DOI: 10.11992/tis.201810002
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.tp.20190527.1407.010.html

公理化模糊共享近邻自适应谱聚类算法

储德润, 周治平
(江南大学 物联网技术应用教育部工程研究中心, 江苏 无锡 214122)

摘 要: 针对传统的谱聚类算法通常利用高斯核函数作为相似性度量, 且单纯以距离决定相似性不能充分表现原始数据中固有的模糊性、不确定性和复杂性, 导致聚类性能降低的问题。提出了一种公理化模糊共享近邻自适应谱聚类算法, 首先结合公理化模糊集理论提出了一种模糊相似性度量方法, 利用识别特征来衡量更合适的数据成对相似性, 然后采用共享近邻的方法发现密集区域样本点分布的结构和密度信息, 并且根据每个点所处领域的稠密程度自动调节参数 σ, 从而生成更强大的亲和矩阵, 进一步提高聚类准确率。实验表明, 相较于距离谱聚类、自适应谱聚类、模糊聚类方法和地标点谱聚类, 所提算法有着更好的聚类性能。

关键词: 机器学习; 数据挖掘; 聚类分析; 模糊聚类; 谱聚类; 公理化模糊集理论; 共享最近邻; 尺度参数
中图分类号: TP18 **文献标志码**: A **文章编号**: 1673−4785(2019)05−0897−08

中文引用格式: 储德润, 周治平. 公理化模糊共享近邻自适应谱聚类算法 [J]. 智能系统学报, 2019, 14(5): 897–904.
英文引用格式: CHU Derun, ZHOU Zhiping. Shared nearest neighbor adaptive spectral clustering algorithm based on axiomatic fuzzy set theory[J]. CAAI transactions on intelligent systems, 2019, 14(5): 897–904.

Shared nearest neighbor adaptive spectral clustering algorithm based on axiomatic fuzzy set theory

CHU Derun, ZHOU Zhiping
(Engineering Research Center of Internet of Things Technology Applications Ministry of Education, Jiangnan University, Wuxi 214122, China)

Abstract: For the traditional spectral clustering algorithm, the Gaussian kernel function is usually used as the similarity measure. However, the similarity of distance cannot fully express the ambiguity, uncertainty, and complexity inherent in the original data, resulting in the reduction of clustering performance. To solve this problem, we propose an axiomatic fuzzy set shared nearest neighbor adaptive spectral clustering algorithm. First, the proposed algorithm uses a fuzzy similarity measurement method based on axiomatic fuzzy set theory to measure more suitable data pairwise similarity by identifying features. Then, the structure and density information of sample point distribution in a dense area is obtained using the method of sharing the nearest neighbor, and the parameter σ is automatically adjusted according to the density degree of each point in the domain, thereby generating a more powerful affinity matrix to further increase the accuracy rate of clustering. Experimental results show that the proposed algorithm has better clustering performance than distance spectral clustering, adaptive spectral clustering, fuzzy clustering, and landmark spectral clustering.
Keywords: machine learning; data mining; clustering analysis; fuzzy clustering; spectral clustering; axiomatic fuzzy set theory; shared nearest neighbor; scale parameter

聚类技术作为机器学习领域中的一种无监督技术, 在检测数据的内在结构和潜在知识方面发挥着重要的作用。在过去的几十年中, 许多聚类方法得到了发展, 如基于分区的方法 (k-means)、基于模型的方法、基于密度的方法、层次聚类方法、模糊聚类方法 (fuzzy c-means) 和基于图的方

收稿日期: 2018−10−03. **网络出版日期**: 2019−05−28.
通信作者: 储德润. E-mail: CDR0727@163.com.

