K-Means 聚类分析

B21031803 蒋平

摘要

随着大数据时代的到来，数据挖掘对于发现数据价值变得至关重要。k-Means 算法在聚类领域占据重要地位，但其面临的主要挑战是难以确定合适的聚类数量 k，以及初始中心点的选择敏感性。本文深入探讨了该算法，并与层次聚类、密度聚类等其他相关算法进行了对比分析，详细探讨了算法的优缺点。研究提出结合肘部法来确定 k 值，并采用 K-Means++ 优化初始化过程，通过轮廓系数和 Calinski-Harabasz 指数等指标进行评估。实验结果表明，K-Means++ 不仅收敛速度快，而且聚类效果显著。在实际应用中，建议结合多种方法以克服现有算法的局限，准确挖掘数据信息，从而增强分析和决策能力。

关键词：聚类分析；K-Means；肘击法；K-Means++

1 介绍

随着大数据时代的来临，海量数据的涌现使得从数据中挖掘有价值的信息愈发重要。数据挖掘技术作为一种从大量数据中提取隐藏模式和知识的有效手段，在各领域得到了广泛的应用。然而，在面对复杂多样的数据结构和特征，如何有效地对数据进行分类和聚类，以发现数据中的潜在规律和群组结构，成为一个亟需解决的问题 [1]。K-Means 聚类算法因其简单有效、易于实现等优点，在数据挖掘领域中占据着重要地位，但其在实际应用中也面临着一些挑战，如聚类簇数 k 的确定、对初始聚类中心的敏感性等。因此，本文旨在深入研究 K-Means 聚类算法，并与其他的一些聚类算法进行比较，以提高算法的性能和聚类效果，为数据挖掘任务提供更有效的解决方案。

2 相关工作

2.1 层次聚类算法

层次聚类 [2] 通过计算数据点之间的距离或相似度，将数据点逐步合并或分裂形成层次化的聚类结构。自底向上的层次聚类开始时将每个数据点视为一个单独的簇，然后合并最相似的簇，直到达到预定的停止条件，如簇的数量达到要求或簇间距离超过阈值。自顶向下的层次聚类则相反，开始时将所有数据点视为一个簇，然后不断分裂簇，直至每个簇只包含一个数据点或满足其它停止条件。

该算法不需要预先指定聚类数量，能够生成层次化的聚类结果，便于观察不同粒度下的数据结构。但计算复杂度高，尤其是对于大规模数据集，时间和空间开销

大，且一旦合成或分裂操作完成，无法回溯，且可能导致局部最优解。

2.2 密度聚类算法

密度聚类 [3] 基于数据点的密度来进行聚类，核心思想是将具有足够高密度的区域划分为簇，低密度区域则被视为簇间的边界或噪声。常见的密度聚类算法有DBSCAN，它通过定义邻域半径和最小点数来确定核心点、边结点和噪声点，然后通过不断扩展核心点的邻域来形成簇。

该类算法可以发现任意形状的簇，会噪声数据不敏感，能够有效地处理数据集中的噪声和离群点。但对参数设置较为敏感，不同的参数设置可能导致截然不同的聚类结果。在高维数据中，密度的定义和计算变得复杂，可能会影响聚类效果。

3 问题说明

本研究的任务主要是对给定的数据集 xclara.csv 进行聚类分析，以发现数据中潜在的数组结构和模式。通过对这些数据进行聚类，能够将相似的数据点归为同一簇，不同簇之间具有较大的差异，从而为后续的数据分析、决策制定等提供依据。

4 算法与解决方案

4.1 K-Means 算法

4.1.1 基本思路

K-Means 是一种常见的聚类分析方法，它通过将数据点划分为 k 个簇，使得每个簇中的数据点尽可能相似，而不同簇之间的数据点尽可能不同，属于机器学习中的无监督学习。该算法 [4] 的基本步骤为：

（a）随机选择 k 个数据点作为初始簇中心，或者是通过某些启发式方法选择。

（b）将每个数据点分配到离它最近的簇中心所对应的簇中。

（c）计算每个簇内所有数据点的均值，并将簇中心更新为这个均值。

（d）重复步骤（b）和（c），直到簇中心的变化非常小或达到最大迭代次数为止。

（e）算法收敛，簇中心不再变化或变化非常小，最终得到的簇划分即为聚类结果。

K-Means 最核心的部分就是先固定中心点，调整每个样本所属的类别来减少损失值；再固定每个样本的类别，调整中心点继续减小损失值。两个过程交替循环，损失值单调递减直至最小值或极小值，中心点和样本划分的类别同时收敛。其本质是基于欧氏距离的数据划分算法，均质和方差大的维度对数据的聚类产生决定性影响，所以在计算过程中需要先对数据进行归一化和统一单位的处理。

