一种改进的模糊聚类算法

**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。

关键词：目标函数；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

FCM聚类算法是模糊聚类算法中最重要的算法之一，该方法通过最小化其目标函数获得样本集的模糊划分或聚类。由于模糊聚类能够适应分离性不是很好的样本数据，得到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，如今已经在信息技术和控制决策领域得到了广泛的应用。但FCM聚类算法仍存在和硬聚类一样的缺点，对初始聚类中心点的敏感、有噪声时分类效果差、容易陷入局部最优等[1]。

由于FCM聚类对初始中心点的敏感，因此在聚类之前选好初始中心点尤为重要。张慧哲[2]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簇的中心，同样影响聚类结果的准确性。谢娟英[3]提出一种选择位于样本分布密集区且相距较远的前K个样本作为初始聚类中心点，选择使聚类误差平方和最小的数据对象更新聚类中心，提高了聚类精确度，减少了迭代次数，但是缺乏一定的理论指导。朱晔等人[4]提出了一种基于密度信息的k-中心点算法，通过获取样本密度信息来选取初始中心点，有效解决了聚类结果对初始中心点选择的依赖性问题。

为了提高聚类效果，研究者越来越关注对目标函数的改进，对目标函数中的距离公式的改进、算法收敛速度的改进、对模糊划分空间的改进等。钱云涛和赵荣椿等[5]在目标函数中融合了图论定义了新的目标函数，提高了算法的鲁棒性。fuHua[6]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。本论文主要研究的是对目标函数中的距离公式进行改正，通过引入样本局部密度修正目标函数来解决簇集大小不同，并且使得改进的算法适用于分布相对均匀的样本。

电影明星是对电影票房有重要贡献的一个群体，不同的明星对票房的影响往往又有很大的不同。分析明星影响力是对电影票房预估的重要手段，可以为影视公司提供重要的决策支持，有的明星对电影票房起到积极的作用，有的明星可以对票房起到消极的作用 [7]。因此，对明星票房影响力进行分类有重要的商业价值。

# 2 FCM聚类算法

## 2.1 初始中心点选择

首先，聚类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

 （1）

算法的结束准则是使目标函数最小。即当不属于簇，则另，当属于簇，。由拉格朗日定理可知，中心点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 FCM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。

 （2）

U表示隶属度矩阵，V表示样本的聚类中心点矩阵，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所有聚类中心的隶属度总和为1。

 （3）

结合（1）和n个（2）根据拉格朗日定理求解的极小值[8]，可得隶属度和。

 （4）

 （5）

在模糊聚类中不同的，可得到不同的分类结果。常见的有两种确定方法，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。

 （6）

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 改进的聚类中心点选择策略

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

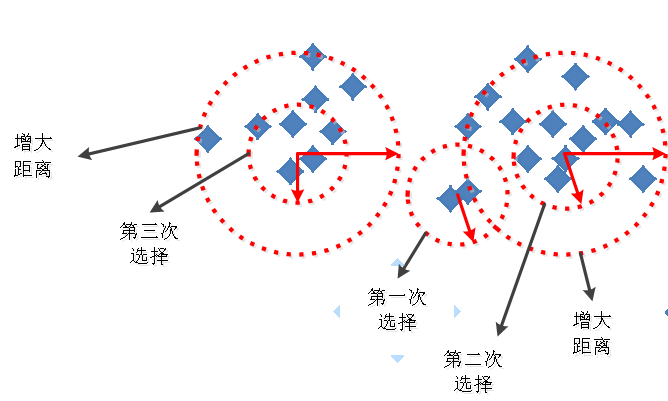


图2 中心点选择流程图

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，另外，这种方法还可以确定簇的个数。

## 3.2 改进的FCM算法

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，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。

因为是根据前一个聚类中心点计算出来，可计算一个密度矩阵。结合公式（3）和公式（8），利用拉格朗日极值定理，可求得隶属度矩阵。

 （9）

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

 （10）

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. 首先计算样本局部密度矩阵； 3. 根据公式（8）和（9）计算样本隶属度； 4. 把隶属度带入到式（5）中，得到聚类中心点； 5. 如果或者，则程序停止，算法结束；否则另，跳转到步骤2）； 6. 输出最终的隶属度矩阵和聚类中心向量； 7. 根据公式（10）对隶属度矩阵进行修正，得到新的隶属度矩阵； 8. 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，评价指标使用聚类结果的错误率，既聚类错误的样本占所有样本的百分比。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4所示。

