基于公理化模糊子集的改进谱聚类算法

赵小强*①②③ 刘晓丽①

①(兰州理工大学电气工程与信息工程学院 兰州 730050)

②(甘肃省工业过程先进控制重点实验室 兰州 730050)

③(兰州理工大学国家级电气与控制工程实验教学中心 兰州 730050)

摘 要：谱聚类算法通常是采用高斯核作为相似性度量，并利用所有可用的特征来构建具有欧氏距离的相似度矩阵，数据集复杂度会影响其谱聚类性能，因此该文提出一种基于公理化模糊子集(AFS)的改进谱聚类算法。首先结合AFS算法，利用识别特征来衡量更合适的数据成对相似性，生成更强大的亲合矩阵；再有效地利用Nyström采样算法，计算采样点间以及采样点和剩余点间的相似度矩阵去降低计算的复杂度；最后通过在不同数据集以及图像分割上进行实验，证明了提出算法的有效性。

关键词：亲和矩阵；谱聚类；公理化模糊子集；Nyström采样算法

中图分类号：TP181 **文献标识码**：A **文章编号**：1009-5896(2018)08-1904-07

DOI: 10.11999/IEIT170904

An Improved Spectral Clustering Algorithm Based on Axiomatic Fuzzy Set

ZHAO Xiaoqiang①②③ LIU Xiaoli①

①(*College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou* 730050, *China*)

②(*Key Laboratory of Gansu Advanced Control for Industrial Processes, Lanzhou* 730050, *China*)

③(*National Experimental Teaching Center of Electrical and Control Engineering, Lanzhou University of Technology, Lanzhou* 730050, *China*)

Abstract: Gaussian kernel is usually used as the similarity measure in spectral clustering algorithm, and all the available features are used to construct the similarity matrix with Euclidean distance. The complexity of the data set would affect its spectral clustering performance. Therefore, an improved spectral clustering algorithm based on Axiomatic Fuzzy Set (AFS) is proposed. Firstly, AFS algorithm is combined to measure the similarity of more suitable data by recognizing features, and the stronger affinity matrix is generated. Then Nyström sampling algorithm is used to calculate the similarity matrix between the sampling points and the sampling points and the remaining points to reduce the computational complexity. Finally, the experiment is carried out by using different data sets and image segmentations, the effectiveness of the proposed algorithm are proved.

Key words: Affinity matrix; Spectral clustering; Axiomatic Fuzzy Set (AFS); Nyström sampling algorithm

1 引言

聚类算法[1–3]已经应用在许多学科研究中，如数据挖掘[4]、文档检索、图像分割[5]和模式识别。传统的聚类算法有K-means算法[6,7]、FCM算法[8]、EM算法等，这些聚类算法在凸球形的数据样本空间上能取得很好的聚类效果，然而，当数据以更复杂或未知的方式出现时，这些方法的聚类往往欠佳，易陷入局部最优。为了解决以上数据聚类问题以及如何收敛于全局最优问题，学者们提出了一种新的聚类算法—谱聚类算法[9](spectral clustering algorithm)，因此，Zelnik-Manor等人[10]提出了Self-Tuning谱聚类算法(STSC), STSC算法把样本点的邻域信息引入到相似矩阵的计算中，通过自适应得到相似矩阵函数；Yang等人[11]提出的密度敏

收稿日期：2017-09-25；改回日期：2018-05-02；网络出版：2018-05-30

*通信作者： 赵小强 xqzhao@lut.cn

基金项目：国家自然科学基金(61763029)，甘肃省基础研究创新群体基金(1506RJIA031)

Foundation Items: The National Natural Science Foundation of China (61763029), The Gansu Province Basic Research Innovation Group Fund (1506RJIA031)

感相似函数的谱聚类，其通过调整不同区域的密度距离定义了一个新的相似性测度函数的谱聚类算法，提高了集群性能的合成。然而，这些改进算法仍然容易受到噪声和不相关因素的影响。

数据相似度的概念往往紧密地联系在一个特定的度量函数上。然而，确定有效谱聚类的相似性一直以来是一个具有挑战性的问题[12]，复杂的数据往往是高维异构，在没有任何先验知识的情况下，盲目测量成对的相似性和构建数据图是容易受到噪声干扰[13]，尤其对现实世界的视觉数据，如图像和视频，由于不可控制的变化较多，如光照变化、背景来源不稳定，以及背景杂波等[14]。为了解决上述问题，Pavan等人[15]提出了一种通过选择最大群体(或最大化平均成对亲和度)来形成紧密邻域的概念算法，希望以较少的样本之间的亲缘关系边缘构造亲合图，在聚类问题上取得了一定的效果；文献[13]利用树层次相似度信息，用非线性亲和建构方式提出了一种随机森林的方法消除噪声特征。本文算法中的图推理方法不是盲目获取所有可用的变量，而是利用识别特征来衡量更合适的数据成对相似性，因此，创建的亲和度矩阵对于嘈杂的真实数据聚类效果更好。