法 (spectral clustering)[1]。由于谱聚类在处理非凸形结构的数据集方面的高效性，谱聚类在图像分割[2-4]、社区检测[5]、人脸识别[6] 等方面得到了广泛的研究和应用。

然而，传统的谱聚类算法在使用高斯核函数构造亲和矩阵时，需要先验信息来设置合适的参数以控制邻域的尺度；并且以距离来度量数据点之间的相似性，没有考虑到整体数据点的变化情况，对于高维数据来说，较难得到更高的聚类精度。近年来有很多学者对谱聚类算法进行了研究。赵晓晓等[7] 结合稀疏表示和约束传递，提出一种结合稀疏表示和约束传递的半监督谱聚类算法，进一步提高了聚类准确率。林大华等[8] 针对现有子空间聚类算法没有利用样本自表达和稀疏相似度矩阵，提出了一种新的稀疏样本自表达子空间聚类方法，所获得的相似度矩阵具有良好的子空间结构和鲁棒性。Chang 等[9] 提出了一种通过嵌入标签传播来改进谱聚类的方法，通过密集的未标记数据区域传播标签。以上方法虽然一定程度上提高了谱聚类算法的聚类性能，但是，在大部分谱聚类算法中，高斯核函数中尺度参数 σ 的选取往往都是通过人工选取，对聚类结果有一定的影响。NJW 算法[10] 对预先给定几个尺度参数 σ 进行谱聚类，最后从这几个聚类结果中选择具有最佳聚类结果的 σ 作为尺度参数，该方法消除了尺度参数 σ 选择的人为因素，但是也增加了计算时间。Ye 等[11] 在有向 KNN 图中考虑了一种基于共享最近邻的鲁棒相似性度量，大大提高了谱聚类的聚类精度。Jia 等[12] 提出了一种基于共享近邻的自校正 p 谱聚类算法，该算法利用共享最近邻来度量数据间的相似性，然后应用果蝇优化算法找到 p-laplacian 矩阵的最优参数 p，从而更好地进行数据分类。王雅琳等[13] 提出一种基于密度调整的改进自适应谱聚类算法，通过样本点的近邻距离自适应得到尺度参数，使算法对尺度参数相对不敏感。

传统的谱聚类以及上述大部分改进的谱聚类算法都是单一的针对距离度量或者尺度参数进行调整，本文从一个新的角度出发，在公理化模糊集 (AFS) 理论的基础上，结合局部密度估计和共享近邻的定义，提出一种基于 AFS 理论的共享近邻自适应谱聚类算法——公理化模糊共享近邻自适应谱聚类算法。利用 AFS 理论提出了一种模糊相似性度量方法，并将其作为谱聚类算法输入的亲和矩阵。同时采用共享近邻的方法发现密集区域样本点分布的结构和密度信息，并且根据每个点所处领域的稠密程度自适应调节参数 σ，与高斯核距离测度相比，本文的解决方案对参数具有较强的鲁棒性，增强了对各种数据集的适应性，减少了噪声数据带来的不良影响。

1　相关算法理论

1.1　谱聚类算法

在谱聚类中，设 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\in\mathbf{R}^d$ 为 n 个具有 d 个特征的样本，数据集可用一个加权无向图来描述，该图由 $|V|$ 个顶点和 $|E|$ 个边组成。对于 $\boldsymbol{v}_i\in V$，$\boldsymbol{v}_i$ 表示一个样本 $\boldsymbol{x}_i$，$\boldsymbol{e}_{ij}$ 表示 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 之间的权重。$\boldsymbol{e}_{ij}$ 通常是由 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的相似性来度量的。通常引入高斯核函数来构造相似矩阵 $\boldsymbol{S}$，其定义为

$$\boldsymbol{S}_{ij}=\begin{cases}\mathrm{e}^{-\frac{\left(\|x_i-x_j\|^2\right)}{2\sigma^2}}, & i\neq j\\ 0, & i=j\end{cases}\tag{1}$$

作为一种简单而有效的聚类准则，归一化割 (Ncut) 在文献 [14] 中提出，其定义为

$$\mathrm{Ncut}(A,B)=\frac{\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y}}\tag{2}$$

式中：$\boldsymbol{D}$ 是对角矩阵；$\boldsymbol{y}$ 是最优分割向量。谱聚类的一般过程为：

1) 构造图 $\boldsymbol{G}$ 的相似度矩阵 $\boldsymbol{S}$,对于给定的 σ，$\boldsymbol{S}$ 是由式 (1) 构造的；

2) 计算对角矩阵 $\boldsymbol{D}$，构造对角矩阵 $\boldsymbol{D}$，$\boldsymbol{D}_{ii}=\sum\limits_{1\leqslant j\leqslant n}\boldsymbol{S}_{ij}$；

