不好把握，我们通过在开始时给定一个适合的数值给 K，通过 K-Means 算法得到聚类中心，根据 K 个聚类的距离情况，合并距离最近的类，聚类中心数减小，为以后的聚类做准备，同时相应的聚类数目也减小，最终得到合适数目的聚类数。并且我们通过一个评判值来确定聚类数找到一个合适的位置停下来，而不继续合并聚类中心。一直重复上述循环，直至评判函数收敛为止，最终获取较优聚类数的聚类[12]结果。第二，

对于非凸性数据集非常难收敛，我们可以选择使用基于密度的聚类算法。第三，隐含类型的数据不平衡，导致聚类效果不佳。第四，采用迭代法，获取的结果出现局部性数据优，但是全局数据不佳的情况。第五，该方法对噪音和异常点比较敏感。综上所述的问题，需要在后续的学习中进行改进，我相信随着数据挖掘技术的成熟，如何分析并获取高精尖数据，在未来值得我们进一步研究。

基金项目

课题：项目编号：2021011074；项目名称：基于大数据技术的就业信息采集与挖掘关键技术研究；项目类别：廊坊市科技支撑计划项目。

参考文献

[1] 钱鑫, 张龙波, 田爱奎, 邓齐志, 汪金苗. 一种面向数据密集型计算环境的聚类算法[J]. 济南大学学报(自然科学版), 2013(1): 11-15.

[2] Idrees, A.K., Al-Qurabat, A., Jaoude, C.A., *et al*. (2019) Integrated Divide and Conquer with Enhanced K-Means Technique for Energy-Saving Data Aggregation in Wireless Sensor Networks., *The* 15*th International Wireless Communications & Mobile Computing Conference* (*IWCMC* 2019), **2019**, 973-978. https://doi.org/10.1109/IWCMC.2019.8766784

[3] 夏长辉. 一种改进的 K-Means 聚类算法[J]. 信息与电脑, 2017(14): 40-42.

[4] 钮永莉, 武斌. 基于改进粒子群和 K-Means 的文本聚类算法研究[J]. 兰州文理学院学报(自然科学版), 2019, 33(4): 44-47.

[5] 杨丹, 朱世玲, 卞正宇. 基于改进的 K-Means 算法在文本挖掘中的应用[J]. 计算机技术与发展, 2019, 29(4): 68-71.

[6] 王康. K-Means 聚类算法的改进研究及其应用[D]: [硕士学位论文]. 大连: 大连理工大学, 2015.

[7] Nayak, S., Panda, C., Xalxo, Z., *et al*. (2015) An Integrated Clustering Framework Using Optimized K-Means with Firefly and Canopies. *Computational Intelligence in Data Mining*, **2**, 333-343. https://doi.org/10.1007/978-81-322-2208-8_31

[8] Yin, J.W., Chen, J.M., Xue, B.L., *et al*. (2013) An Enhancing K-Means Algorithm Based on Sorting and Partition. *International Journal of Database Theory and Application*, **22**, 387-408.

[9] Whang, Y. and Cui, P. (2017) An Efficient K-Means Parallel Algorithm Based on MapReduce. *Journal of Liaoning Technical University* (*Natural Science Edition*), **36**, 1204-1211.

[10] 韩存鸽, 刘长勇. 一种改进的 K-Means 算法[J]. 闽江学院学报, 2019, 40(5): 49-54+90.

[11] 韩琮师, 张高毓, 张熙, 等. 基于改进的 K-Means 算法在套餐精准营销中的研究[J]. 信息技术与信息化, 2021(5): 132-133.

[12] 刘文佳, 张骏. 一种改进的 K-Means 聚类算法[J]. 现代商贸工业, 2018(19): 196-198.