一种基于局部密度与距离的改进的FCM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：针对模糊C-均值（FCM）聚类算法对初始中心选择的敏感、数据分离性差和簇的大小不同导致的分类不准确，提出了一种改进的FCM算法。新算法结合距离和局部密度较好选择了初始中心点，并根据样本局部密度特性改进了目标函数，弥补了FCM聚类算法的不足。实验结果证明新算法与标准FCM算法相比，迭代次数更少，准确度也有较大提高。把算法应用到明星票房的影响力的聚类分析中，实验结果表明了改进算法的有效性。

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”，聚类是一个古老的问题。传统的聚类方法是把每个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[1]。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，提出了模糊聚类的概念和方法[2]。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。FCM聚类算法已经在医疗图像分割领域有了广泛的应用[3]。

Birds of a feather flock together, clustering is an age-old question. Traditional clustering techniques divide every data-object to a certain class. In these methods, membership degree is either 1 or 0 [1]. But in the real world, many categories have no formal division, for example, a movie star who is called as A-lister. It is a vague notion about A-lister, not only because it’s difficult to get rigid boundary to divide, but the opinion varies with different individuals. But it is in line with our cognition in life. To get around this problem, people apply fuzzy set theory to the study of clustering algorithm, the concept and method of fuzzy clustering proposed [2]. Fuzzy clustering through membership function ensures the degree of each data affiliated with each cluster, rather than rigidly categorizing data-object into a certain cluster. There are a lot of fuzzy clustering algorithm was proposed, such as FCM (fuzzy c-means algorithm) algorithm. FCM algorithm has a wide range of applications in medical image segmentation field [3].

But the FCM algorithm also has some defects. When the data has a poor separability,the disadvantage of random selection of initial centers become more prominent, not only the number of iterations increasing, but also  the accuracy of clustering greatly decreasing. Moreover, When the amount of data between clusters is. Out of the above considerations, paper puts forward an improved FCM clustering algorithm，which takes a full consideration about the selection of initial centers. And, because of poor separability and the difference of the size of clusters, paper improved the use of local density in fuzzy clustering, enhancing the ability of getting the optimal solutions.

但FCM算法也存在一些缺陷。当数据的分离性不够好时，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，同时为了解决因数据分离性差、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

# Fuzzy Partitional Clustering Algorithms

FCM clustering algorithm was first proposed by Dunn in 1973, and Bezdek developed and promote it [5] [6]. FCM clustering algorithm is a kind of stochastic search iterative algorithm, which can be described as the minimum performance index function. The core idea is that the similarity of different objects divided into the same cluster is to be maximum, and the similarity among different clusters is to be minimal, with value between 0 and 1 for each sample points to determine the degree of membership belong to different clusters.

For the purpose of these requirements, the following definitions apply:

Definition 1 Sample Set X. is sample set, n standing for the number of sample and standing for multi-dimensional vector.

Definition 2 Fuzzy membership matrix . . c stands for the number of center points, and stands for membership degree of the element in the clustering center in the data collection.

Definition 3 Matrix of clustering center . ，.

Definition4The objective function of FCM clustering algorithm. In order to get the optimal fuzzy classification, Bezdek defined an objective function .

 （1）

m stands for fuzzy index, and m>1. When m=1, fuzzy clustering degenerates to HCM clustering. The higher the value of m, the higher of fuzzy degree. stands for the distance of between and .

Definition 5 Constraints. The degree of membership of each sample to all clusters is 1.

