# 《数据挖掘技术及应用》
# 课程报告

姓　　名：刘军　　学　　号：20212753

学　　院：计算机与数学学院　　年　　级：2021 级

专业班级：计算机科学与技术三班　　指导老师：周健老师

报告评语：

成绩：

# 基于 K-means 和 DBSCAN 算法的聚类分析及参数选择研究

## 目录



## 摘要

本文介绍了数据挖掘领域中聚类分析的概念及其重要性，并重点讨论了 K-means 和 DBSCAN 两种经典聚类算法。K-means 算法通过迭代寻找最优的簇划分方案，适用于对大型数据集进行高效分类，但对初始值敏感且仅适用于凸形数据集。DBSCAN 算法将簇定义为密度相连的点的最大集合，能够发现任意形状的聚类并处理噪声点，但在处理大数据集时计算复杂度高且对参数的选择敏感。论文还介绍了选择 K-means 和 DBSCAN 算法中关键参数的方法，如手肘法和 k-距离曲线。这些方法有助于在实际应用中确定合适的参数值，提高聚类算法的效果。

# 1 绪论

随着大数据时代的到来，以“大量化”、“快速化”、“多样化”、“价值低密度”为特征的数据充斥在人们的生活中[1]。而数据挖掘（Data Mining）作为计算机技术的分支，是人们运用算法从大量数据中提取隐藏信息的过程，并且数据挖掘是结合了人工智能、统计学、模式识别和机器语言等的交叉学科，是人们进行数据提取的重要工具。聚类分析是数据挖掘领域一个非常重要的分支，被学者们研究有长达几十年的历史，它能够挖掘大数据背后的内在逻辑和信息，对相似的数据进行归类和聚合，使得人们能够掌握数据背后的意义，从而进行优化决策。由于其在数据挖掘中的重要性及与其他研究方向的交叉性，聚类分析深受各个领域专家和学者们的青睐。聚类是数据挖掘、机器学习等研究方向的重要分支之一，在识别数据的内在规律和内在结构方面具有非常重要的作用[2]，聚类分析还应用于模式识别、图像分割、机器视觉、数据压缩等[3]。K-means 作为基于划分的聚类中的典型，DBSCAN 算法作为密度聚类的典型，在日常研究中应用非常之多，非常之广，在整个聚类分析中占有绝对的重要性，因此深入了解研究这两种聚类方法的优缺点对于后续算法的改进和发展尤其重要。

# 2 聚类与聚类算法类别

## 2.1 聚类分析的定义

聚类(Clustering)是将数据划分成群组的过程。研究如何在没有训练的条件下把对象划分为若干类。通过确定数据之间在预先制定的属性上的相似性来完成聚类任务，这样最相似的数据就聚集成簇(Cluster)。

## 2.2 聚类算法的类别

现有的聚类技术大致可以分为如下五大类：基于划分的方法(Partitioning Method)，基于层次的方法(Hierarchical Method)，基于密度的方法(Density-based Method)，基于网格的方法(Grid-based Method)和基于模型(Model-based Method)的方法。

基于划分的聚类方法:该类算法基于点的相似性在单个分区中基于距离来划分数据集。其缺点是需要用户预定义一个参数 k,而它通常具有不确定性。代表性的划分算法包括:K-means、K-methods、K-modes、PAM、CLARA、CLARANS 和 FCM。

基于层次的聚类算法:该类算法将数据划分成不同的层次,并提供了可视化。其基于相似性或距离将数据自底向上或自顶向下进行分层划分,其划分结果表示为一

种层次分类树。它的主要缺点是:一旦完成了某个划分阶段,就无法撤销。其代表性算法有:BIRCH、CURE、ROCK 和 Chameleon。

基于密度的聚类算法:该类算法能够以任意一种方式发现簇。簇定义为由低密度区域分开的密集区域。基于密度的聚类算法不适用于大型的数据集。其代表性算法包括:DBSCAN、OPTICAL DBCLASD 和 DENCLUE,它们常用来过滤噪音。

