# 一种优化的语义条件模糊聚类

李洪波 李仁璞 张志旺 周春姐
(鲁东大学信息与电气工程学院 烟台 264025)

**摘 要** 在条件模糊聚类的基础上,提出利用公理化模糊集的成员隶属度函数量化用户语义、确定外部条件的方法。引入调节因子新概念,以调节基于语义的成员隶属度和基于欧拉距离的模糊隶属度对聚类结果的影响,并最终建立了语义条件聚类和经典模糊聚类的统一框架。给出了语义聚类的评价指标——语义强度期望,以找到距离目标语义最近的聚类。为使条件模糊聚类的聚类准确性更高,对原始数据进行了谱变换,尔后进行语义条件聚类。利用 Iris 数据集,对标准模糊聚类、语义条件聚类和语义条件聚类的谱优化 3 个算法进行了多指标综合实验比较。实验结果表明,语义条件聚类能够发现最贴近用户给出的语义的聚类。

**关键词** 条件聚类,公理化模糊集,成员隶属度函数,调节因子,语义强度期望
**中图法分类号** TP391 **文献标识码** A

## Optimized Semantic Conditioned Fuzzy C-Means

LI Hong-bo LI Ren-pu ZHANG Zhi-wang ZHOU Chun-jie
(School of Information and Electrics Engineer, Ludong University, Yantai 264025, China)

**Abstract** On the basis of Conditioned Fuzzy C-Means, it was proposed that a foreign condition is determined by a computational user semantic. The user semantic was computed by the membership function based Axiomatic Fuzzy Sets. Further, a new concept, Adjusted Factor, was introduced to adjust the impact of the membership based on semantic and that one based on Euclidean Distance on clustering results, and one uniformed coherence framework of Fuzzy C-Means and Conditioned Fuzzy C-Means was built up. In addition, Semantic Strength Expectation was brought forward in order to assess the clustering quality. Furthermore, in order to raise the clustering accuracy, Semantic Conditioned Fuzzy C-Means was processed after the raw data was transformed into spectral data. Finally, based on multiple assessment indexes, FCM, Semantic Conditioned Fuzzy C-Means and its Spectral Optimization were tested on Iris data set. Experiment results show that the cluster that is closest to user semantic is able to be found by Semantic Conditioned Fuzzy C-Means.

**Keywords** Conditional clustering, Axiomatic fuzzy sets, Coherence membership function, Adjusted factor, Semantic strength expectation

## 1 引言

聚类是数据挖掘、数据识别、机器学习中的一项重要数据分析技术,在识别数据内在结构方面具有极其重要的作用。聚类旨在把一个数据集分割成若干类,使得同一类中的对象具有较强的相似性,而非同类之间的对象具有较弱的相似性[1,2]。

在聚类的不同风景图中,工作于模糊集框架[3]内的算法取得了重要而且独特的位置,原因很简单:被看作基本信息粒的模糊集是以人为中心的。为外部提示所引导的聚类能够更精确、更快速地挖掘出符合用户需求的信息粒。这种外部提示是将聚类放在一定的上下文中进行,称之为条件模糊聚类[4,5]。

条件模糊聚类的外部提示,是将特定的领域知识及上下文信息以"约束"的形式表达,并嵌入到聚类过程中[3]。因为利用了领域知识及上下文信息,使得聚类算法获得了更多的启发式信息,从而减少了其搜索过程中的"盲目性",降低了对噪声的敏感度,提高了效率和聚类质量,也更符合用户需求。

条件聚类的关键问题是如何接受用户语义信息,进行具有实际意义的启发式聚类,使得聚类的准确性和有效性指标得到更好的提高,即如何形成上下文的问题。我们的思路是运用 AFS 模糊逻辑[6],将原始数据的客观性和人类的主观性相统一,指导上下文信息的形成。实际就是将 AFS 模糊逻辑和条件聚类有机结合,利用基于 AFS 的模糊隶属函数计算条件隶属度,并给出语义聚类的评价指标,以解释聚类结果。

## 2 条件模糊聚类

### 2.1 条件模糊集和目标函数

领域知识及上下文信息是一个非常大的概念,目前不存在一种统一的表达形式。除了"类标号"表达指导形式外,还

到稿日期:2012-12-09 返修日期:2013-03-28 本文受山东省自然科学基金面上项目(ZR2010GM013)资助。
李洪波(1969—),男,硕士,副教授,主要研究方向为商务智能应用、云服务工程开发,E-mail:fast_run_man@126.com。