本文提出的算法中，存储相似度矩阵需要的空间复杂度是$O(N^2)$，而对Laplace矩阵特征分解，需要的时间复杂度通常为$O(N^3)$，因此，降低谱聚类算法的复杂度并提高聚类处理速度就非常重要。目前，Nyström采样算法是降低复杂度的有效算法[16]，本文利用Nyström采样算法降低复杂度，并应用到改进谱聚类算法中。

2　AFS算法

AFS(Axiomatic Fuzzy Set)算法[17, 18]不是使用常见的欧式距离作为度量，而是通过模糊隶属函数来获取数据结构，而样本之间的距离由隶属度表示，能够建立距离的测量区分特征子空间，为处理现实数据中存在的噪声提供了有效方法。然而，在AFS聚类中，相似矩阵$\boldsymbol{S}=(s_{ij})_{N\times N}$不一定满足模糊传递条件。通常当且仅当它们之间的相似度大于或等于预定阈值α时，对象被认为与另一对象相似。因此，传递条件表明，对于任何3个对象i，j和k，如果对象i类似于对象$k(s_{ik}\geq\alpha)$，并且对象k类似于对象$j(s_{kj}\geq\alpha)$，对象i也就类似于对象$j(s_{ij}\geq\alpha)$。由于传递条件对于聚类是必不可少的，由$\mathbf{TC}(\boldsymbol{S})=(t_{ij})_{N\times N}$表示。$\mathbf{TC}(\boldsymbol{S})$被定义为最小对称和传递矩阵。通常，通过搜索整数$k$来获得$\mathbf{TC}(\boldsymbol{S})$，使得$\left(\boldsymbol{S}^k\right)^2=\boldsymbol{S}^k$。对于给定的阈值$\alpha$，对象可以被分割成不同的集群，然而，每个阈值$\alpha$导致一个特定的聚类结果，因此，评估标准对于获得清晰的结果是必要的，特别是在模糊聚类中。此外，相似性矩阵可能不是自反的(例如$s_{ii}=1$并不总是保持)，这意味着某些样本不能以某个α(当$s_{ii}<\alpha$)聚类。因此，对于AFS聚类(例如，提取在以前的聚类过程中未聚类的样本)需要重新聚类过程。上述过程也非常耗时。

3　基于AFS的改进谱聚类算法

3.1　在AFS空间中建立距离测量

由隶属函数确定的AFS模糊集及其逻辑运算，通过考虑模糊性和随机性，AFS算法可以基于数据库内的信息来创建隶属函数和模糊逻辑。同时，由从概率空间绘制的观测数据确定的隶属函数和模糊逻辑运算与由概率空间中表示的概率分布确定的隶属函数和模糊逻辑运算一致，AFS算法的主要思想是将数据转换为模糊隶属函数，并实现其逻辑运算，然后从AFS空间而不是原始特征空间提取信息。本文算法在AFS空间中建立距离测量，而不是通常使用的欧氏距离来更好地提取数据结构。

设样本集为$X=\{x_1,x_2,\cdots,x_n\}\subseteq R$，其中$x_i=\{x_{i1},x_{i2},\cdots,x_{is}\}$，$X$上的隶属度函数$F=\{f_1,f_2,\cdots,f_s\}$，$x_{ij}=f_j(x_i)$表示样本$x_i$在属性$f_j$上的值，其中$i=1,2,\cdots,n$，$j=1,2,\cdots,s$。

$\boldsymbol{M}=\{m_{1,1},m_{1,2},\cdots,m_{1,r_1},m_{2,1},m_{2,2},\cdots,m_{2,r_2},\cdots,m_{s,1},m_{s,2},\cdots,m_{s,r_s}\}$表示简单模糊概念的集合，属性$f_i$上取$r_i$个模糊概念，$m_{i,1},m_{i,2},\cdots,m_{i,r_i}$是与属性$f_i$相关联的模糊概念“大”，“中”，“小”等。

若m是非空集。集合EM*定义为

$$\mathrm{EM}^*=\left\{\sum_{i\in I}\left(\prod_{m\in\boldsymbol{A}_i}m\right)\middle|\boldsymbol{A}_i\subseteq\boldsymbol{M},i\in I\right\}\tag{1}$$

其中，I为非空索引集。

因此，EM*可以通过与等效关系相关联的EM*来定义，实际上，每个模糊集可以被唯一地分解：

$$\xi=\sum_{i\in I}\left(\prod_{m\in\boldsymbol{A}_i}m\right)\tag{2}$$

这里，$\boldsymbol{A}_i$是$\boldsymbol{M}$的子集。

设x是一个集合，$\boldsymbol{M}$是一个集合的模糊项。对于$\boldsymbol{A}\subseteq\boldsymbol{M}$，$x\in\boldsymbol{A}$，可以写成

$$\boldsymbol{A}^{\geq}(x)=\{y\in x\,|x\geq_m y，\text{对于}\forall m\in\boldsymbol{A}\}\subseteq X\tag{3}$$