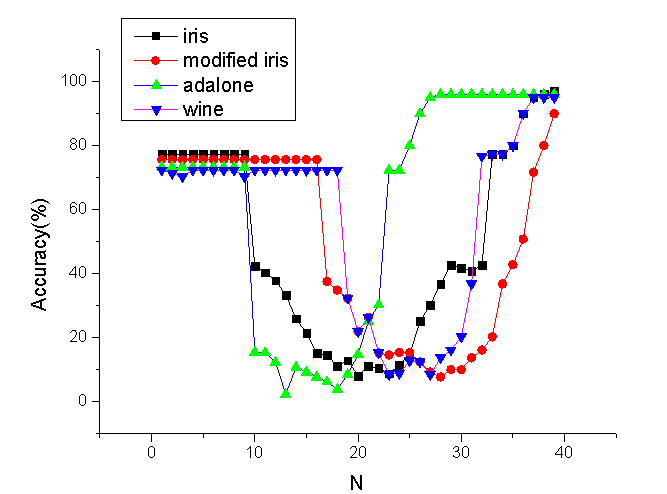


图4不同数据集的聚类错误率

由图4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，聚类中心点不超过三个，说明有一个簇集很大，几乎包含所有的样本。原因是当N较小时阈值半径较大，几乎所有的样本都聚类到一个中心点。在数据集Wine的实验中，当N=12时，错误率在四组实验中错误率最低为。

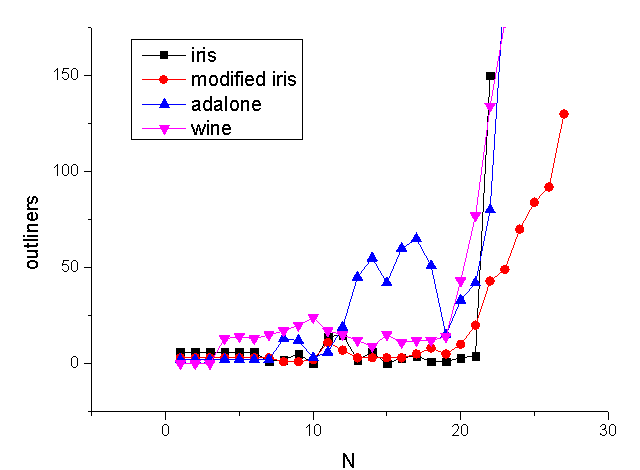


图5 不同数据集离群点个数

结合图4和图5所示当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，经过多次选择之后，大部分点都被认为是离群点，所以离群点个数随着N的增大越来越多。

图6 iris数据集中心点个数

如图6所示，当N在8到23之间时，聚类中心点为3个，结合图4这在个区间聚类的错误率也是较低。在修改之后的Iris实验中，添加的3个离群点也被正确的识别出来，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，而不是被错误的分类到其他簇集。

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 | 19.22 | 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 | 20.90 | 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 | 13.14 | 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

**实验1** 人造数据

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，包含148条样本，被标记为3类（1，2，3），每类分别有18、50和80条数据。算法迭代终止阈值， 实验结果，如图7到图9。