有关于问题的结构方面的信息、启发性指导规则、数据对象间的相互关系，或是上述表达形式的组合。

全局级条件(global conditions)是参与聚类的数据对象都起作用。例如，只对消费金额超过 1000 元的客户进行聚类。条件模糊聚类基于全局级条件，通过聚焦于一部分原始数据来建立类，使聚类模块化；同时，采用经典的模糊聚类算法 FCM 思想[3]，把具有 $N$ 个元素的数据集 $X$ 分为 $C$ 个模糊组，用隶属度确定每个数据点属于某个聚类的程度，并求每组的聚类中心，使得非相似性指标的价值函数值达到最小。将模糊聚类放在这种全局条件的上下文环境中进行，称之为基于上下文的聚类，简称条件模糊聚类(Conditioned Fuzzy C-Means，以下简称 CFCM)[5]。

聚类问题的形式化表示必须合并进上下文的约束，以减少聚类的盲目性。上下文可以看作定义在上下文空间上的模糊集或模糊关系。给定数据集 $X$，赋予上下文 $F$，意味着对于每一个 $x_k$，$k=1,2,\cdots,N$，给出一个与它相关的上下文值(隶属度)，用 $f_k$ 表示它的隶属度。换言之，我们形成了用于后面聚类过程的对象对$(x_k,f_k)$。考虑用于描述顾客的例子，我们将每次交易记录下来，从而给出以下的形式数据：

$$\{(x_1,z_1),(x_2,z_2),\cdots,(x_N,z_N)\} \tag{1}$$

通过顾客的消费额，给出一个上下文模糊集："高消费"，用 $F$ 表示，如图 1 所示。

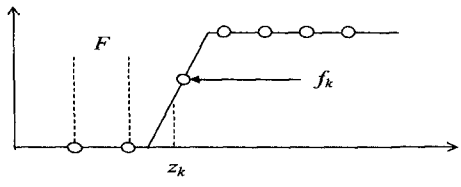


图 1　通过所给的上下文模糊集(高消费)变换 $z_k$，构造带有约束 $f_k$ 的上下文模糊集

$$\{(x_1,f_1),(x_2,f_2),\cdots,(x_N,f_N\} \tag{2}$$

这种形式的数据集成为聚类过程的起始点。在对聚类问题进行公式化描述时，只考虑上下文值。目标函数 $Q$ 的标准形式为：

$$Q=\sum_{k=1}^{N}\sum_{i=1}^{C}u_{ik}^{m}\parallel x_k-v_i\parallel^2 \tag{3}$$

式中，$v_i$ 是原型；$U$ 为划分矩阵，它们都是需要优化的成分；$\parallel\ \parallel$ 是数据 $x_k$ 和原型 $v_i$ 间的距离(通常采用欧几里德(Euclidean)函数)；$m$ $(m>1)$为模糊化因子，有助于控制类的形状，在接近 0 或 1 的隶属与那些具有中间值的隶属度之间产生一个平衡。作用在隶属度上的限制被修改后合并了 $f_k$，由式(4)的约束条件给出：

$$\sum_{i=1}^{C}u_{ik}=f_k,k=1,2,\cdots,N \tag{4}$$

其它的两个约束条件与标准 FCM 相同，如下所示：

$$u_{ik}\in[0,1],i=1,2,\cdots,C;k=1,2,\cdots,N \tag{5}$$

$$0<\sum_{k=1}^{N}u_{ik}<N,i=1,2,\cdots,C \tag{6}$$

综上所述，基于上下文的模糊聚类最优化问题表达形式如下(Pedrycz，1996；Pedrycz 和 Sosnowski，2000)：

$$Q=\min_{U\in U(F),v_1,v_2,\cdots,v_c}\sum_{k=1}^{N}\sum_{i=1}^{C}u_{ik}^{m}\parallel x_k-v_i\parallel^2 \tag{7}$$

根据拉格朗日乘子法，得到以下表达式：

$$V=\sum_{i=1}^{C}u_{ik}^{m}\parallel x_k-v_i\parallel^2-\lambda(\sum_{i=1}^{C}u_{ik}-f_k),k=1,2,\cdots,N \tag{8}$$

对 $V$ 最小化，得划分矩阵迭代公式：

$$u_{ik}=\frac{f_k}{\sum_{j=1}^{C}\left(\frac{\parallel x_k-v_i\parallel}{\parallel x_k-v_j\parallel}\right)^{2/m-1}},k=1,2,\cdots,N \tag{9}$$

原型迭代公式：

$$v_i=\frac{\sum_{k=1}^{N}u_{ik}^{m}x_k}{\sum_{k=1}^{N}u_{ik}^{m}},i=1,2,\cdots,C \tag{10}$$

### 2.2　CFCM 算法

**算法 1**　CFCM 算法

输入：数据集 $X=\{x_1,x_2,\cdots,x_N\}$，数据集的条件隶属度 $F=\{f_1,f_2,\cdots,f_N\}$，聚类数 C，模糊化因子 m，阈值 ε；