4.1.2 手肘法确定 k 值

手肘法 [5] 是一种在聚类分析中确定合适聚类数 k 的有效方法，其主要是通过计算误差平方和 SSE 值来进行判断。

在聚类过程中，SSE 是衡量聚类质量的一个重要指标，它的计算方式是针对每个数据点，计算其到所属簇中心的距离的平方，并对所有的这些距离平方进行求和。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 SSE 值自然会逐渐变小。当 k 小于真实聚类数时，由于 k 的增大会大幅度增加每个簇的聚合程度，故 SSE 的下降幅度会很大；而当 k 到达真实聚类数时，此时数据点已经被较为合理地划分到各个簇中，再继续增加 k 所得到的聚合程度回报会迅速变小，因此 SSE 的下降幅度会骤减，然后随着 k 值超过真实聚类数后继续增大，SSE 值的变化会趋于平缓，因为此时进一步增加聚类数已经无法有效地提高数据的聚合程度，只是在不断地对数据进行过度细分，而这种细分对降低 SSE 值的作用微乎其微。

数据集 xclara.csv 中数据 SSE 与 k 的关系如下：

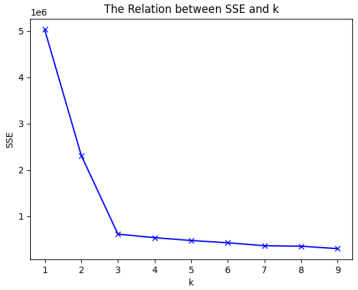
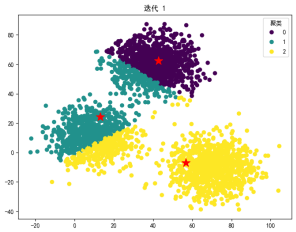


图 1: SSE 与 k 值的关系

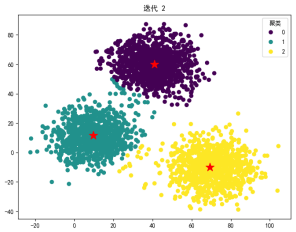
由图 1可知真实聚类数 k=3。

4.1.3 可视化结果

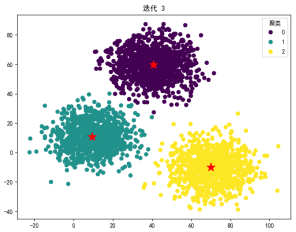
取 k=3，并设置最大迭代次数为 20，记录每次迭代的聚类可视化结果如下：



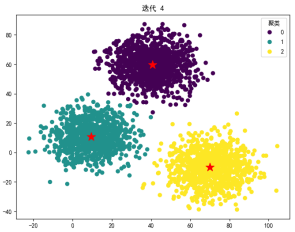
(a) K-Means 迭代 1



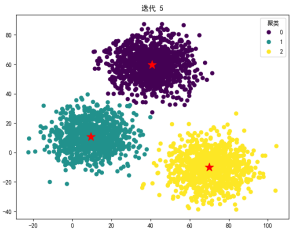
(b) K-Means 迭代 2



(c) K-Means 迭代 3



(d) K-Means 迭代 4



(e) K-Means 迭代 5

图 2: K-Means 算法不同迭代次数的聚类结果可视化

4.2 改进的 K-Means 算法 K-Means++

K-Means++ 算法 [6] 是对 K-Means 随机初始化质心的方法的优化，其优化策略如下.

（a）从数据集中随机选择一个样本作为初始聚类中心 $c1$;

（b）计算每个样本与当前已有聚类中心之间的最短距离为 $D(x)$，一般采用欧几里得距离等距离度量方式来衡量样本点与聚类中心之间的远近关系;

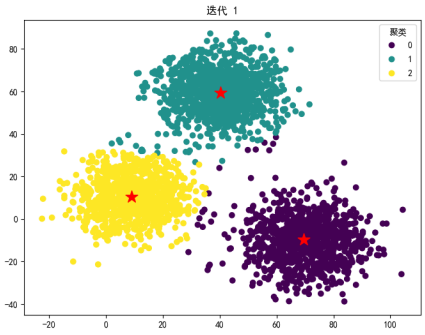
（c）计算每个样本被选为下一个聚类中心的概率 $P(x)$ 并选取概率最大的作为聚类中心.