3) 计算归一化拉普拉斯矩阵，拉普拉斯矩阵 $\boldsymbol{L}$ 为 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$，$\boldsymbol{L}$ 被归一化为

$$\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2};$$

4) 对 $\boldsymbol{L}$ 进行特征分解，得到前 k 个特征向量，并构建特征向量空间；

5) 利用经典的聚类算法如 k-means 对特征向量空间中的特征向量进行聚类。

1.2　AFS 理论

AFS 理论是刘晓东[15-18] 在 1998 年提出的一种基于 AFS 代数和 AFS 结构的模糊理论，AFS 理论放弃使用距离度量来计算数据之间的相似性关系，而是将观测数据转化为模糊隶属函数，并实现其逻辑运算。然后，可以从 AFS 空间而不是原始特征空间中提取信息。在 AFS 空间中利用模糊关系来构建数据之间的相似性度量。采用模糊隶属度来表示数据之间的距离关系，增

强了在处理现实数据中对各种数据集的适应性, 为处理离群点提供了有效方法。

在文献 [15] 中, 根据 EI 代数的定义, 对于任意概念集合 $A\subseteq M$, $\prod\limits_{m\in A} m$ 表示 A 中模糊概念的集合, 为了更好地提取数据结构, 清楚地说明可将 AFS 理论结合以下场景:

设 $X=\{x_1,x_2,\cdots,x_6\}$ 是 6 个人的样本集合及其特征 (属性), 由真实数字 (年龄、身高、体重)、布尔值 (性别)、额定值 (黑色素) 和顺序关系 (发黑、发白) 描述; $M=\{m_1,m_2,\cdots,m_6\}$ 是样本 X 上的模糊概念集。M 中的每一个元素都被看作描述事物的简单概念, 它们相应的语言标签被定义为: m_1(老年人)、m_2(高的人)、m_3(重的人)、m_4(头发黑色素更多的)、m_5(男性)、m_6(女性)。则 $A_1=\{m_1,m_6\}\subseteq M$, $\prod\limits_{m\in A_1} m=m_1m_6$ 是一个新的复杂概念“老年女人”; $A_2=\{m_1,m_3\}\subseteq M$, $\prod\limits_{m\in A_2} m=m_1m_3$ 表示“重的老年人”; $A_3=\{m_3\}\subseteq M$, $\prod\limits_{m\in A_3} m=m_3$ 表示“高的人”。对于 $\gamma=m_1m_6+m_1m_3+m_2$, 它的概念为“老年女人”或者“重的老年人”或者“高的人”, 是一个复杂概念的集合。一个新的模糊集可以被写成:

$$\sum_{i=1}^{3}\left(\prod_{m\in A_i} m\right)=\prod_{m\in A_1} m+\prod_{m\in A_2} m+\prod_{m\in A_3} m \tag{3}$$

这些基于简单概念的 EI 代数运算生成的概念被认为复杂的概念。

因此, 概念的逻辑表达可以用 $\sum\limits_{i\in I}\left(\prod\limits_{m\in A_i} m\right)$ 表示 $(A_i\subseteq M, i\in I)$。若 m 是非空集, 集合 E 定义为

$$E=\left\{\sum_{i\in I}\left(\prod_{m\in A_i} m\right)\middle| A_i\subseteq M, i\in I\right\} \tag{4}$$

式中 I 为非空索引集。

其中, 每个模糊集可以被唯一地分解:

$$\xi=\sum_{i\in I}\left(\prod_{m\in A_i} m\right) \tag{5}$$

式中 A_i 是 M 的子集。

2　所提算法

2.1　在模糊空间建立距离度量

本文提出的亲和矩阵构造方法是建立在 AFS 理论基础上的, 该过程允许我们在发现的判别模糊子空间中表示不同模糊项的样本。这些子空间由模糊隶属函数选择, 消除了不明显或噪声特征。因此, 它们被认为能够改善内部相似和减少相互相似。此外, 利用 AFS 中定义的模糊隶属度和逻辑运算, 放宽了欧氏假设对数据距离推断的影响。更具体地说, 本文使用一个样本的隶属度属于另一个样本的描述, 用模糊集表示为距离度量。在最初的 AFS 聚类[19-21] 基础上, 在 AFS 空间上构建相似性度量。