输出：模糊划分矩阵 U(F)，原型 $V=\{v_i|i=1,2,\cdots,C\}$。

过程：

(1)选取欧拉距离函数，(随机地)初始化模糊划分矩阵 U(F)，并满足条件式(4)、式(5)和式(6)；

(2)根据式(10)，计算聚类原型 $v_i$，i=1,2,…,C；

(3)根据式(9)，计算新的模糊划分矩阵 U(F)，并比较连续两次迭代的划分矩阵，若它们之间的距离 $\parallel Q(iter+1)-Q(iter)\parallel<\varepsilon$，则终止计算；否则转入(2)。

## 3　AFS 模糊逻辑

AFS 理论[6]是将模糊集的思想公理化，研究如何把蕴含在原始数据中的内在规律和模式转化到模糊集及其逻辑运算中的一种新的模糊数学研究方法。它依靠原始数据，利用 AFS 代数和其上的一个逆序对合运算来建立模糊逻辑系统，用拓扑分子格来刻画人类概念间的抽象关系，较好地揭示了靠经验和直觉描述复杂的模糊概念以及确定相应模糊概念的隶属函数的内在机理，使得隶属函数和模糊逻辑系统的建立更具客观性、严密性和统一性。AFS 理论[6]对 1965 年 L. A. Zadeh 提出的模糊集理论[7]及国内外许多学者在隶属函数和模糊逻辑方面所做的大量研究工作进行了较大改进。

AFS 模糊逻辑理论给出了 AFS 代数、AFS 结构和认知域 3 个新的数学对象。AFS 代数是建立 AFS 理论的数学基础，AFS 结构是复杂关系的数学抽象，认知域使模糊概念的 AFS 描述更直观和简便。

### 3.1　AFS 代数

AFS 代数包含 EI、EII，…，$EI^n$、$E^{\#}I$，这些 AFS 代数研究模糊概念的格值表示。

EI 代数 EM(expanding one set M)是由属性集 $M$ 生成的，并且 EM 中的每一个元素都有确切的语义。为研究模糊概念和逻辑运算的内在本质，首先引入符号 $EM^*$ 和 AFS 代数。设 $M$ 为一个属性或概念集合，集合 $EM^*$ 的定义如下：

$$EM^*=\{\sum_{i\in I}(\prod_{m\in A_i}m)\,|\,A_i\subseteq M,i\in I,I\text{ 为任意非空指标集}\} \tag{11}$$

每个 $\sum_{i\in I}(\prod_{m\in A_i}m)$ 是集合 $EM^*$ 的一个元素，其中对于属性集 $A_i\subseteq M$ 来说，符号 $\prod_{m\in A_i}m$ 表示属性集 $A_i$ 中各个元素的"交"，所表达的语义为"并"；符号 $\sum_{i\in I}$ 表示概念的"并"，所表达的语义为"或者"。

文献[6]在 $EM^*$ 上定义了等价关系。

**定义 1** $M$ 是一个非空集合。$EM^*$ 上的二元关系定义如下：

$$\forall \sum_{i\in I}(\prod_{m\in A_i} m), \sum_{j\in J}(\prod_{m\in B_j} m)\in EM^*, [\sum_{i\in I}(\prod_{m\in A_i} m)]R[\sum_{j\in J}(\prod_{m\in B_j} m)]$$

$$\Leftrightarrow$$

$$(\mathrm{i})\ \forall A_i(i\in I), \exists B_h(h\in J) \text{使得 } A_i\supseteq B_h; \tag{12}$$

$$(\mathrm{ii})\ \forall B_j(j\in J), \exists A_k(k\in I) \text{使得 } B_j\supseteq A_k。 \tag{13}$$

显然，$R$ 是一个等价关系。不失一般性，下面用 EM 表示商集 $EM^*/R$，此时，$\sum_{i\in I}(\prod_{m\in A_i} m)=\sum_{j\in J}(\prod_{m\in B_j} m)$ 意味着 $\sum_{i\in I}(\prod_{m\in A_i} m)$ 和 $\sum_{j\in J}(\prod_{m\in B_j} m)$ 在等价关系 $R$ 下等价，即它们表示的语义等价。

文献[6]证明了如下定义在 EM 上的运算，$(EM,\vee,\wedge)$ 是完全分配格，满足 CD1 或 CD2，即分子格。

$$[CD1]\bigwedge_{i\in I}(\bigvee_{j\in J_i} a_{ij})=\bigvee_{f\in\prod_{i\in I}J_i}(\bigwedge_{i\in I} a_{if(i)}) \tag{14}$$

$$[CD2]\bigvee_{i\in I}(\bigwedge_{j\in J_i} a_{ij})=\bigwedge_{f\in\prod_{i\in I}J_i}(\bigvee_{i\in I} a_{if(i)}) \tag{15}$$

