基于K-means聚类算法在分类上的应用以及改进

张镇鸿

陈栋

邱俊文

摘要　聚类分析是数据挖掘领域最重要的研究方向之一，通过数据分析能够发现有用的信息，其广泛地应用于市场研究、数据分析、模式识别、图像处理、人工智能等领域。在商业智能应用中，聚类分析能够帮助数据挖掘人员分析顾客的购买模式，刻画不同顾客群体的特征，从顾客消费信息数据库发现特殊的顾客。在生物学应用中，聚类分析可以用于推测物种的类别信息，根据功能的相关性对基因进行分门别类，从而可以获得对种群原有结构的认识。聚类分析还有助于在识别卫星监测数据库中对使用相关的区域进行识别，根据房屋的价值、类型和具体地理位置对城市中存在的房屋分组识别。与使用其它数据挖掘方法不同，用户在运用聚类分析算法之前并不知道数据集的内容和类别等特征信息，也即聚类分析不需要基于先验知识，是一种无监督的机器学习。

关键词　聚类分析　用户分类

1. 引言

1.1 关于聚类分析

聚类分析是数据挖掘领域最重要的研究方向之一，通过数据分析能够发现有用的信息，其广泛地应用于市场研究、数据分析、模式识别、图像处理、人工智能等领域[1]-[2]。在商业智能应用中，聚类分析能够帮助数据挖掘人员分析顾客的购买模式，刻画不同顾客群体的特征，从顾客消费信息数据库发现特殊的顾客。在生物学应用中，聚类分析可以用于推测物种的类别信息，根据功能的相关性对基因进行分门别类，从而可以获得对种群原有结构的认识。聚类分析还有助于在识别卫星监测数据库中对使用相关的区域进行识别，根据房屋的价值、类型和具体地理位置对城市中存在的房屋分组识别。与使用其它数据挖掘方法不同，用户在运用聚类分析算法之前并不知道数据集的内容和类别等特征信息，也即聚类分析不需要基于先验知识，是一种无监督的机器学习。

1.2 聚类分析研究动态

聚类分析作为数据挖掘研究的一大主题，针对聚类分析的研究已有多年，其中，基于距离的聚类分析是学者们研究的主要内容。基于K-medoids算法、K-Means算法及其它聚类算法的聚类挖掘工具广泛应用于众多统计分析软件中或者系统中[3]。

在机器学习学习领域中，机器学习主要分为有监督学习和无监督学习，聚类分析是属于无监督学习的范畴，而分类是属于有监督学习的范畴。聚类和分类有所不同，聚类是不依赖于训练样本的分类信息，而分类是依赖于训练样本的分类信息的。

为实现更好的运用，将聚类方法与其它领域的方法相结合，以弥补该方法在数据挖掘中的缺陷，使聚类方法的性能更加优越。通常，常与聚类方法相结合的方法有：蚁群算法、遗传算法和免疫算法等[4]。