$$P(x) = \frac{D(x)^2}{\sum_{x \in X} D(x)^2} \tag{1}$$

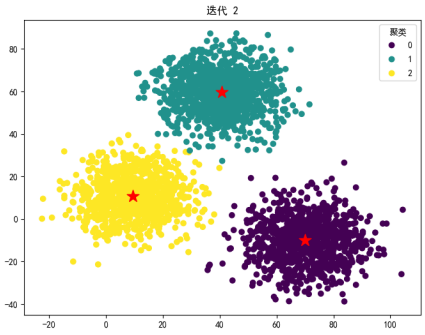
（d）重复步骤（b）和（c）直至选出 k 个聚类中心。

（e）在确定了 k 个初始聚类中心之后，使用这些初始点作为聚类中心，按照一般的 K-Means 算法的流程进行后续操作操作。

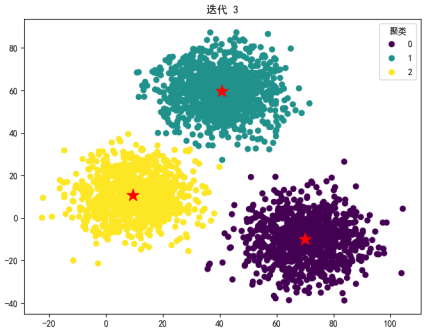
使用 K-Means++ 算法每次迭代的聚类可视化如下所示：



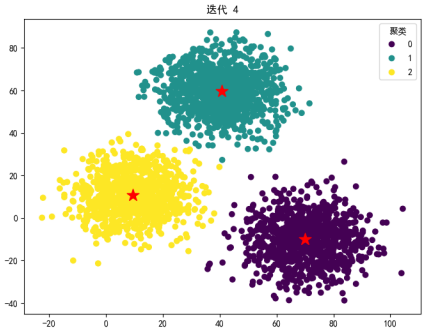
(a) K-Means++ 迭代 1



(b) K-Means++ 迭代 2



(c) K-Means++ 迭代 3



(d) K-Means++ 迭代 4

图 3: K-Means++ 算法不同迭代次数的聚类结果可视化

5 评估

5.1 评估初始 k 值的选定

5.1.1 轮廓系数

轮廓系数是一种用于评估聚类质量的重要指标，它全面且深入地考虑了每个数据点在聚类结构中的位置关系 [7]，尤其与同一簇内其它点的相似度和与最近簇的相似度这两个关键方面。

$$s = \frac{b-a}{max(a,b)} \tag{2}$$

其中 a 是样本与同一聚类中其它样本的平均距离，该值在一定程度上反映了样本点在所属簇内的紧密程度，若 a 值较小说明该样本点与其它点较为接近，聚类的内聚性较好。b 是样本与最近聚类中所有样本的平均距离，体验了样本点与其它簇的分离程度，若 b 值相对较大，意味着该样本点与其它簇的样本点距离较远，聚类的分离性较好。整个样本的轮廓系数是所有样本轮廓系数的平均值。

轮廓系数 s 的值在-1 到 1 之间，值越接近 1 表示聚类效果越好，即不仅在各自所属的簇内紧密聚集，而且与其它簇之间有明显的区分，内聚性和分离性都很好。s 值越接近 0 表示聚类效果一般，此时样本点在簇内的紧密程度与其它簇的分离程度相对较为均衡，没有明显的有时或劣势，可能需要进一步分析数据或调整据类参数来提高聚类质量。s 为负值则表明数据点可能被错误地聚类到某个簇中。

5.1.2 Calinski-Harabasz 指数

Calinski-Harabasz 指数 [8] 又称为方差比准则，它的核心原理在于精确地衡量簇间离散度与簇内离散度地比例关系，以次来对聚类的效果进行评估。CH 值越高表示聚类效果越好。

$$CH = \frac{B_k/(k-1)}{W_k/(n-k)} \tag{3}$$

其中 B_k 是簇间离散度，W_k 是簇内离散度，k 是簇的数量，n 是样本的数量。

W_k 是各个簇内样本到簇中心的平方距离之和，表示同一簇内样本之间的相似程度，c_i 是簇中心。当 W_k 值较小时，说明同一簇内的样本点紧密地围绕在中心周围，样本之间的相似性较高，聚类的紧凑型较好。

$$W_k = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2 \tag{4}$$

B_k 是各个簇中心到整体数据集中心的平方距离之和，从宏观层面反应不同簇之间的差异程度。若 B_k 值较大，意味着各个簇在整体空间的分布较为分散，不同之间的差异较为显著，是聚类效果良好的一个重要体现。

$$B_k = \sum_{i=1}^{k} |C_i| \cdot \|c_i - \bar{x}\|^2 \tag{5}$$

5.1.3 不同 k 值时的 s 和 CH 值

根据公式 2 和 3 得到不同 k 值时的轮廓系数 s 和 Calinski-Harabasz 指数 CH 值如表 1所示。

表 1: 不同 k 值时的 s 和 CH 值

k	2	3	4	5	6
s	0.509	0.695	0.531	0.545	0.437
CH	3033.99	10836.60	829.98	7026.49	6493.15