FCM算法是由Dunn于1973年第一个提出的，并且由Bezdek将其发展和推广[5][6]。FCM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。其思想是使得被划分到同一簇的对象之间相似度最大，而不同簇之间的相似度最小，每个样本点用值在0和1间的隶属度来确定其属于各个簇的程度。

为了便于理解，给出相关定义如下：

定义1 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2 模糊分类矩阵。，c表示中心点个数，表示第个样本对聚类中心点的隶属度。

定义3 聚类中心点坐标矩阵。，。

定义4 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[7]。表示样本i和样本k之间的距离。

定义5 约束条件。每个样本对所有簇的隶属度总和为1。

 （2）

定义6 隶属度和聚类中心点坐标。结合公式（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

 （3）

以及聚类中心点

 （4）

其中表示聚类迭代次数。

定义7 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经验结合实际需求确定阈值，另一种根据统计量确定[9]，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类的聚类中心坐标。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。

Definition 6 Degree of membership and coordinate of clustering center. Combining formulation (1) with a formulation (2) of size n, we can get minimization of according to Lagrange theorem. The degree of membership is

 （3）

And the clustering center is

 （4）

: The number of iterations of the algorithm.

Definition 7 An optimal threshold. After the membership degree matrix is obtained, choosing a different value of, the different classification results will be obtained. Two ways to determine the optimal threshold, one to determine the threshold value is by expert experience combined with the actual demand; another is by statistic F. F is defined as the following:

 （5）

: The number of samples of of cluster. : Sample tag of cluster. : Center coordinates of cluster. The great the value of F is, meaning the distance between cluster and cluster great, the better the classification is, so calculate the maximum value F to make the best classification results.

在FCM算法中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

The improved FCM clustering function is mainly to improve the selection of initial centers and the objective function. The improvement of the selection of initial centers solve the instability of random selection and converges to a local minimum easily.

改进的FCM聚类算法主要是对初始中心点选择以及对目标函数的改进。初始中心点选择的改进避免了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

## 3.1 selection of initial centers based on local density

In this paper, the basic idea of the center of the initial selection strategy based on the density is to make the distance between clustering center bigger, and make density of each cluster greater than defined threshold which is the average density of the sample set. So in the clustering algorithm we can calculate the clustering center in multiple feasible regions.

本文提出的基于密度的初始中心点选择策略其基本思想是尽量使各类初始中心点之间的距离较大，同时尽量使各个簇的密度大于所设定的阈值即样本集的平均密度。这样在接下来的聚类算法中就可以在多个可行域内对聚类中心进行求取。

初始中心点选择的基本步骤是：根据样本集X生成距离相异度矩阵，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D中最小的值对应的两个样本点分别为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中的次最小值，以此类推，直到所有的点都是被设为已访问。

图1为二维数据初始中心点选择示意图。为了防止在聚类过程中因为半径过大导致有的簇过大，使得簇中有的样本到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，半径R和R+r之间的样本集到的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。

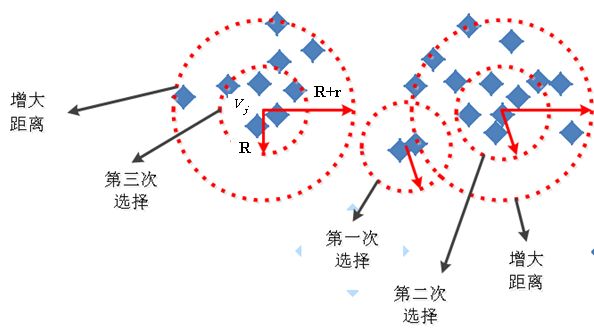


图1 中心点选择示意图

Fig.1 The schematic of selection of centers

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。这样不仅保证了每一类数据的紧密型同时也确保了各个簇之间的分离度。新的目标函数为

 （6）

其中，，表示局部密度矩阵。

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，同时这些簇的密度不同，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，那么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。加入密度之后，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，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

**The Improved FCM Algorithm**

The implementation procedure of Improved FCM Algorithm is shown as following:

改进的FCM算法在聚类中的实现过程如下：

|  |
| --- |
| Input: sample data set |
| Output: membership degrees matrix ，center coordinates matrix |
| main steps:   1. Add flag for each data in X, and flag=false; 2. Calculate average density of samples ; Select threshold radius R, and r=R; 3. Calculate distance dissimilarity matrix ; 4. Select smallest numeric values in D; set the center which is the average of two samples about this value as dot and calculate the density  within radius r; 5. If <, set flag=true about samples within radius r; if flag=true about all samples, executing 7）. If still exist flag=true in X, executing 4）, if >=, executing 6）; 6. Set radius R=R + r, calculate local density between R-r and R, and then update local density matrix . If  and , set flag=true between R-r and R. If flag=true , execute 7）; if still exist flag=false in X, executing 6）; If it doesn’t meet and, output , update center coordinates matrix , and then execute 5）; 7. Output center coordinates matrix and local density matrix ; 8. Confirm threshold , the maximum number of iteration T and the number of iterations =0; 9. If =0, plug into equation（7） to calculate degree of membership ; If ，use as center and calculate local density matrix , and then plug and into equation（7）degree of membership ; 10. Plug degree of membership into equation（4）, and then after getting new center coordinates , update ; 11. If or , and then stop the application, otherwise set , and then initialize flag=false in X, executing 9）; 12. Output degree of membership matrix and Matrix of clustering centers； 13. Calculate based on formula（5）, acquire the final classification results. |

图2 改进的FCM算法在聚类中的实现过程

传统的模糊C -均值（FCM）聚类是一种基于梯度下降的优化算法，该方法对初始化较敏感，且易陷入局部极小。

改进的初始中心点选择策略避免了随机选择中心点造成的聚类算法结果的不稳定，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处理分离性不好的数据，同时还能较好的发现簇大小不同且相距较近的簇集，但因为要计算局部密度，算法时间复杂度也相应增加。

The improved selection of initial centers strategy avoids random selection resulting in instability of clustering result. This also brings in some time overhead, because random selection of traditional FCM algorithm leads to too many times iterative computations. However, because the result of improved selection of initial centers strategy is closer to the real centers, the iteration times of FCM algorithm reduces significantly at last. The improved FCM algorithm changes the way the local density is calculated, so algorithm not only can solve poor separation of data, but also can find those clusters the distance is closer between clusters and the size of clusters is different. But the computation burden also largely increases, because of calculating local density.

# 4 实验设计与结果分析

# 4 Experimental design and Resultant analysis

To verify the accuracy and usefulness of the method which this thesis puts forward, one verifies the reasonable of the selection center strategy, another verifies the availability of improved FCM algorithm in the practical application.

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算法在实际应用中的有效性。

## 4.1 初始中心点选择实验

**4.1 selection of initial centers experiment**

Experimental data derives from databases in the UCI repository which is designed for machine learning and data mining to provide test data. It includes IRIS data, Abalone data and Wine [10]. The instances of IRIS data is 150, which is marked as 3 classes (Iris-setosa, Iris-versicolo, Iris-virginica) and each class includes 50 instances, and each instance includes 4 attributes. In order to find different sizes of clusters, delete 25 instances in Iris-setosa and make sure Iris-setosa not too sparse. The instances of Abalone data is 4177, which is marked as 3 classes (F，I，M) and each class includes 1307, 1342 and 1527 instances, and each instance includes 8 attributes. The instances of Wine data are 178, which is marked as 3 classes (1, 2, 3) and each class includes 59, 70 and 47 instances, each instance includes 13 attributes.

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。为了能发现大小不同的簇，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，错误率即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，一个简单选择策略是令，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值的改进中心点选择策略的实验结果如图3所示。

In order to verify the effect of improved selection of centers strategy, evaluating indexes are error rate of clustering results and iterative times, and error rate is the percentage of samples classified into clusters wrong. In order to determine optimal threshold , a simpler alternative is and D is distance matrix. Through continuously adjusting N, choose the optimal value of N. For different data set, the experiment is shown in figure 3.