其中，对于$m\in\boldsymbol{M}$，“$x\geq_m y$”意味着x属于m的程度大于等于y属于m的程度。$\boldsymbol{A}^{\geq}(x)$是x中其属性集合$\prod_{m\in\boldsymbol{A}}m$小于或等于$x$的所有元素的集合。$\boldsymbol{A}^{\geq}(x)$由

模糊集$\boldsymbol{A}$的语义和观测数据集的概率分布决定。

对于模糊集$\xi\in\mathrm{EM}$，对于μ_ξ：$X\to[0,1]$。$\{\mu_\xi(x)\,|\xi\in\mathrm{EM}\}$为AFS算法模糊逻辑的一组相关隶属函数。$\rho_v$称为模糊项$v$的权重函数，其中，$v$为$X$上的模糊项，$X\to R^+=[0,\infty)$。则可以计算相关隶属函数为

$$\mu_\xi(x)=\sup_{i\in I}\inf_{\gamma\in\boldsymbol{A}_i}\frac{\sum\limits_{u\in\boldsymbol{A}^{\geq}(x)}\rho_\gamma(u)\,N_u}{\sum\limits_{u\in X}\rho_\gamma(u)\,N_u},\quad\forall x\in X \tag{4}$$

其中，N_u是u的样本数。

本文算法对于度量空间中每个样本x，找到一个模糊子集$\zeta_x=\prod_{m\in M}m$，使得ζ_x有效地表示x，这个集合称为“x的描述”。在这里，模糊隶属函数被用作特征的度量，如果属于m_{ij}的x_k的隶属度大于某个阈值，m_{ij}足以区分x_k与其它阈值。使用以上的模糊项来定义一组：

$$B_x^\varepsilon=\{m\in M\,|\mu_m(x)\geq\max\{\mu_m(x)\}-\varepsilon\} \tag{5}$$

其中，ε是表示误差阈值的最小正数，其是经验性地设定，B_x^ε是可以表示x所有可能的模糊项的集合。x可以描述为

$$\zeta_x=\bigwedge_{m\in B_x^\varepsilon}m \tag{6}$$

其中，$\wedge$是AFS算法中的模糊连接逻辑运算。上述过程能在区分模糊子空间中用不同模糊项表示样本，通过模糊隶属函数选择这些子空间，消除噪声。此外，通过模糊隶属度和在AFS中定义的逻辑运算的数据距离推理，即使用属于由模糊集表示的另一个描述的样本的隶属度作为距离度量。对于两个样本X_i和X_j，它们之间的距离定义为

$$\boldsymbol{D}_{ij}=1-\min\left\{\bar{\mu}_{\zeta_{X_j}}(X_i),\,\bar{\mu}_{\zeta_{X_i}}(X_j)\right\} \tag{7}$$

$$\bar{\mu}_{X_i}(X_j)=\left\{m_k\in\zeta_{X_i}\left|\frac{\sum\limits_{k=1}^{N}\mu_{m_k}(X_j)}{N}\right.\right\} \tag{8}$$

其中，$\mu_{m_k}(X_j)$是属于模糊项m_k的X_j的隶属度，如式(6)所示，m_k表示属于ζ_{X_i}的模糊项，$\bar{\mu}_{X_i}(X_j)$表示属于X_i描述的X_j的平均隶属度。这种距离测量不是盲目地使用整个特征空间，而是考虑由一对样本共享的独特特征，通过减少无用特征或噪声来更好地表达数据中的真实结构的成对距离。

因此，本文将距离测量的方法应用到谱聚类算法中，得到新的相似度函数：

$$\boldsymbol{W}_{ij}=\exp\left(-\frac{{\boldsymbol{D}_{ij}}^2}{2\sigma_i\sigma_j}\right) \tag{9}$$

其中，$\sigma_i=d(s_i,s_K)$表示样本点s_i到第K个最近样本点之间的距离，$K=7$[10]。

3.2 Nyström采样算法

Nyström采样算法将N个数据点分成两部分，其中一部分为n个数据样本点，为随机抽样所得，另一部分为剩余的$N-n$个样本点，谱聚类相似矩阵$\boldsymbol{W}$就可以写成

$$\boldsymbol{W}=\begin{bmatrix}\boldsymbol{A}&\boldsymbol{B}\\\boldsymbol{B}^{\mathrm{T}}&\boldsymbol{C}\end{bmatrix} \tag{10}$$

其中，$\boldsymbol{A}\in\mathbb{R}^{n\times n}$，$\boldsymbol{A}(i,j)$为第$i$个随机抽样点与第$j$之间的相似度矩阵，$\boldsymbol{B}\in\mathbb{R}^{(N-n)\times n}$，$\boldsymbol{B}(i,k)$为第$i$个抽样点与第$k$个剩余点间的相似度矩阵，$\boldsymbol{C}\in\mathbb{R}^{(N-n)\times(N-n)}$，$\boldsymbol{C}(s,k)$为第$k$个剩余点与第$s$个剩余点间的相似度矩阵，令$\overline{\boldsymbol{U}}$为$\boldsymbol{W}$的近似特征向量，由Nyström扩展可得

