模糊聚类算法及应用

MOHU JULEI SUANFA JI YINGYONG

蔡静颖 著

冶金工业出版社
Metallurgical Industry Press

模糊聚类算法及应用

蔡静颖　著

北　京
冶 金 工 业 出 版 社
2015

内 容 提 要

本书主要针对模糊聚类算法中最经典的 FCM 算法进行了系统分析，并对原始算法进行了改进，将经典的 FCM 算法和改进的 FCM 算法应用于图像识别、数据聚类和软件测试等不同领域。全书共分 7 章，第 1 章介绍了聚类分析发展背景和基础概念；第 2 章介绍了模糊理论基础知识及模糊聚类分析的方法和应用；第 3 章介绍了模糊 c - 均值算法的理论知识和研究现状以及目前存在的问题；第 4 章介绍了马氏距离的基本原理和处理方法；第 5 章介绍了马氏距离在模糊聚类中的应用；第 6 章介绍了基于优化 KPCA 特征提取的 FCM 算法；第 7 章介绍了 FCM 算法在软件测试的等价类划分方法中的应用。

本书可供从事模式识别教学研究的师生、科研人员参考，也可供从事数据挖掘、图像识别、软件测试等工作的相关人员学习。

图书在版编目(CIP)数据

模糊聚类算法及应用/蔡静颖著. —北京：冶金工业出版社，2015.8

ISBN 978-7-5024-7015-9

Ⅰ.①模… Ⅱ.①蔡… Ⅲ.①聚类分析—算法 Ⅳ.①O212.4

中国版本图书馆 CIP 数据核字(2015)第 191385 号

出 版 人 谭学余
地　　址 北京市东城区嵩祝院北巷 39 号 邮编 100009 电话 (010)64027926
网　　址 www.cnmip.com.cn 电子信箱 yjcbs@cnmip.com.cn
责任编辑 夏小雪 美术编辑 彭子赫 版式设计 孙跃红
责任校对 郑 娟 责任印制 李玉山
ISBN 978-7-5024-7015-9
冶金工业出版社出版发行；各地新华书店经销；北京百善印刷厂印刷
2015 年 8 月第 1 版，2015 年 8 月第 1 次印刷
148mm×210mm；4.375 印张；129 千字；131 页
27.00 元

冶金工业出版社 投稿电话 (010)64027932 投稿信箱 tougao@cnmip.com.cn
冶金工业出版社营销中心 电话 (010)64044283 传真 (010)64027893
冶金书店 地址 北京市东四西大街 46 号(100010) 电话 (010)65289081(兼传真)
冶金工业出版社天猫旗舰店 yjgycbs.tmall.com

前　言

数据库技术的不断发展及数据库管理系统的广泛应用使得各组织机构积累了海量数据，为了从中提取有用信息，更好地利用这些数据资源，人们提出了数据挖掘技术。数据挖掘技术将传统的数据分析方法与处理大量数据的复杂算法相结合，是目前信息领域和数据库技术的前沿研究课题。

聚类分析技术是数据挖掘的主要方法，它将数据划分成有意义或有用的组（簇），在众多的聚类分析算法中，模糊聚类算法是当前研究的热点。其中最受欢迎的是基于目标函数的模糊聚类方法，该方法把聚类归结成一个带约束的非线性规划问题，通过优化求解获得数据集的模糊划分和聚类。该方法设计简单，解决问题范围广，还可以转化为优化问题而借助经典数学的非线性规划理论求解，并易于在计算机上实现。因此，随着计算机的应用和发展，基于目标函数的模糊聚类算法成为新的研究热点。

在基于目标函数的聚类算法中，FCM 类型算法的理论最为完善、应用最为广泛。FCM 类型的算法最早是从“硬”聚类目标函数的优化中导出的。为了借助目标函数法求解聚类问题，人们利用均方逼近理论构造了带约束的非线性规划函数，从此类内平方误差和（Within – Groups Sum of Squared Error，WGSS）J_1 成为聚类目标函数的普遍形式。为极小化该目标函数而采取 Pikard 迭代优化方案就是著名的硬 c – 均值（Hard c – means，HCM）算法。后来 Dunn 把 WGSS 函数 J_1 扩展到 J_2，即类内加权平均误差和函数。Bezdek 又引入了一个参数 m，把 J_2 推广到一个目标函数的无限簇，即形成了人们所熟知的 FCM 算法。从此，奠定了 FCM 算法在模糊聚类中的地位。

