基于数据间内在关联性的自适应模糊聚类模型

唐成龙 [1] 王石刚 [1]

摘 要 提出了一种新的模糊聚类模型 (Fuzzy C-means clustering model, FCM), 称为自适应模糊聚类 (Adaptive FCM, AFCM). 和现有的大多数模糊聚类方法不同的是, AFCM 考虑了数据集中全体数据的内在关联性, 模型中引入了自适应度向量 $\boldsymbol{W}$ 和自适应指数 p. 其中, $\boldsymbol{W}$ 在迭代过程中是自适应的, p 是一个给定参数. $\boldsymbol{W}$ 和 p 共同作用调控聚类过程. AFCM 同时输出三组参数: 模糊隶属度集 U, 自适应度向量 $\boldsymbol{W}$, 以及聚类原型集 V. 本文给出了两组数据实验验证 AFCM 的性能. 第 1 组实验验证 AFCM 的聚类性能, 以 FCM 为比较对象. 实验表明 AFCM 可以得到更好的聚类质量, 而且通过合理选择自适应指数 p, AFCM 和 FCM 在时间复杂性上保持同一水平. 第 2 组实验检验了 AFCM 的离群点挖掘性能, 以目前常用的基于密度的 LOF 为比较对象. 实验表明 AFCM 算法具有极大的计算效率优势, 且 AFCM 得到的离群点是全局的, 反映的是离群点和整个数据集的关系, 离群点涵盖的信息也更丰富. 文章指出, AFCM 在挖掘大数据集和实时数据中的离群点应用方面, 以及获得高质量的聚类结果的应用方面, 特别在聚类的同时需要挖掘离群点的应用方面具有独特的优势.

关键词 模糊聚类, 离群点挖掘, 自适应聚类策略, 自适应度, 自适应指数

DOI 10.3724/SP.J.1004.2010.01544

Adaptive Fuzzy Clustering Model Based on Internal Connectivity of All Data Points

TANG Cheng-Long[1] WANG Shi-Gang[1]

Abstract This paper proposes a new kind of fuzzy C-means clustering model (FCM), which is named as adaptive fuzzy clustering (AFCM). Different from most current fuzzy clustering methods, the AFCM considers the internal connectivity of all data points. An adaptive degree vector $\boldsymbol{W}$ and an adaptive exponent p are introduced into the model to jointly influence the clustering process. The AFCM simultaneously outputs three categories of parameters: fuzzy membership degree matrix U, adaptive degree vector $\boldsymbol{W}$, and cluster prototype matrix V. Two groups of numerical experiments, Group 1 and Group 2, were executed to evaluate the AFCM. Group 1 demonstrates the clustering performance of the AFCM, with FCM being its counterpart, and the results showed that the AFCM can obtain better clustering quality, meanwhile its time complexity can hold the same level as that of the FCM by choosing the available p. Group 2 checks the ability of the AFCM in mining the outliers, with the density-based LOF being its counterpart and the results showed that the AFCM has considerable advantages in computing efficiency, and that the outliers minded by the AFCM are global, and reflect the relationship between the outliers and the whole data set. It is pointed out that the AFCM possesses the unique advantages when mining the outliers of the large-scale or dynamic data sets, and clustering the data set for better clustering results, especially when it is necessary to simultaneously fulfill both tasks of clustering and mining outliers.

Key words Fuzzy clustering, outliers mining, adaptive clustering approach, adaptive degree, adaptive exponent

人工智能的研究和应用涉及到众多的子领域[1]. 本文讨论其中的两个: 基于聚类的模式识别和离群点挖掘. 对于一个给定数据集, 基于聚类的模式识别是指采用聚类的方法将该数据集划分为若干模式 (类). 而离群点挖掘, 顾名思义, 是指挖掘数据集中异常数据, 也称离群数据[2]. 当前, 一方面, 基于聚类的模式识别和离群数据挖掘有着密切的关联. 在进行模式识别前, 数据集中的离群点通常需要事先处理, 处理的方法有替换离群数据或者直接将之剔除, 目的是消除或者减弱离群数据对模式识别结果的负面影响. 另一方面, 在对待离群数据上, 二者又有着明显的区别. 对于基于聚类的模式识别来说, 离群数据通常被认为是 "有害" 的, 此时所采取的主要措施是围绕如何消除或减轻离群数据的 "有害" 性展开, 离群点的判定通常是作为附属品出现的. 而在离群点挖掘研究领域, 离群点本身成为 "焦点", 离群点是 "挖掘", 而不是简单的 "判定". 另外, 在多数情况下, 这两个领域内所使用的研究方法差异明显. 在离群点挖掘领域, 当前主要有基于统计、密度、距离、特征偏差的方法[2], 基于聚类的离群点挖掘方法虽

收稿日期 2009-11-11 录用日期 2010-05-28
Manuscript received November 11, 2009; accepted May 28, 2010
国家自然科学资金资助 (50875159)
Supported by National Natural Science Foundation of China (50875159)
1. 上海交通大学机械及动力工程学院 上海 200240
1. School of Mechanical and Dynamical Engineering, Shanghai Jiao Tong University, Shanghai 200240

有报道, 但有限. 在数据分析的实际应用中, 有这样一种需求: 对于某个给定的数据集, 已知其含有一定量的离群点, 对该数据集的分析同时包括三个任务: 第一, 对该数据集模糊聚类, 将数据集划分为若干类后, 确定各个数据点的类隶属关系; 第二, 构建分类器, 分类器的构建基于聚类所获得的聚类原型集; 第三, 离群点挖掘, 包括判定离群点并发掘离群点所蕴含的丰富信息. 当前这三个任务主要是在不同的领域内解决, 例如, 第一、二个任务是基于聚类的模式识别和分类的任务, 第三个任务是数据挖掘和知识发现领域内的研究内容. 在实际应用中, 这三个任务往往是需要同时解决的. 将这三个任务作为一个整体同时解决, 目前报道的比较少, 缺乏成熟的方法.

聚类分析是现代数据分析的重要工具之一. 聚类分析的基本思想是将一个数据集中全体数据按照 "相似" 或者 "不相似" 的原则划分为若干个类. 类内数据尽可能相似, 而类间数据尽可能不相似. 聚类分析有 "刚性" 和 "模糊" 之分. 相比于刚性聚类, 模糊聚类分析丰富了数据结构的描述手段. 模糊聚类的任务在数学上可以抽象为一个目标函数求最优解的问题. 当前广泛研究和应用的模糊聚类大多是基于均值的模糊聚类模型 (Fuzzy C-Means clustering model, FCM)[3]. FCM 的目标函数值取决于聚类的原型和数据的模糊隶属度. 用 $J_{\rm FCM}$ 表示 FCM 的目标函数, $J_{\rm FCM}$ 的定义如式 (1), 模糊隶属度 u_{ij} 的约束条件见式 (2).

$$J_{\rm FCM}(X,U,V)=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}d_{ij}^{2} \tag{1}$$

$$\text{s.t.}\quad u_{ij}\in[0,1],\ \sum_{j=1}^{n}u_{ij}>0,\ \sum_{i=1}^{c}u_{ij}=1 \tag{2}$$

上式中, X, U 和 V 分别称为数据集, 模糊隶属度集和聚类原型集. u_{ij} 称为数据 $\boldsymbol{x}_j$ 和类原型 $\boldsymbol{v}_i$ 之间的模糊隶属度, d_{ij} 称为数据 $\boldsymbol{x}_j$ 和类原型 $\boldsymbol{v}_i$ 之间的距离, m 称为模糊指数, n 和 c 分别为 X 中数据和 V 中类原型的个数.

由于很难直接求解此类目标函数的最优值, 往往使用交互优化 (Alternative optimization, AO) 的求解策略, 即交替进行迭代过程, 直至目标函数收敛到指定精度内的最优值. 迭代过程中涉及到两个更新式: 模糊隶属度和聚类原型. 这两个迭代更新式分别见式 (3) 和 (4).

$$u_{ij}=\frac{d_{ij}^{-\frac{2}{m-1}}}{\sum\limits_{i=1}^{c}d_{ij}^{-\frac{2}{m-1}}} \tag{3}$$

$$v_i=\frac{\sum\limits_{j=1}^{n}u_{ij}^{m}\boldsymbol{x}_j}{\sum\limits_{j=1}^{n}u_{ij}^{m}} \tag{4}$$

上式中, 距离测度通常是欧氏距离, 适应于不同的需要, 也有其他的距离测度, 例如马氏距离等. 本文实验部分所采用的距离均为马氏距离, 下文不再提及. FCM 方法采用的是概率型约束条件, 即任一个数据和 c 个聚类原型之间的模糊隶属度的和是 1, 即某一个数据以概率的形式属于某个类. 这个约束条件用于产生模糊隶属度的更新等式. 然而, FCM 得到的模糊隶属度并不总是符合模糊隶属度原本应具有的直观概念, 即隶属度越大, 属于相应的类的概率越大. 当 FCM 用于含噪数据环境时, 存在着明显的缺点.

