一种改进的模糊聚类算法在明星票房影响力上的应用

**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。

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

“物以类聚，人以群分”，所以聚类是一个古老的问题。传统的聚类方法是把每个需要处理的数据对象严格地归属于某一个类，在这类方法中，隶属度或是1或是0。但在现实世界中有很多划分并没有严格的界限，例如说一个明星是“一线明星”，这是一个模糊概念，不仅很难给出一个严格的界限划分条件，而且看法也会因人而异。但这也更符合我们生活中的认知。为了解决这类问题，人们将模糊集合理论应用到聚类算法研究实践中。模糊聚类是通过隶属函数来确定每个数据隶属于各个簇的程度，而不是将一个数据对象硬性地归类到某一簇中。目前已有很多关于模糊聚类的算法被提出，如FCM（fuzzy c-means algorithm）算法 [3]。FCM聚类算法已经在医疗图像分割领域有了广泛的应用[6]。

但FCM算法也有很多缺陷。当数据的分离性不好时，随机选择初始中心点的缺点更加突出，不仅迭代次数增多而且聚类结果的准确性也大大降低。另外，当簇和簇之间的数据量相差较大，且不同簇之间的距离相距较近时，容易导致小簇集的中心点不断向大簇集靠近，影响实验结果的准确性。基于上述考虑，论文提出了一种改进的FCM聚类算法，对算法的初始中心点选择进行充分的考虑，同时为了解决数据分离性差、簇的大小不同导致的分类不准确，论文改进了局部密度在模糊聚类中的使用，提高了算法的寻优能力。

# 2 基于划分的FCM聚类算法

FCM算法是由Dumm第一个提出的，并且由Bezdek将其发展和推广。FCM模糊聚类算法是一种迭代优化算法，可以描述为最小化指标函数。

定义一：样本集合X。集合为被分类样本集合，n表示样本数，样本可以是多维向量。

定义二：模糊分类矩阵。，c表示分类数，表示第个样本对聚类中心点的隶属度。

定义三：聚类中心点坐标矩阵。，i， m表示样本维度。

定义四：目标函数。为了得到最优模糊分类，Bezdek定义一个目标函数。

 （1）

m表示模糊指数，且m>1，当时，模糊聚类退化为硬聚类，m越大，模糊程度越高，通常取值m=2[2]。表示样本i和样本k之间的距离。

定义五：约束条件。每个样本对所有簇的隶属度总和为1。

 （2）

定义六：隶属度和聚类中心点坐标。结合（1）和n个（2）根据拉格朗日定理求解的极小值[6]，可得隶属度

 （3）

以及聚类中心点

 （4）

其中表示聚类迭代次数。

定义七：最佳阈值。求得隶属度矩阵之后，选择不同的，可得到不同的分类结果。常见的有两种确定最佳阈值的方法，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，如公式（5）。

 （5）

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。

普通FCM算法的隶属度依赖于中心点坐标，而最终的中心点坐标受初始中心点坐标选择的影响，因此恰当的初始中心点坐标对聚类结果有较大的影响。当数据分离性较差、簇的大小不同且相距较近时，基于欧式距离求解隶属度矩阵不适用这类数据分类。因此，论文提出了改进的初始中心点的选择策略，使得FCM算法在初始化隶属度矩阵时，不是随机选择数字进行初始化。同时又提出了改进的FCM算法，改进局部密度的计算方式使得FCM算法能够较好的对这类数据进行分类。

# 3 改进的FCM聚类算法

改进的FCM聚类算法主要是对初始中心点选择的改进以及对目标函数的改进。改进的初始中心点选择的目的是避免随机选择造成的聚类算法的不稳定性以及收敛到局部极小的情况。

## 3.1 基于局部密度的初始中心点选择

初始中心点选择的基本步骤是：根据样本集X生成相异度矩阵，选择其中最大值max作为直径，计算所有样本的平均密度。选择距离最近的两个样本点a和b，然后计算两个点的中心点，选定阈值半径r，把半径r内的局部密度和相比较，如果局部密度大于平均密度，增大一个r，此时半径为2r，计算r和2r内的局部密度，直到局部密度小于平均密度，输出中心点坐标，并把半径内的样本设置为已访问。再次选择D中的次最大值，以此类推，直到所有的点都是被设为已访问。

为了防止在聚类过程中因为半径过大导致有的簇过大，使得簇中有的样本和其他中心点距离远远小于该样本到所属簇的中心点距离，在增大半径r时，增大的面积内的样本集合到距离比到其他中心点的最小距离大2倍， 表示除了j点之外的所有中心点，即，则停止继续增大阈值半径，输出中心点坐标。