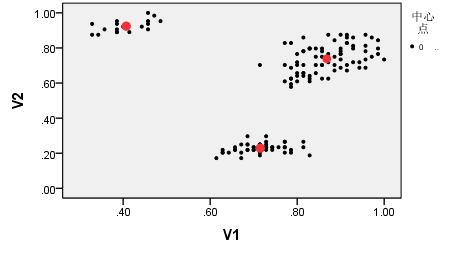


图4 原始数据中心点坐标图

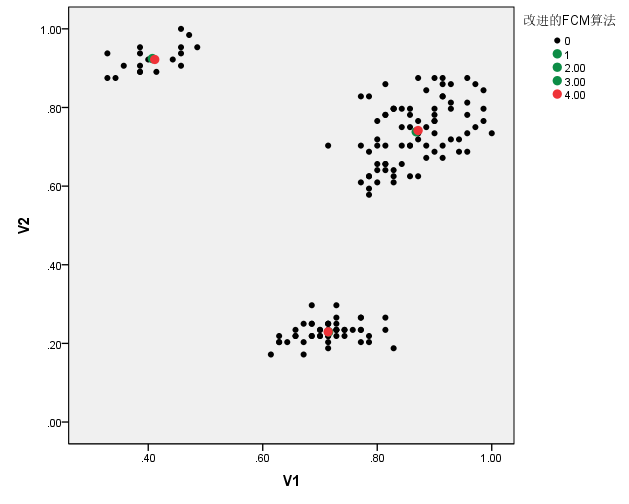
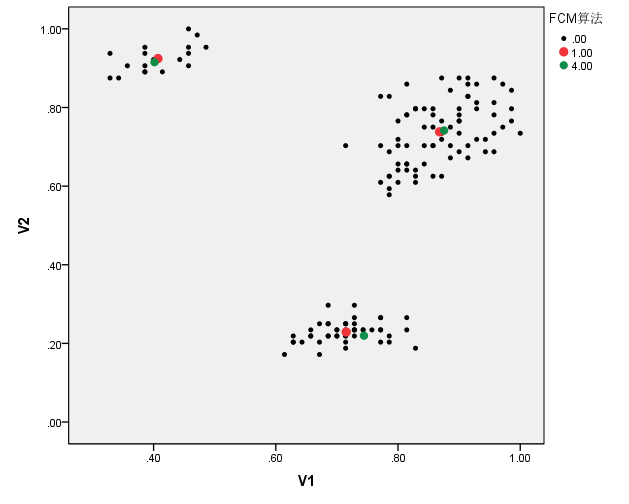


图5 FCM算法中心点坐标图 图6 改进的FCM算法中心点坐标图

由图6可知，改进的FCM算法的聚类中心点的坐标和原始坐标几乎重合，根据表2原始中心点和改进FCM的误差值非常小，而原始的FCM算法的中心点1和中心点3的坐标和原始中心点坐标非常接近，但中心点2和原始中心点坐标相差较大。改进的FCM算法比原始的FCM算法的聚类效果要好。

表2 中心点坐标

|  |  |  |  |
| --- | --- | --- | --- |
|  | 原始中心点坐标 | FCM算法 | 改进的FCM算法 |
| 中心点1 | （0.407，0.924） | （0.399，0.920） | （0.401，0.915） |
| 中心点2 | （0.715，0.229） | （0.744，0.220） | （0.714，0.220） |
| 中心点3 | （0.870，0.738） | （0.873，0.742） | （0.875，0.742） |

**实验二** 4.1中4组数据集

为了进一步检验FCM算法效果，实验中主要评价指标包括：均方差MSE

 （10）

其中为真实的聚类中心，为所计算的聚类中心；另外一个指标是分类正确率，也就是分类正确的样本占所有样本的百分比。数据集采用上文用的4组数据Iris、修改之后的iris、Wine、Abalone，聚类结果如表3所示。

表3 聚类结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 方法 | MSE | 分类正确率% | 迭代次数 |
| Iris | FCM算法 | 0.139943 | 84.3589 | 11 |
| 改进的FCM算法 | 0.132665 | 87.4258 | 4 |
| 修改之后的iris | FCM算法 | 0.144415 | 80.3548 | 5 |
| 改进的FCM算法 | 0.124710 | 82.1592 | 7 |
| Wine | FCM算法 | 0.115213 | 89.0102 | 22 |
| 改进的FCM算法 | 0.105275 | 91.6362 | 5 |
| Abalone | FCM算法 | 0.332467 | 73.4170 | 20 |
| 改进的FCM算法 | 0.204017 | 82.9382 | 9 |

根据表3可知，改进的FCM算法的MSE比普通的FCM算法要低，分类正确率也高。iris和Wine数据集聚类相差不多，因为这两个数据集簇和簇之间相差较远，并且簇的密度较高，没有离群点。修改之后的iris数据集，比修改之前的FCM均方差增大，原因是在iris里面增加了离群点。Abalone数据集，分类正确率明显提高，原因是Abalone分布相对均匀，簇和簇之间较近，同时簇的大小不同，导致原始的FCM算法的分类正确率较低。因为初始中心点选择方法的效果较好，使得改进之后的算法迭代次数明显降低，虽然修改之后的iris数据集的迭代次数比改进的要好，但普通的FCM算法的初始中心点选择是随机的，在实验中会发生随机选择的初始中心点和聚类结果的中心点比较接近。