首先建立隶属度函数, 需要定义以下有序关系: 设 X 是一个样本集合, M 是 X 上的一组模糊集合。对于任意 $A_i\subseteq M$, $x\in X$, 可以写成:

$$A_i\geqslant(x)=\{y\in X|x\geqslant_m y, \text{for}\ \forall m\in A_i\}\subseteq X \tag{6}$$

式中: $m\in M$; “$x\geqslant_m y$”表示 m 代表 x 的模糊关系大于或等于 m 代表 y 的模糊关系; $A_i\geqslant(x)$ 是 x 中所有模糊关系小于或等于 $\prod\limits_{m\in A_i} m$ 的 x 样本中的元素的集合。$A_i\geqslant(x)$ 是由 A_i 模糊集的语义和观测数据集的概率分布决定的。

对于模糊集合 $\xi\in E$, $\mu_\xi: X\to[0,1]$。根据文献 [17], $\{\mu_\xi(x)|\xi\in E\}$ 是 AFS 模糊逻辑系统 $(E,\vee,\wedge)$ 的一组相关隶属度函数, 则满足条件:

1) 对于 $\alpha,\beta\in E$, 如果 $\alpha\leqslant\beta$ 在系统 $(E,\vee,\wedge)$ 内, 并且对于任意 $x\in X$, $\mu_\alpha(x)\leqslant\mu_\beta(x)$ 都成立;

2) 对于 $x\in X$, $\eta=\sum\limits_{i\in I}\left(\prod\limits_{m\in A_i} m\right)\in E$, 如果对于任意 $i\in I$, $A_i\geqslant(x)=\varnothing$, 则 $\mu_\eta(x)=0$;

3) 对于 $x,y\in X$, $A_i\subseteq M$, $\eta=\prod\limits_{m\in A_i} m\in E$, 如果 $A_i\geqslant(x)\subseteq A_i\geqslant(x)$, 则 $\mu_\eta(x)\leqslant\mu_\eta(y)$; 如果 $A_i\geqslant(x)=X$, 则 $\mu_\eta(x)=1$。

确定相关性隶属函数首先要确定 X 上的测度。下面给出了实现该测度的具体内容。设 γ 是关于 X 的一个模糊项。文献 [17] 中 A_i 的权重函数被定义为:

设 $\rho_\gamma: X\to R^+=[0,\infty)$, 如果 ρ_γ 满足条件:

1) $\rho_\gamma(x)=0\Leftrightarrow x<_m x, x\in X$;

2) $\rho_\gamma(x)\geqslant\rho_\gamma(y)\Leftrightarrow x\geqslant_m y, x,y\in X$。

则称 ρ_γ 为模糊项 γ 的权重函数。

设 (Ω,f,P) 是一个概率测度空间, M 是 X 上的一组模糊集合, ρ_γ 是模糊项 $\gamma\in M$ 的权重函数, $X\subseteq\Omega$ 是概率空间 (Ω,f,P) 上的一个有限观测样本集。如果对于任意的 $m\in M$, $x\in\Omega$; $m\geqslant(x)\in f$, 则 $\{\mu_\xi(x)|\xi\in E\}$ 是 $(E,\vee,\wedge)$ 的一组相关隶属度函数, 条件是每个模糊集 $\xi=\sum\limits_{i\in I}\left(\prod\limits_{\gamma\in A_i}\gamma\right)\in E$ 的隶属函数被定义为

$$\mu_\xi(x)=\sup_{i\in I}\inf_{\gamma\in A_i}\frac{\sum\limits_{u\in(A_i\geqslant(x))}\rho_\gamma(u)}{\sum\limits_{u\in X}\rho_\gamma(u)},\quad\forall x\in X \tag{7}$$

$$\mu_\xi(x)=\sup_{i\in I}\inf_{\gamma\in A_i}\frac{\int_{A_i\geqslant(x)}\rho_\gamma(t)\,\mathrm{d}P_t}{\sum\limits_{\Omega}\rho_\gamma(t)\,\mathrm{d}P_t},\quad\forall x\in\Omega \tag{8}$$