本文提出了一种新模型, 旨在解决上述问题. 提出了一种新的模糊聚类模型, 即自适应模糊均值聚类 (Adaptive FCM, AFCM). 这一新方法的核心思想是: 数据集中各个数据本质上是 "各异" 的, 而数据之间是内在关联的. AFCM 在对数据集进行聚类分析时, 引入了表达数据之间的相互关联的约束条件, 表现为在 AFCM 模型中引入了一个自适应度向量 (一个数据对应一个自适应度) 和一个作用于该自适应度的自适应指数. 自适应度向量和自适应指数共同作用, 达到调控聚类过程, 由此实现 "各异" 地处理各个数据的思想. 由于 AFCM 考虑了各个数据之间的差异性, 因此具有很强的处理离群数据的能力. AFCM 一方面通过有针对性地处理离群点来提高聚类质量, 另一方面也可以直接判定出离群点及发掘离群点所蕴含的丰富信息, 而且二者是同时实现的. 相比之下, 当前绝大多数模糊均值聚类模型由于没有考虑各个数据之间的关联性, 在聚类过程中, 各个数据被认为是不关联的, 以一种 "等同" 的方式处理, 因而存在各种问题. 本文给出了 AFCM 完整的理论模型, 并结合数据实验, 验证了 AFCM 在获得高质量聚类结果和离群点挖掘上的能力, 特别是二者需要同时解决时, AFCM 所具有的独特优点.

本文后续组织如下: 第 1 节回顾了现有模糊聚类和离群点挖掘方面的有关工作. 第 2 节是理论部分, 提出了完整的自适应模糊聚类模型 AFCM. 第 3 节是实验部分, 通过两个真实数据集测试和验证了 AFCM 的性能. 实验分为两组, 第 1 组实验验证了 AFCM 的聚类质量以及算法的时间复杂性, 比较对象是当前广泛使用的模糊均值聚类算法 (FCM); 第 2 组实验验证了 AFCM 在离群点挖掘方面的能力和时间复杂性, 比较对象是基于密度的 LOF 方法. 第 4 节讨论了两组新参数: 自适应度向量和自适应

指数. 第 5 节总结了研究结果.

1 相关工作

1.1 模糊均值聚类

为了克服 FCM 对异常数据敏感的缺点, Krishnapuram 等提出了 PCM 聚类模型[4]. PCM 将聚类问题引入到可能性理论的框架内. 和 FCM 模型不同的是, PCM 下数据集的划分是一种可能性划分, 而不是概率性划分, 即模糊隶属度 u_{ij} 此时解释为一个数据点属于某个类的可能性, 称为典型度, 用 t_{ij} 表示. PCM 通过放松典型度的约束条件来弱化数据集中异常数据对聚类结构的不利影响, PCM 模型有一些成功应用. 但是正如 Barni 等[5] 指出的, PCM 在放松约束条件换来对离群点不敏感的优点的同时, 却带来了对初始凝聚点很敏感的缺陷, 更为严重的是, PCM 有聚类原型趋同的弊病.

为了解决 FCM 对离群点敏感及 PCM 的聚类原型趋同的缺点, Pal 等提出了一种模糊可能性均值聚类算法, 称为 FPCM[6]. FPCM 算法可以同时输出模糊隶属度 u_{ij} 和典型度 t_{ij}, 以及每个类的原型. FPCM 模型中的模糊隶属度的定义及约束条件同 FCM 模型的模糊隶属度. 典型度的定义和 PCM 中的典型度是一样的, 区别是 FPCM 下的典型度有 "和为 1" 的约束条件. 与 FCM 以及 PCM 不同的是, FPCM 输出三组参数: 模糊隶属度集 $U(c\times n)$, 典型度集 $T(c\times n)$ 以及聚类原型集 $V(c\times q)$, 其中 q 是数据的维数.

尽管 FPCM 可以同时解决 FCM 的离群点敏感以及 PCM 的聚类原型趋同的弊病, 但是当数据集中的数据量比较大时出现了新问题. 因为 FPCM 模型中典型度的约束条件是全体数据点到某个聚类原型的典型度的和为 1, 即在此 "和为 1" 的约束条件式中共有 n 个典型度因子. 当 n 很大的时候, 典型度的值会变得很小, 以至于失去了 "典型" 的原意. 而且在算法的计算中, 如果典型度的值很小, 会出现迭代不收敛等问题. 为了解决这个问题, 2005 年 Pal 等在 FPCM 的基础上, 又提出了一种所谓的可能性模糊均值聚类模型, 简称为 PFCM[7]. 类似于 FPCM, PFCM 也同时输出模糊隶属度和典型度. 此处, PFCM 模型中的模糊隶属度的定义和约束条件同 FCM 或者 FPCM 模型的模糊隶属度. 而典型度的定义和 PCM 中的典型度定义是一样的, 即解除了全体典型度 "和为 1" 的约束条件.

实际上, PFCM 模型是 PCM 和 FCM 的综合体, PFCM 没有 FCM 的对离群点敏感的缺点, 没有 PCM 的聚类原型趋同的缺点以及 FPCM 应用于大数据集时的缺点. 但是, PFCM 引入了众多需用户自行定义的参数, 使得模型变得复杂了许多. 特别是如何合理地定义这些参数是新问题, 这些问题使得该模型的推广性变差.

PCM, FPCM 以及 PFCM 模型的提出主要是为了解决聚类过程中离群点的影响问题, 这一点和本文提出的自适应聚类模型有相似之处. 但这些模型并不回答究竟哪些点是离群点这一问题. 因此, 这些模型主要用于聚类, 不适用于离群点挖掘. 除了这些模型之外, 在模糊聚类新模型的开发研究中, 典型工作还有 Kang 等[8] 提出的一种改进的聚类模型以及应用. 对目标函数的自身修改, 主要是为了满足具体的应用, 例如减少离群点对聚类结果的负面影响, 或者提高诸如图象处理的质量. 在图像处理上, Chuang 等[9] 提出了一种所谓的 SFCM 算法, 该算法将图像局部信息引入到目标函数的表达式中, Cai 等[10] 在 Chuang 工作的基础上, 又提出了改进的算法 FAFCM.

以上模型可认为都是在 FCM 模型的基础上加以改进得到的. 故以 FCM 为代表展开以下的分析. 由式 (1) 可知, FCM 目标函数是 n 个因子的和, 每个数据点对应一个因子, 将每个因子加上前缀 "1", 则 FCM 目标函数可以改写成为式 (5) 的形式. 注意到 FCM 模型下, 由于 FCM 没有给出数据之间的约束条件, n 个数据之间是相互独立的.

$$J_{\mathrm{FCM}}(X,U,V)=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}d_{ij}^{2}=\underbrace{1\cdot\sum_{i=1}^{c}u_{i1}^{m}d_{i1}^{2}}_{x_1}+\underbrace{1\cdot\sum_{i=1}^{c}u_{i2}^{m}d_{i2}^{2}}_{x_2}+\cdots+\underbrace{1\cdot\sum_{i=1}^{c}u_{in}^{m}d_{in}^{2}}_{x_n}\tag{5}$$

由式 (5) 可知, FCM 模型中, 任一数据点到聚类原型之间的模糊平方距离和的系数都是 "1", 这说明在 FCM 算法中, 离群数据和正常数据是一视同仁的, 即便事实上数据集中存在着离群点. 而从数据的本质来看, 各个数据的本质是 "各异" 的. 一视同仁地对待所有数据点所带来的缺点也解释了 FCM 及其扩展算法对离群点敏感的原因, 因为 FCM 算法没有给出反映数据间关联性的约束, 不能反映出数据的 "各异" 性, 当然也看不出离群点的 "离群" 性.

另外, 从 FCM 求最优解的过程来看, 在将给定数据集划分为若干类前, 聚类的数目 c 和初始凝聚点需要事先给定, 之后聚类的结果很大程度地取决

于模糊指数 m. 即对于 FCM 模型, m 是调节聚类结果的唯一渠道. 然而, 关于 m 的取值范围, Pal 等[11] 进行了深入研究, 认为 m 的合理取值范围是 1.5～2.5, 通常取为 2. 这表明, 这个渠道事实上是很窄的.

聚类方面的相关工作总结如下: 现有的众多模糊聚类方法均可认为是由 FCM 模型扩展得到的, 它们都面临一个共同问题, 即如何处理离群点. 尽管 FCM, FPCM 和 PFCM 等模型在回答这个问题方面做了很多有价值的工作, 但是仍然存在着亟待解决的新问题. 本文的出发点也是回答如何处理离群点的问题, 但是本文提出的是一个全新思路, 即考虑了数据间的内在关联性, 与现有方法显著不同. 本文提出聚类方法, 既为了得到更好的聚类质量, 也为了可以直接对数据集中的离群点进行判定和挖掘, 将聚类和离群点挖掘合二为一.

