一种优化聚类中心点选择的模糊聚类算法及在影响力聚类中的应用

**单位**：上海大学 计算机工程与科学学院，上海 宝山 200444

摘要：基于距离的FCM聚类算法对初始中心的选择较为敏感，易陷入局部最优，影响算法的稳定性。针对这个问题，提出了一个种改进的中心点的选择策略，结合距离和密度在中心点选择初就可以判断离群点。根据聚类中心的分离特性改进了目标函数，解决了距离公式不适应分布相对均匀的样本数据，最后通过修正隶属度矩阵，优化了边界点的归属。

关键词：聚类分析；聚类中心点；FCM聚类算法；隶属度；

**中图分类号：**查中图分类号手册 **文献标识码：**

# 1 引言

影响力分析是关于明星分析的重要内容之一，目前关于各种影响力的分析越来越多，在线社交网络影响力分析、期刊学术影响力分析、链路预测、推荐系统等。但是目前关于明星票房影响力是基于明星出演票房的均值来评价，这种方式仅仅考虑明星参演电影票房不符合实际情况。本文借鉴其他影响力分析方法，把改进的聚类算法应用到明星影响力分析上。目前已经有把聚类应用到微博用户影响力分析上，并取得了不错的效果。

聚类分析是数据挖掘领域应用十分广泛的算法之一，聚类就是根据某种准则对样本进行分类，每种类别在某些特征上具有一定的相似度，而不同类别之间相似度相差较大，主要是发现数据内部的有规律的信息。目前，在图像处理上，利用聚类对图像进行分割得到目标区域；在市场研究上，利用聚类调查品牌在市场定位中的应用。聚类发展至今，针对不同的样本数据，演化出了各种方法，比较常见的有基于密度的方法、基于层次的方法、基于网格的方法等。【基于领域的K中心点聚类算法，最好根据不同的点选择不同的文献】。在众多聚类中，由于模糊聚类能够适应分离性不是很好的样本数据，得到了样本数据属于各个类别的不确定性程度，能更好的反映客观世界。但模糊聚类（常见的如：FCM）算法仍存在和其他聚类一样的缺点，对初始聚类中心点的敏感、有噪声时分类效果差、边界数据归属、容易陷入局部最优等【文献】。然而，很难能够找到一种聚类能解决所有问题，当面对不同提的问题时，选择不同的聚类算法才是解决问题的关键。

最早的k-means算法是由MacQueen在1967年提出。随后，提出的k-medoids算法减少了噪声点对聚类的影响，clarans算法能够适应大规模数据集的聚类等等。Xin Wang【Semi–supervised K-Means Clustering by Optimizing Initial Cluster Centers】提出的改进的CSK-means算法把监督信息融合进k-means算法优化初始聚类中心点，较好的提高了聚类精确度，但缺点是过度依赖监督信息准确性，容易产生一致的类中心。fuHua【An Improved Automatic FCM Clustering Algorithm】提出的改进的FCM聚类算法，引入了偏差系数和距离系数实现了最优聚类，减弱了离群点的影响，提高了聚类的健壮性。张慧哲【基于初始聚类中心选取的改进FCM聚类算法】提出聚类中心选择原则是尽量使各类初始聚类中心之间的距离大于所设定的阈值，能够有效解决孤立点对聚类的影响，但是缺点是阈值是人工不断调整，而且如果选取的初始中心点不在类的中心，同样影响聚类结果的准确性。

# 2 FCM聚类算法

## 2.1 硬聚类

目前，针对硬聚类的研究很多是关于初始中心点的选择、目标函数的选择或者加入其他的算法进行融合。很多学者认为聚类划分过程是一个NP优化问题，在获取最优结果时，往往容易陷入局部最优。因此，不同样本数据，选择不同的中心点选择方法。有的样本分布呈凸型结构，这种情况往往选择距离公式作为目标函数，有的样本数据呈不规则形状，这种情况往往选择基于密度的函数作为目标函数。

基于k-means算法的思想，在实际应用中存在两个主要的缺点[6]，使得算法无法得到有效的聚类结果。首先，k-means算法需要事先给定聚类结果的簇数k，但在实际应用的时候，一般情况下并不知道数据集划分为多少个簇是最合适的，并且各个领域的情况也是不同的，很多时候通过领域专家的经验知识给出k值。另外，鉴于k-means算法对k值的敏感性，不同的k值可能会生成完全不同的聚类结果，严重影响聚类结果的质量。针对k-means算法的第二个缺点，有许多种改进的算法，因为它已经是k-means算法的一个最主要缺点。由于聚类结果对初始聚类中心值的依赖，选择不同的初始聚类中心，聚类结果就可能不同，导致算法的聚类结果非常不稳定。因此，k-means算法仅仅通过随机方法选择初始聚类中心，很难获得稳定和优质的聚类结果，其实当随机选择的聚类中心之间距离比较远，也就是相异性很大时，可以获得高质量的聚类结果。