基于网格的聚类算法:该类算法的过程分为 3 个阶段,首先,将空间划分为矩形方格以获取一个具有相同大小的方格的网格; 然后,删除低密度的方格;最后,将相邻的高密度的方格进行结合以构成簇。其代表性的算法有:CRIDCLUS、STING、OptiGrid、CLICK 和 WaveCluster。

基于模型的聚类算法:该类算法基于多元概率分布规律,可以测量划分的不确定性,其中,每个混合物代表一个不同的簇。该类算法对大数据集的处理很慢。该类算法的代表性算法有 EM、COBWEB、CLASSIT 和 SOM。

# 3 两种聚类算法

K-means 作为基于划分的聚类中的典型，DBSCAN 算法作为密度聚类的典型，深入了解研究这两种聚类方法的理论和优缺点对于后续算法的改进和发展尤其重要。

## 3.1 K-means 算法

聚类属于非监督学习，K-means 聚类算法是最基础常用的聚类算法。它的基本思想是，通过迭代寻找 K 个簇(Cluster)的一种划分方案，使得聚类结果对应的损失函数最小[4][5]。

k-means 算法接受输入量 k，然后将 n 个数据对象划分为 k 个聚类以便使所获得的聚类满足：同一聚类中的对象相似度较高，而不同聚类对象中的对象相似度较小。聚类相似度是利用各聚类中对象的均值所获得一个“中心对象”（引力中心）来计算的。

k-means 算法描述：

输入：聚类个数 k，以及包含 n 个数据对象的数据库

输出：满足方差最小标准的 k 个聚类

处理流程：

Step1 从 n 个数据对象任意选择 k 个对象作为初始聚类中心；

Step2 根据簇中对象的平均值，将每个对象重新赋给最类似的簇；

Step3 更新簇的平均值，即计算每个簇中对象的平均值；

Step4 循环 Step2 到 Step3 直到每个聚类不再发生变化为止。

K-means 算法的优点与不足[6]

优点：能对大型数据集进行高效分类,其计算复杂性为 O(t Kmn),其中,t 为迭代次数,K 为聚类数,m 为特征属性数,n 为待分类的对象数, 通常,K,m,t<<n.在对大

型数据集聚类时,K-means 算法比层次聚类算法快得多。
不足：通常会在获得一个局部最优值时终止;仅适合对数值型数据聚类;只适用于聚类结果为凸形(即类簇为凸形)的数据集。
K-means 聚类算法还有一个难点是关于聚类簇 k 的选择，在实际研究中，通常不会给出相应的聚类簇，由于 K-means 算法必须事先给定聚类数，此时用 K-means 算法将无法进行聚类。
下面介绍一种关于聚类簇 k 的选择的方法——手肘法[7]
手肘法的中心思想是，随着聚类数 k 的增大，算法对于数据样本的划分会逐步精细，从而每个聚类簇的聚合程度逐渐提高，误差平方和 E 自然而然逐渐变小，而当 k 到达真实聚类数时，再增加 k 只能小幅度甚至不能增加每个簇的聚合程度了，此时 E 的减小很缓慢，最后趋于平稳，这时 E 和 k 的关系就像一个手肘的形状，此时这个手肘的肘部对应的 k 值就是数据的真实聚类数 k。因此上述选择 K-means 算法聚类数的方法也被称为“手肘法”。