1.2 离群点挖掘

离群点挖掘是数据挖掘和知识发现研究领域的基本任务之一, 其目的是发掘数据集中异常对象的不同属性. 在离群点挖掘研究领域, 离群点本身成为研究的 "焦点", 这一点是不同于前面所讨论的模糊聚类模型的, 在第 1.1 节的讨论中, 离群点被认为是有害的, 挖掘离群点是放在从属地位的. 在现实的许多情况下, 例如, 网络入侵监测、信用卡防欺诈、复杂工业过程中参数的异常波动等, 罕见的事件往往比正常事件更值得关注. 从数学语言描述, 罕见的事件即是一种离群点.

离群点挖掘任务通常描述成所谓 "Top-k" 的原则[2], 即给定含有 n 个数据点的数据集以及预期的离群点数目 k, 与剩余的数据相比, 发现显著相异或者不一致的前 k 个数据, 并将这些数据判定为离群点. 离群点挖掘中的一个基本问题是, 什么样的数据被认为是显著相异或者不一致的.

Han 等[2] 将目前最广泛使用的离群点挖掘方法划分为四种, 即基于统计分析[12]、密度[13]、距离[14−15] 以及特征偏差的方法, 它们主要采用两种策略: 一种是二进制的定性判定, 即某个数据点是或不是离群点, 如基于统计和距离的方法; 另一种是给每个数据附以一个离群程度的定量判定, 表达一个数据在多大程度上是一个离群点, 如基于密度的方法. 在以上四种主要方法中, 对基于密度和基于距离的方法研究和应用的最多. 关于基于密度的方法, 典型代表是 Breunig 等[13] 提出的局部离群因子模型, 称为 LOF 模型. LOF 模型计算每个数据的离群程度, 然后根据 "Top-k" 原则, 发掘出 k 个离群点. 在基于距离的方法研究上, Ghoting 等[14] 提出了一种基于距离的快速离群点检测方法. Weng 等[15] 提出了一种挖掘时间序列数据集中离群点的方法, 通过采用扩展的 Frobenius 范数来计算距离, 并由此来确定离群点的存在.

除了上述四种主要的离群点挖掘方法外, 也有聚类分析方法用于离群点挖掘的报道. 下面以 NC (噪声聚类) 模型[16] 为代表说明. NC 模型中引入了一个所谓的 "噪声" 类, 将判定为噪声的数据全部划入到这个噪声类中去. NC 模型采用了一种距离度量来定义离群点, 该距离称为噪声距离. 某个数据属于该噪声类的模糊隶属度或者概率, 随着这些数据点到其他正常类的距离的增大而增大. NC 聚类的主要目的是为了减少噪声数据对正常类原型的不利影响, 其关注点并不是放在准确识别离群点上. 然而, 在许多应用中, 离群包含有重要信息, 而这些信息是希望被有效挖掘的. 显然, NC 模型不适合此类应用. 现有的模糊聚类方法用于离群点挖掘时, 它的主要关注点仍然是 "聚类", 由聚类结果附带判定离群点, 而不是从获得最佳的离群点挖掘的角度来优化求解的. 而且, 在大多数情况下, 离群点的判定标准是隐含的. 因此, 当前模糊聚类的方法用于离群点挖掘存在着不足.

当前离群点挖掘有以下缺点: 1) 所挖掘的信息不够, 例如, 对某个数据点仅作一个二元判定, 显然离群点所蕴含的丰富信息没有充分得到挖掘. 2) 现有方法下, 离群点的物理意义有时难以解释. 例如, 在 LOF 算法下, 离群点表达的是某个数据的局部信息, 不能反映某个数据和整个数据集的关系. 3) 现有的方法存在着计算效率低的问题, 例如, 对于使用最广泛的基于密度的方法, 其在计算每个数据点的参数时, 都要扫描整个数据集, 对于大数据集, 或者动态数据集, 计算效率低成为阻碍其运用的主要问题.

总的来说, 无论是现有的模糊聚类方法, 还是离群数据挖掘方法, 均存在着明显的不足. 理论研究和实际应用都对开发新的方法有需求, 特别是在同时解决这两个领域的问题时.

2 考虑数据内在关联的自适应模糊聚类模型

本节首先给出了 AFCM 目标函数的定义式, 给出相关迭代参数的更新表达式. 之后分析了 AFCM 如何用于聚类和离群点挖掘, 给出了计算流程. 最后分析了 AFCM 的主要特点.

2.1 自适应模糊聚类 AFCM

2.1.1 AFCM 模型的目标函数

所提出的自适应模糊聚类的目标函数见式 (6), 并给出两个优化的约束条件, 分别见式 (7) 和 (8).

$$J_{\mathrm{AFCM}}(X,U,V,\boldsymbol{W})=\sum_{j=1}^{n}w_j^p\sum_{i=1}^{c}u_{ij}^m d_{ij}^2 \qquad (6)$$

$$\text{s.t.} \quad \sum_{j=1}^{n} u_{ij} > 0, \quad \sum_{i=1}^{c} u_{ij} = 1 \tag{7}$$

$$\prod_{j=1}^{n} w_j = 1 \tag{8}$$

J_{AFCM} 称为 AFCM 模型的目标函数. 一组新的参数 w_j 引入到 J_{AFCM} 中. w_j 称为数据 x_j 的自适应度, w_j 反映了数据 $\boldsymbol{x}_j$ 对目标函数 J_{AFCM} 的影响程度. 所有的 w_j 构成一个向量, 称为自适应度向量, 用 W 表示. 式 (6) 中还引入了一个新的参量 p ($p \neq 0$), p 称为自适应指数, 是一个预先给定的常数, 用来调节自适应向量的取值. 式 (8) 表明全体自适应度 w_j 施加了 "乘为 1" 这样的约束条件, 这个约束条件描述了数据间的内在关联性. 式 (6) 和式 (1) 中其他相同的符号具有相同的定义.

和 FCM 模型类似, J_{AFCM} 求最优解也采用交互迭代的策略. 自适应度向量 W 在给出处置后迭代更新, 因此需要给出自适应度 w_j 的更新表达式.

聚类目标函数求最优是一种非线性优化问题, 通常可以用拉格朗日极值法来求解. 在考虑两类约束条件的前提下, 优化函数可以写为

$$J_{\text{AFCM},\phi_1,\phi_2}(X,U,V,\boldsymbol{W}) = \sum_{j=1}^{n} w_j^p \sum_{i=1}^{c} u_{ij}^m d_{ij}^2 + \phi_1\left(\sum_{i=1}^{c} u_{ij} - 1\right) + \phi_2\left(\prod_{j=1}^{n} w_j - 1\right) \tag{9}$$

此处, ϕ_1 和 ϕ_2 是两个拉格朗日乘法算子, 分别对应于模糊隶属度和自适应度的约束条件.

在式 (9) 的两侧分别对 u_{ij} 和 w_j 求偏导数, 得到

$$\begin{cases} \dfrac{\partial J_{\text{AFCM},\phi_1,\phi_2}}{\partial u_{ij}} = m \cdot w_j^p \cdot u_{ij}^{m-1} d_{ij}^2 + \phi_1 \\ \dfrac{\partial J_{\text{AFCM},\phi_1,\phi_2}}{\partial w_j} = p \cdot w_j^{p-1} \cdot \sum\limits_{i=1}^{c} u_{ij}^m d_{ij}^2 + \phi_2\left(\prod\limits_{k=1,k\neq j}^{n} w_k\right) \end{cases} \tag{10}$$

令式 (10) 中两个偏导函数均为零, 并变换等式, 得到 ϕ_1 和 ϕ_2 的表达式为

$$\phi_1 = -m \cdot w_j^p \cdot u_{ij}^{m-1} d_{ij}^2 \tag{11}$$

$$\phi_2 = \frac{-p \cdot w_j^{p-1} \cdot \sum\limits_{i=1}^{c} u_{ij}^m d_{ij}^2}{\prod\limits_{k=1,k\neq j}^{n} w_k} \tag{12}$$

在求 J_{AFCM} 最优解的过程中, 有三组参数需要更新, 它们是模糊隶属度矩阵 $U(c\times n)$, 自适应度向量 $W(1\times n)$ 和聚类原型矩阵 $V(c\times q)$. 下文逐一给出各自的迭代更新表达式.

