

改进基于加权密度的 KWDM 聚类算法

1. 摘要

由于本案例是基于 KWDM 算法的改进，所以首先介绍一下，文献[1]提出的 KWDM 算法。KWDM 算法是对 K-means 算法的优化，它克服了 K-means 算法对聚类中心选择的随机性，防止聚类结果陷入局部最优，同时在欧式距离的基础上加入权值，进一步加强了数据中不同属性的区分程度，减少了离群点的影响，优化了初始聚类中心的选择，提高了聚类结果的准确率。

本案例首先基于不同赋权方法给欧氏距离赋权比较 KWDM 算法的结果，发现在不同赋权方法（信息熵赋权、构建权重矩阵、根据指标类型利用夹角余弦赋权法三种赋权方法）下，基于构建权重矩阵的准确率（即文献[1]的方法）低于基于信息熵赋权和夹角余弦赋权的准确率。然后对基于加权密度的 KWDM 算法的选择聚类中心点的方法进行改进，在选择密度最大的样本作为第一个聚类中心点的前提下，选择综合密度和距离两个指标的优劣进行第二个聚类中心点的选取。基于评价结果，可以得出改进后的算法有更高的准确率，改进有效果。

2. 改进措施

改进尝试（1）

分别使用信息熵赋权法、构建权重矩阵和夹角余弦赋权法三种赋权方法对欧式距离进行加权后，再基于 KWDM 算法进行聚类，比较其准确率。

改进尝试（2）

对 KWDM 算法进行改进，KWDM 算法是利用加权密度法排除离群点，选出聚类中心点，通过最大最小距离准则在聚类中心点集中选取聚类中心，最后利用簇内样本距离与簇间样本距离的比值来确定 K 值。基于类别为 2 的情况下，KWDM 算法将密度最大的样本点作为第一个聚类中心，将距离样本点最远的样本点作为第二个聚类中心，考虑到第二个聚类中心可能为离群点的情况，本案例将密度和距离的综合指标作为选取第二个聚类中心的依据，即确保第二个聚类中心距离第一个聚类中心有较远的距离的同时还要确保第二个聚类中心的密度不能太小，以避免将离群点错误作为聚类的中心，影响聚类效果。

3. 数据采集

本案例通过西南证券采集 2021 年 4 月 10 日食品饮料板块共 96 支股票和广告传媒板块共 67 支股票的涨幅、涨速、换手、量比、振幅、强弱度、均涨幅和实体涨幅 8 个指标的数据，共 163 个样本点，8 个属性。

4. R 程序运行结果

改进尝试 1 的结果如下图：

改进尝试1:			
聚类方法		准确率	
K-means			0.5337423
基于构建权重矩阵赋权的加权欧氏距离的KWDM(文献方法)			0.5582822
基于夹角余弦赋权的加权欧氏距离的KWDM			0.5828221
基于信息熵赋权的加权欧氏距离的KWDM			0.5828221

图 1 基于不同赋权方法改进 KWDM 算法

改进尝试 2 的结果如下图：

改进尝试2:	准确率	改进KWDM算法后的准确率
基于构建权重矩阵赋权的加权欧氏距离的KWDM	0.5582822(聚类中心:144,139)	0.5889571(聚类中心:144,69)
基于夹角余弦赋权的加权欧氏距离的KWDM	0.5828221(聚类中心:19,139)	0.6134969(聚类中心:19,80)
基于信息熵赋权的加权欧氏距离的KWDM	0.5828221(聚类中心:1,139)	0.595092(聚类中心:1,77)

图 2 改进 KWDM 算法准确率与原 KWDM 算法准确率对比图

5. 结论与展望

- (1) 改进尝试 1 中，可以得出，基于夹角余弦赋权或信息熵赋权的加权欧氏距离的 KWDM 算法比文献[1]中的权重矩阵赋权的加权欧氏距离的 KWDM 算法的准确率更高。
- (2) 改进尝试 2 中，可以得出，在不同赋权方法下，改进后的 KWDM 算法都能够比原 KWDM 算法的准确率更高，表现更好，所以改进有效果。其中基于夹角余弦赋权的加权欧氏距离的改进 KWDM 算法比其他两种赋权方法的改进 KWDM 算法的准确率更高，其次是基于信息熵赋权的加权欧氏距离的改进 KWDM 算法，最后是基于权重矩阵赋权的改进 KWDM 算法。
- (3) 对于密度大小和距离大小没有固定的值，所以可能改进算法稳定性不强，可以根据不同数据的特性，研究如何确定最佳的密度下限和距离下限。

6. 相关文献

[1]马克勤,杨延娇,秦红武,耿琳,王丕栋.结合最大最小距离和加权密度的 K-means 聚类算法[J].计算机工程与应用,2020,56(16):50-54.

[2]颜子寒,张正军,王雅萍,金亚洲,严涛.基于加权马氏距离的改进深度嵌入聚类算法[J].计算机应用,2019,39(S2):122-126.

[3]刘群,冯长焕.修正的加权欧氏距离的实证分析[J].兰州文理学院学报(自然科学版),2016,30(02):16-22.