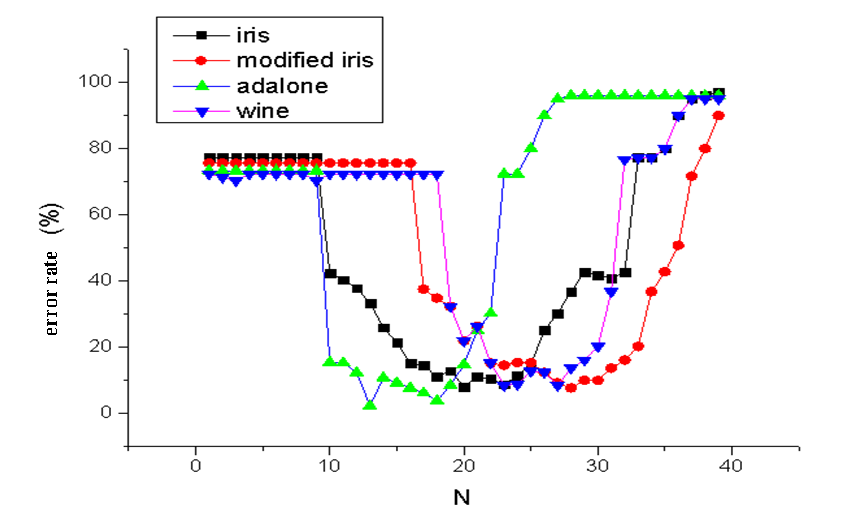


图3不同数据集的聚类错误率

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。

Shown in Figure 4, when N is greater than a certain value, clustering error rate surpasses 70% and tends towards stability. Centers are less than three, it shows that certain cluster is larger and almost includes all samples. The reason is that when N is small and radius of the threshold is large, almost all samples are classified into one center. In the experiment about Wine data, when N=12, lowest error rate is in these four experiments.

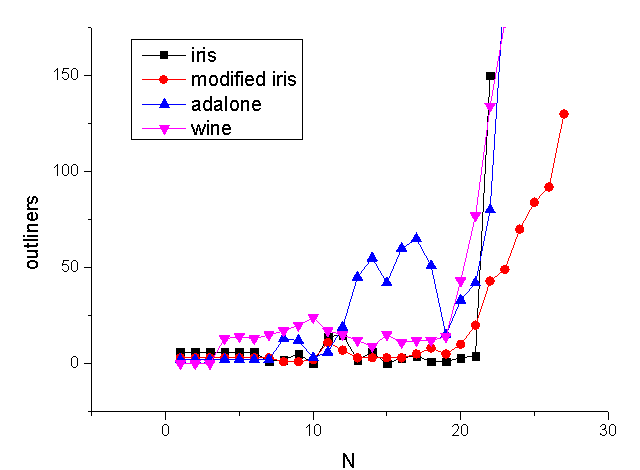


图4 不同数据集离群点个数

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较大时而阈值半径很小，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。

As shown in figure 4 and figure 5, when N is greater than a certain value, clustering error rate surpasses 70% and keeps increasing. This illustrates that when N is large and radius of the threshold is small, only two samples leave behind within radius. After many times selection, most of samples are considered as outlines, so the number of outlines gets more and more with N increasing.

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，在这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，但是识别错的点往往也被认为是离群点，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初始中心点选择的性能，对比实验采用随机选择初始中心点，聚类算法都采用标准的FCM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的值进行实验，实验结果如表1所示。

Shown in Figure 4, when N is between 8 and 23, the number of clusters is three, combining with figure 4 and in this range error rates of clustering is lower. In the experiment about modified IRIS, 3 outliers are also correctly identified. Although the improved algorithm discerns outlines accurate, some samples are also identified as outliers, but not classified into other clusters.

In order to verify selection of initial centers based on distance and density performance, use random selection of initial centers as a comparison experiment and both clustering algorithms adopt conventional FCM algorithm. The number of clusters c is three. From the above experiment results, the range of values for N should between 9 and 40, so, randomly selects N to experiment. Experimental result is shown in Table 1.