2.1.2 模糊隶属度 $U(u_{ij})$ 更新等式

变换式 (11) 可以得到含有 ϕ_1 的模糊隶属度 u_{ij} 的表达式 (13), 下一步工作是利用模糊隶属度的约束条件, 得到不含 ϕ_1 的 u_{ij} 的表达式为

$$u_{ij} = \left(\frac{\phi_1}{-m \cdot w_j^p}\right)^{\frac{1}{m-1}} d_{ij}^{-\frac{2}{m-1}} \tag{13}$$

考虑到 u_{ij} 表达的是数据 $\boldsymbol{x}_j$ 和聚类原型 v_i 之间的模糊隶属关系, 将 x_j 到全体聚类原型的模糊隶属度相加, 得到式 (14).

$$\sum_{i=1}^{c} u_{ij} = \left(\frac{\phi_1}{-m \cdot \ w_j^p}\right)^{\frac{1}{m-1}} \sum_{i=1}^{c} d_{ij}^{-\frac{2}{m-1}} \tag{14}$$

由模糊隶属度的约束条件可知, 式 (14) 的左侧为 1, 继续变换式 (14), 得到

$$\left(\frac{\phi_1}{-m \cdot w_j^p}\right)^{\frac{1}{m-1}} = \frac{1}{\sum\limits_{i=1}^{c} d_{ij}^{-\frac{2}{m-1}}} \tag{15}$$

综合式 (13) 和 (15), 得到不含 ϕ_1 的 u_{ij} 的表达式为

$$u_{ij} = \frac{d_{ij}^{-\frac{2}{m-1}}}{\sum\limits_{i=1}^{c} d_{ij}^{-\frac{2}{m-1}}} \tag{16}$$

从式 (16) 和式 (3) 可以看出, 在 AFCM 下, u_{ij} 的迭代式和 FCM 下 u_{ij} 的迭代式是相同的. 这表明在 AFCM 下, 模糊隶属度并没有改变 FCM 下模糊隶属度所表达的某个数据和聚类原型之间的模糊隶属关系. 这一点很重要, 因为 AFCM 是在 FCM 的基础上诞生的, 在引入了自适应度这一概念的同时, 并不希望改变模糊隶属度的本来定义.

2.1.3 自适应度 $W(w_j)$ 更新等式

变换式 (12), 并注意到等式的右边没有 w_j 项, 在等式的两侧同时乘以 w_j, 见式 (17). 考虑到全部自适应度的 "乘为 1" 这一约束条件, 式 (17) 的左侧为 1, 则得到 ϕ_2 的表达式, 见式 (18). 继续变换式 (18), 得到含有 ϕ_2 的 w_j 的表达式, 见式 (19). 下一步的工作是在 w_j 的表达式中消除 ϕ_2.

$$\phi_2\left(\prod_{k=1k\neq j}^{n} w_k\right) w_j = \left(-p \cdot w_j^{p-1} \cdot \sum_{i=1}^{c} u_{ij}^m d_{ij}^2\right) w_j \tag{17}$$

$$\phi_2 = -p \cdot w_j^p \cdot \sum_{i=1}^{c} u_{ij}^m d_{ij}^2 \tag{18}$$

$$w_j = \left[\frac{-\varphi_2}{p \sum\limits_{i=1}^{c} u_{ij}^m d_{ij}^2} \right]^{\frac{1}{p}} \tag{19}$$

式 (19) 表明, 对于任一数据 x_j, 均有一个乘法算子 ϕ_2 的表达式, 下标的取值为 1 到 n. 则对于所有数据, 式 (19) 均成立, 也即可以得到 n 个类似的等式. 将所有这些等式的左侧和左侧相乘, 右侧和右侧相乘, 显然有式 (20). 进一步转换式 (20), 并考虑到自适应度的 "乘为 1" 约束, 并逐渐推导, 最后得到不含 ϕ_2 的 w_j 的表达式, 见式 (24).

$$\phi_2^n = \prod_{j=1}^{n} \left[-p w_j^p \sum_{i=1}^{c} u_{ij}^m d_{ij}^2 \right] \tag{20}$$

$$\Rightarrow \phi_2^n = (-p)^n \left(\prod_{j=1}^{n} w_j \right)^p \prod_{j=1}^{n} \left[\sum_{i=1}^{c} u_{ij}^m d_{ij}^2 \right] \tag{21}$$

$$\Rightarrow \phi_2^n = (-p)^n \prod_{j=1}^{n} \left[\sum_{i=1}^{c} u_{ij}^m d_{ij}^2 \right] \tag{22}$$

$$\Rightarrow \phi_2 = p \left[\prod_{j=1}^{n} \left(\sum_{i=1}^{c} u_{ij}^m d_{ij}^2 \right) \right]^{\frac{1}{n}} \tag{23}$$

$$w_j = \left[\frac{\left[\prod\limits_{j=1}^{n} \left(\sum\limits_{i=1}^{c} u_{ij}^m d_{ij}^2 \right) \right]^{\frac{1}{n}}}{\sum\limits_{i=1}^{c} u_{ij}^m d_{ij}^2} \right]^{\frac{1}{p}} \tag{24}$$

式 (24) 用于在交互迭代过程中更新 w_j.

2.1.4　聚类原型 $\boldsymbol{V(v_i)}$ 更新等式

在 AFCM 模型下, 参考 FCM 算法直接给出聚类原型的迭代等式, 见式 (25). 此处 $\bar{v}_i$ 称为 AFCM 下的第 i 个聚类原型.

$$\bar{v}_i = \frac{\sum\limits_{j=1}^{n} \bar{u}_{ij}^m \boldsymbol{x}_j}{\sum\limits_{j=1}^{n} \bar{u}_{ij}^m} \tag{25}$$

此处, $\bar{u}_{ij}$ 称为自适应模型下的模糊隶属度, 其计算式为

$$\bar{u}_{ij}^m = w_j u_{ij}^m \tag{26}$$

在式 (25) 和 (26) 的基础上, 新的聚类原型迭代式可以写为

$$\bar{v}_i = \frac{\sum\limits_{j=1}^{n} w_j u_{ij}^m \boldsymbol{x}_j}{\sum\limits_{j=1}^{n} w_j u_{ij}^m} \tag{27}$$

特别要说明的是, 式 (26) 或 (27) 中没有加入自适应指数 p. 但并不说明聚类的原型和 p 无关, 相反, 聚类原型的计算和自适应度 w_j 也是密切有关的, 只是二者关系是隐性的. 原因解释如下: 变换式 (24), 得到式 (28), 显然其右侧是个和 p 无关的量, 即此时 w_j^p 是作为一个整体出现的. 如果在式 (26) 或 (27) 中也用了 w_j^p, 而不是 w_j, 则由于 w_j^p 是作为一个整体参与到目标函数最优值计算的, 反而起不到 p 的调节作用.

$$w_j^p = \frac{\left[\prod\limits_{j=1}^{n} \left(\sum\limits_{i=1}^{c} u_{ij}^m d_{ij}^2 \right) \right]^{\frac{1}{n}}}{\sum\limits_{i=1}^{c} u_{ij}^m d_{ij}^2} \tag{28}$$

w_j^p 虽然是一个整体, 选择不同的 p, 则 w_j 是不相同的, 从而间接地影响了聚类原型的计算结果, 见式 (27), 这种影响是通过迭代实现的, 所以称为隐性影响.

2.2　基于 AFCM 模型的离群点挖掘

本文的第 2 节指出, 在离群点挖掘研究领域中, 定义什么样的点为离群点是一个首先要解决的问题. 当 AFCM 用于挖掘离群点时, 数据 x_j ($j = 1, \cdots, n$) 到全体聚类原型之间的模糊距离的定量测度被用来作为离群点的判定依据.

将 $\sum_{i=1}^{c} u_{ij}^m d_{ij}^2$ 称为数据 x_j 的模糊平方距离和, 并用 $FSDS(\boldsymbol{x}_j)$ 表示. $FSDS(x_j)$ 表达了 x_j 和全部聚类原型之间的模糊距离关系. $FSDS(x_j)$ 值越大, 则说明 x_j 是离群点的可能性越大. 当 J_{AFCM} 收敛到最优值时, 每个数据点均有一个 $FSDS$ 值. 将全部 n 个 $FSDS$ 值按照从大到小的次序排列, 并根据 "Top-k" 原则, 最大的 k 个 $FSDS$ 值所对应的数据判定为离群点.

另外, 变换式 (28), 并用 $FSDS(\boldsymbol{x}_j)$ 代替 $\sum_{i=1}^{c} u_{ij}^m d_{ij}^2$, 得到下式

$$w_j^p \cdot FSDS(\boldsymbol{x}_j) = \prod_{j=1}^{n} \left[(FSDS(\boldsymbol{x}_j))^{\frac{1}{n}} \right] \tag{29}$$

在 J_{AFCM} 收敛到其最优值后, 式 (29) 的右侧是一个确定的值, 不必关心这个值究竟是多少, 所需关心的是, 对于任何数据, 式 (29) 的左侧都是相同的, 也即所有数据的 w_j^p 和 $FSDS(\boldsymbol{x}_j)$ 的乘积相同. 显然, 最大的 k 个 $FSDS$ 值, 对应的是最小的 k 个

w_j^p. 因此可以根据 w_j^p 的大小来判定离群点, 这样的判定更直观. 定义:

$$O_{\mathrm{AFCM}}(\boldsymbol{x}_j)=\frac{FSDS(\boldsymbol{x}_j)}{\max\limits_{j=1,\cdots,n}[FSDS(\boldsymbol{x}_j)]}=\frac{\min\left[w_j^p\right]}{w_j^p} \tag{30}$$

$O_{\mathrm{AFCM}}(\boldsymbol{x}_j)\in(0,1]$ 称为 x_j 的全局离群度. $O_{\mathrm{AFCM}}(\boldsymbol{x}_j)$ 越大, 则 x_j 是离群点的可能性越大. 最大的 k 个 $O_{\mathrm{AFCM}}(\boldsymbol{x}_j)$ 值所对应的 k 个数据可以判定为离群点. 之所以称为全局离群点, 是因为 $O_{\mathrm{AFCM}}(\boldsymbol{x}_j)$ 表达的是 $\boldsymbol{x}_j$ 和全部聚类原型, 即整个数据集的关系. 这和基于密度的局部离群点 LOF 是明显不同的.