如果对于每个 $\gamma\in M$, $\rho_\gamma(x)$ 在 Ω 上是连续的且 $|X|$ 是从概率空间 (Ω,f,P) 中随机抽取的一组样

本集，则式 (7) 所定义的隶属函数等于式 (8) 所定义的隶属函数。

根据以上可知，在 E 中模糊概念的隶属函数可以由 $\rho_\gamma(x)$ 和 $A_i\geqslant(x)$ 来确定，这既考虑了模糊性又考虑了随机性。根据文献 [17]，对于任意模糊概念 $\xi=\sum\limits_{i\in I}\left(\prod\limits_{\gamma\in A_i}m\right)\in E$，对于任意 $x\in X$，ξ 的隶属函数被定义为

$$\mu_\xi(x)=\sup_{i\in I}\left\{\frac{|A_i\geqslant(x)|}{|X|}\right\} \tag{9}$$

$$\mu_\xi(x)=\sup_{i\in I}\left\{\frac{\left|y\in X\middle|x\geqslant_{\rho_m}y,\forall m\in A_i\right|}{|X|}\right\} \tag{10}$$

式中 ρ_m 是一个权重函数。

接着设样本集为 $X=\{x_1,x_2,\cdots,x_n\}\subseteq\mathrm{R}$，$X$ 上的一个特征集合 $F=\{f_1,f_2,\cdots,f_l\}$，$M=\{m_{i,j}|1\leqslant i\leqslant l,1\leqslant j\leqslant k_i\}$ 表示一组相对于特征 f_i 的选择的 k_i 个模糊概念的集合，$m_{i,1},m_{i,2},\cdots,m_{i,k_i}$ 是其中与特征 f_i 相对应的一组简单模糊概念。在新的测度空间中，对于每一个样本 x，发现一个显著的模糊子集 $\zeta_x=\prod\limits_{m\in M}m$，以便 ζ_x 有效地表示 x，而不是整个模糊集。这里，如果 x_k 属于 $m_{i,j}$ 的隶属度大于某一阈值，则模糊隶属函数用作特征的度量，$m_{i,j}$ 足够好地将 x_k 与其他值区分开。在这里定义：

$$B_x^\varepsilon=\{m\in M|\mu_m(x)\geqslant\max\{\mu_m(x)\}-\varepsilon\} \tag{11}$$

式中：ε 表示误差阈值[20]，在这里设 $\varepsilon=0.3$，通过设定误差阈值避免了使用模糊隶属函数的不明显特征，使样本的表示方法能够更好地表达数据中的底层语义结构，排除了噪声对数据样本特征的干扰，若对误差阈值运用自适应策略，将明显增加算法的计算复杂度。考虑本文算法的总体性能表现，实验中参考文献 [20] 中的经验数值进行设定。B_x^ε 表示关于数据样本 x 的全部的模糊项的集合。然后将 ζ_x 描述为

$$\zeta_x=\mathop{\wedge}_{m\in B_x^\varepsilon}m \tag{12}$$

式中 $\wedge$ 是 AFS 代数中的模糊隶属关系运算。通过这样表示，所有理想的模糊项都结合在一起作为样本表示。然后，通过 AFS 代数和 AFS 理论中的模糊运算，结合模糊隶属度的关系，通过逻辑运算的数据距离推理，将模糊集表示的另一个描述样本的隶属度用作距离度量。根据 AFS 聚类[19-21]，对于两个数据样本 X_i 和 X_j，它们之间的距离定义为：

$$D_{ij}=1-\min\left\{\bar{\mu}_{\zeta_{X_i}}(X_i),\bar{\mu}_{\zeta_{X_j}}(X_j)\right\} \tag{13}$$

$$\bar{\mu}_{\zeta_{X_i}}(X_j)=\left\{m_k\in\zeta_{X_i}\middle|\frac{\sum\limits_{k=1}^{N}\mu_{m_k}(X_j)}{N}\right\} \tag{14}$$