表1 随机选择和改进的基于距离和密度的算法

Table 1 Random selection algorithm and improved algorithm based on distance and density

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 | 随机选择 | 9.44 | 24.00 | 11 |
| 改进的基于距离和密度算法 | 19.22 | 34.00 | 8 |
| Abalone | 随机选择 | 12.81 | 25.00 | 17 |
| 改进的基于距离和密度算法 | 20.90 | 33.00 | 7 |
| Wine | 随机选择 | 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率都要比随机选择高很多，但是最小错误率对于改进算法是可知的，算法是稳定的，而随机选择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，改进的选择策略的最小错误率远远小于随机选择。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。

Although the iterative times and average error rate of improved algorithm is higher than the randomly selected, minimum error ratio are knowable to improved algorithm and this algorithm is unconditionally stable, in addition, the optimal clustering to random selection is unknown for clustering results relying on the centers of random selection. The minimum error ratio of improved algorithm is much less than random selection. So, the improved algorithm not only guarantees the stability and avoids trapping in local optimum, but also reduces iteration times. Moreover, when we don't know the distribution of the sample and can be sure the number of clusters C, algorithm can also predict the distribution of the sample.

## 4.2 Improved FCM Algorithm

**First Experiment** UCI data experiment

In order to verify the effects of improved FCM algorithm, mainly evaluating indexes used in the experiment includes MSE and Purity. MSE is defined as following:

 （8）

: The real cluster centers. : The calculated cluster centers. MSE: the distance between the calculated cluster centers and the real cluster centers.

Purity is defined as following:

 （9）

Purity: The percentage of correctly classified samples of all samples. : All clusters,. : cluster of . X: sample set. : Sample of . N: The number of samples.

Data sets are mentioned above IRIS, Wine, and Abalone and modified IRIS. The experimental results show as following Table 3.

## 4.2 改进的FCM算法

**实验一** UCI数据实验

为了检验改进FCM算法的效果，实验中使用的主要评价指标包括：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，其中是聚类的集合，表示第c个簇，X表示样本集，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

Table 1 Clustering results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| data sets | methods | MSE | Purity (%) | iteration times |
| IRIS | FCM algorithm | 0.139943 | 84.3589 | 11 |
| Improved FCM algorithm | 0.132665 | 87.4258 | 4 |
| modified IRIS | FCM algorithm | 0.144415 | 80.3548 | 5 |
| Improved FCM algorithm | 0.124710 | 82.1592 | 7 |
| Wine | FCM algorithm | 0.115213 | 89.0102 | 22 |
| Improved FCM algorithm | 0.105275 | 91.6362 | 5 |
| Abalone | FCM algorithm | 0.332467 | 73.4170 | 20 |
| Improved FCM algorithm | 0.204017 | 82.9382 | 9 |

According to Table 3, the MSE of improved FCM algorithm is lower than traditional FCM, and classification accuracy is higher. The clustering results of IRIS are close to Wine, because the distance is far from each other and the clusters’ density is large. To the modified IRIS, MSE is higher than unmodified IRIS, because outliers are put in modified IRIS. To Abalone, the classification accuracy increase significantly, because Abalone distributes relatively even, the distance is close to each other, and the clusters’ size is different resulting in classification accuracy of traditional FCM algorithm low. Because of better effective selection initial centers, makes iteration times of improved algorithm decrease obvious.

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数据集的聚类结果相差不多，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，导致传统的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。

**Second Experiment** the Data of stars’ Box-office Influence

**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。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。根据初始中心点坐标，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，要么是电影评分略低。

# 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，普通FCM算法仅仅考虑距离，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。通过对明星数据的聚类分析，发掘了一些有价值的信息。

# 参考文献

1. 姜伦, 丁华福. 关于模糊C-均值(FCM)聚类算法的改进[J]. 计算机与数字工程, 2010, 38(2):4-6.
2. 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.
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. Liu B, Wang L, Liu H Y. FCM Segmentation Algorithm Research for CT Cerebrovascular Medical Image Based on Sober Extraction Algorithm[J]. Applied Mechanics & Materials, 2012, 263-266:2203-2206.
5. 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.
6. James C. Bezdek†. Cluster Validity with Fuzzy Sets[J]. Journal of Cybernetics, 1973, 3(3):58-73.
7. Bezdek J C. Pattern recognition with fuzzy objective function algorithms /[M].NY:Plenum, 1981.
8. 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.
9. 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.
10. FISHER R A． Iris Plants Database//http：//www.ics.uci.edu/~mlearn /MLRepository．Html．Authorized license。