在判定了哪些点是离群点之后, 由于 AFCM 首先是一种聚类方法, 离群点具有 "类" 的信息, 这种信息往往是受到关注的, 从而实现了 "挖掘" 的目的.

2.3 AFCM 模型的计算流程

如上所述, AFCM 模型引入了数据之间的内在关联性, 可以同时实现高质量的聚类和离群数据挖掘. 图 1 给出了 AFCM 计算的流程图. 图中有两个框架, 上框为聚类, 下框为离群点挖掘.

3 数据实验

本节通过一系列数据实验评估 AFCM 在聚类和离群点挖掘方面的性能. 实验分为两组, 第 1 组实验评估 AFCM 在模糊聚类方面的性能, 比较对象是 FCM. 第 2 组实验评估 AFCM 在离群点挖掘方面的性能, 比较对象是 LOF. AFCM 主要在处理含噪, 且大规模数据集时具有明显的优势. 本文实验选择了两个数据集, 一个数据的规模较小, 且数据集被认为是 "干净的", 即里面没有明显的异常数据存在; 另一个数据集的 n 很大, 且已知明显存在离群数据. 另外, 在处理大规模数据集或者动态数据集时, 算法的时间复杂性也不容忽视. 在上述实验中, 给出了 AFCM 分别在聚类和离群点挖掘时各自的参照方法的时间复杂性分析.

按照实验要求, 两个真实的数据集, 一个是知名的公共数据集 IRIS, 另一个来源于作者参加的一个工程实际问题, 称为 FL6C. 下文将对这两个数据做详细介绍. 第 1 组实验中, 采用了两个数据集为分析对象, 第 2 组实验仅采用了 FL6C 为分析对象. 这是因为第 2 组实验主要分析离群点的挖掘能力, IRIS 通常被认为是一个 "干净" 的数据集, 而 FL6C 数据集含有明显的异常数据.

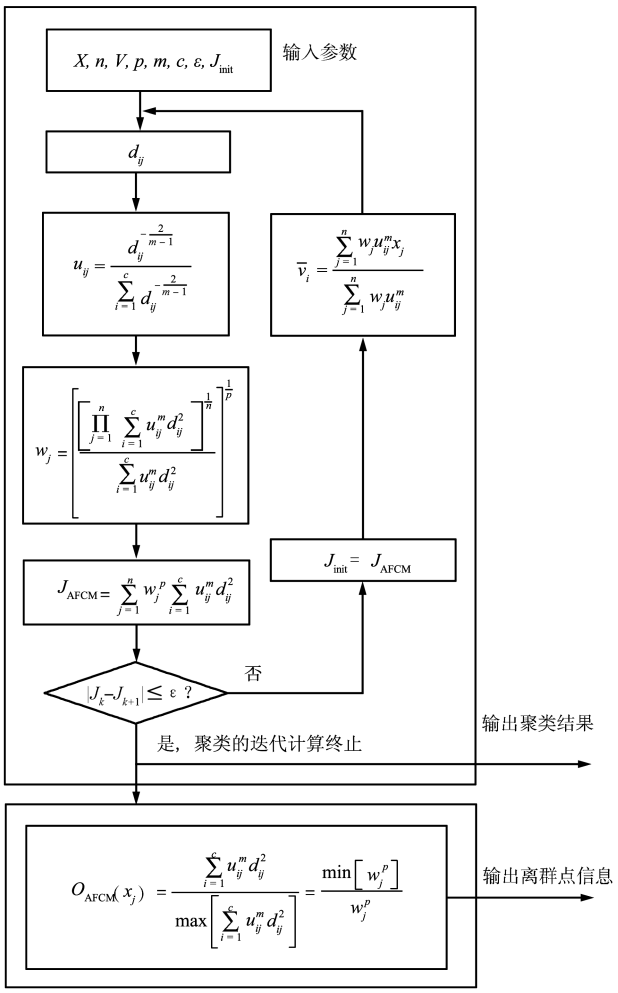

图 1 AFCM 算法的计算流程图

Fig. 1 Clustering and outlier mining flowchart of the AFCM

3.1 两个真实数据集

IRIS 是一个公共数据集, 广泛应用于新聚类方法的评估. IRIS 数据集包括三组数据, 共 150 个数据对象, 数据的特征数是 4.

FL6C 来源于一个实际生产过程. 作者参与的一个项目和极薄带钢的轧制有关, 智能控制方法应用于其中. 其中的一个任务是实时识别带钢平坦度的模式, 同时发掘异常的板形数据并作出响应. 在一个实际生产线中, 以 1 秒为周期采集了 2 646 组带钢平坦度的高维数据, 然后通过六阶勒让德多项式将该高维数据简化, 六个勒让德多项式系数作为带钢平坦度数据的特征, 构建了一个 $2\,646\times 6$ 的数据集, 称之为 FL6C.

3.2 第 1 组实验: 聚类分析

在新的聚类方法研究中, 对聚类性能的评估通

常包括三个方面的内容: 1) 聚类的有效性问题. 聚类的有效性问题通常在一个具体的算法内部讨论. 即, 对于一个聚类算法, 开发出各种有效性函数, 通过参数的组合来计算有效性指标值, 从中确定最优的 c 值. 2) 聚类的质量分析. 即已知数据集划分为 c 类, 怎样的划分质量更好. 聚类的质量分析主要关注的是聚类所得到的原型. 因为聚类是一个无监督的学习方法, 最理想的划分事先并不知道. 因此, 分析聚类的质量优劣时, 通常是两个算法之间的比较, 而且比较所得出的评估结论也是相对的. 3) 算法的计算效率问题. 一个算法如果计算效率很低, 即便可以得到好的聚类结果, 也不能认为是好算法. 计算效率用算法运行所需要的时间或者迭代次数等来进行评估.

本文所提出的 AFCM 首先是一种新的聚类方法, 但本文不讨论聚类的有效性问题, c 是事先给定的. 聚类的质量分析和计算效率分析是 AFCM 聚类性能分析关注的两个重点, 是第 1 组实验要解决的问题. 在分别得到 FCM 和 AFCM 的聚类的结果之后, 下一步工作是分析哪种方法得到的结果更优以及计算效率的比较.

3.2.1　聚类质量评价函数

在新的聚类方法研究中, 通常用公共数据集, 如著名的 UCI 数据集, 来检验新方法的性能. 对于公共数据集而言, 它的聚类原型集以及数据的类属性通常是事先给出的, 此时聚类的结果可以和这些给定值进行比较. 然而, 也有不同的情况, 例如 FL6C 数据集, 事先并不知道聚类原型矩阵和数据的类属性. 对这些数据进行分析时, 有必要提出评价方法来比较和评估两种不同聚类方法的聚类性能.

本文提出了用一组评价函数 CMP, SPT 和 EVA 来评估聚类的质量. 其中, CMP 称为聚类的紧致度, SPT 称为聚类的分离度, EVA 是二者的比例, 称为有效度.

1) 聚类的紧致度 CMP:

$$CMP = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^{m}\left(\boldsymbol{x}_j - \bar{\boldsymbol{v}}_i\right)^2 \tag{31}$$

数据集划分为 c 类之后, CMP 反映的是类内数据紧致程度, CMP 值越小, 表明聚类紧致性越好.

2) 聚类的分离度 SPT:

$$SPT = \min_{i\neq k}\left\|\bar{\boldsymbol{v}}_i - \bar{\boldsymbol{v}}_k\right\|^2 \tag{32}$$

SPT 反映的是类间数据的分离程度, SPT 值越大, 表明聚类分离性越好.

3) 聚类的有效度 EVA:

$$EVA = \frac{CMP}{SPT} \tag{33}$$

显然, CMP 或 SPT 仅给出了聚类划分的部分信息, 二者的比率 EVA 更能全面反映聚类质量. EVA 越小, 则聚类的总体质量越好. 本文中, 给出了这三个函数值, 在三个指标的判定不一致时, EVA 作为最终的优劣判定依据.

实际上 EVA 函数和 X-B 有效性函数[17] 的定义是一致的. 虽然 X-B 指标也可以用于评价不同算法下聚类的质量. 但是在大多数应用中, X-B 有效性指数主要用于聚类的有效性分析, 即针对某个具体的聚类算法, 计算一系列 X-B 有效性指标值, 由此确定最佳的 c. 但正如前文所述, 本文不讨论聚类有效性问题, 此外, 在评估聚类性能时, 紧凑性和分离性有时也是受到关注的. 基于上述原因. 本文采用三种指标来评估聚类质量.