**定理 1**[9] $M$ 是一个非空集合，则 $(EM,\vee,\wedge)$ 在如下定义的二元运算 $\vee$、$\wedge$ 下形成一个完全分配格：

对任意的 $\forall \sum_{i\in I}(\prod_{m\in A_i} m), \sum_{j\in J}(\prod_{m\in B_j} m)\in EM^*$，

$$[\sum_{i\in I}(\prod_{m\in A_i} m)]\vee[\sum_{j\in J}(\prod_{m\in B_j} m)]=\sum_{k\in I\cup J}(\prod_{m\in C_k} m) \tag{16}$$

$$[\sum_{i\in I}(\prod_{m\in A_i} m)]\wedge[\sum_{j\in J}(\prod_{m\in B_j} m)]=\sum_{i\in I,j\in J}(\prod_{m\in A_i\cup B_j} m) \tag{17}$$

式中，$I\cup J$ 是 $I$ 和 $J$ 的不交并。若 $k\in I$，则 $C_k=A_k$，若 $k\in J$，则 $C_k=B_k$。

$(EM,\vee,\wedge)$ 被称为 $M$ 上的 EI(Expanding one set M)代数。值得注意的是，用少数几个模糊概念和分明概念生成的 EM 可以表示非常多的概念，$\vee$、$\wedge$ 是这些模糊概念的并、交运算，并且在 EM 中的每个元素都有确切的语义。对于 $\alpha=\sum_{i\in I}(\prod_{m\in A_i} m)$，$\beta=\sum_{j\in J}(\prod_{m\in B_j} m)$，$\alpha\leqslant\beta\Leftrightarrow\alpha\vee\beta=\beta\Leftrightarrow\forall A_i(i\in I), \exists B_h(h\in J)$，使得 $A_i\supseteq B_h$。

文献[6]给出了逻辑运算“′”(非的定义)，其定义如式(18)所示。

$$(\sum_{i\in I}(\prod_{m\in A_i} m))'=\bigwedge_{i\in I}(\sum_{m\in A_i} m') \tag{18}$$

如果 $m'$ 代表概念 $m\in M$ 的非，则对于任意模糊概念 $\zeta\in EM$，$\zeta'$ 表示 $\zeta$ 的非。代数系统 $(EM,\vee,\wedge,')$ 称为 AFS 模糊逻辑系统。

### 3.2 AFS 结构

AFS 结构是一个三元组 $(M,\tau,X)$，它用于表示论域 $X$ 和属性集 $M$ 之间复杂关系的数学抽象。

**定义 2**[6] 设 $\zeta$ 是论域 $X$ 上的一个属性或概念，$\zeta$ 与 $X$ 是一个二元关系 $R_\zeta$(即 $R_\zeta\subseteq X\times X$)相对应，其中 $x,y\in X$，$(x,y)\in R_\zeta\Leftrightarrow x$ 以某种程度属于 $\zeta$ 且强于或等于 $y$ 属于 $\zeta$ 的程度。

**定义 3**[6] 设 $R$ 为 $X$ 上的二元关系，如果 $R$ 满足：

1. 如果 $(x,y)\in R$，则 $(x,x)\in R$；

2. 如果 $(x,x)\in R$ 且 $(y,y)\notin R$，则 $(x,y)\in R$；

3. 如果 $(x,y),(y,z)\in R$，则 $(x,z)\in R$；

4. 如果 $(x,x)\in R$ 且 $(y,y)\in R$，则或者 $(x,y)\in R$ 或者 $(y,x)\in R$。

则称 $R$ 为 $X$ 上的弱偏好关系。与弱偏好关系对应的概念称为简单概念，否则称为复杂概念。

AFS 结构是一个三元组 $(M,\tau,X)$，它导出 EM 中模糊概念隶属度和 EM 中模糊概念的 Zadeh 模糊集表示。

**定义 4**[6] 设 $X,M$ 为两个集合，$2^M$ 是 $M$ 的幂集，$\tau:X\times X\to 2^M$。如果 $\tau$ 满足：

$$AX1:\forall(x_1,x_2)\in X\times X,\tau(x_1,x_2)\subseteq\tau(x_1,x_1) \tag{19}$$

$$AX2:\forall(x_1,x_2),(x_2,x_3)\in X\times X,\tau(x_1,x_2)\cap\tau(x_2,x_3)\subseteq\tau(x_1,x_3) \tag{20}$$