本书重点分析了 FCM 算法和马氏距离的基本原理，从而利用

马氏距离的优点来弥补FCM算法中存在的缺陷，并从两个方面对FCM算法进行了改进。

首先，经典的模糊c－均值（FCM）算法是基于欧氏距离的，它只适用于球型结构的聚类，且在处理属性高相关的数据集时，分错率增加。针对这个问题，提出了一种新的聚类算法（FCM－M），它将马氏距离替代模糊c－均值中的欧氏距离，并在目标函数中引进一个协方差矩阵的调节因子，利用马氏距离的优点，有效地解决了FCM算法中的缺陷，并利用特征值、特征矢量及伪逆运算来解决马氏距离中遇到的奇异问题。

其次，经典的模糊c－均值算法认为样本矢量各特征对聚类结果贡献均匀，没有考虑不同的属性特征对模式分类的不同影响，且在处理属性高相关的数据集时，该算法分错率增加。针对这些问题，提出了一种基于马氏距离特征加权的模糊聚类算法，利用自适应马氏距离的优点对特征加权处理，从而对高属性相关的数据集进行更有效的分类。

本书还将FCM算法和KPCA方法结合，利用KPCA进行特征提取，然后利用FCM算法进行数据聚类分析。将FCM算法应用于软件测试中是作者未来研究的重点，本书主要介绍了将FCM算法应用于等价类划分方法中，每个应用在本书中都做了详尽介绍。

作者在编写本书过程中参考了大量同行著作及论文资料，在此向相关作者表示感谢！

本书的编写和出版得到牡丹江师范学院省级重点创新预研项目（NO. SY2014001）、牡丹江师范学院青年一般项目（NO. QY2014002）的支持。

由于作者学识有限，FCM算法在软件测试中的应用还处于研究初期，书中难免存在不足之处，恳请读者批评、指正。

著 者

2015年5月

目　录

1 绪　论

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。通过聚类，人们能够识别密集的和稀疏的区域，近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。

1.1 聚类分析的概述

聚类分析也称为无监督学习，或无教师学习，或无指导学习，因为和分类学习相比，聚类样本没有标记，需要由聚类学习算法来自动确定。聚类分析就是研究如何在没有训练的条件下把样本划分为若干类。

聚类（Clustering）是对物理的或抽象的样本集合分组的过程，是将数据划分成有意义或有用的组或子集，也称为簇。簇（Cluster）是数据样本的集合，聚类分析使得每个簇内部的样本之间的相关性比其他簇中样本之间的相关性更紧密，即簇内部的样本之间具有较高的相似度，而不同簇的样本之间具有较高的相异度。样本间的距离通常是描述样本间相似度的度量指标。

聚类分析作为数据挖掘的一个重要功能，聚类分析能作为一个独立的工具来获得数据的分布情况，观察每个类的特点，集中对某

些特定的类做进一步的分析。此外，聚类分析也可以作为其他算法（如关联分析和分类）的预处理步骤，这些算法再在生成的类上进行处理，这样可以大大提高这些算法的执行效率。因此，聚类分析已经成为数据挖掘领域中一个非常活跃的研究课题，已经开发了许多有效的聚类算法，新的算法还在不断涌现。

聚类算法具有以下特点：

（1）处理不同字段类型的能力。算法要求能处理不同类型的字段的能力。目前很多聚类算法是处理数值型字段，但实际应用中也会出现要求对其他类型字段，如字符型字段的数据进行聚类。

（2）可伸缩的能力。数据挖掘（Data Mining）是在大型数据存储库中，自动地发觉有用信息的过程。数据挖掘技术用来探查大型数据库，发现先前未知的有用模式。所以数据挖掘主要是研究大型的数据库，因此可伸缩性是一个基本要求。可伸缩性是指算法能处理大数据量的数据库样本，例如 WEB 数据库中上亿条记录的数据库样本。这就要求算法的时间复杂度最好是多项式时间，时间复杂度不至于过高。目前的聚类算法在小数据集上都是有效的，但随着网络的发展，大型数据库得到广泛应用。现有的聚类算法在处理这些大型数据库数据时，结果可能出现错误或偏差。

（3）处理高维数据的能力。大型数据库一般都含有若干个维或属性。较早出现的聚类算法通常是解决低维的数据集，对于处理高维数据时准确率会降低。所以，针对高维数据的聚类算法是新的课题，尤其是在高维空间中，数据的分布通常是稀疏又倾斜的，而且形状也是无规则的。目前，已经出现了一些针对高维数据的聚类算法。