3.2.2　聚类结果分析

1) 聚类结果

分别用 FCM 和 AFCM 对 IRIS 和 FL6C 进行聚类划分, 所得到的 CMP, SPT 及 EVA 结果分别见表 1 和表 2. 两个数据集的聚类数据分别为 $c = 2-4$ 和 $c = 8-16$, 模糊指数 m 均为 2, 两种情况下目标函数停止迭代的精度 ε 均为 0.0001, 而且两种方法下迭代的初值相同, 均由刚性聚类算法产生. 另外 AFCM 算法中所需要的自适应指数 p 也在表 1 和表 2 中分别列出, p 如此取值的原因详见第 5 节分析.

表 1　FCM 和 AFCM 下 IRIS 聚类质量指标对比

Table 1　Comparison of evaluating values under FCM and AFCM, for IRIS

	FCM			AFCM			
c	CMP	SPT	EVA	CMP	SPT	EVA	p
2	2.0526	2.1914	0.9366	2.0762	3.054	0.6798	1
3	1.3140	2.6592	0.4941	1.3141	2.6852	0.4894	−10
4	0.9894	2.7224	0.3634	0.9966	3.5352	0.2819	−1

2) 聚类质量分析

对于两个数据集, 不论 c 的取值是多少, 均有以下的结论 (见表 1 和表 2):

AFCM 下的 CMP 值大于 FCM 下的 CMP 值, 说明 FCM 得到的聚类的类内紧致性要好; AFCM 下的 SPT 要大于 FCM 下的 SPT 值, 说明 AFCM 得到的聚类类间分离性更好; AFCM 下的 EVA 均小于 FCM 下的 EVA 值, 说明 AFCM 得到的聚类的综合质量更好.

总之, 在 AFCM 模型中, 引入的自适应度和自适应算法对聚类结果的紧致性影响不大, 但对分离性影响很大, 综合聚类质量可判定, AFCM 要优于 FCM.

表 2 FCM 和 AFCM 下 FL6C 聚类质量指标对比

Table 2 Comparison of evaluation indices under FCM and AFCM, for FL6C

	FCM			AFCM			
c	*CMP*	*SPT*	*EVA*	*CMP*	*SPT*	*EVA*	p
8	0.8454	0.0925	9.1412	0.8484	0.2952	2.8740	−1
9	0.7956	1.2158	0.6544	0.7959	1.2731	0.6252	−2
10	0.6779	0.2556	2.6516	0.6793	0.5443	1.2481	−2
11	**0.6680**	**1.3361**	**0.4999**	**0.6686**	**1.4372**	**0.4652**	**1**
12	0.6064	0.0400	15.1795	0.6082	0.2420	2.5135	−1
13	0.5515	0.2355	2.3421	0.5531	0.3550	1.5578	−1
14	0.5385	0.0373	14.4223	0.5401	0.2937	1.8391	−1
15	0.4976	0.9587	0.5190	0.4977	0.9834	0.5061	−2
16	0.4706	0.0946	4.9737	0.4732	0.3430	1.3796	−1

3.2.3 AFCM 和 FCM 用于聚类时的时间复杂性比较

时间复杂性通常用两个指标来评估: 迭代次数和总耗费时间. 本部分的时间复杂性比较采用了这两个指标. IRSI 的计算结果见表 3, FL6C 的计算结果见表 4. 实验用的计算机的主要参数为: CPU 主频 2.4 GHz, 内存 1 GB. 在实验中发现, 重复运算算法, 总迭代次数是固定的, 算法运行 1 次的总耗费时间采用了 5 次相同实验耗费时间的均值.

表 3 FCM 和 AFCM 分析 IRIS 时的时间复杂性比较

Table 3 Time complexities of FCM and AFCM, for IRIS

	FCM		AFCM		
c	迭代次数	平均用时 (s)	迭代次数	平均用时 (s)	p
2	6	0.0022	11	0.0037	1
3	9	0.0068	14	0.0083	−10
4	11	0.0066	14	0.0061	−1

对 AFCM 和 FCM 两种算法的时间复杂性已进行了测试和比较. 需要指出的是, 比较分析的本意不是为了精确地比较究竟哪个模型运行得更快, 而是通过实验可知, AFCM 算法并没有带来严重的计算复杂问题. 分析表 3 和表 4 的计算结果, 从迭代次数来判定, 除了 FL6C, $c = 8$ 和 12 时, AFCM 的迭代次数要少于 FCM 的迭代次数外, 其他情况下, 均表明 AFCM 需要更多的迭代次数, 这说明 AFCM 的聚类质量提高是有代价的, AFCM 的时间复杂性增高. 但表 3 和表 4 也表明, AFCM 和 FCM 在时间复杂性方面可以认为是同一水平的, 甚至在个别场合下, AFCM 的时间复杂性要低. 时间复杂性变化不大的原因是: 由于 $(\boldsymbol{W}, p)$ 的引入, 一方面, 确实每一个迭代步的计算时间延长, 但另一方面, 也正是因为 $(\boldsymbol{W}, p)$ 的引入, 使得算法的收敛速度发生变化. 综合以上两方面原因知, AFCM 可以保持和 FCM 同一水平的时间复杂性, 甚至效率更高, 这才是此处实验所要表达的思想.

表 4 FCM 和 AFCM 分析 FL6C 时的时间复杂性比较

Table 4 Time complexities of FCM and AFCM, for FL6C

	FCM		AFCM		
c	迭代次数	平均用时 (s)	迭代次数	平均用时 (s)	p
8	21	0.1324	19	0.1273	−1
9	17	0.0803	21	0.1641	−2
10	18	0.0933	27	0.2172	−2
11	16	0.0983	28	0.2144	1
12	22	0.1306	19	0.1528	−1
13	20	0.1316	32	0.2772	−1
14	30	0.2094	31	0.2826	−1
15	10	0.0799	14	0.1621	−2
16	19	0.1544	27	0.2794	−1

3.3 第 2 组实验: 离群点挖掘分析

当前, 密度法是离群点挖掘应用的主要方法之一, 典型代表是 LOF 模型. 对 LOF 的详细介绍, 可参考文献 [2, 13]. 本组实验将 LOF 作为 AFCM 在离群数据挖掘方面的比较参照物, 比较了二者用于离群点挖掘的特性以及此时的计算效率.

LOF 使用了一个数据的局部离群点的概念来表达一个数据是离群数据的量化程度. 如果数据 x_j 不是一个离群点, 则 LOF (x_j) 的值趋近于 1. LOF (x_j) 值越大, 则这个数据是离群点的可能性越大. 此时 "Top-k" 的原则仍然适用, k 个最大的 LOF 值对应的 k 个数据判定为离群点.

3.3.1 离群点挖掘结果分析

本组实验以 FL6C 为研究对象, 评估 AFCM 用于离群点挖掘的特性. FL6C 的特点是, 数据集中明显含有离群点, 且数据个数多. 在聚类之前, 给出算法所需要的基本参数的取值为: $c = 10$, $m = 2$, $p = -2$. 并且假定数据集含有 10 % 的离群点, 即含有 265 个离群点. 将离群点的挖掘结果绘制于图 2. 图 2 包含以下信息: 首先, 任何一个数据, 不论是否为离群点, 都被划分属于某个类; 其次, 图 2 中有一条实线, 称为离群点判定线, 线以上的点判定为离群点, 也即 265 个数据的 O_{AFCM} 值在这条线之上, 表明判定线的纵坐标是第 116 个大的 O_{AFCM} 值.

图 3 是采用 LOF 模型分析 FL6C 数据集中离群点的结果. 图中的虚线具有和图 2 中类似的定义, 也即虚线的纵坐标是第 116 大的 LOF 值. 由于 LOF 值不是归一化的值, 为便于分析比较, 将所有的 LOF 值除以最大的 LOF 值, 得到 LOF 的相对值, 仍称为相对 LOF 值, 用 RelLOF 表示, 见图 3.

显然这种处理不影响 LOF 对离群点的判定.

比较图 2 和图 3 可知, 在离群点判定上, AFCM 和 LOF 的判定结果大体上是相似的, 但也存在一些离群点判定不一致的情况, 这是由于二者对离群点的不同定义造成的. LOF 是基于密度的局部判定, 而 AFCM 是基于聚类的全局判定. 实际上, 由于二者对离群点的定义本质上的不同, 严格意义讲不具有可比性, 究竟是 LOF 还是 AFCM 更适合挖掘离群点, 还需视实际运用而定. 但是, 当对离群点的信息要深入挖掘, 例如要揭示离群点和数据集之间的全局关系时, AFCM 提供了一种切实可行的手段. 从图 2 和图 3 可以看出, AFCM 发掘的离群点显然比 LOF 发掘的离群点包含的信息更多, 即 AFCM 下的离群点含有类的信息, 信息是全局的, 这是 LOF 不具有的特点.

3.3.2 AFCM 和 LOF 用于离群点挖掘时的时间复杂性比较

本组实验的另一部分内容是检验 AFCM 和 LOF 用于离群点挖掘时的时间复杂性. 即第 3.3.1 节的数据实验记录算法所耗费的时间, 实验重复了 5 次. AFCM 在 $c = 10$, $m = 2$, $p = -2$ 时, 平均耗时 0.0933 秒, 而 LOF 耗时高达 141.66 秒. 计算结果表明和 LOF 方法相比, AFCM 在计算效率上具有无可比拟的优势, 特别是在挖掘大数据集或者动态集中的离群点时, 这个计算高效的优点很有意义.