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}} \tag{11}$$

$$\boldsymbol{\Lambda}=\begin{bmatrix}\lambda_1&&&\\&\lambda_2&&\\&&\ddots&\\&&&\lambda_n\end{bmatrix} \tag{12}$$

其中，λ_i为$\boldsymbol{A}$的特征值，$i=1,2,\cdots,n$，如λ_i均大于零，则$\boldsymbol{A}$为正定矩阵，反之则不是。

$$\overline{\boldsymbol{U}}=\begin{bmatrix}\boldsymbol{U}\\\boldsymbol{B}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\end{bmatrix} \tag{13}$$

则当$\hat{\boldsymbol{W}}$为近似的$\boldsymbol{W}$时，有

$$\begin{aligned}\hat{\boldsymbol{W}}&=\bar{\boldsymbol{U}}\boldsymbol{\Lambda}\bar{\boldsymbol{U}}^{T}=\begin{bmatrix}\boldsymbol{U}\\\boldsymbol{B}^{T}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\end{bmatrix}\boldsymbol{\Lambda}\begin{bmatrix}\boldsymbol{U}^{\mathrm{T}}&\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^{\mathrm{T}}&\boldsymbol{B}\end{bmatrix}\\&=\begin{bmatrix}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}&\boldsymbol{B}\\\boldsymbol{B}^{\mathrm{T}}&\boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{B}\end{bmatrix}=\begin{bmatrix}\boldsymbol{A}&\boldsymbol{B}\\\boldsymbol{B}^{\mathrm{T}}&\boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{B}\end{bmatrix}\end{aligned} \tag{14}$$

由式(14)可知，Nyström扩展采样算法用$\boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{B}$逼近$\boldsymbol{C}$。由于$n\ll N$，$N-n$会很大，而采用Nyström扩展采样算法逼近，避免了使用剩余点间的相似度矩阵，就会在很大程度上减低问题的空间和时间的复杂度。

定理 1 如果$\boldsymbol{A}$是正定矩阵，定义：

$$\boldsymbol{P}=\boldsymbol{A}+\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{-1/2} \tag{15}$$

其中，$\boldsymbol{A}^{1/2}$为$\boldsymbol{A}$的对称正定的平方根，将$\boldsymbol{P}$对角化得

$$\boldsymbol{P}=\boldsymbol{U}_{\boldsymbol{P}}\boldsymbol{\Lambda}{\boldsymbol{U}_{\boldsymbol{P}}}^{\mathrm{T}} \tag{16}$$

则 $\hat{\boldsymbol{W}}$ 的特征向量为

$$\boldsymbol{V}=\begin{bmatrix}\boldsymbol{A}\\ \boldsymbol{B}^{\mathrm{T}}\end{bmatrix}\boldsymbol{A}^{-1/2}\boldsymbol{U}_{\boldsymbol{P}}\boldsymbol{\Lambda}_{\boldsymbol{P}}{}^{-1/2} \tag{17}$$

即 $\hat{\boldsymbol{W}}=\boldsymbol{V}\boldsymbol{\Lambda}_{\boldsymbol{P}}\boldsymbol{V}^{\mathrm{T}}$, $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{I}$。

定理 2　如果 $\boldsymbol{A}$ 是非正定矩阵，令

$$\bar{\boldsymbol{U}}_{\boldsymbol{P}}^{\mathrm{T}}=\boldsymbol{U}_{\boldsymbol{P}}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{P}}^{-1}\boldsymbol{U}_{\boldsymbol{P}}^{\mathrm{T}}\boldsymbol{B} \tag{18}$$

定义：$\boldsymbol{S}=\bar{\boldsymbol{U}}_{\boldsymbol{P}}\boldsymbol{\Lambda}^{1/2}$，对 $\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}$ 进行对角分解，可得

$$\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}=\boldsymbol{Q}\boldsymbol{\Lambda}_{\boldsymbol{Q}}\boldsymbol{Q}^{\mathrm{T}} \tag{19}$$

则 $\hat{\boldsymbol{W}}$ 的特征向量为

$$\boldsymbol{V}=\boldsymbol{S}\boldsymbol{Q}\boldsymbol{\Lambda}_{\boldsymbol{Q}}^{-1/2} \tag{20}$$

即 $\hat{\boldsymbol{W}}=\boldsymbol{V}\boldsymbol{\Lambda}_{\boldsymbol{Q}}\boldsymbol{V}^{\mathrm{T}}$, $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{I}$。