**实验三** 明星票房影响力数据

实验数据来源于微博和时光网。实验目标的是想根据明星的相关数据对明星在票房上的影响力进行分类。实验数据一共200条，演员每条数据包含7个属性，属性为演技评分、评价人数、电影平均评分、电影平均评价人数、电影平均评分、最近主演的3部电影平均票房、粉丝数，特征选择是根据信息论中的信息熵的大小来确定。从实际角度考虑出发，明星票房影响力分为4类较为合适，第一类对上映电影票房有很大影响的明星，第二类对电影票房有正作用的明星，第三类对电影票房没有影响的明星，第四类是对电影票房起副作用的明星。部分数据如表4所示。

表4 部分明星数据

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | Score | Evaluate | Avg\_filmScore | Avg\_userCount | Avg\_atCount | Avg\_officebox | Fans |
| 鹿晗 | 5.7 | 6735 | 6.47 | 4998.67 | 2756.33 | 250000000 | 175369 |
| 杨幂 | 5.5 | 5687 | 5.8 | 5593.84 | 2165.71 | 170000000 | 645201 |
| 姜文 | 9 | 5186 | 7.47 | 6478.94 | 2671.18 | 550000000 | 715782 |
| 范冰冰 | 6.6 | 4722 | 6.53 | 5886.08 | 1372.6 | 166666667 | 552010 |
| 赵薇 | 7.5 | 4591 | 6.98 | 6736.78 | 1827.61 | 433333333 | 625841 |
| 黄晓明 | 4.9 | 3345 | 6.35 | 5852.64 | 1561.29 | 216666667 | 478882 |
| 章子怡 | 6.7 | 3123 | 6.62 | 11106.06 | 1858.81 | 43940000 | 287873 |
| 黄渤 | 8.8 | 3084 | 6.92 | 9669.29 | 2519.45 | 1000000000 | 312581 |
| 刘亦菲 | 7.3 | 2771 | 6.23 | 5051.65 | 1695.65 | 78650000 | 452810 |
| 邓超 | 7.5 | 2331 | 6.8 | 8427.88 | 2379.88 | 500000000 | 775362 |

采用改进的初始中心点选择策略，确定初始中心点坐标。结合初始中心点坐标，使用改进的FCM算法，另算法迭代终止阈值。算法共迭代44次，分类系数为0.30144。隶属度矩阵部分数据如表5所示。为了验证明星分类结果的准确率，对实验的200个明星采用专家标记分类，第一类影响力较高的明星32个，第二类影响力一般的明星85个，第三类影响力可忽略的明星61个，第四类影响力起副作用的明星22个。

表5 明星隶属度

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 鹿晗-3 | 0.2272-0 | 0.3996-1 | 0.2243-0 | 0.1489-0 | 陈坤-1 | 0.1798-0 | 0.1946-0 | 0.2834-0 | 0.4022-1 |
| 杨幂-4 | 0.2480-0 | 0.2151-0 | 0.2741-0 | 0.2628-0 | 邓超-1 | 0.1392-0 | 0.2062-0 | 0.2366-0 | 0.4180-1 |
| 姜文-1 | 0.1571-0 | 0.1012-0 | 0.1789-0 | 0.5629-1 | 孙红雷-2 | 0.0983-0 | 0.1056-0 | 0.5564-1 | 0.2396-0 |
| 汤唯-2 | 0.1821-0 | 0.2296-0 | 0.3025-1 | 0.2859-0 | 白百何-2 | 0.1256-0 | 0.3165-1 | 0.1633-0 | 0.3946-1 |
| 范冰冰-2 | 0.1823-0 | 0.2189-0 | 0.3142-1 | 0.2846-0 | 李冰冰-2 | 0.1299-0 | 0.1919-0 | 0.2357-0 | 0.4426-1 |
| 赵薇-2 | 0.2246-0 | 0.2040-0 | 0.2904-0 | 0.2811-0 | 徐峥-1 | 0.1017-0 | 0.1482-0 | 0.2758-0 | 0.4744-1 |
| 周迅-2 | 0.1924-0 | 0.2048-0 | 0.3168-1 | 0.2960-0 | 包贝尔-4 | 0.3812-1 | 0.3067-1 | 0.1186-0 | 0.1935-0 |
| 黄晓明-2 | 0.1568-0 | 0.2404-0 | 0.3701-1 | 0.2328-0 | 葛优-2 | 0.0098-0 | 0.2615-0 | 0.3664-1 | 0.3622-1 |
| 章子怡-2 | 0.2997-0 | 0.1627-0 | 0.2758-0 | 0.2619-0 | 景甜-4 | 0.6025-1 | 0.2694-0 | 0.1087-0 | 0.0194-0 |
| 黄渤-1 | 0.1442-0 | 0.1221-0 | 0.2181-0 | 0.5157-1 | 王宝强-1 | 0.0884-0 | 0.2039-1 | 0.3001-0 | 0.6077-1 |
| 刘亦菲-3 | 0.3244-1 | 0.1584-0 | 0.1555-0 | 0.2618-0 | 张涵予-1 | 0.1207-0 | 0.0976-0 | 0.3638-1 | 0.4179-1 |

