# 浅析K-means聚类算法

摘要：聚类（Clustering）算法是无监督学习的重要算法，该方法可以通过基于数据的无监督训练得到一个较为稳定的值。其目标是找到混合样本集中存在的内在的群组关系，即对对象集合分割分块或分组为不同的子集或类，使得类内相似性高，类间相似性差。K-means算法作为一种典型的硬聚类算法，具有高效的运算速度和良好的可扩展性。本文将从K-means的数学基础、代码实现以及算法优化对K-means算法进行简单分析。

关键词：无监督学习；聚类；K-means算法

# K-means算法的数学基础

## 距离公式

闵可夫斯基距离

待插入公式

P=1；

P=2；

P=3；

聚类（Clustering）是一种可以取代直方图的特征提取方法，该方法可以通过基于数据的无监督训练得到更稳定的值。其目标是找到混合样本集中存在的内在的群组关系，即对对象集合分割分块或分组为不同的子集或类，使得类内关联性高，类间关联性差。

常用的聚类方法有K-means聚类、均值漂移聚类、基于密度的聚类方法（DBSCAN）、基于高斯混合模型的最大期望聚类（EM）、凝聚层次聚类和图团体检测等方法。

其中，K-means主要用于初步测试，在数据量足够大的情况下，得到的数据结果已经较为准确。K-means方法是将数据集中的n个d维特征向量分成k组，使各分组内具有最小组内平方和：

//todo 待插入公式

其中，k是一个超参数，无法基于数据直接学习出来，只能通过人工选择定义。

其具体提取方法如图三流程图所示：

其优点是快速且易于实现，同时K-means算法的缺点也很明确，如对“噪声”和异常点较敏感，同时由于K-means算法本质上是一个贪心算法，常常得到局部最优的结果，难以达到全局最优。所以该算法中k值的选取和初始中心的选取尤为重要，k值的选取可以通过多值尝试取聚类指标最优或提升转折点来选取，初始中心的选择可以通过多次全随机取最优、最远选取或K-means++半随机等方法来进行选择。

# K-means算法的代码实现

# K-means算法的优化

参考文献