模糊 C 均值聚类算法及实现

摘要：模糊聚类是一种重要数据分析和建模的无监督方法。本文对模糊聚类进行了概述，从理论和实验方面研究了模糊 c 均值聚类算法，并对该算法的优点及存在的问题进行了分析。该算法设计简单，应用范围广，但仍存在容易陷入局部极值点等问题，还需要进一步研究。
关键词：模糊 c 均值算法；模糊聚类；聚类分析

Fuzzy c-Means Clustering Algorithm and Implementation

Abstract: Fuzzy clustering is a powerful unsupervised method for the analysis of data and construction of models.This paper presents an overview of fuzzy clustering and do some study of fuzzy c-means clustering algorithm in terms of theory and experiment.This algorithm is simple in design,can be widely used,but there are still some problems in it,and therefore,it is necessary to be studied further.
Key words: fuzzy c-Mean algorithm；fuzzy clustering；clustering analysis

1 引言

20 世纪 90 年代以来，随着信息技术和数据库技术的迅猛发展，人们可以非常方便地获取和存储大量的数据。但是，面对大规模的数据，传统的数据分析工具只能进行一些表层的处理，比如查询、统计等，而不能获得数据之间的内在关系和隐含的信息。为了摆脱“数据丰富，知识贫乏”的困境，人们迫切需要一种能够智能地、自动地把数据转换成有用信息和知识的技术和工具，这种对强有力数据分析工具的迫切需求使得数据挖掘技术应运而生。

将物理或抽象对象的集合分组成由类似的对象组成的多个类的过程称为聚类。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其它簇中的对象相异。

聚类是一种重要的数据分析技术，搜索并且识别一个有限的种类集合或簇集合，进而描述数据。聚类分析作为统计学的一个分支，已经被广泛研究了许多年。而且，聚类分析也已经广泛地应用到诸多领域中，包括数据分析、模式识别、图像处理以及市场研究[1]。通过聚类，人们能够识别密集的和稀疏的区域，因而发现全局的分布模式，以及数据属性之间的有趣的相互关系。在商务上，聚类能帮

助市场分析人员从客户基本信息库中发现不同的客户群，并且用购买模式来刻画不同的客户群的特征。在生物学上，聚类能用于推导植物和动物的分类，对基因进行分类，获得对种群中固有结构的认识。聚类在地球观测数据库中相似地区的确定，汽车保险单持有者的分组，及根据房屋的类型、价值和地理位置对一个城市中房屋的分组上也可以发挥作用。聚类也能用于对 Web 上的文档进行分类，以发现信息。基于层次的聚类算法文献中最早出现的 Single-Linkage 层次聚类算法是 1957 年在 Lloyd 的文章中最早出现的，之后 MacQueen 独立提出了经典的模糊 C 均值聚类算法，FCM 算法中模糊划分的概念最早起源于 Ruspini 的文章中，但关于 FCM 的算法的详细的分析与改进则是由 Dunn 和 Bezdek 完成的。

聚类分析是多元统计分析的一种，也是非监督模式识别的一个重要分支，在模式分类、图像处理和模糊规则处理等众多领域中获得最广泛的应用。它把一个没有类别标记的样本集按某种准则划分为若干个子集(类) ，使相似的样本尽可能的归为一类，而将不相似的样本尽量划分到不同的类中。硬聚类把每个待辨识的对象严格地划分到某类中，具有非此即彼的性质，模糊聚类由于能够描述样本类属的中介性，能够客观地反映现实世界，已逐渐成为聚类分析的主流[2-3]。在众多的模糊聚类算法中，模糊 c 均值聚类算法(FCM)应用最为广泛。它按照某种判别准则，将数据的聚类转化为一个非线性优化问题，并通过迭代来进行求解，目前已成为非监督模式识别的一个重要分支。

数据挖掘中的聚类分析主要集中在针对海量数据的有一效和实用的聚类方法研究，聚类方法的可伸缩性，高维聚类分析，分类属性数据聚类和具有混合属性数据的聚类，非距离模糊聚类等。因此，数据挖掘对聚类分析有其特殊的要求;可伸缩性，能够处理不同类型属性，强抗噪性，高维性，对输入顺序不敏感性，可解释性和可用性等。

本文正是在此背景下对数据挖掘中的聚类分析进行论述，并着重研究了 FCM 算法。

2 模糊聚类算法

2.1 模糊聚类算法概述

模糊聚类算法是一种基于函数最优方法的聚类算法，使用微积分计算技术求

最优代价函数。在基于概率算法的聚类方法中将使用概率密度函数，为此要假定合适的模型，模糊聚类算法的向量可以同时属于多个聚类，从而摆脱上述问题。在模糊聚类算法中，定义了向量与聚类之间的近邻函数，并且聚类中向量的隶属度由隶属函数集合提供。对模糊方法而言，在不同聚类中的向量隶属函数值是相互关联的。硬聚类可以看成是模糊聚类方法的一个特例。