在谱聚类算法中，要对相似度矩阵进行归一化处理。文献[19]中提出了节点度：

$$\begin{aligned}d&=\hat{\boldsymbol{W}}\boldsymbol{1}\\&=\begin{bmatrix}\boldsymbol{A}\boldsymbol{1}_n+\boldsymbol{B}\boldsymbol{1}_m\\ \boldsymbol{B}^{\mathrm{T}}\boldsymbol{1}_n+\boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{1}_m\end{bmatrix}=\begin{bmatrix}\boldsymbol{a}_r+\boldsymbol{b}_r\\ \boldsymbol{b}_t+\boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{b}_r\end{bmatrix}\end{aligned} \tag{21}$$

其中，$m=N-n$，$\boldsymbol{a}_r,\boldsymbol{b}_r\in\mathbb{R}^m$ 分别表示矩阵 $\boldsymbol{A},\boldsymbol{B}$ 的行之和，$\boldsymbol{b}_t\in\mathbb{R}^n$ 为矩阵 $\boldsymbol{B}$ 的列之和，$\boldsymbol{1}$ 表示元素均为1的列向量，利用节点度将 $\boldsymbol{A},\boldsymbol{B}$ 归一化处理：

$$\begin{aligned}&\frac{\boldsymbol{A}_{ij}}{\sqrt{d_id_j}}\to\boldsymbol{A}_{ij},\quad i,j=1,2,\cdots,n\\&\frac{\boldsymbol{B}_{ij}}{\sqrt{d_id_{j+n}}}\to\boldsymbol{B}_{ij},\quad i,j=1,2,\cdots,n\end{aligned} \tag{22}$$

3.3　基于AFS的改进谱聚类算法的步骤

输入：数据集 $X=\{x_i|i=1,2,\cdots,N\}$，Nyström采样算法的随机采样数为 n，聚类数目为 k；

输出：聚类产生 k 个类；

步骤1　为每个特征 f_i 构造模糊项 m_{ij}；

步骤2　用式(4)计算样本隶属函数 $\mu_{m\in M}(x)$；

步骤3　利用式(5)找到模糊项 B_x^{ε}；

步骤4　利用式(6)构建每个样本的描述 ζ_x；

步骤5　通过式(7)计算成对距离 $\boldsymbol{D}_{ij}$；

步骤6　用公式(9)构建亲和矩阵 $\boldsymbol{W}_{ij}$；

步骤7　利用Nyström采样算法，随机采样 n 个数据点，然后从 $\boldsymbol{W}$ 中选出相应的元素，再构造抽样点间的相似度矩阵 $\boldsymbol{A}$ 和抽样点与剩余点间的相似度矩阵 $\boldsymbol{B}$；

步骤8　利用式(21)计算节点度 d，然后根据式(22)对 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 进行归一化处理；

步骤9　在 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 进行归一化处理后，根据定理1以及定理2计算出 $\boldsymbol{W}_{ij}$ 的正交特征向量 $\boldsymbol{V}$；

步骤10　计算矩阵 $\boldsymbol{V}$ 前 k 个最大的特征向量，记为：$\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k$，设矩阵 $\boldsymbol{V}_k=[\boldsymbol{v}_1\ \boldsymbol{v}_2\cdots\ \boldsymbol{v}_k]$；

步骤11　对矩阵 $\boldsymbol{V}$ 的每一行进行归一化处理得到矩阵 $\boldsymbol{Y}_{ij}=\dfrac{\boldsymbol{v}_{ij}}{\left(\sum_j\boldsymbol{v}_{ij}^2\right)^{1/2}}$；

步骤12　使用K-means算法对矩阵 $\boldsymbol{Y}_{ij}$ 进行聚类。

4　实验分析

4.1　性能指标

为了评估该算法的性能，本文算法与其它聚类方法使用以下两个性能指标比较聚类结果：聚类误差(Clustering Error, CE)[20]和归一化互信息(Normalized Mutual Information, NMI)。

CE被广泛用于评价聚类性能，CE的计算为

$$\mathrm{CE}=1-\frac{\sum_{i=1}^{n}\delta\left(t_i,\mathrm{map}\left(r_i\right)\right)}{n} \tag{23}$$

其中，t_i 和 r_i 分别是 x_i 的真实类标签和获得的聚类索引，$\delta(x,y)$ 是三角函数。

NMI是测量算法聚类性能的另一个广泛使用的度量标准，如式(24)[21]：

$$\mathrm{NMI}=\frac{\sum_{l}^{c}\sum_{h=1}^{c}n_{l,h}\lg\left(\dfrac{nn_{l,h}}{n_l\hat{n}_h}\right)}{\sqrt{\left(\sum_{l=1}^{c}n_l\lg\dfrac{n_l}{n}\right)\left(\sum_{h=1}^{c}\hat{n}_h\lg\dfrac{\hat{n}_h}{n}\right)}} \tag{24}$$

其中，n_l 表示簇 $c_l(1\le l\le c)$ 中包含的数据数，$\hat{n}_h$ 是属于第 h 类的数据数 $(1\le h\le c)$，$n_{l,h}$ 表示簇 c_l 和第 h 类之间的交集中的数据数。NMI越大，性能越好。