根据实验结果可知，大部分明星都是隶属于一类，有少部分的明星隶属于两类或者不属于任何一类。明星名字后面的数字表示标记分类，隶属度矩阵中“-”前面的数字表示隶属度，后面的“1”表示隶属于某个类别，“0”表示不属于这个类别。正确率如表6所示。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 每类总数 | 错误数 | 不属于任一类 | 隶属两个类 | 正确率（%） |
| 第一类 | 32 | 3 | 0 | 2 | 92.625 |
| 第二类 | 85 | 17 | 4 | 6 | 80.000 |
| 第三类 | 61 | 13 | 3 | 5 | 79.689 |
| 第四类 | 22 | 4 | 2 | 2 | 81.818 |
| 总计 | 200 | 37 | 9 | 15 | 81.500 |

由上表可知，有9个明星没有分类，15个明星隶属于两个或两个以上分类，13个明星分到其他类中，其中第一类的分类正确率最高。根据实验结果可知，第一类明星的演技评分、电影评分、票房、粉丝数都较高；第二类的明星的票房、粉丝数较高，但是演技和电影评分中等；第三类明星所有的特征属性都是中等；第四类明星的演技评分、电影评分和票房都较低，但是粉丝数都较高。没有分类的明星例如杨幂的特征属性演技和电影评分都较低，粉丝较多，票房中等。隶属两个类的明星和第一类相比，要么是票房略低，或者电影评分略低。

**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，同时引入密度和距离，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，但也增加了算法的复杂度。对于明星票房影响力的分析，挖掘了有一定价值的信息，但根据专家经验验证实验结果，是由主观验证客观，接下来需要寻找更好的验证方法。

# 参考文献

1. 陈晓勇, 顾晖, 彭志娟. 数据挖掘中K-均值聚类算法的缺陷及工作效率改进的实验研究[J]. 科学技术与工程, 2013, 13(34):10359-10363.
2. 张慧哲, 王坚. 基于初始聚类中心选取的改进FCM聚类算法[J]. 计算机科学, 2009, 36(6):206-209.
3. 谢娟英, 郭文娟, 谢维信. 基于邻域的K中心点聚类算法[J]. 陕西师范大学学报:自然科学版, 2012, 40(4).
4. 朱晔, 冯万兴, 郭钧天,等. 一种改进的k-中心点聚类算法及在雷暴聚类中的应用[J]. 武汉大学学报:理学版, 2015, 61(5):497-502.
5. Qian Y T, Zhao R C. Robust clustering based on global data distribution and local connectivity matrix[C]// IEEE International Conference on Intelligent Processing Systems. IEEE, 1997:1629-1633 vol.2.
6. Yu F, Xu H, Wang L, et al. An Improved Automatic FCM Clustering Algorithm[C]// International Workshop on Database Technology and Applications, Dbta 2010, Wuhan, Hubei, China, November 27-28, 2010, Proceedings. 2010:1-4.
7. Liu Y. Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue[J]. Journal of Marketing, 2006, 70(3):74-89.
8. Wang X, Wang C, Shen J. Semi–supervised K-Means Clustering by Optimizing Initial Cluster Centers [J]. Lecture Notes in Computer Science, 2011, 6988:178-187.