（4）发现任意簇形状的能力。大多数聚类算法是基于距离度量的，例如使用欧几里得距离的 K－均值算法。这一类算法所发现的聚类通常是一些球状，并且密度、大小相近的簇。但是，实际存在的数据集可能是任意形状的，并且密度、大小也不尽相同。因此，在实际应用中，聚类算法应具有发现任意簇形状的能力。

（5）处理异常数据的能力。数据集中经常包含异常数据，例如：缺失值、孤立点、未知的错误信息等。如果聚类算法对这些数据异

常敏感，就会导致错误的聚类结果。所以，在处理孤立点时，需要尽量排除或降低孤立点对聚类结果的影响。但针对一些特殊应用，如商业欺诈的数据分析，又要求对数据的孤立点极其敏感。因此，这类的应用要求聚类算法在执行过程中，合理发现孤立点。目前，针对孤立点研究也出现了一些聚类算法。

（6）对输入参数的弱依赖性。大部分聚类算法都要求用户输入特定的参数，例如聚类数目。聚类分析的结果通常都对这些参数很敏感，参数的变化也可能导致变化很大的聚类结果产生。但是处理高维数据时，这些参数是很难确定的，参数的选取是用户的一个难题。因此，优秀的聚类算法应该具有对输入参数的弱依赖性。

（7）聚类结果的分析能力。最终面对用户的是聚类结果，所以一个好的聚类算法应该提供给用户一个易于理解、易于解释、易于分析应用的聚类结果。领域知识如何影响聚类分析算法的设计是很重要的一个研究方面。

1.2 聚类分析的基础概念

聚类分析的输入可以用一组有序对（X，s）或（X，d）表示，这里 X 表示一组样本，s 和 d 分别是度量样本间相似度或相异度（距离）的标准。聚类系统输出时对数据的区分结果，即 $C=\{C_1, C_2, \cdots, C_k\}$，其中 $C_i(i=1, 2, \cdots, k)$ 是 X 的子集，且满足如下条件：

$$C_1 \cup C_2 \cup \cdots \cup C_k = X$$

$$C_1 \cap C_2 = \Phi \quad (i \neq j)$$

C 中的成员 C_1，C_2，…，C_k 称为类或者簇。每一个类可以通过一些特征来描述。通常有如下几种表示方式：

（1）通过类的中心或类的边界点表示一个类。

（2）使用聚类树中的结点图形化地表示一个类。

（3）使用样本属性的逻辑表达式表示类。

用类的中心表示一个类是最常见的方式，当类是紧密的或各向分布同性时用这种方法非常好，然而，当类是伸长的或各向分布异性时，这种方式就不能正确地表示它们了。

1.2.1 聚类算法的主要类型

簇的集合通常称为聚类。聚类方法主要有以下几种类型：

（1）基于划分的聚类算法。对于一个给定的 N 个样本或元组的数据集，采用目标函数最小化的策略，通过迭代将数据构造成 K 个分组，每个划分块为一个簇，这就是划分方法。划分聚类（Partitional Clustering）简单地将数据对象集划分成不重叠的子集（簇），使得每个数据对象恰在一个子集中。对于给定的 K，算法首先给出一个初始的分组方法，以后通过反复迭代的方法改变分组，使得每一次改进之后的分组都较之前一次好。所谓的“好”的标准是同一分组中的记录越紧越好，而不同分组中的记录越远越好。划分方法满足两个条件：

1）每个分组至少包含一个样本。

2）每个样本比属于且仅属于某一个分组。

常见的基于划分的聚类算法有 K－均值方法、K－中心点方法和在这两种方法上的改进算法，如模糊 c－均值算法，这些算法在下一节会详细介绍。

（2）基于层次的聚类算法。如果允许簇具有子簇，则我们就会得到一个层次聚类（Hierarchical Clustering）。这实际上是将簇本身逐步分组，使得在每一层，组内聚类样本之间比不同组的样本之间更为相似。这种方法对给定的数据集进行层次的分解，直到某种条件满足为止。具体又可分为“自底向上”和“自顶向下”两种方案。

分层聚类技术可以从小到大分层次创建聚类，反映了将信息按不同程度总结和概括起来的一种方法。

层次聚类的过程可以组织成一棵树，除叶节点外，树中每一个节点（簇）都是其子女（子簇）的并，而树根是包含所有对象的簇。

层次聚类按数据分层建立簇，形成一棵以簇为节点的树，称为聚类图。如果按自底向上层次分解，则称为凝聚（Agglomerative）的层次聚类；如果按自顶向下层次分解，就称为分裂（Divisive）的层