4.2　UCI数据集的实验结果与分析

为了验证本文算法的聚类效果，本文分别选取了UCI数据集中的以下8类不同数据集作为测试样本，表1为这8类数据集的特征，分别为数据总数、类的个数以及维数。

表 1　数据集特征

数据集	数据总数	类数	维数
Iris	150	3	4
Heart	270	2	13
Sonar	208	2	60
Wine	178	3	13
Protein	552	8	77
Hepatitis	155	2	19
Segmentation	2310	7	19
Pen digits	10992	10	16

本文分别采用4种聚类算法进行对比：SC, STSC, AFS和本文算法。在SC, STSC以及本文算法中，都使用了K-means算法对矩阵的前K个最大特征值对应的特征向量进行聚类，聚类前对数据集进行归一化处理，对特征向量独立运行30次，其结果为所有数据点到聚类中心距离的平方和的最小值。本文算法中，从原始数据集中随机采样100个数据点作为数据采样点，表2为独立运行30次4种算法的平均值。

由表2可以得出，本文算法对比其它算法在多个数据集上都有所提高。在Iris, Wine数据集中，本文算法比STSC算法相对聚类误差率略高，由于这两个数据集中只有150个样本和178个样本，实际上差异只有1个或2个样本；AFS聚类对Protein, Heart等多集群的数据集，聚类误差率较高，是因为AFS根据每个集群的边界选择最佳的数据分区，随着群集数量的增加，寻找边界的难度也越来越大，而本文算法在Heart, Sonar, Protein, Hepatitis, Segmentation, Pen digits数据集上，相对于SC算法、STSC算法以及AFS算法，聚类误差(CE)均得到了改善。因此，本文算法降低了聚类误差率。

NMI测量的聚类性能如表3所示，本文算法优于其它算法，在处理Sonar数据集时，STSC的聚类性能相对较弱。由于STSC和本文算法都使用邻域信息，因此可知与基于整个特征空间的欧氏距离相比，采用具有特征特征子空间的模糊隶属度作为距离度量，在亲和度图构造方面能使聚类性能更好。

表 2 数据集的CE(%)

数据集	SC	STSC	AFS	本文算法
Iris	10.71	7.46	9.72	7.63
Heart	20.96	22.13	30.63	12.42
Sonar	44.53	46.83	38.52	33.60
Wine	2.92	2.91	3.54	3.13
Protein	54.70	55.67	55.12	48.87
Hepatitis	30.76	38.73	32.34	23.20
Segmentation	22.08	21.35	31.17	18.63
Pen digits	25.37	24.25	–	22.16

表 3 数据集的NMI(%)

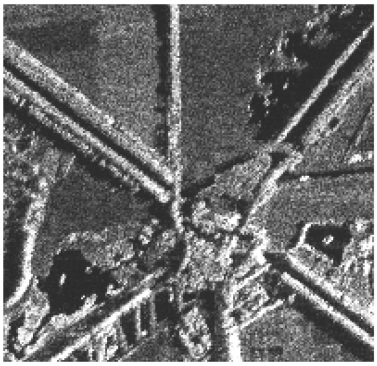
数据集	SC	STSC	AFS	本文算法
Iris	75.87	78.63	78.06	85.49
Heart	28.54	26.23	18.45	40.33
Sonar	7.32	1.83	15.47	22.38
Wine	89.30	89.34	85.67	87.96
Protein	54.43	48.24	36.62	65.80
Hepatitis	13.75	4.78	3.57	17.42
Segmentation	65.58	66.72	58.56	72.24
Pen digits	60.53	61.48	–	66.52

4.3 图像分割

为了进一步验证本文算法的有效性，将其应用到图像分割。本文算法中，图像提取采用基于小波能量的特征提取方法[22]，从原始数据集中随机采样100个像素点作为采样点。

图1(a)为一副256×256像素的合成孔径雷达(Synthetic Aperturo Radar, SAR)图像，图2(a)为谱聚类算法分割结果，图3(a)为本文算法分割结果。表4是SAR图像分割性能对比表，由于使用了像素的灰度信息，并且谱聚类算法中的欧氏距离对噪声相对较为敏感。而本文算法利用模糊隶属函数导出关联图，来代替使用欧氏距离，有效地消除了噪声。由表4可知本文算法在有效缩短运行时间的同时还提高了图像分割的精度和效率，因此本文算法对于SAR图像的分割优于谱聚类算法。

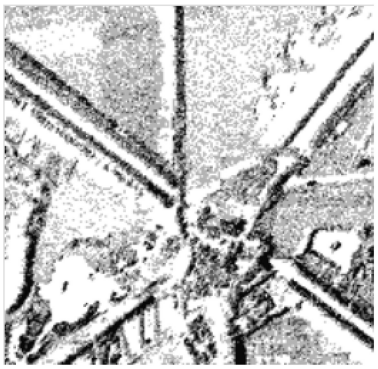
图1(b)为一副350×258像素的图像，包括3类地物：树、河流和陆地；图2(b)为谱聚类算法分割结果；图3(b)为本文算法分割结果。由表5可知，本文算法相比谱聚类算法缩短了运行时间，降低了误