2. K-means算法聚类过程

2.1 主要思想

k-means算法以k为参数，把n个对象分成k个簇，使簇内具有较高的相似度，而簇间的相似度较低。

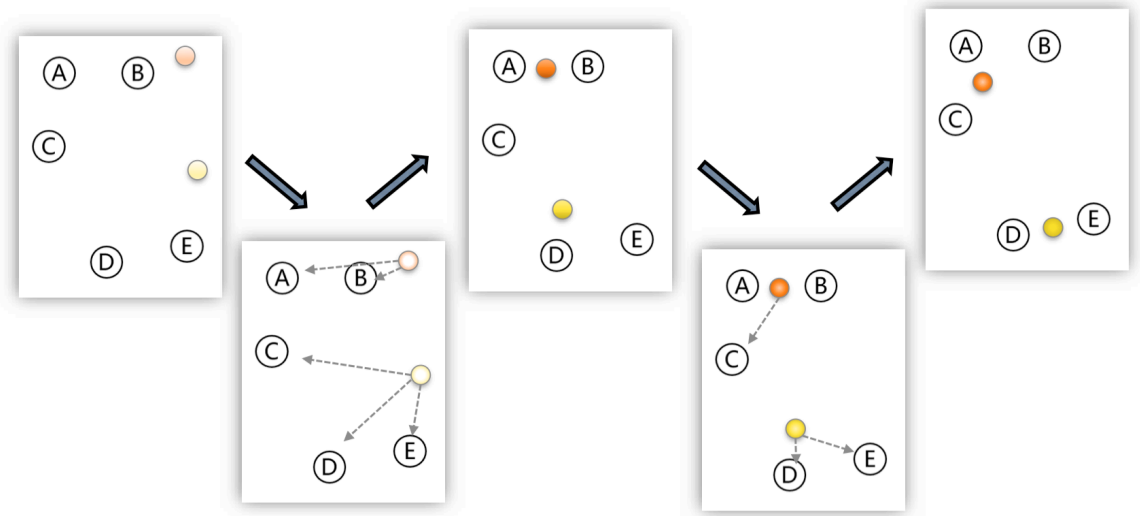
2.2 处理过程

1. 随机选择k个点作为初始的聚类中心；

2. 对于剩下的点，根据其与聚类中心的距离，将其归入最近的簇；

3. 对每个簇，计算所有点的均值作为新的聚类中心；

4. 重复2、3直到聚类中心不再发生改变。



2.3 主要应用

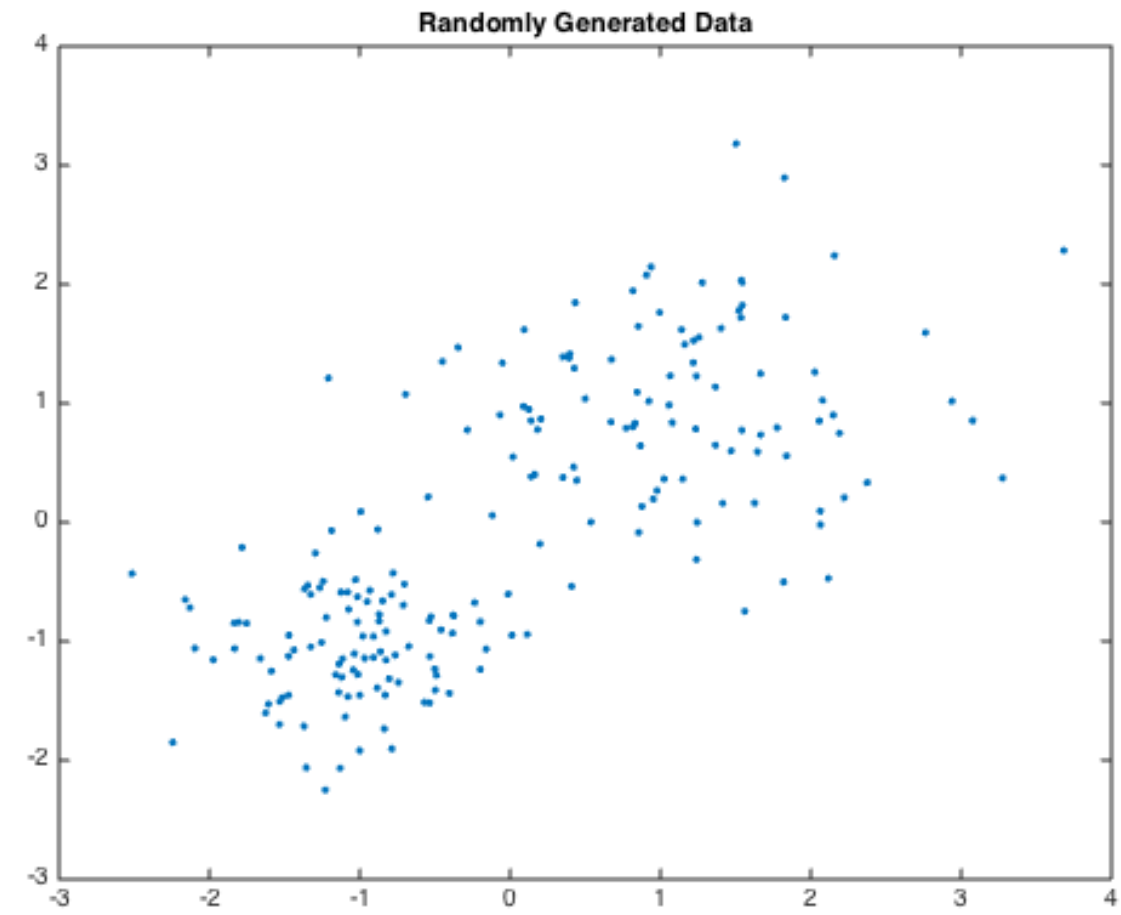
K-means将应用于用户细分，通过层次分析法建立用户的价值体系，量化用户价值；将用户划分成不同的类，由此来有效的开展客户管理。本文运用的数据挖

掘的方法，从实际情况出发，通过一系列可操作的用户价值评价指标来细分用户。

2.4 实际操作(MATLAB)