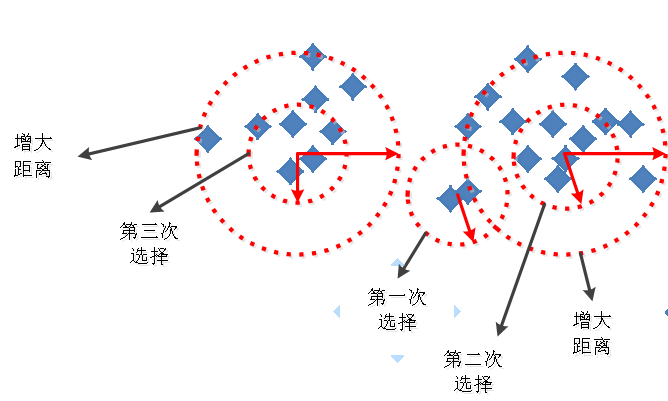


图1 中心点选择流程图

图1为二维数据初始中心点选择过程。改进的中心点选择策略是有目的选择中心点，避免了随机选择。因为要计算局部密度，算法的时间复杂度要比随机选择要高，但较好的选择初始中心点也能减少后来的FCM算法的迭代次数。

## 3.2 基于距离和局部密度的隶属度计算

FCM算法过程是基于目标函数方法寻找最优分类，本文提出了一种目标函数，综合考虑样本到聚类中心的距离和样本局部密度对聚类结果的影响。这样可以保证每一类数据的紧密型的同时确保了各个簇之间的分离度。新的目标函数为

 （6）

该局部密度的思想是如果样本a到多个簇中心点的距离相近，但这些簇的密度不同，对于其中密度较大的簇A，表明样本a和簇A的关联性更高，相应的隶属度也应该更大。另外当多个簇大小不同且距离较近时，簇与簇之间的样本如果仅仅考虑距离，聚类时小簇的中心点会向大簇集偏移，如果加入密度，如果大簇的密度比小簇的密度小，样本对大簇隶属度变小，因此可以避免小簇中心点偏移。

的计算方式：根据初始中心点选择中所述的局部密度的计算方式，可得局部密度矩阵。如果局部密度小于平均密度，则令=1，聚类时该样本只受距离影响；如果密度大于平均密度，令。

改进的目标函数加入了局部密度。它描述的是在计算隶属度的过程中，样本局部密度在聚类过程中的影响。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簇的中心点距离有关。加入局部密度后，如果密度越大，隶属度越大，密度越小，则隶属度越小。

结合公式（2）和公式（6），利用拉格朗日极值定理，可求隶属度矩阵。

 （7）

## 3.3 改进的FCM算法

改进的FCM算法在聚类中的实现过程如下：

|  |
| --- |
| 输入：样本数据 |
| 输出：隶属度矩阵，聚类中心点坐标矩阵 |
| 主要步骤：   1. 计算所有样本的平均密度；选定阈值半径R，令r=R； 2. 计算相异度矩阵，为矩阵增加标志位flag=false； 3. 选择D中最小的值，把两个点的中心点作为圆点，计算半径r内的密度； 4. 如果<，令r内的点flag=true，如果flag都为true，则执行6），如果仍有false执行3）；如果>=，执行5）； 5. 令半径，计算R-r和R之间的局部密度。如果且，令R-r和R之间的点flag=true，如果flag都为true，则执行6），如果仍有false执行5）；如果不满足且，输出，执行4）； 6. 输出初始中心点坐标矩阵和局部密度矩阵； 7. 确定阈值，算法最大迭代次数，算法迭代次数=0； 8. 如果=0，把和代入到公式（7）计算样本隶属度；如果，以中的点为圆心，计算局部密度矩阵；结合和代入到公式（7）计算样本隶属度； 9. 把隶属度带入到公式（4）中，得到新的聚类中心点坐标，更新； 10. 如果或者，则程序停止，算法结束；否则令，跳转到步骤9）； 11. 输出隶属度矩阵和聚类中心点坐标矩阵； 12. 根据公式（5）求得，得到分类结果。 |

图2 改进的FCM算法在聚类中的实现过程

改进的初始中心点选择策略避免了随机选择中心点缺陷造成聚类算法结果不稳定，虽然也带来了一定的时间开销，但传统的FCM聚类算法随机选择中心点导致算法迭代多次，而改进的中心点选择策略由于中心点选择较为接近真实中心点，最后使得FCM算法的迭代次数也明显减少。改进之后的FCM算法改变了局部密度的计算方式，算法不仅能够较好的解决分离性不好的数据，同时还能较好的发现簇大小不同且相距较近的样本集，但因为要计算局部密度，算法时间复杂度也相应增加。

# 4 实验设计与结果分析

为了验证论文所提方法的正确性，本文设计了两组实验，一组用来验证中心点选择策略的合理性，另一组用来验证改进的FCM算法的实际应用中的有效性。

## 4.1 初始中心点选择实验

实验数据来源于专为机器学习和数据挖掘提供实验测试数据的加州大学UCI数据库。其中采用Iris数据集、Abalone数据集以及Wine数据集[10]。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。