(a) SAR 图像

(b) 树图像

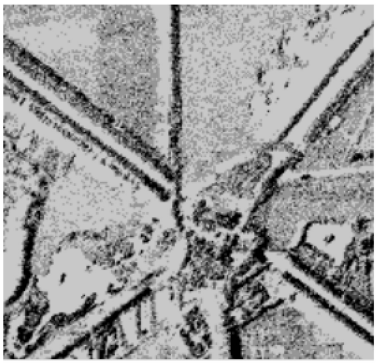
图 1 原图



(a) SAR 图像

(b) 树图像

图 2 谱聚类算法分割结果



(a) SAR 图像

(b) 树图像

图 3 本文算法分割结果

表 4　SAR图像分割性能对比表

	谱聚类算法	本文算法
运行时间(S)	30.62	3.27
误分率(%)	9.53	5.34

表 5　树图像分割性能对比表

	谱聚类算法	本文算法
运行时间(S)	16.39	4.25
误分率(%)	6.87	2.13

分率，因此本文算法图像分割明显优于谱聚类算法。

其性能用运行时间和误分率表示，其中误分率的计算方法为：$\mathrm{RI}=(n_z/N)\times 100\%$，其中，$n_z$为错误分割的像素点数，$N$为图像的总像素点数。

4.4　算法复杂度的分析

本文算法利用Nyström采样算法降低数据处理的复杂度，其复杂度分析如表6所示。其中，N为样本点的总数，n为随机采样数，T为K-means算法的迭代次数。由此可见，本文算法的计算复杂度为$O(n^2N)$，而未利用Nyström采样算法时的计算复杂度为$O(N^3)$，因此，利用Nyström采样算法有效降低了数据处理的复杂度。

表 6　复杂度分析

	计算步骤	复杂度
若矩阵$\boldsymbol{A}$正定	计算矩阵$\boldsymbol{A}$	$O(n^2)$
	计算矩阵$\boldsymbol{B}$	$O(n(N-n))$
	对$\boldsymbol{A}$矩阵分解	$O(n^3)$
	求解矩阵$\boldsymbol{p}$	$O(n^2(N-n))$
	矩阵分解	$O(n^3)$
	求解矩阵$\boldsymbol{Y}$	$O(n^2N)$
若矩阵$\boldsymbol{A}$非正定	求解矩阵$\boldsymbol{S}$	$O(n^2N)$
	对矩阵$\boldsymbol{S}$对角分解	$O(n^3)$
	求解矩阵$\boldsymbol{Y}$	$O(n^2N)$
K-means算法进行聚类		$O(nK^2T)$

5　结论

本文通过改进AFS算法的模糊理论定义数据相似性，并通过模糊隶属函数导出关联图，来代替使用欧氏距离，从而减少噪声的干扰。此外，本文通过亲和图获取和结合分布在区分特征子空间中的样本，并通过每个样本的模糊描述识别邻近度。再利用Nyström采样算法去降低计算的复杂度，通过在不同类型数据集以及图像实验证明其有效性。与其他先进算法相比，实验结果显示出了其优越性和可行性。

参 考 文 献

[1] JAIN A. Data clustering: a review[J]. *ACM Computing Surveys*, 1999, 31(3): 264–323. doi: 10.1145/331499.331504.

[2] XU Rui. Survey of Clustering Algorithms[M]. New Jersey: IEEE Press, 2005: 645–678. doi: 10.1109/TNN.2005.845141.

[3] RAMON-GONEN R and GELBARD R. Cluster evolution analysis: Identification and detection of similar clusters and migration patterns[J]. *Expert Systems with Applications*, 2017, 83: 363–378. doi: 10.1016/j.eswa.2017.04.007.

[4] WITTEN I H and FRANK E. Data Mining: Practical Machine Learning Tools and Techniques[M]. Massachusetts: Morgan Kaufmann, 2005: 81–82. doi: 0120884070, 9780120884070.

[5] 夏平, 任强, 吴涛, 等. 融合多尺度统计信息模糊C均值聚类与Markov随机场的小波域声纳图像分割[J]. 兵工学报, 2017, 38(5): 940–948. doi: 10.3969/j.issn.1000-1093.2017.05.014.
XIA Ping, REN Qiang, WU Tao, *et al*. Sonar image segmentation fusion of multi-scale statistical information FCM clustering and MRF model in wavelet domain[J]. *Acta Armamentaria*, 2017, 38(5): 940–948. doi: 10.3969/j.issn.1000-1093.2017.05.014.