4 自适应度和自适应指数的讨论

4.1 自适应度

自适应模糊聚类过程中, 每个数据均引入了一个自适应度. 在给出满足约束条件的自适应度初值之后, 在目标函数的交互迭代过程中, 自适应度是自动更新的, 也称为是自适应的. AFCM 方法的巧妙之处在于, 提出了一种描述数据集中所有数据之间的关联关系, 即将所有自适应度的乘积约定为 1, 见式 (15).

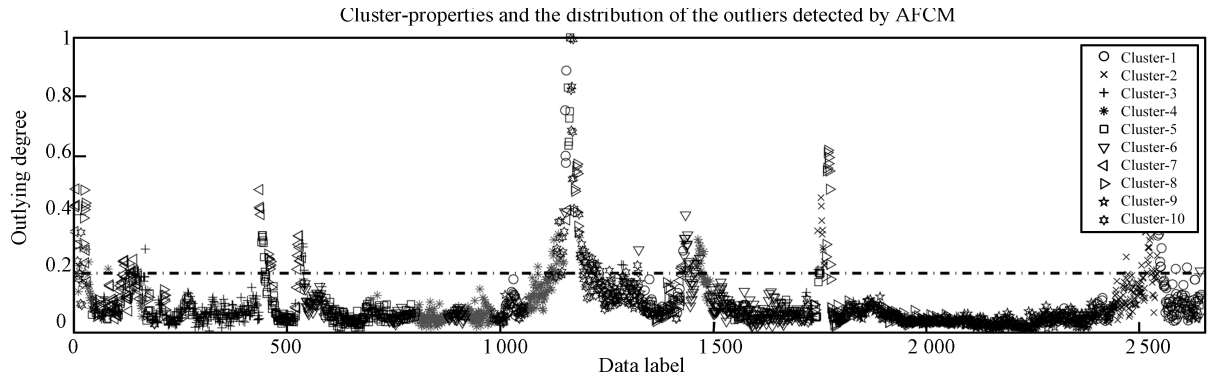

图 2　AFCM 模型下, FL6C 中 265 个离群点的分布图以及离群点的类属性图

Fig. 2　265 outliers distribution and the "cluster-belongings" of outliers under AFCM, for FL6C

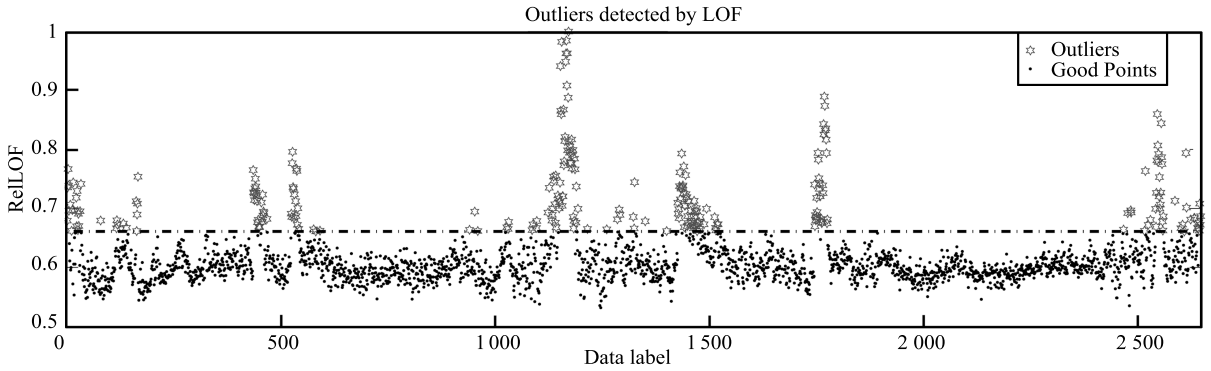

图 3　LOF 模型下, FL6C 中 265 个离群点的分布图

Fig. 3　265 outliers distribution and the "cluster-belongings" of outliers under LOF, for FL6C

在现有众多聚类策略中, “和为 1” 的约束形式更为常见. 例如, 在 FCM, FPCM 以及 PFCM 模糊聚类方法中, 模糊隶属度的约束条件均采用了 “和为 1”. 下文解释采用 “乘为 1” 的约束的原因.

在构建聚类优化问题的约束条件时, 根据约束等式中因子的数目, 通常有三种: 聚类原型的个数 c, 数据的维数 q, 以及数据点的个数 n. 与 n 相比, c 或 q 通常非常小. 因此当采用 “和” 的约束形式时, 如果约束等式中的因子个数为 c 或者 q 时, 则因子的均值还不至于过小. 但是, 当约束等式中约束因子的个数是 n 时, 特别是对于大型数据集, 此时约束因子的均值将会变得非常小, 这会带来意想不到的问题, 例如迭代计算变得发散, 导致不能运算. 有一个有趣的例子, Pal 等于 1997 年提出了 FPCM. FPCM 没有对噪声敏感和聚类原型趋同的缺点, 但在处理大规模数据集时, 由于其典型度采用了 “和为 1” 的约束条件, 而典型度的个数为 n, 这使得典型度的均值很小, FPCM 算法甚至不能进行下去. 为了解决这个问题, Pal 等于 2005 年又提出了 PFCM 模型, 其主要解决的是 n 过大时的问题. 本文提出的自适应聚类方法, 引入的自适应度的数目也为 n, 为避免类似的问题, 采用了 “积为 1” 的约束形式, 并取得令人满意的结果. 以下给出一个自适应度 w_j 取值的实例.

用 AFCM 对 FL6C 数据集进行聚类分析, 计算条件: $c = 10$, $m = 2$, $p = -2$, 得到 n 个自适应度值, 绘制于图 4, 自适应度的统计信息如下: 最大为 3.6064, 最小为 0.2831, 均值为 1.0666, 显然没有自适应度过小的缺点.

4.2 自适应指数

自适应指数 p 是一个很重要的参数, 通过调控自适应度的大小来调控聚类过程. p 的不同取值影响聚类算法的两个方面: 1) 聚类的质量, 即用 EVA 值来衡量的聚类的质量高低; 2) 算法的收敛速度. 二者并不总是一致的, 例如, 对聚类质量要求高时, 可以选择满足这个条件的 p 值, 这时, 计算效率可能会下降. 而在 AFCM 用于离群点挖掘时, 可以采用其他的 p 值, 使得算法的计算效率提高.

目前, 在回答如何选择 p 的问题上尚没有严格的数据推导. 在 AFCM 的实际应用中, 建议选用一系列的 p 来计算, 然后根据具体的应用, 从中选择一个最优的 p. 本文对自适应指数 p 的取值范围做了数据实验, 建议 p 的取值条件为: $1 \le |p| \le 10$. 实验的数据集是 FL6C. 选择最优的 p 的依据是以获得最好的聚类质量为主, 即 EVA 值最小. 即在每一种 c 的情况下, 选择不同的 p, 然后分别计算 EVA 值, 最小的 EVA 值所对应的 p 值为最优. 表 5 列出了 FL6C 在 p 和 c 的不同组合下的 EVA 值.

表 5 中用粗体字标出给定 c 的情况下, 不同的 p 对应的 EVA 值, 此时响应的 EVA 最小, 这解释了在第 4.2.2 节中对 FL6C 聚类时 p 选值的原因. 第 4.2.2 节中对 IRIS 的取值有类似的分析.

5 结论

本文提出了一种新的自适应的模糊聚类模型, 称为 AFCM. 其主要创新之处在于模糊聚类过程中增加了反映数据之间关联性的约束条件, 引入了一组自适应度向量 $\boldsymbol{W}$ 和一个用以调节 $\boldsymbol{W}$ 的自适应指数 p.

现有的模糊均值聚类主要通过 (U, m) 这一渠道调节聚类过程, AFCM 在此基础上, 又引入了一个新的调节渠道 $(\boldsymbol{W}, p)$, 实现了对数据集中不同数据不同处理的思想, AFCM 模型中提出了 n 个自适应度的 “积为 1” 的约束条件, 解决了数据集的规模过大时, 采用 “和” 的约束形式时常见的约束因子过小的缺点.