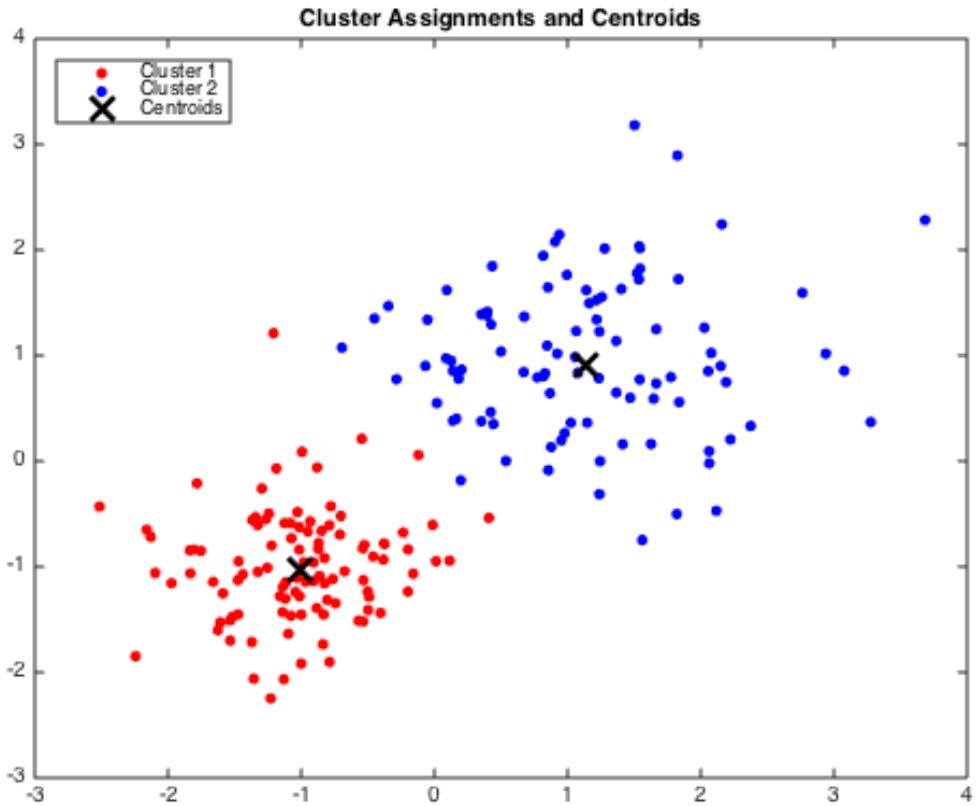
2.4.1 若干随机点聚类

```
>> rng default;
X = [randn(100,2)*0.75+ones(100,2);
     randn(100,2)*0.5-ones(100,2)];
figure;
plot(X(:,1),X(:,2),'.');
title 'Randomly Generated Data';
>>
```



ommand Window

```
>> rng default;
X = [randn(100,2)*0.75+ones(100,2);
    randn(100,2)*0.5-ones(100,2)];
figure;
plot(X(:,1),X(:,2),'.');
title 'Randomly Generated Data';
>> opts = statset('Display','final');
[idx,C] = kmeans(X,2,'Distance','cityblock','Replicates',5,'Options',opts);
Replicate 1, 4 iterations, total sum of distances = 201.533.
Replicate 2, 6 iterations, total sum of distances = 201.533.
Replicate 3, 4 iterations, total sum of distances = 201.533.
Replicate 4, 4 iterations, total sum of distances = 201.533.
Replicate 5, 3 iterations, total sum of distances = 201.533.
Best total sum of distances = 201.533
>> figure;
plot(X(idx==1,1),X(idx==1,2),'r.','MarkerSize',12)
hold on
plot(X(idx==2,1),X(idx==2,2),'b.','MarkerSize',12)
plot(C(:,1),C(:,2),'kx',...
     'MarkerSize',15,'LineWidth',3)
legend('Cluster 1','Cluster 2','Centroids','Location','NW')
title 'Cluster Assignments and Centroids'
hold off
>>
```



2.4.2 学生成绩分类

```
ed*
score=[100 90 80 90 100;90 90 70 75 90;80 70 60 70 90;90 90 90 90 90;100 95 98 99 100;
    55 60 50 68 80;60 60 70 67 48;89 83 75 94 100;60 55 60 55 50]
%第i列代表第i科的成绩
%第i行代表第i个同学

%Idx存储的是每个点的聚类标号
%C代表3个聚类的每一个科目的具体数值
%sumD存储的是类间所有点与该类质心点距离之和
%D存储的是每个点与所有质心的距离
```