## 3.2 DBSCAN 算法

DBSCAN 是一个比较有代表性的基于密度的聚类算法。与划分聚类方法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在有“噪声”的空间数据库中发现任意形状的聚类。
DBSCAN 算法基本思想是先选择一个数据点作为起始点,计算它的邻域内所有点的密度。如果密度大于等于一个预先设定的阈值,那么这个点就是一个核心点;如果密度小于该阈值,那么这个点就是一个噪声点。对于每个核心点,使用一个邻域半径 eps 来计算邻域内所有点的密度。如果密度大于等于该阈值,那么这些点就可以组成一个簇。 对于每个边界点,如果它在一个核心点的邻域内,那么它就可以被归为该核心点所在的簇中。如果它不在任何一个核心点的邻域内,那么它就是一个噪声点。最终,所有被归为簇的数据点就是聚类结果,所有噪声点就是噪声结果。
DBSCAN 算法描述
输入：包含 n 个数据对象的数据库，半径 ε，最少数目 MinPts
输出：所有达到密度要求的簇
处理流程：
Step1 从数据库中抽取一个未处理的点；
Step2 IF 抽出的点是核心点 THEN 找出所有从该点密度可达的对象，形成一个簇；
Step3 ELSE 抽出的点是边缘点（非核心对象），跳出本次循环，寻找下一个点；
Step4 循环 Step1 到 Step3 直到所有点都被处理。
DBSCAN 算法的显著优点不需要定性簇的定义和对任意数据都能进行有效的处理，并且摆脱了分层聚类和划分聚类仅对球形簇敏感的缺点，在此基础上可以快速有效处理噪声点以及对任意形状的空间聚类进行发现。
DBSCAN 算法的弱点在于当空间维数数据量很大时，需要查询数据库中的每一个对象，因此增加了计算的复杂度，造成了 I/O 操作的频繁性。并且该算法依赖于数据对象的分布，对于密度不均的空间聚类以及聚类相差比较大的聚类，其聚类质量不理想
DBSCAN 基于一组“邻域”参数(ε, MinPts)来刻画样本分布的紧密程度，结果完全

依附于参数(ε,MinPts),因此对于参数(ε,MinPts)的选择尤其重要。对于这两个参数具体数值的选取，同样有如下选择技巧。

ε：ε 值可以通过绘制 k-距离曲线得到，当 k-距离曲线图中有明显拐点时，此时有明显拐点的位置对应的参数就是 ε。若参数设置过小，会导致数据集中大部分数据不能参与聚类；若参数设置过大，会使很多簇合并到同一个簇类中，此时聚类效果很差。

MinPts:通常让 MinPts≥dim+1,这里 dim 表示数据集的维度。

## 总结

本文系统地介绍了数据挖掘领域中的聚类分析以及两种经典的聚类算法：K-means 和 DBSCAN。首先，从大数据时代的背景出发，说明了数据挖掘在处理大量、快速、多样化、价值低密度的数据中的重要性。然后介绍了聚类分析的定义以及聚类算法的五大类别：基于划分、基于层次、基于密度、基于网格和基于模型。接着，重点讨论了 K-means 和 DBSCAN 算法的原理、优缺点以及参数选择方法。K-means 算法通过迭代寻找 K 个簇的划分方案，适用于对大型数据集进行高效分类，但对初始值敏感且只适用于凸形数据集。DBSCAN 算法将簇定义为密度相连的点的最大集合，能够发现任意形状的聚类并处理噪声点，但在处理大数据集时计算复杂度高且对参数的选择敏感。本文还介绍了选择 K-means 和 DBSCAN 算法中关键参数的方法，如手肘法和 k-距离曲线。这些方法有助于在实际应用中确定合适的参数值，提高聚类算法的效果。

## 参考文献


[1] 海沫.大数据聚类算法综述[J].计算机科学,2016,43(S1):380-383.

[2] 孙吉贵,刘杰,赵连宇.聚类算法研究[J].软件学报,2008(01):48-61.

[3] 项冰冰，钱光超.聚类算法研究综述[J].电脑知识与技术(学术交流),2007(12):1500-1501.

[4] 方诗乔,胡佩玲,黄莹莹,等.K-means 算法的优化及应用[J].现代信息科技,2023,7(06):111-115.DOI:10.19850/j.cnki.2096-4706.2023.06.028.

[5] 周爱武,于亚飞.K-means 聚类算法的研究[J].计算机技术与发展,2011(2):4

[6] Marques JP,Written;Wu YF,Trans.Pattern Recognition Concepts,Methods and Applications.2nd ed.,Beiing:Tsinghua University Press,2002.51-74(in Chinese).

[7] 成晓敏. 两种聚类算法的比较及其应用[J]. 湖北职业技术学院学报, 2024, 27 (01): 91-96.