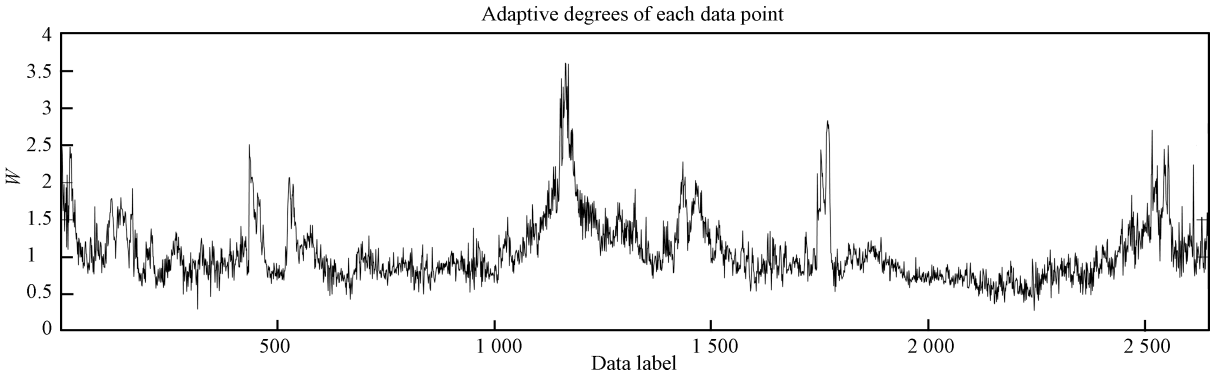

图 4 自适应取值的分布

Fig. 4 Values of adaptive degrees

表 5 对 FL6C, 不同的 c 和 p 组合, AFCM 得到的 EVA 值

Table 5 Values of EVA under the combinations of c and p, for FL6C

	c								
p	8	9	10	11	12	13	14	15	16
−10	7.3394	0.6480	2.4069	0.5029	13.5491	2.2419	13.8844	0.5157	5.2036
−9	7.2027	0.6472	2.3728	0.5032	13.3809	2.2280	13.7576	0.5153	5.2608
−8	7.0388	0.6461	2.3294	0.5036	13.1549	2.2107	13.5594	0.5147	5.3330
−7	6.8399	0.6448	2.2725	0.5041	12.8420	2.1888	13.2441	0.5140	5.4266
−6	6.5901	0.6430	2.1948	0.5047	12.3940	2.1017	12.7339	0.5132	5.5519
−5	6.2718	0.6405	2.0838	0.5056	11.7248	2.0657	11.8906	0.5121	5.7253
−4	5.8516	0.6369	1.9168	0.5069	10.6840	2.0136	10.4930	0.5105	5.9693
−3	5.2783	0.6311	1.6550	0.5090	9.0106	1.9325	8.2618	0.5084	6.2662
−2	4.4654	0.6252	1.2481	0.5125	6.3357	1.8005	5.1680	0.5061	5.9544
−1	2.8740	0.7233	0.4613	0.5180	2.5135	1.5578	1.8391	0.5195	1.3796
1	8.3915	8.3102	8.0972	0.4652	4.8367	4.4514	4.8052	0.5620	3.4118
2	27.1642	16.7474	3.7468	0.4833	8.0244	3.0690	6.5337	0.5493	3.4931
3	21.4565	0.6927	3.3526	0.4892	10.1611	2.7616	8.0666	0.5385	3.7220
4	16.0573	0.6754	3.1947	0.4920	11.5341	2.6302	9.2171	0.5334	3.8957
5	14.0337	0.6721	3.1033	0.4937	12.3973	2.5576	10.0574	0.5304	4.0236
6	12.8707	0.6695	3.0410	0.4948	12.9422	2.5106	10.6743	0.5286	4.1166
7	12.1280	0.6676	2.9958	0.4955	13.3074	2.4779	11.1409	0.5272	4.1901
8	11.6111	0.6661	2.9608	0.4961	13.5585	2.4538	11.4943	0.5263	4.2483
9	11.2355	0.6650	2.9330	0.4965	13.7369	2.5311	11.7750	0.5255	4.2953
10	10.9466	0.6641	2.9105	0.4969	13.8675	2.5147	11.9993	0.5249	4.3318

AFCM 模型中需人工设定的参数少, 除了 p 和 m 外, $\boldsymbol{W}$ 和 U 均是在给出初值后迭代更新的, 模型的应用不需要额外的知识和经验. 其中, p 是一个重要的参数, 可以通过合理选择 p, 获得高质量的聚类结果或者改变算法的收敛速度, 后者在挖掘大规模或者动态含噪数据集中的离群点时很有意义.

AFCM 同时输出三组参数: 模糊隶属度集 U, 自适应度向量 $\boldsymbol{W}$ 和聚类的原型集 V, 分别用于数据的类属性判定、离群点挖掘和分类器的构建. 在当前, 能够同时实现以上三个目的, AFCM 的优点是独有的.

AFCM 下得到的离群点是全局的, 表达的是离群点和整个数据集之间的关系. 此外, 因为 AFCM 是一种聚类方法, 因此 AFCM 得到的离群点含有更丰富的离群点的信息, 例如离群点的类属特性.

最后, 在算法的时间复杂性方面, 当 AFCM 用于聚类时, 其和现有的大多数聚类算法相当; 而当用于离群点挖掘时, 其与目前常用的基于密度的离群数据挖掘相比, 具有无可比拟的优势.

AFCM 研究的后续工作包括自适应指数 p 的选择方法; 另外本文没有讨论模糊指数 m 的作用, 在本文中, m 的取值均设为 2. 模糊指数 m 和自适应指数 p 之间的联合作用, 对聚类过程有哪些重要影响, 目前尚未开展研究, 这是今后工作着重要解决的.

References

1 Cai Zi-Xing, Xu Guang-You. *Artificial Intelligence: Principles and Application (Third Edition)*. Beijing: Tsinghua Press, 2004. 16−21
(蔡自兴, 徐光祐. 人工智能及其应用 (第三版). 北京: 清华大学出版社, 2004. 16−21)

2 Han J, Kamber M [Author], Fan Ming, Meng Xiao-Feng [Translator]. *Data Mining: Concepts and Techniques*. Beijing: China Machine Press, 2007
(Han J, Kamber M [著], 范明, 孟小峰 [译]. 数据挖掘: 概念与技术. 北京: 机械工业出版社, 2007)

3 Bezdek J C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981

4 Krishnapuram R, Keller J M. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1993, **1**(2): 98−110

5 Barni M, Cappellini V, Mecocci A. Comments on "a possibilistic approach to clustering". *IEEE Transactions on Fuzzy Systems*, 1996, **4**(3): 393−396

6 Pal N R, Pal K, Bezdek J C. A mixed C-means clustering model. In: Proceedings of the 6th IEEE International Conference on Fuzzy Systems. Barcelona, Spain: IEEE, 1997. 11−21

7 Pal N R, Pal K, Keller J M, Bezdek J C. A possibilistic fuzzy C-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, 2005, **13**(4): 517−530

8 Kang J Y, Min L Q, Luan Q X, Li X, Liu J Z. Novel modified fuzzy C-means algorithm with applications. *Digital Signal Processing*, 2009, **19**(2): 309−319

9 Chuang K S, Tzeng H L, Chen S, Wu J, Chen T J. Fuzzy C-means clustering with spatial information for image segmentation. *Computerized Medical Imaging and Graphics*, 2006, **30**(1): 9−15

10 Cai W L, Chen S C, Zhang D Q. Fast and robust fuzzy C-means clustering algorithms incorporating local information for image segmentation. *Pattern Recognition*, 2007, **40**(3): 825−838

11 Pal N R, Bezdek J C. On cluster validity for the fuzzy C-means model. *IEEE Transactions on Fuzzy Systems*, 1995, **3**(3): 370−379

12 Chatzis S, Varvarigou T. Factor analysis latent subspace modeling and robust fuzzy clustering using T-distributions. *IEEE Transactions on Fuzzy Systems*, 2009, **17**(3): 505−517

13 Breunig M M, Kriegel H P, Ng R T, Sander J. LOF: identifying density-based local outliers. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, USA: ACM, 2000. 93−104

14 Ghoting A, Parthasarathy S, Otey M E. Fast mining of distance-based outliers in high-dimensional dataset. *Data Mining Knowledge Discovery*, 2008, **16**(3): 349−364

15 Weng X Q, Shen J Y. Detecting outlier samples in multivariate time series dataset. *Knowledge-Based Systems*, 2008, **21**(8): 807−812

16 Dave R N. Characterization and detection of noise in clustering. *Pattern Recognition Letters*, 1991, **12**(11): 657−664

17 Xie X L, Beni G. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991, **13**(8): 841−847

唐成龙 上海交通大学机械及动力工程学院博士研究生, 高级工程师. 主要研究方向为人工智能、数据挖掘以及冷轧板带钢轧制和后处理技术. 本文通信作者.
E-mail: tangchenglong@baosteel.com
(**TANG Cheng-Long** Senior engineer, Ph. D. candidate at the School of Mechanical and Dynamical Engineering, Shanghai Jiao Tong University. His research interest covers artificial intelligence, data mining, and rolling and post-treating technologies of steel strip productions. Corresponding author of this paper.)

王石刚 上海交通大学机械及动力工程学院教授. 主要研究方向为机器视觉和模式识别, 机器人, 复杂机电系统的设计及控制.
E-mail: wangshigang@sjtu.edu.cn
(**WANG Shi-Gang** Professor at the School of Mechanical and Dynamical Engineering, Shanghai Jiao Tong University. His research interest covers machine vision and pattern recognization, robotics, and design and control of complex mech-electrical system.)