在进行k-means聚类时，首先要指定聚类的簇数，并选择初始中心点，然后将数据中其他的样本按照和簇中心之间的距离大小划分和样本最近的簇中。设为特征空间上的一个集合，n表示样本个数，s表示样本特征个数。表示聚类中心点集合，c表示中心点个数。选择欧式距离作为相似性度量，首先计算样本到初始中心点的距离平方和

算法的结束准则是在使得目标函数最小。即当不属于簇，则另，当属于簇，。由拉格朗日定理可知，也就是求每个簇中样本的平均值。

基于距离的聚类算法适用于数据簇内密集、簇的大小相差不大、簇间稀疏的样本集。但如果初始点选择不当很容易陷入局部最优，中心点的选择是使聚类算法不稳定的一个重要原因，而且还需要预先确定聚类中心点个数。本论文通过结合距离和密度不断调整聚类中心点，有效的避免了算法陷入局部最优，使得中心点的选择较为稳定。

## 2.2 模糊C均值算法

模糊聚类的一个重要特点是每个样本聚类中心点的归属不是绝对的1或0，这更符合我们生活中的认知和规律。例如，以下雨来说，特大暴雨、暴雨、大雨、中雨、小雨，不同的人有不同的看法，于是就有了模糊集的概念，利用模糊相似矩阵进行划分。表示第个样本对聚类中心点的隶属度。

U表示隶属度矩阵，V表示样本到聚类中心点矩阵，表示模糊指数，当时，模糊聚类退化为硬聚类，m越大，模糊程度越高。在求解的最小值时，有一个约束条件即每个样本对每个聚类中心的隶属度总和为1

结合（1）和n个（2）根据拉格朗日定理求解的极小值，可得

 （5）

在模糊聚类中不同的，可得到不同的分类结果。常见的包括两种确定方法，一种以根据经验丰富的专家结合实际需求确定阈值，另一种根据统计量确定。

对于的分类数为c，第j类的样本数为，表示第j类样本标记，为第j类的聚类中心向量。统计量，遵从自由度为，的分布。值越大，说明簇和簇之间的距离越大，簇和簇之间的差异越大，分类效果越好，因此通过求的最大值来使得分类结果最好。

本论文引入密度到距离公式中，通过修正距离公式使算法更具有普适性。当样本对于不同簇的隶属度非常相似时，同一个样本可能划分到不同的簇，针对这种边界点，本论文结合特征权值的学习对边界点进行划分，使算法能够适用于分布较为均匀的样本集。

# 3 改进的FCM聚类算法

## 3.1 聚类中心点的选择策略

簇的中心点选择的基本思想：首先计算数据集中样本间的距离，生成相异度矩阵D，选择矩阵D中最小值和最大值。计算整个样本的平均密度和最近两个点的中心点v，选择初始阈值半径r，计算在半径内的平均密度，如果平均密度小于，则返回到选择距离次最小，此时选择的点很有可能是孤立点；如果平均密度大于或等于，距离增大一个，计算增大空间的密度，直到密度小于。接下来选择剩余点的最小值，依据上述步骤依次选择直到选择选择所有的点则结束。在选择的过程中，通过不断的调整距离最近的点可以让中心点更接近真实中心点，通过不断增加距离可以把和中心点关联度高的点选择进来。

为了修正在聚类过程中因为半径阈值过大导致有的类簇过大，使得有的样本和其他中心点距离远远小于该样本所属中心点，在增大半径阈值的时候计算和其他聚类中心的距离，当最外层样本到该样本中心点比到其他聚类中心点的最小距离大2倍则停止继续增大。

|  |
| --- |
| 基于改进的中心点选择策略步骤 |
| 输入：样本数据 |
| 输出：C个簇集 |
| 主要步骤：   1. 计算样本数据间的距离，生成相异度矩阵D，选择矩阵D中最大的值MAX； 2. 计算当前所有样本的平均密度，选择初始半径为，N为常数； 3. 选择D中最小的数值，并计算样本i和j均值作为中心点； 4. 计算增加空间的平均密度，如果<，返回到（3），并把这两个点作为噪声点处理；如果>=，半径增加一个r，计算新增加的点v和其他中心点距离，如果，则返回（3）；否则，返回到（4）；   （5） 直到所有的点被选择则结束。 |