2.2 模糊聚类算法的分类

模糊聚类分析算法大致可分为三类[4]：

1）分类数不定，根据不同要求对事物进行动态聚类，此类方法是基于模糊等价矩阵聚类的，称为模糊等价矩阵动态聚类分析法。

2）分类数给定，寻找出对事物的最佳分析方案，此类方法是基于目标函数聚类的，称为模 c 均值聚类。

3）在摄动有意义的情况下，根据模糊相似矩阵聚类，此类方法称为基于摄动的模糊聚类分析法。

3 模糊 c 均值（FCM）聚类算法

3.1 算法描述

模糊 c 均值聚类算法的步骤还是比较简单的，模糊 c 均值聚类（FCM），即众所周知的模糊 ISODATA，是用隶属度确定每个数据点属于某个聚类的程度的一种聚类算法。1973 年，Bezdek 提出了该算法，作为早期硬 c 均值聚类（HCM）方法的一种改进。

FCM 把 n 个向量 x_i（i=1,2,…,n）分为 c 个模糊组，并求每组的聚类中心，使得非相似性指标的价值函数达到最小。FCM 与 HCM 的主要区别在于 FCM 用模糊划分，使得每个给定数据点用值在 0，1 间的隶属度来确定其属于各个组的程度。与引入模糊划分相适应，隶属矩阵 U 允许有取值在 0，1 间的元素。不过，加上归一化规定，一个数据集的隶属度的和总等于 1：

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,..., n \qquad (3.1)$$

那么，FCM 的价值函数（或目标函数）就是：

$$J(U,c_1,...,c_c)=\sum_{i=1}^{c}J_i=\sum_{i=1}^{c}\sum_{j}^{n}u_{ij}^{m}d_{ij}^{2}\quad,\qquad(3.2)$$

这里 u_{ij} 介于 0，1 间；c_i 为模糊组 I 的聚类中心，$d_{ij}=\|c_i-x_j\|$为第 I 个聚类中心与第 j 个数据点间的欧几里德距离；且 $m\in[1,\infty)$是一个加权指数。

构造如下新的目标函数，可求得使（3.2）式达到最小值的必要条件：

$$\begin{aligned}\overline{J}(U,c_1,...,c_c,\lambda_1,...,\lambda_n)&=J(U,c_1,...,c_c)+\sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c}u_{ij}-1)\\&=\sum_{i=1}^{c}\sum_{j}^{n}u_{ij}^{m}d_{ij}^{2}+\sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c}u_{ij}-1)\end{aligned}\qquad(3.3)$$

这里λ_j，j=1 到 n，是（3.1）式的 n 个约束式的拉格朗日乘子。对所有输入参量求导，使式（3.2）达到最小的必要条件为：

$$c_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j}{\sum_{j=1}^{n}u_{ij}^{m}}\qquad(3.4)$$

和

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}\qquad(3.5)$$

由上述两个必要条件，模糊 c 均值聚类算法是一个简单的迭代过程。在批处理方式运行时，FCM 用下列步骤确定聚类中心 c_i 和隶属矩阵 U[1]：

步骤 1：用值在 0，1 间的随机数初始化隶属矩阵 U，使其满足式（3.1）中的约束条件

步骤 2：用式（3.4）计算 c 个聚类中心 ci，i=1,…,c。

步骤 3：根据式（3.2）计算价值函数。如果它小于某个确定的阀值，或它相对上次价值函数值的改变量小于某个阀值，则算法停止。

步骤 4：用（3.5）计算新的 U 矩阵。返回步骤 2。

上述算法也可以先初始化聚类中心，然后再执行迭代过程。由于不能确保 FCM 收敛于一个最优解。算法的性能依赖于初始聚类中心。因此，我们要么用另外的快速算法确定初始聚类中心，要么每次用不同的初始聚类中心启动该算法，多次运行 FCM。

设被分类的对象的集合为：X = { x_1 ， x_2 ，…，x_N}，其中每一个对象 x_k 有 n 个特性指标，设为 $x_k = (x_{1k}, x_{2k}, \cdots, x_{nk})^T$，如果要把 X 分成 c 类，则它的每一个分类结果都对应一个 c×N 阶的 Boolean 矩阵 $U=[u_{ik}]_{c\times N}$，对应的模糊 c 划分空间为：

$M_{fc} = \{ U \subset R^{cN} \mid u_{ik} \in [0, 1], \forall i, \forall k; \sum_{k=1}^{N} u_{ik} = 1, \forall k; 0 < \sum_{k=1}^{N} u_{ik}, \forall i\}$