式中：$\mu_{m_k}(X_j)$ 表示数据样本模糊项 m_k 的 X_j 的模糊隶属度；m_k 表示模糊项集合 ζ_{X_i} 中的每个模糊项；$\bar{\mu}_{\zeta_{X_i}}(X_j)$ 表示模糊项集合 ζ_{X_i} 的数据样本 X_j 的平均隶属度。

在 AFS 理论的基础上，为了更好地提取数据结构，在 AFS 空间中建立了距离测量，公式为

$$A(x_i,x_j)=\exp\left(-\frac{{D_{ij}}^2}{2\sigma^2}\right) \tag{15}$$

式中 D_{ij} 为式 (13) 所示基于公理化模糊集理论的距离定义。

2.2　所提算法

在 2.1 小节中虽然利用 AFS 理论在谱聚类算法中构建了新的距离度量，即 $A(x_i,x_j)=\exp\left(-{D_{ij}}^2/2\sigma^2\right)$，但是高斯核函数中 σ 是一个人工指定的参数，为每个数据集指定一个合适的参数 σ 是一件很复杂的事，需要花费大量的时间和精力。本文将数据点的领域信息加入相似度的计算中，并结合共享近邻的思想，在 AFS 理论距离测量的基础上定义了一个能够自适应得到尺度参数 σ 的高斯核函数——基于 AFS 理论的共享近邻自适应高斯核函数，其定义为

$$A(x_i,x_j)=\begin{cases}\exp\left(-\dfrac{{D_{ij}}^2}{\sigma_i\sigma_j(\mathrm{SNN}(x_i,x_j)+1)}\right), & i\neq j\\ 0, & i=j\end{cases} \tag{16}$$

式中 σ_i 的取值由数据点 x_i 的 K 个最近邻确定，通常为 $K=7$[22]。根据文献 [22] 可知，K 的选择与尺度参数无关，并且是嵌入空间数据维数的函数，通过 K 共享近邻的方法能够根据数据样本本身之间的关系自适应的获得更加适合的尺度参数，避免了选取尺度参数给算法带来的不确定性。其计算方法为

$$\sigma_i=\frac{1}{K}\sum_{m=1}^{K}d(x_i,x_m) \tag{17}$$

式中：σ_i 表示样本点 x_i 和其 K 个最近邻距离的平均值；同理 σ_j 表示样本点 x_j 和其 K 个最近邻距离的平均值。$\mathrm{SNN}(x_i,x_j)$ 表示样本点 x_i 和 x_j 在 K 邻域内所共有的邻居个数。$\mathrm{SNN}(x_i,x_j)$ 反映了 x_i 和 x_j 点的局部密度，可以用来提高数据点间的相似性，有助于样本点正确的划分。

2.3　所提算法流程

算法　公理化模糊共享近邻自适应谱聚类算

法 (AFSSNNSC)

输入　数据集 $X=\{x_1,x_2,\cdots,x_n\}$, 目标聚类数 k;

输出　聚类实际产生 y 类。

1) 首先为数据点的每个特征 f_i 构造模糊关系模糊项 $m_{i,j}$;

2) 使用式 (10) 计算每个数据点的隶属度函数 $\mu_{m\in M}(x)$;

3) 通过式 (11) 找出模糊项的集合 B_x^{ε};

4) 通过 $\zeta_x=\underset{m\in B_x^{\varepsilon}}{\wedge} m$ 构建每个样本的描述 ζ_x;

5) 根据 ζ_x 利用式 (13) 计算成对距离 D_{ij};

6) 通过式 (17) 计算 σ_i 和 σ_j;

7) 利用式 (16) 构建亲和矩阵 $\boldsymbol{A}$;

8) 根据 $\boldsymbol{D}_{ii}=\sum\limits_{1\leqslant j\leqslant n}\boldsymbol{S}_{ij}$ 计算对角矩阵 $\boldsymbol{D}$;

9) 计算归一化拉普拉斯矩阵, 拉普拉斯矩阵 $\boldsymbol{L}$ 由 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$ 被归一化为

$$\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$$

10) 对 $\boldsymbol{L}$ 进行特征分解, 得到前 k 个特征向量, 并构建特征向量空间;