图1 中心点选择策略

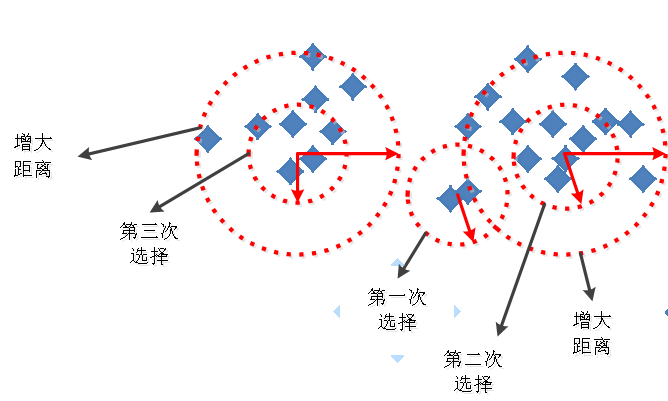


图2 中心点选择流程图

选择距离最近的两个，避免随机选择初始点带来的影响。如果选择的是一个簇集的边界点，通过密度判断这个中心点的选择是否恰当，通过不断增加距离半径，把其他的点包含进来。通过这种选择策略不仅能选择出大的簇集还能选择出小的簇集，也能判断点是不是孤立点。

## 3.2 模糊聚类

上述是结合距离和密度作为选择策略，能够发现大小不同的簇。为了在模糊聚类中实现同样的效果，采用是上述类似的方式，因此需要修改FCM算法中的距离公式

 （7）

对所有聚类中心点进行如下操作。以聚类中心点为圆心r为半径，如果在增大r的时候，增大面积的密度小于平均密度，并且样本还没有找到，则另=1；如果在增大的过程中有所求样本点，则另，可以在确定初始中心点之前就可以确定，首先需要对进行归一化处理，确保。修正后的目标函数为

 （8）

在原始的FCM算法中，越大，隶属度越小，越小，隶属度越大，所以隶属度只和点到类簇的中心点距离有关和密度没有关联。如果密度越大，越大，隶属度越大，密度越小，则隶属度越小，修改之后目标函数，更具有普适性。

在得到最终的隶属度矩阵后，有的簇集相对较大，这时边界点对簇集的隶属度已经相对较弱。为了优化边界点的归属，以边界点为中心点求在半径内包含的样本，求得半径内样本所属簇个数，然后计算在各个簇中所占比例。重新对簇和簇之间的边界点归属进行修正，选择方法采用公式

（9）

根据的大小调整样本对所有簇的隶属度，即最大的隶属度赋值给最大，把次最大的隶属度赋值给次最大的，依次类推，可得到新的隶属度矩阵。

基于上述分析，改进的FCM算法在聚类中的实现过程如下：

|  |
| --- |
| 改进的FCM算法执行步骤 |
| 输入：样本数据 |
| 输出：隶属度矩阵，聚类中心点 |
| 主要步骤：   1. 通过上述中心点选择策略，确定中心点c，阈值，迭代次数，最大迭代次数，0根据模糊理论一般选择模糊指数m=2，初始化隶属度矩阵； 2. 根据公式（3）和（8）计算样本隶属度； 3. 把隶属度带入到式（5）中，得到聚类中心点； 4. 如果或者，则程序停止，算法结束；否则另，跳转到步骤（2）； 5. 输出最终的隶属度矩阵和聚类中心向量； 6. 根据公式（9）对隶属度矩阵进行修正，得到新的隶属度矩阵； 7. 根据公式（6）求得，得到分类结果。 |

图3 改进的FCM算法在聚类中的实现过程

# 4 实验结果与分析

实验环境如下：Pentium(R) Dual-Core CPU E5700 @ 3.00GHXz 3.00GHz 4.00 GB RAM，操作系统：window 10，Program：Intellij idea。 为了验证论文所提方法的正确性，实验数据来源于加州大学UCI数据库，UCI数据库是专门为机器学习和数据挖掘提供实验测试数据。初始中心点选择的实验采用Iris数据集、Abalone数据集以及Wine数据集，常用的数据集Iris共有 150条数据，每个样本包括4个属性，被标记为3类（Iris-setosa，Iris-versicolo，Iris-virginica），每类分别有50条数据。为了测试改进方法能够发现孤立点，在数据集中添加三条干扰数据，为了能发现样本中的大簇集和小簇集，删除类别Iris-setosa中的25条，同时确保Iris-setosa不易过于稀疏。Abalone数据集共有4177条数据，每个样本包括8个属性，被标记为3类（F，I，M），每类分别有1307、1342、1527条数据。Wine数据集包括178条数据，每个样本包括13个属性，共分为3类（1，2，3），每类包括59、70和47条属性。