则 $(M,\tau,X)$ 被称为一个 AFS 结构，称 $X$ 为论域，称 $M$ 为属性集，称 $\tau$ 为结构。

在实际应用中，如果 $M$ 是 $X$ 上的简单概念集合，$R_m$ 是定义 4 所述的偏好关系，则可如下定义 $\tau$：

$$\tau(x,y)=\{m\mid m\in M,(x,y)\in R_m\}\in 2^M \tag{21}$$

### 3.3 基于 AFS 理论的隶属函数

$(M,\tau,X)$ 是一个 AFS 结构。对于 $A\subseteq M$，$x\in X$，首先定义

$$A^\tau(x)=\{y\mid y\in X,\tau(x,y)\supseteq A\} \tag{22}$$

**定义 5**[6] 设 $(M,\tau,X)$ 是一个 AFS 结构，$S\subseteq 2^X$ 是一个在 $X$ 上的 $\sigma$-代数，$m$ 是在 $S$ 上的测度 $0\leqslant m(A)\leqslant 1$，对任意的 $A\in S$。对于模糊概念 $\eta=\sum_{i\in I}(\prod_{m\in A_i})\in EM$，如果 $\forall x\in X$，$\forall i\in I$，$A_i^\tau(x)\in S$，则 $\eta$ 的隶属函数定义如下：

$$u_\eta(x)=\sup_{i\in I}\{m(A_i^\tau(x))/m(X)\}\in[0,1] \tag{23}$$

在本文中，$m(A)=|A|$($|A|$ 为集合 $A$ 的元素个数)。式(23)所定义的模糊集隶属函数完全由 $M$ 中简单概念的弱偏好关系确定。这种方法特别适合仅有序关系的模糊数据。在这一过程中，只需要知道数据在属性上的序关系，而不需要具体的数值就可以将模糊概念的隶属函数确定出来。

## 4 AFS 语义驱动的条件模糊聚类及其谱优化模糊

首先在算法 2 中给出语义驱动的条件模糊聚类。接着在算法 3 中对原始数据进行谱变换，尔后进行条件聚类，以改进聚类的有效性。算法结构图如图 2 所示。

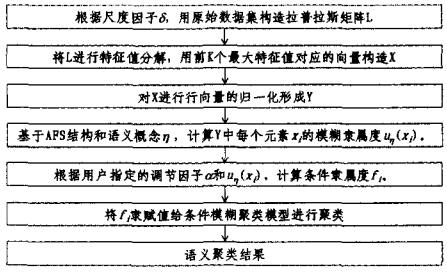


图 2 算法结构图

### 4.1 AFS 语义驱动的条件模糊聚类

设样本集合 $X$ 上的 AFS 结构为 $(M,\tau,X)$，$X=\{x_i\mid 1\leqslant i\leqslant n\}$，语义驱动的条件模糊聚类算法的核心思想是用数据的全部属性集合 $M$，基于用户给出的模糊概念 $\eta$，对于 $\forall x_i(1\leqslant i\leqslant n)$，利用式(23)的 $u_\eta(x_i)$ 作为条件隶属度的计算函数，即

$$f_i=u_\eta(x_i) \tag{24}$$

从而得到 $F=\{f_i|1\leqslant i\leqslant n\}$。

式(24)易产生数量众多的样本因 $u_\eta(x_i)$过低而积聚到一类的现象。即使该样本离其它聚类中心的欧拉距离很近,也会因 $f_i$ 过低而使欧拉距离不起实际作用。为此,做下面的变换,以拉大因 $u_\eta(x_i)$而过小的所有 $f_i$ 值,如式(25)所示。

$$f_i=(u(x_i)+a)/\max_j(u(x_j)+a)(j=1,2,\cdots,n) \quad (25)$$

式(25)中的 $a$ 被称为调节因子,以调节用欧拉距离公式计算出的模糊隶属度在实际聚类中所起作用的大小。$a$ 为 0 时将欧拉距离在聚类过程中所起的作用降到最低,语义聚类结果最强。

**性质 1** SC-FCM 的兼容性

当式(25)中 $a\to\infty$时,$f_i$ 的值为 1,此时 SCFCM 蜕变成标准 FCM。换言之,标准 FCM 是 SCFCM 的一种特殊情况,即调节因子 $a$ 取值无穷大时的情况。此时,语义影响趋近 0。因此,SC-FCM 是兼容标准 FCM 的,SC-FCM 比标准 FCM 适应面更广。

综上,语义驱动的条件模糊聚类(Semantic-driven Conditioned Fuzzy C-Means,下文简称 SCFCM)如算法 2 所示。

**算法 2** SCFCM 算法

输入:数据集 $X=\{x_1,x_2,\cdots,x_n\}$,模糊概念 $\eta$,调节因子 a,聚类数 C,模糊化因子 m,阈值 $\varepsilon$;

输出:模糊划分矩阵 U(F),原型 $V=\{v_i|i=1,2,\cdots,C\}$。

过程:

(1)根据式(25)计算 F;