Command Window

```
>> score=[100 90 80 90 100;90 90 70 75 90;80 70 60 70 90;90 90 90 90 90;100 95 98 99 100;...
    55 60 50 68 80;60 60 70 67 48;89 83 75 94 100;60 55 60 55 50]
score =
   100    90    80    90   100
    90    90    70    75    90
    80    70    60    70    90
    90    90    90    90    90
   100    95    98    99   100
    55    60    50    68    80
    60    60    70    67    48
    89    83    75    94   100
    60    55    60    55    50
>> [Idx,C,sumD,D]=kmeans(score,3,'dist','sqEuclidean','rep',4)
Idx =
     3
     3
     1
     3
     3
     1
     2
     3
     2
C =
   67.5000   65.0000   55.0000   69.0000   85.0000
   60.0000   57.5000   65.0000   61.0000   49.0000
   93.8000   89.6000   82.6000   89.6000   96.0000
sumD =
   1.0e+03 *
    0.4645
    0.1365
    1.1584
D =
   1.0e+03 *
    2.9722    6.3232    0.0615
    1.4172    3.8582    0.4225
    0.2323    2.3432    1.5055
    2.8222    5.1033    0.1055
    4.9303    8.1402    0.4091
    0.2323    1.2663    4.1669
    1.6792    0.0683    4.9921
    2.0362    5.2812    0.1597
    1.6022    0.0683    6.1635
```

3. K-means算法的局限

3.1 K-means算法的收敛

K-Means算法有许多可能的收敛条件。搜索可能终止于划分误差在重新分配时不再变化时，这表明这个划分可能是局部最优的。另一个终止条件可以是预先定义好的迭代次数。K-Means算法在样本点为N，每个样本点的属性为m维，聚类的类别数为K，进行T次迭代的时间复杂性为O(T*K*N*M)[5]。

3.2 K-means算法的缺点

由3.1可知，K-means算法收敛速度相当快，故K-means算法在数据挖掘中得以大量地运用，然而，K-Means聚类算法也有着自身的局限性。该算法只能应用于簇的均值可以定义的情况下，而且大多数情况只能用于凸状，而对于非凸的聚类不适用，而如果一组数据中，存在噪音数据较多，也会大大增大K-means算法的聚类误差。K-means算法对于初始中心的选择是随机的，所以在迭代过程中，极易陷入局部最优[6]的情况。

4. K-means算法的改进

4.1 K-means算法初始聚类中心的改进

由D. Arthur等人提出的K-means++针对K-means算法中的第一步做了改进[7]。可以直观地将这改进理解成这K个初始聚类中心相互之间应该分得越开越好。

4.1.1 K-means++算法聚类流程

1. 从数据集中随机选取一个样本作为初始聚类中心c_1；

2. 首先计算每个样本与当前已有聚类中心之间的最短距离，用$D(x)$表示；接着计算每个样本被选为下一个聚类中心的概率$\frac{D(x)^2}{\sum_{x\in X} D(x)^2}$。最后，按照轮盘法选择出下一个聚类中心；

3. 重复第2步直到选择出K个聚类中心；

4. 剩余步骤与2.2中2–4的步骤相同。

4.2 聚类中心数的优化

K-means和K-means++的聚类中心数K是固定不变的。而ISODATA算法在运行过程中能够根据各个类别的实际情况进行两种操作来调整聚类中心数K：

Ⅰ分裂操作，对应着增加聚类中心数；

Ⅱ合并操作，对应着减少聚类中心数。

4.2.1 ISODATA算法流程

1. 预期的聚类中心数目Ko：虽然在ISODATA运行过程中聚类中心数目是可变的，但还是需要由用户指定一个参考标准。

2. 每个类所要求的最少样本数目Nmin：用于判断当某个类别所包含样本分散程度较大时是否可以进行分裂操作。如果分裂后会导致某个子类别所包含样本数目小于Nmin，就不会对该类别进行分裂操作。

3. 最大方差Sigmα：用于衡量某个类别中样本的分散程度。当样本的分散程度超过这个值时，则有可能进行分裂操作。

4. 两个类别对应聚类中心之间所允许最小距离dmin：如果两个类别靠得非常近，则需要对这两个类别进行合并操作。是否进行合并的阈值就是由dmin决定。

5. 总结

数据挖掘中的聚类算法正在走一条综合了机器学习、数据挖掘、模式识别、物理等领域的研究成果，不断创新发展的道路。 正如文献[4]指出的一样：聚类算法的聚类结果有一定的不可预见性，在实际应用中应根据数据类型选择合适的聚类算法，以取得最佳的聚类效果。

参考文献

[1]周卫星,廖欢.基于K均值聚类和概率松弛法的图像区域分割[J].计算机技术与发展,2010,20(2):68-70

[2]Jiawei han,Micheline Kamber.Data Mining Concepts and Techniques[M].China Machine Press,2007,3

[3]Ketchen D J,Shook C L. The application of cluster analysis in strategic management research: an analysis and critique[J]. Strategic management journal, 1996, 17(6): 441-458.

[4]孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.

[5]Selim S Z, Ismail M A.K-means-type algorithms: a generalized convergence theorem and characterization of local optimality[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1984 (1): 81-87.

[6]朱颢东,钟勇,赵向辉.一种优化初始中心点的K-Means文本聚类算法.郑州大学学报（理学版）,2009,6,41(2):30-30

[7]David Arthur and Sergei Vassilvitskii k-means++: The Advantages of Careful Seeding