11) 利用经典的聚类算法如 k-means 对特征向量空间中的特征向量进行聚类。

3　实验与结果分析

3.1　实验环境及性能指标

在 UCI、USPS 手写数字的相同数据集上, 采用本文提出的方法和文献 [10] 的 NJW 谱聚类 (SC)、文献 [21] 的 AFS 聚类算法 (AFS)、文献 [22] 的 self-tuning spectral clustering(STSC) 算法、基于 K 均值的近似谱聚类 (KASP)[23]、基于 Nystrom 近似谱聚类 (Nystrom)[24] 和基于地标点谱聚类算法 (LSC-R, LSC-K)[25] 进行对比实验。本文算法实验是在 MATLAB 2014b, 计算机的硬件配置为 Intel i7-4770 CPU 3.40 GHz、16 GB RAM 的平台下进行。为了评估所提算法的聚类性能, 本文使用聚类误差 (CE)[26] 和归一化互信息 (NMI)[27]2 种性能指标对本文算法与其他聚类方法的聚类结果进行了比较。其中 CE 被广泛用于评价聚类性能, CE 越低聚类性能越好。NMI 也是一种广泛使用的评估算法聚类性能的测量标准, NMI 越大性能越好。

3.2　UCI 数据集实验结果与分析

为了验证所提出算法的有效性, 本文将所提的算法和其他方法应用于 UCI 数据库中的 11 个基准数据集作为测试样本, 表 1 为这 11 类数据集的特征, 分别是数据总量、维数以及类的个数。

表 1　UCI 实验数据集的特性

Table 1　Characteristics of the UCI experimental datasets

数据集	数据总量	维数	类数
Heart	270	13	2
Hepatitis	155	19	2
Sonar	208	60	2
Wobc	699	9	2
Wdbc	569	30	2
Iris	150	4	3
Wine	178	13	3
Protein	552	77	8
Libras	360	90	15
LetterRec	20 000	16	26
Covtype	581 012	54	7

基于聚类误差 (CE) 的实验结果如表 2 所示, 由表 2 可知所提的方法在大部分实验数据集上均获得了优于对比算法的 CE 值; 所提出的方法在 Heart、Hepatitis、Wdbc、Protein 和 Libras 数据集上的 CE 分别为 15.27%、27.14%、6.03%、51.12%、52.31%, 相比较 AFS 算法而言均改进了 10% 以上, 在其他 5 个数据集上的 CE 相比较 AFS 也均有所提高。证明了利用谱理论对相似矩阵进行划分比之前提出的传递闭包理论好得多, 考虑到传递闭包方法的验证循环, 所提出的方法也相对更快、更容易实现。在 Iris、Wine 数据集中, 所提算法的 CE 分别为 7.42% 和 2.89%, 相对聚类错误率略高于 STSC 算法。因为这两个数据集中只有 150 个样本和 178 个样本, 但是差异实际上只有一两个样本, 但相对于总体而言, 所提算法 CE 普遍低于其他算法在各数据集上的结果, 仍具有较好的优越性; 与基于距离度量的方法相比所提算法在给出的所有数据集中都显示出了优越性, 在 Sonar 数据集上更加改进 5% 以上, 本文算法与基于 Nystrom 近似谱聚类方法相比在所有数据集上均有 1% 以上的优势。本文算法与基于地标点的谱聚类方法 LSC-R 和 LSC-K 相比也展现出较好的聚类性能。这是因为通过模糊隶属函数代替距离度量数据之间的相似性, 利用模糊语义结构解释数据之间的复杂的相互关系, 增强了算法的鲁棒性。对于 Protein、Libras 等多聚类数据集, AFS 的聚类错误率偏高, 因为 AFS 聚类需要根据每个集群的边界选择最好的数据聚类分区。随着集群规模数量的不断增加, 将很难去清晰地找到边界, 这样聚类误差也会随之增高。总体而言, 与对比文献方法相比, 所提算法的 CE 值在所有实验数据集上均得到了改善, 降低了算法的聚类错误率。