(2)选取欧拉距离函数,随机地初始化模糊划分矩阵 U(F),并满足条件式(4)、式(5)和式(6);

(3)根据式(10),计算聚类原型 $v_i,i=1,2,\cdots,C$;

(4)根据式(9),计算新的模糊划分矩阵 U(F),并比较连续两次迭代的划分矩阵,若它们之间的距离 $\|Q(iter+1)-Q(iter)\|<\varepsilon$,则终止计算;否则转入(2)。

### 4.2 SCFCM 的谱优化 SO-SCFCM

谱聚类建立在谱图理论基础上。与传统的聚类算法相比,它具有能在任意形状的样本空间上聚类且收敛于全局最优解的优点[8-10]。该算法首先根据给定的样本数据集定义一个描述数据对间相似度的相似矩阵,并且计算矩阵的特征值和特征向量,再选择合适特征向量聚类不同的数据点。谱聚类算法的本质是将聚类问题转化为图的最优划分问题,是一种点对聚类算法。SCFCM 的谱优化(a Spectral Optimization of Semantic driven Conditioned Fuzzy C-Means,以下简称 SO-SCFCM)的处理思想是将谱聚类思想应用于前期对样本的处理,使聚类算法能对任意形状的样本进行优化聚类。

谱聚类算法的一般流程如下[8]:

(1)构造相似性矩阵 $S\in R^{n\times n}$,矩阵中元素 $S_{ij}=\exp(-\|s_i-s_j\|/2\sigma^2)$,且当 $i=j$ 时,$S_{ij}=0$;

(2)构造矩阵 $W$ 为度矩阵,度矩阵主对角线上的元素 $W(i,i)$为相似矩阵 S 的第 $i$ 行元素之和,其他元素均为 0,然后构造拉普拉斯矩阵 $L=W^{-1/2}SW^{-1/2}$;

(3)对拉普拉斯矩阵 $L$ 进行特征值分解,找出前 $k$ 个最大特征值所对应的特征向量 $x_1,x_2,\cdots,x_k$,然后构造矩阵 $X=[x_1,x_2,\cdots,x_k]\in R^{n\times k}$,其中特征向量按列存储;

(4)对 $X$ 的行向量再进行归一化,记归一化后的矩阵为 $Y,Y_{ij}=X_{ij}/(\sum_j X_{ij}^2)^{1/2}$;

(5)将 $Y$ 的每一行看作为 $R^k$ 空间中的样本(样本数量为 $n$,样本维数为 $k$),再选用已有聚类算法聚类归一化矩阵 $Y$;

(6)最后,把最初的样本点 $S_i$ 划分为第 $j$ 聚类当且仅当矩阵 $Y$ 的第 $i$ 行被划分为第 $j$ 聚类。

**算法 3** SO-SCFCM 算法

输入:$X=\{x_1,x_2,\cdots,x_N\}$,模糊概念 $\eta$,调节因子 a,聚类数 C,模糊因子 m,阈值 $\varepsilon$,尺度参数 $\sigma$;

输出:模糊划分矩阵 U、原型集 V。

过程:

(1)计算数据对间的欧式距离矩阵 D,

$$D_{ij}=\|x_i-x_j\|=\sqrt{\sum_p(x_{ip}-x_{jp})^2} \quad (26)$$

(2)构造矩阵 S(表示数据对间的相似性),$S_{ij}=\exp(-\|x_i-x_j\|/2\sigma^2)$,且当 i=j 时,$S_{ij}=0$;

(3)构造矩阵 W 为度矩阵,度矩阵主对角线上的元素 W(i,i)为相似矩阵 S 的第 i 行元素之和,其他元素均为 0,然后构造拉普拉斯矩阵 $L=W^{-1/2}SW^{-1/2}$;

(4)对拉普拉斯矩阵 L 进行特征值分解,找出前 k 个最大特征值所对应的特征向量 $x_1,x_2,\cdots,x_k$,然后构造矩阵 $X=[x_1,x_2,\cdots,x_k]\in R^{n\times k}$,其中特征向量按列存储;

(5)对 X 的行向量再进行归一化,记归一化后的矩阵为 $Y,Y_{ij}=X_{ij}/(\sum_j X_{ij}^2)^{1/2}$;

(6)根据模糊概念 $\eta$ 和调节因子 a 按式(25)计算数据集 X 的条件隶属度集合 F;

(7)将 Y 的每一行看作为 $R^k$ 空间中的样本(样本数量为 n,样本维数为 k),然后将这些样本用 SCFCM 聚类算法根据条件参数集 F、聚类数 C 和收敛阈值 $\varepsilon$ 进行聚类;