[6] TREVOR H, ROBERT T, and FRIEDMAN J J H. The Elements of Statistical Learning[M]. New York: Springer, 2001: 460–514. doi: 10.1198/jasa.2004.s339.

[7] 李武, 赵娇燕, 严太山. 基于平均差异度优选初始聚类中心的改进K-均值聚类算法[J]. 控制与决策, 2017, 32(4): 759–762. doi: 10.13195/j.kzyjc.2016.0274.
LI Wu, ZHAO Jiaoyan, and YAN Taishan. Improved K-means clustering algorithm optimizing initial clustering centers based on average difference degree[J]. *Control and Decision*, 2017, 32(4): 759–762. doi: 10.13195/j.kzyjc.2016.0274.

[8] CHEN Weijie and GIGER M L. A fuzzy c-means (fcm) based algorithm for intensity inhomogeneity correction and segmentation of MR images[C]. IEEE International Symposium on Biomedical Imaging, Nano to Macro Marriott Crystal Gateway, Arlington, 2005, 2: 1307–1310. doi: 10.1109/ISBI.2004.1398786.

[9] 蔡晓妍, 戴冠中, 杨黎斌. 谱聚类算法综述[J]. 计算机科学, 2008, 35(7): 14–18. doi: 10.3969/j.issn.1002-137X.2008.07.004.
CAI Xiaoyan, DAI Guanzhong, and YANG Libin. Survey on spectral clustering algorithms[J]. *Computer Science*, 2008, 35(7): 14–18. doi: 10.3969/j.issn.1002-137X.2008.07.004.

[10] ZELNIK-MANOR L and PERONA P. Self-tuning spectral clustering[C]. Advances in Neural Information Processing Systems, Vancouver, 2004: 1601–1608.

[11] YANG Peng, ZHU Qingsheng, and HUANG Biao. Spectral clustering with density sensitive similarity function[J]. *Knowledge-Based Systems*, 2011, 24(5): 621–628. doi: 10.1016/j.knosys.2011.01.009.

[12] JAIN A K. Data Clustering: 50 Years Beyond K-means[M]. Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2008: 651–666. doi: 10.1007/978-3-540-87479-9_3.

[13] ZHU Xiatian, CHEN Change Loy, and GONG Shaogang. Constructing robust affinity graphs for spectral clustering[C]. Computer Vision and Pattern Recognition. IEEE, Ohio, 2014: 1450–1457. doi: 10.1109/CVPR.2014.188.

[14] GONG Shaogang, CHEN Change Loy, and XIANG Tao. Security and Surveillance[M]. Visual Analysis of Humans. Springer London, 2011: 455–472. doi: 10.1007/978-0-85729-997-0_23.

[15] PAVAN M and PELILLO M. Dominant sets and pairwise clustering[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2006, 29(1): 167–172. doi: 10.1109/TPAMI.2007.10.

[16] 丁世飞, 贾洪杰, 史忠植. 基于自适应 采样的大数据谱聚类算法[J]. 软件学报, 2014(9): 2037–2049. doi: 10.13328/j.cnki.jos.004643.
DING Shifei, JIA Hongjie, and SHI Zhongzhi. Spectral clustering algorithm based on adaptive Nyström sampling for big data analysis[J]. *Journal of Software*, 2014(9): 2037–2049. doi: 10.13328/j.cnki.jos.004643.

[17] LIU Xiaodong and REN Yan. Novel artificial intelligent techniques via AFS theory: Feature selection, concept categorization and characteristic description[J]. *Applied Soft Computing*, 2010, 10(3): 793–805. doi: 10.1016/j.asoc.2009.09.009.

[18] LIH Xiaodong, WANG Xianchang, and PEDRYCZ W. Fuzzy clustering with semantic interpretation[J]. *Applied Soft Computing*, 2015, 26: 21–30. doi: 10.1016/j.asoc.2014.09.037.

[19] BELONGIE S, FOWLKES C, FAN C, *et al*. Spectral partitioning with indefinite kernels using the Nyström extension[C]. European Conference on Computer Vision. Springer-Verlag, Denmark, 2002: 531–542. doi: 10.1007/3-540-47977-5_35.

[20] WU Mingrui. A local learning approach for clustering[C]. International Conference on Neural Information Processing Systems, Hong Kong, China, 2006: 1529–1536.

[21] STREHL A and GHOSH J. Cluster ensembles-aknowledge reuse framework for combining multiple partitions[J]. *The Journal of Machine Learning Research*, 2003, 3(3): 583–617. doi: 10.1162/153244303321897735.

[22] HU Zhaoling, GUO Dazhi, and SHENG Yehua. Extracting textural in formation of satellite SAR image based on wavelet decomposition[J]. *Journal of Remote Sensing*, 2001, 5: 423–427.

赵小强：男，1969年生，博士生导师，教授，主要研究方向为数据挖掘、故障诊断、图像处理、污水处理、生产调度等.

刘晓丽：女，1992年生，硕士生，研究方向为数据挖掘.