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和迭代次数，错误率既聚类错误的样本占样本总数的百分比。改进的选择策略需要确定距离半径，为了确定最佳，一个简单选择策略令，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3所示。

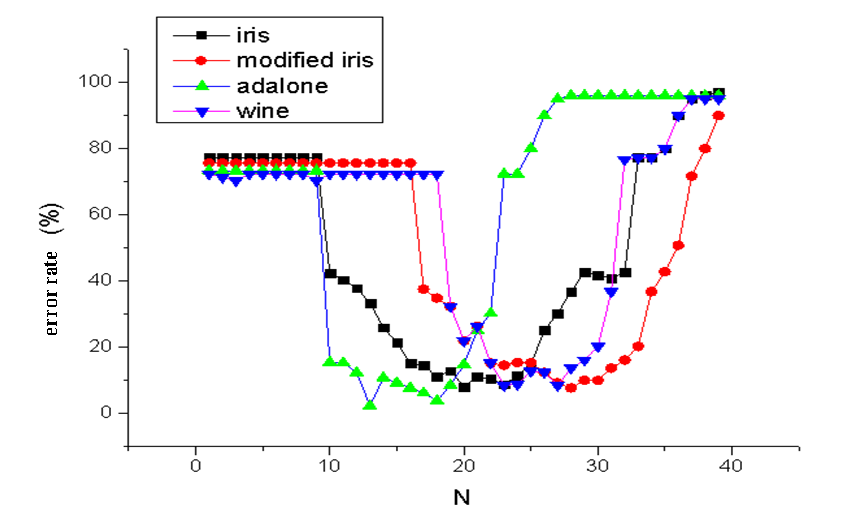


图3不同数据集的聚类错误率

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。

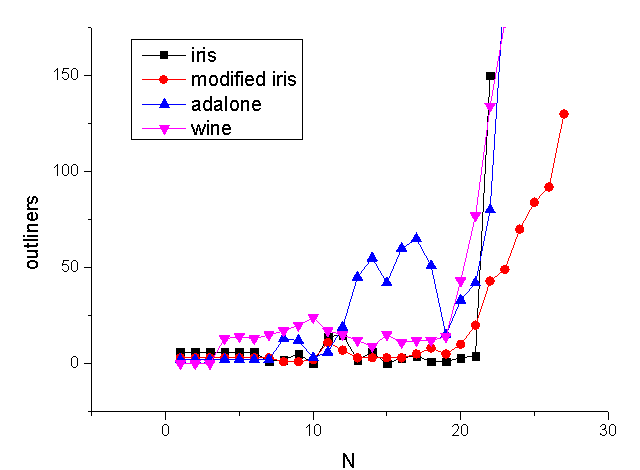


图4 不同数据集离群点个数

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5 iris数据集中心点个数

如图5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

**实验一** UCI数据实验

为了检验FCM算法效果，实验中主要评价指标包括：均方差MSE和Purity方法。MSE定义如下

 （8）

其中为真实的聚类中心，为所计算的聚类中心，MSE表示计算得出的聚类中心点和真实聚类中心点的距离。

Purity定义如下

 （9）

Purity方法表示分类正确的样本占样本总数的比例，其中是聚类中心点的集合，表示第j个样本，N表示样本总数。

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1 聚类结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 方法 | MSE | Purity (%) | 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。

**实验二** 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2所示。

表2 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，令算法迭代终止阈值。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3 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 |

由表3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4所示。

表 4 分类准确率

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

由表4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。

# 5. 结论

本文解决了随机选择初始中心的缺点，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，所有的簇大小一样，但并不适用于明星数据集。改进之后的算法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，是主观验证客观，缺乏客观性。

# 参考文献

1. Liu Y. Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue[J]. Journal of Marketing, 2006, 70(3):74-89.
2. Dellarocas C, Zhang X, Awad N F. Exploring the value of online product reviews in forecasting sales: The case of motion pictures[J]. Journal of Interactive Marketing, 2007, 21(4):23-45.
3. 孙吉贵，刘杰，赵连宇．聚类算法研究[J]．软件学报，2008，19(1)：48-61。
4. J. C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters[J]. Journal of Cybernetics, 1973, 3(3):32-57.
5. James C. Bezdek†. Cluster Validity with Fuzzy Sets[J]. Journal of Cybernetics, 1973, 3(3):58-73.
6. Bezdek J C. Pattern recognition with fuzzy objective function algorithms /[M].NY:Plenum, 1981.
7. 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.
8. 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.
9. 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.
10. FISHER R A． Iris Plants Database//http：//www.ics.uci.edu/~mlearn /MLRepository．Html．Authorized license。