(8)最后,把最初的样本点 $x_i$ 划分为第 j 聚类当且仅当矩阵 Y 的第 i 行被划分为第 j 聚类。

## 5 聚类的评价指标及其实验结果比较分析

### 5.1 语义聚类的评价指标

(1)聚类的评价指标

聚类结果的分析与评价常常使用紧致度、分离度和有效度 3 个指标[11]。因此,用评价函数 *SPT*,*CMP* 和 *EVA* 来评价聚类的质量。

**定义 6**(聚类的分离度,*SPT*)

$$SPT=\min_{i\neq k}\|\bar{v}_i-\bar{v}_k\|^2 \quad (27)$$

*SPT* 反映的是类与类之间数据的分离程度,*SPT* 的值越大,说明聚类的分离性越好,即聚类的质量越高。

**定义 7**(聚类的紧致度,*CMP*)

$$CMP=\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{C}u_{ij}^m(x_j-v_i)^2 \quad (28)$$

当数据集被划分为 $C$ 个类之后,可以用 *CMP* 来反映类内数据的紧密程度,*CMP* 的值越小,表明类的内聚性越好,聚类的效果越好。

**定义 8**(聚类的有效性,*EVA*)

$$EVA=\frac{CMP}{SPT} \quad (29)$$

*CMP* 和 *SPT* 分别给出聚类的类内紧致性和类间的分离性,前者越小越好,后者越大越好,这不能全面地反映聚类的

整体质量。而聚类的有效性是聚类的紧致度和分离度的比值，更能全面地反映出算法的聚类结果的质量。显然，聚类有效性的值越小，聚类的总体质量越高。

综上所述，在同时给出这 3 个函数的函数值且 3 个指标的判定结果不一致时，以聚类有效性作为分析聚类结果好坏的主要指标。

**定义 9**(响应时间，*RT*)

响应时间(Response Time)指算法从开始执行到结束花费的时间，以秒为单位。

**定义 10**(语义强度，*SS*) 语义强度(Semantic Strength)指在给定概念下聚类内部各样本的模糊隶属度的平均值，具体如下：

$$SS(C_i)=\frac{\sum_{x_j\in C_i}u_\eta(x_j)}{\|C_i\|} \tag{30}$$

式中，$C_i$ 为第 $i$ 个聚类，元素 $x_j$ 为属于该聚类的样本，$\|C_i\|$ 为聚类 $C_i$ 中数据元素的个数。*SS* 的值越高，表明聚类越接近模糊语义概念 $\eta$，聚类语义越强。反之，*SS* 的值越小，表明聚类离模糊语义概念 $\eta$ 越远，聚类语义越弱。

**定义 11**(语义强度期望，*SSE*) 语义强度期望(Semantic Strength Expectation)指各聚类的准确率及其对应的语义强度之积中的最大值。

$$SSE=\underset{j\in I}{MAX}(E\times SS(C_j)),I=1,2,\cdots,C \tag{31}$$

式(31)的 $E$ 指准确率。*SSE* 的值越高，表明聚类越接近模糊语义概念 $\eta$，聚类的总体语义效果越好。反之，*SSE* 的值越小，表明聚类离模糊语义概念 $\eta$ 越远，聚类的总体语义效果越弱。

### 5.2 Iris 数据集上的实验结果分析

Iris 原始数据集由 150 个样本、4 个属性构成，可用 150×4 的矩阵表示。该矩阵的第一列为萼片长度，第二列为萼片宽度，第三列为花瓣长度，第四列为花瓣宽度。样本的原始数据分为 3 类：$x_1,x_2,\cdots,x_{50}$ 属于 Iris Setosa；$x_{51},x_{52},\cdots,x_{100}$ 属于 Iris Versicolour；$x_{101},x_{102},\cdots,x_{150}$ 属于 Iris Virginica。

下面给出验证 FCM、SCFCM 和 SO-SCFCM 算法所用的输入参数：

(1)$X=\{x_1,x_2,\cdots,x_{150}\}$。

(2)$M=\{m_1,m_2,m_3,m_4\}$，$m_1$ 为萼片长，$m_2$ 为萼片宽，$m_3$ 为花瓣长，$m_4$ 为花瓣宽。$\eta=A_1=m_1m_2m_3m_4$，表示萼片长、并且萼片宽、并且花瓣长、并且花瓣宽。