次聚类。每一层表示把数据划分为观测不相交的簇的特定分组。整个分层结构表示分组的一个有序序列。

1）凝聚：凝聚的曾经聚类采用自底向下的分层策略。它从底部开始，把每一个点作为一个单独的簇，有与记录数量一样多的簇，其中每一簇仅仅包含一条记录。在每一层递归地将两个选定的簇合并为一个簇，对簇进行适当合并，相互之间接近的簇合并在一起形成下一个较大的簇。重复地合并两个最靠近的簇，直到产生单个的、包含所有点的簇（即终止条件），该簇是在层次体系中位列最高层。其中某些技术可以用基于图的聚类解释，而另一些可以用基于原形的方法解释。

2）分裂：分裂的层次聚类采用自顶向下的策略，该策略正好是与凝聚聚类技术相反。分裂聚类技术开始时将所有样本置于同一个簇中，然后，试着将该簇分成更小的簇，在每一层递归地将当前层中的一个簇分裂为两个新簇，直到满足某个终止条件。选取分裂产生具有最大组相异度的两个新组。

在上述两种聚类技术中，凝聚聚类技术更常用来做聚类分析，根据其原理也产生了很多聚类算法。

（3）基于密度的聚类算法。通常聚类算法都是基于不同的距离的，这样的聚类算法只能局限于发现“类圆形”聚类，为了克服这样的缺点，出现了基于密度的聚类算法，它与其他的聚类算法的一个根本区别是：它不是基于各种各样的距离的，而是基于密度的。由于数据集中可能出现球形、线形、延展形等多种形状的簇，所以一个好的聚类算法，应具有能够发现任意形状簇的能力。基于密度的聚类算法的基本思想是，只要一个区域中的点的密度大过某个阈值，就把它加到与之相近的聚类中去。

基于密度的方法主要有两类，即基于连通性的算法，如 DBSCAN，DBCLASD 和基于密度函数的算法，如 DBNCLUE 等。

（4）基于网格的聚类算法。基于网格的聚类算法的基本思想是：首先将数据集划分成有限个单元的网格结构，所有的处理都是以单个单元为样本的。这样处理的一个明显的优势是处理速度很快，通常，它只与把数据划分成多少个单元有关，而与目标数据库中记录

的个数无关。典型的基于网格的算法有：CLIQUE 算法、STING 算法等。

（5）基于模型的聚类算法。基于模型的聚类算法的目标是优化给定数据与某些数学模型之间的拟合。它给每一个聚类假定一个模型，然后去寻找能够很好地满足这个模型的数据集。基于模型的聚类方法主要分为统计学方法和神经网络方法等。

目前，基于统计学的聚类方法主要有 COBWEB 算法、CLASSIT 算法和 AutoClass 算法。

在神经网络方法中，每个簇被描述为一个样本。该算法主要包括竞争学习神经网络和自组织特征映射神经网络。神经网络的聚类算法的缺点是处理时间较长，具有较高的数据复杂性，较适于大型的数据库。

（6）基于图的聚类算法。如果数据用图表示，其中节点就是对象，而边代表对象之间的联系，则簇可以表示为连通分支，即互相连通但不与组外对象连通的对象组。基于图的簇中一个重要应用是基于邻近的簇，其中两个样本是相连的，在基于邻近的簇中，每个样本到该簇某个样本的距离比到簇中其他样本的距离更近。

1.2.2 聚类分析的相似度和相异度

相似度和相异度是聚类分析中两个重要的概念。

两个对象之间的相似度是这两个对象相似程度的数值度量。因而，两个对象越相似，它们之间的相似度就越高。通常，相似度的取值在 0（不相似）和 1（完全相似）之间取值。

两个对象的相异度是这两个对象差异程度的数值度量。对象越相似，它们的相异度就越低。通常，距离用作相异度的同义词。一般来说，相异度的取值在［0，1］之间，但有时也会在［0，∞］之间取值。

1.2.2.1 相似度

如果映射 $s: x \times y \rightarrow R$ 满足 $\forall i, j, h$：

（1）$s(X_i, X_j) \geqslant 0$；

（2）$s(X_i, X_j) = s(X_j, X_i)$；

(3) $s(X_i,X_j) \leqslant s(X_h,X_{h)}$;

(4) $s(X_i,X_j) = s(X_j,X_j)$;