在此空间上，模糊 c 均值算法如下：

Repeat for l = 1 ， 2……

Step 1：compute the cluseter prototypes(means)：

$$p_i^{(l)} = \frac{\sum_{k=1}^{N} (u_{ik}^{(l-1)})^m x_k}{\sum_{k-1}^{n} (u_{ik}^{(l-1)})^m}, 1 \le i \le c$$

Step 2：compete the distance：

$$(d_{ik})^2 = (x_k - p_i^{(l)})^T A (x_k - p_i), 1 \le i \le c, 1 \le k \le n$$

Step 3：Update the partition matrix：

For $1 \le k \le N$

If $(d_{ik})^2 > 0$ for all i=1，2，…，c

$$u_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{c} (d_{ikA} / d_{jkA})^{2/(m-1)}}$$

Otherwise

$u_{ik}^{(l)}$=0 if $d_{ikA} > 0$，and $u_{ik}^{(l)} \in [0, 1]$ with $\sum_{i=1}^{c} u_{ik}^{(l)} = 1$

Until $\| U^{(l)} - U^{(l-1)} \| < \varepsilon$

3.2 实验

采用著名的 iris 数据集对算法进行测试实现，其中样本总数 m=150，样本属性数 n=4，设定的划分内别 k=3。

运算次数为 10 次的输出结果：

```
输出第10次运行的聚类结果:
**************************************************
第0类样本:
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0
正确样本数: right[0]=50
错误样本数: 0
聚类类别号: sx=0

第1类样本:
2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1
1 1 1 1 1 1
正确样本数: right[1]=46
错误样本数: 4
聚类类别号: sx=1
```

```
第2类样本:
2 1 2 2 2 2 1 2 2 2 2 2 2 1 2 2 2 2 2 1 2 1 2 1 2 2 1 2 2 2 2 2 2 1 1 2 2 2 1 2 2 2 1 2
2 2 2 2 2 1
正确样本数: right[2]=38
错误样本数: 12
聚类类别号: sx=2

聚类正确率: 89.3333%
//////////////////////////////////////////////////////////////////////
10次运行,平均聚类正确率: 89.3333%
平均目标函数: 0.160435
```

能对数组实现分类，但是分类正确率不是很理想。

3.3 FCM 算法优缺点

通过实验和算法的研究学习，不难发现 FCM 算法的优缺点[5-8]：

首先，模糊 c 均值泛函 J_m 仍是传统的硬 c 均值泛函 J_1 的自然推广。J_1 是一个应用很广泛的聚类准则，对其在理论上的研究已经相当的完善，这就为 J_m 的研究提供了良好的条件。

其次，从数学上看，J_m 与 R_s 的希尔伯特空间结构(正交投影和均方逼近理论)有密切的关联，因此 J_m 比其他泛函有更深厚的数学基础。

最后，FCM 聚类算法不仅在许多邻域获得了非常成功的应用，而且以该算

法为基础，又提出基于其他原型的模糊聚类算法，形成了一大批 FCM 类型的算法，比如模糊 c 线(FCL) ，模糊 c 面(FCP) ，模糊 c 壳(FCS) 等聚类算法，分别实现了对呈线状、超平面状和“薄壳”状结构模式子集(或聚类) 的检测。

4 结语

模糊 c 均值算法因设计简单，解决问题范围广，易于应用计算机实现等特点受到了越来越多人的关注，并应用于各个领域。但是，自身仍存在的诸多问题，例如强烈依赖初始化数据的好坏和容易陷入局部鞍点等，仍然需要进一步的研究。

参考文献：

[1] A K Jain，M N Murty，P J Flynn. Data Clustering：A Review，ACM Computing Surveys[J]，1999，31(3)：264-323.

[2] Spragins J. Learning without a teacher [J].IEEE Transactions of Information Theory ，2005 ，23 (6) ：223 -230.

[3] Babusk R. FUZZYAND NEURAL CONTROL[M] . Netherlands ： Delft University of Technology ，2001.

[4] Theodoridis S.Pattern Recongnition [M].Second Edition.USA ： Elsevier Scinece ，2003.

[5] 高新波.FCM 聚类算法中模糊加权指数 m 的优选方法[J].模糊系统与数学，2005，19（1）：143-148

[6] 朱剑英，应用模糊数学方法的若干关键问题及处理方法[J].模糊系统与数学，1992，11（2）：57-63

[7] 高新波.模糊聚类分析及其应用[M].西安：西安电子科学出版社，2004

[8] 刘蕊洁，张金波，刘锐.模糊 c 均值聚类算法[J].重庆工学院学报，2007-21-1