由表 1中数据可知，当聚类数 k 值为 3 时，对目标数据集的聚类效果最好。

5.2 比较 K-Means 和 K-Means++ 法

多次运行 K-Means 和 K-Means 方法，得到迭代次数图如图 11 所示。

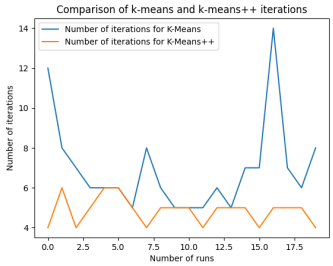


图 4: k-Means 与 k-Means 迭代的比较

显然，K-Means++ 算法在迭代过程展现了明显的优势，能够以更快的速度达到收敛状态，有效地减少所需的迭代次数。同时，进一步比较图 2 和图 7 中第一次迭代的聚类可视化呈现。

可以发现 K-Means++ 算法的初始阶段，其聚类中心的分布更加均匀，这种均匀性一定程度上反映了 K-Means++ 算法独特的初始化策略，该策略通过精心挑选初始点，使得这些点在数据空间中的分布更为合理，能够更好地反应数据的内在结构，从而为后续的迭代过程奠定了良好的基础，最终得到的聚类结果也更加稳定。相比之下，K-Means 算法由于采取随机选择初始聚类中心的方式，其初始分布往往缺乏这种均衡性，进而可能导致需要更多的迭代次数来调整和优化聚类结果。

6 结论

K-Means 作为一种经典的聚类算法，以其简单直观和相对快速的运算特点，广泛地应用于数据聚类任务。在实际的数据分析场景中，它能够在一定程度上对数据进行有效分组，帮助人们快速洞察数据的潜在结构和分布规律。

但 K-Means 算法也存在着一些不可忽视的局限性。其中，聚类种类数 k 值需要提前指定，且初始化中心随机产生，可能导致不同的聚类结果，进而影响数据聚类效果，也更容易陷入局部最优解的困境中。而 K-Means++ 是对 K-Means 算法随机初始化质心的方法的优化，在初始阶段，它为基于数据点之间的距离关系，优先选择那些距离已有初始点较远的数据点作为新的数据点作为新的初始中心，从而确保了初始质心在数据空间中的分布更加均匀和合理。这样的优化不仅减少了

算法达到收敛所需的迭代次数，提高了计算效率，同时也显著降低了陷入局部最优解的可能性，使得聚类结果更加可靠和稳定，进而提升了 K-Means 聚类的整体性能。

在实际应用时，为了充分发挥聚类算法的优势并获得最佳的聚簇效果，可以将 K-Means 算法与 K-Means++、手肘法等多种手段相结合。如利用手肘法来初步确定较为合理的聚类数 k 值的范围，再结合 K-Means 算法的优化初始化方式来运行 K-Means 算法，对数据进行聚类分析。通过类似这样的组合策略，可以再一定程度上克服单一算法的局限性，更准确地挖掘数据中的信息，为后续的数据分析、决策制定等提供有力的支持，满足不同领域和场景下对数据聚类的多样化需求。

参考文献

[1] 孟增辉. 聚类算法研究. Diss. 河北大学.

[2] 贺玲, 吴玲达, and 蔡益朝. " 数据挖掘中的聚类算法综述." 计算机应用研究 24.1(2007):4.

[3] 孙吉贵, 刘杰, and 赵连宇. " 聚类算法研究." 软件学报 1(2008):14.

[4] Wagstaff, Kiri , et al. "Constrained K-means Clustering with Background Knowledge." Eighteenth International Conference on Machine Learning Morgan Kaufmann Publishers Inc. 2001.

[5] 张玉琨. 基于 K-Means 聚类分析的电商学生客户细分研究 [J]. 商场现代化, 2022, (08): 33-35.

[6] Arthur, David , and S. Vassilvitskii . "K-Means++: The Advantages of Careful Seeding." Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007 ACM, 2007.

[7] 朱连江, 马炳先, and 赵学泉. " 基于轮廓系数的聚类有效性分析." 计算机应用 12(2010):4.

[8] Baarsch J , Celebi M E .Investigation of Internal Validity Measures for K-Means Clustering[J].Lecture Notes in Engineering & Computer Science, 2012, 2195(1):471-476.DOI:10.1111/j.1365-4632.2010.04135.x.