(3)聚类数 $C=3$，模糊因子 $m=2$，阈值 $\varepsilon=0.00001$。

(4)尺度参数 $\omega=0.25$。

将 SO-SCFCM、SCFCM 和 FCM 进行 10 次实验，取其均值进行比较，比较结果如表 1 和表 2 所列。

表 1 在 Iris 数据集 3 种算法的聚类质量指标实验结果

| 算法 | 调节因子 a | 错聚数 | 准确率 E | 聚类质量评价 STP | CMP | EVA | RT |
|---|---|---|---|---|---|---|---|
| SO-SCFCM | 2 | 9 | 94% | 1.3497 | 0.0412 | 0.0305 | 1.2630 |
| | 1.75 | 10 | 93.3% | 1.3484 | 0.0388 | 0.0287 | 1.2362 |
| | 0.3 | 9 | 94% | 1.2738 | 0.0102 | 0.0080 | 1.1230 |
| SCFCM | 2 | 15 | 90% | 1.7373 | 0.2411 | 0.1388 | 1.2640 |
| | 1.75 | 16 | 89.3% | 1.3241 | 0.1733 | 0.1309 | 1.1860 |
| FCM | ∞ | 16 | 89.3% | 1.7164 | 0.4034 | 0.2350 | 0.983 |

表 2 在 Iris 数据集 3 种算法的语义强度期望指标实验结果

| 算法 | 调节因子 a | SS $C_1$ | $C_2$ | $C_3$ | E | SSE |
|---|---|---|---|---|---|---|
| SO-SCFCM | 2 | 0.0735 | 0.0324 | 0.2784 | 94% | 0.261696 |
| | 1.75 | 0.2807 | 0.0752 | 0.0324 | 93.3% | 0.2618931 |
| | 0.3 | 0.0846 | 0.0324 | 0.3104 | 94% | 0.291776 |
| SCFCM | 2 | 0.0324 | 0.0867 | 0.3486 | 90% | 0.3252438 |
| | 1.75 | 0.0324 | 0.0867 | 0.3486 | 89.3% | 0.327684 |
| FCM | ∞ | 0.0834 | 0.0324 | 0.3403 | 89.3% | 0.3038879 |

由表 1 可以看出，在同样的模糊语义概念 $\eta$ 下，SO-SCFCM 算法因引入了谱聚类算法思想而使之能处理包括凹数据集在内的任意形状的数据集，使其聚类的有效性指标显著优于 SCFCM 和 FCM，而且其聚类的准确性也是最高的。从响应时间来看，FCM 的运行时间是最低的，其次为 SCFCM，最后为 SO-SCFCM。进一步地，SO-SCFCM 不但错聚数是最低的，而且，SO-SCFCM 中的错聚元素也出现在其它两个方法中。

从表 2 可以看出，SCFCM 的聚类语义强度期望值较大，而 SO-SCFCM 的聚类语义强度期望值较小。在这 3 种聚类算法中，语义强度期望值所在的类均为 Virginica 类。

**结束语** 本文先将条件聚类、基于 AFS 的模糊隶属度函数和经典的 FCM 集成在一个统一的聚类框架中，然后引入谱聚类思想处理非凸样本集合，使聚类结果既满足用户的指导，又得到更准确的聚类结果。

值得指出的是调节因子 $a$ 是用户指定的一个值，理论上该值越小，模糊语义指导力越强，其值越大，模糊语义指导力越弱，到无穷时语义指导消失，蜕变成经典的 FCM。调节因子 $a$ 取何值时准确率又高，语义指导能力又强，有待进一步研究。其次，聚类数 $K$ 对语义强度期望的影响值得关注。最后，谱聚类 $\delta$ 参数的选择是假设的，基于语义将 $\delta$ 参数、模糊因子 $m$ 以及调节因子 $a$ 进行联动研究也是本文的延伸。

## 参 考 文 献


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

[2] Jain A，Murty M，Flynn P. Data clustering：a review[J]. ACM Computing Surveys，1999，31(3)：264-323

[3] Bezdek J C. Pattern recognition with fuzzy objective function algorithms[M]. New York：Plenum，1981：52-103

[4] Leski J M. Generalized weighted conditional fuzzy clustering [J]. IEEE Trans. On Fuzzy Systems，2003，11(6)：709-715

[5] Pedrycz W. Conditional Fuzzy C-Means[J]. Pattern Recognition Letters，1996，17：625-632

[6] Liu X D，Pedrycz W. Axiomatic Fuzzy Set Theory and Its Applications[J]. Springer-Verlag Berlin Heidelberg，2009(4)：111-166

[7] Zadeh L A. Fuzzy Sets[J]. Information And Control，1965(8)：338-353

[8] 蔡晓妍，戴冠中，杨黎斌. 谱聚类算法综述[J]. 计算机科学，2008(07)：14-18

[9] Guo C，Zhao H. Community structure discovery method based on the Gaussian kernel similarity matrix[J]. Physica A：Statistional Mechonics and its Applications，2012，391(6)：2268-2278

[10] 郭崇慧，苏木亚. 基于独立成分分析的时间序列谱聚类方法[J]. 系统工程理论与实践，2011，31(10)：1921-1931

[11] 李洪波. 基于减法聚类和快速紧密性函数的 SF-FCM[J]. 控制与决策，2011，26(7)：1074-1078