则称 s 为相似度。

常用的相似度测量为夹角余弦。设样本 X_i 和 X_j 之间的夹角为 $\theta(X_i, X_j)$，显然，$\theta(X_i, X_j) \in [0, +180]$。夹角余弦为：

$$\cos[\theta(X_i,X_j)] = \frac{\sum_{k=1}^{m} x_{ik}x_{jk}}{\left(\sum_{k=1}^{m} x_{ik}^2 \sum_{k=1}^{m} x_{jk}^2\right)^{\frac{1}{2}}}$$

显而易见，当角度为0时，夹角余弦为1（最大），其相似度为1（最相似）；夹角越大，其夹角余弦值越小，相似度也越小。

1.2.2.2 相异度

一个聚类分析过程的质量取决于对度量标准的选择，因此必须仔细选择度量标准。在通常情况下，聚类算法不是计算两个样本间的相似度，而是用特征空间中的距离作为度量标准来计算两个样本间的相异度。对于某个样本空间来说，距离的度量标准可以是度量的或半度量的，以便用来量化样本的相异度。相异度的度量用 $d(x, y)$ 来表示，通常称相异度为距离。当 x 和 y 相似时，距离 $d(x, y)$ 的取值很小；当 x 和 y 不相似时，$d(x, y)$ 就很大。

距离为相异度测量，距离为0时，相异度为0（最不相异）；距离越大，相异度越大。设 $d(X_i, X_j)$ 为样本 X_i 和 X_j 之间的距离。距离函数 $d(X_i, X_j)$ 应满足如下条件：

(1) $d(X_i, X_j) = 0$，当且仅当 $X_i = X_j$；

(2) 非负性：$d(X_i, X_j) \geqslant 0$；

(3) 对称性：$d(X_i, X_j) = d(X_j, X_i)$；

(4) 三角不等式：$d(X_i, X_j) \leqslant d(X_i, X_k) + d(X_k, X_j)$。

最常用的距离度量方法有欧几里得距离、切比雪夫距离、明可夫斯基距离、马氏距离和余弦距离等。

(1) 欧几里得距离。欧几里得距离定义：

$$d(X_i,X_j) = \| X_i - X_j \|^2 = \left(\sum_{k=1}^{m} w_k | x_{ik} - x_{jk} |^2\right)^{\frac{1}{2}}$$

（2）切比雪夫距离。切比雪夫距离定义：

$$d(X_i,X_j) = \| X_i - X_j \|_\infty = \max_{k\in\{1,2\cdots,m\}} | x_{ik} - x_{jk} |$$

（3）马氏距离。马氏距离是一种有效的计算两个未知样本集的相似度的方法，它表示数据的协方差距离。与欧氏距离不同的是它考虑到各种特性之间的联系并且是尺度无关的（Scale - invariant），即独立于测量尺度。

设样本为 $(x_i, y_i) \in R^n \times R^1 (i = 1, 2, \cdots, m)$，共有 m 个样本，x_i 是 n 维特征矢量，$y_i \in \{-1, 1\}$ 表示 x_i 的类标号。令 X 代表 $m \times n$ 的输入矩阵，每行为一个样本，则样本的均值、自相关矩阵和协方差矩阵可用矩阵表示为：

$$\mu = E\{X\} = X^T\left(\frac{1}{m}\right)_{m\times 1}$$

$$S = \left(\frac{1}{m}\right)X^TX, \Sigma = E\{(X-\mu)^T\} = \left(\frac{1}{m}\right)X^TX - \mu\mu^T$$

其中 $\left(\frac{1}{m}\right)_{m\times 1}$ 代表元素均为 $\frac{1}{m}$ 的 m 维列矢量。样本 x_i 到样本总体 X 的马氏距离定义为：

$$d^2(x_i,X) = (x_i - \mu)^T\Sigma_i^{-1}(x_i - \mu)$$

1.3 聚类分析算法

聚类分析是一个活跃的研究领域，已经有大量的、经典的和流行的算法涌现。常用的聚类分析算法有基于划分的聚类算法、基于层次的聚类算法、基于密度的聚类算法、基于网格的聚类算法以及基于模型的聚类算法等。

1.3.1 聚类算法性能的衡量指标

通常情况下，对于同一个问题可以用几种不同的算法来解决，衡量算法性能主要从以下几方面进行衡量：

（1）可伸缩性。如果一个算法既适用于小型数据集合，又在大型数据库上进行聚类不会导致有偏差的结果，则称这个算法具有高度的可伸缩性。