## 4.1 中心点选择实验

为了检测改进中心点选择策略的效果，使用聚类结果的错误率，离群点数量以及中心点数量来评价算法策略。改进的选择策略需要确定距离半径，为了确定最佳，一个简单选择策略另，D为样本距离矩阵，通过不断调整N，选择出最优N值。采用不同的数据集，对不同N值的改进的中心点选择策略的实验结果如图所示。

图1 聚类错误率

图2 iris数据集的离群点

图3 iris数据集的中心点

由图1图2可知，当N小于某一个值时，聚类错误率高达70%以上并趋于稳定，聚类中心点有不超过三个，但有一个簇集很大，几乎包含所有的样本，说明N较小时阈值半径较大，所有的样本都聚类到一个中心点。当N大于某一个值时，聚类错误率同样高达70%以上并一直增加，说明当N较大时阈值半径很小，每次半径内只有初始选择的两个点，很多点被认为是离群点，如图3所示。在Wine的实验中，当N=12时，错误率为，错误率远远低于其他算法。当N属于某一范围内时，聚类中心点为3个，有的样本被错误的识别为离群点，这类点往往距离簇中心点很远。在修改之后的Iris实验中，3个离群点被识别出来，改进后的算法对于离群点的识别虽然也很准确，但是识别出错的点往往也是被认为是离群点。当N较大时算法的迭代次数也明显增多，算法执行时间明显增多，而当N在某一范围内时，出错率普遍较低，同时算法迭代次数也是相对较低，所以选择。

为了测试基于距离和密度的中心点选策略的性能，对比实验采用k-中心点算法。样本类别已知为3类，所以选择聚类数目。上述实验可知，N值的取值范围应该在9和40之间，因此，随机从这里面选择N的值，k-中心点算法进行9次实验，实验结果如表【1】所示。

表1 k-中心算法和改进的基于距离和密度的算法

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 算法 | 迭代次数 | 平均聚类错误率% | 最小错误率% |
| 修改的Iris | k-中心点算法 | 9.44 | 24.00 | 11 |
| 改进的基于距离和密度算法 | 19 | 34.00 | 8 |
| Abalone | k-中心点算法 | 12.81 | 25.00 | 17 |
| 改进的基于距离和密度算法 | 20 | 33.00 | 7 |
| Wine | k-中心点算法 | 10.23 | 29.00 | 15 |
| 改进的基于距离和密度算法 | 13 | 35.00 | 2 |

虽然改进的算法迭代次数和平均聚类错误都要比k-中心要高很多，但是最小错误率对于改进算法是可知的，算法是稳定的，而k-中心算法对于最优聚类是未知的，聚类结果依赖初始中心点的选择。对于三组实验，改进算法的最小错误率远远小于k-中心算法。所以改进算法不仅保证了算法的稳定性，避免算法陷入局部最优解，同时减少了迭代的次数。另外，当不知道样本的分布情况时，无法确定分类数C，算法同样也能预测样本分布，所以算法更具有普适性。

## 4.2 改进的FCM算法

论文使用了上述四组数据集，

根据上文获得初始聚类中心点，聚类中心点根据簇中样本求其均值。为了更直观的分析改进的FCM算法的效果，对比实验采用传统的FCM算法，实验采用人造数据，数据为二维数据。另算法迭代终止阈值，如表【】所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

明星数据集共有1400个样本，每个样本包含【】属性，明星数据集没有标注所属类别，由于实际对明星票房影响力进行分析时，无法确定哪个明星确定属于哪一类。实验结果的准确性只能通过专家评估，或者以问卷的形式进行调查统计，然后对比实验结果，计算误差。无论是专家评估还是调查统计，这个方式本身就存在一定的误差。

**结束语** 本文解决了随机选择初始中心的缺点，能够很好的发现离群点，得到的初始聚类中心点较为准确，与真实的样本分布较为一致，这也使得FCM算法收敛速度很快，效果良好。改进后的FCM算法同时考虑了距离和密度两个因素，k-means聚类仅仅考虑距离，所有的簇集大小一样，不符合我们真实的样本，引入密度，使得样本不仅仅可能属于距离近的簇，同时也可能属于密度大的簇。通过特征权值的学习、再聚类解决了临界点归属问题，但也增加了算法的复杂度。

参考文献

[1]