表 2　UCI 数据集上的 CE 比较

Table 2　Comparison of CE on the UCI datasets　%

数据集	KASP	SC	STSC	AFS	Nystrom	LSC-R	LSC-K	本文算法
Heart	19.47	19.63	21.14	29.62	21.64	17.18	16.25	**15.27**
Hepatitis	37.48	29.72	38.77	44.18	36.74	31.36	29.43	**27.14**
Sonar	40.25	42.38	42.85	37.24	40.57	39.26	38.27	**35.61**
Wobc	3.31	3.45	3.37	2.78	3.24	3.12	2.84	**2.72**
Wdbc	8.58	9.54	7.27	18.13	8.32	6.85	6.53	**6.03**
Iris	9.57	10.05	**7.31**	9.66	9.34	8.78	8.62	7.42
Wine	3.62	3.46	**2.82**	3.47	3.74	3.36	3.15	2.89
Protein	56.53	53.74	56.28	64.65	56.35	54.37	53.58	**51.12**
Libras	57.56	55.46	53.48	62.76	58.82	56.27	55.63	**52.31**
LetterRec	45.42	47.83	46.56	48.35	45.27	43.85	43.26	**41.37**
Covertype	53.81	54.76	54.36	52.25	53.68	52.85	52.24	**51.15**

在归一化互信息 (NMI) 中测量的聚类性能如表 3 所示。所提出的算法的聚类结果 NMI 与其他方法的 NMI 相比都得到了改善，尤其在 Heart 和 Protein 数据集上，所提算法相对于 KASP、SC、STSC、AFS 和 Nystrom 对比算法而言 NMI 均提高了 5% 以上。只是在 Wine 数据集上，所提算法的 NMI 为 87.86%，与其他算法相当，但在整个对比表格中为最好的聚类性能。由于所选 Covertype 数据集是一个从地图变量预测森林覆盖类型的数据集，它们都主要是在荒野地区发现的，所以覆盖类型在实际地理上是非常接近的，相对于其他数据集而言，这个数据集数据特性更加复杂。所以在 Covertype 数据集下所有算法的 NMI 都普遍较低，但是所提算法获得了比其他算法更好的聚类效果。

从实验结果可以看出，STSC 不是很稳定，它在 Hepatitis 和 Sonar 数据集上的 NMI 情况都非常差，由于在 STSC 和本文算法中都考虑到了数据之间的相互关系，利用到了数据邻居的近邻作用，所以可以从中得出结论，与考虑到数据样本关系之间的传统距离度量作为相似性度量相比，采用具有数据样本模糊关系的模糊隶属度作为距离度量，在相似性度量上更具有鲁棒性。总体而言，所提算法相较于对比算法都具有明显的改善。

表 3　UCI 数据集上的 NMI 比较

Table 3　Comparison of NMI on the UCI datasets　%

数据集	KASP	SC	STSC	AFS	Nystrom	LSC-R	LSC-K	本文算法
Heart	32.61	28.51	25.83	17.62	30.37	35.45	37.11	**38.18**
Hepatitis	14.84	14.55	4.84	3.23	14.68	15.16	15.33	**15.86**
Sonar	14.53	7.56	1.67	17.22	12.15	16.48	17.23	**19.09**
Wobc	78.57	77.12	80.04	73.86	78.34	80.49	81.13	**81.76**
Wdbc	65.69	63.33	61.45	60.31	64.57	67.14	67.62	**68.97**
Iris	79.73	77.85	79.21	78.57	78.67	80.24	80.75	**81.46**
Wine	87.59	87.32	86.93	85.56	87.46	87.61	87.69	**87.86**
Protein	56.17	54.42	46.25	35.63	56.83	59.43	60.85	**62.18**
Libras	65.48	63.72	64.91	38.14	64.63	66.37	66.88	**68.07**
LetterRec	40.16	35.19	37.67	34.53	39.12	37.34	39.63	**41.27**
Covertype	7.44	6.87	7.19	6.54	7.42	8.31	9.02	**9.83**

3.3　USPS 数据集实验结果与分析

选择两个典型谱聚类算法 SC 和 STSC 与所提方法在广泛使用的 USPS 数据库中的手写数字数据集进行对比实验。该数据集包含美国邮政总局通过扫描信封中的手写数字获得的数字数据。原始扫描的数字是二进制的，大小和方向不同。本文使用的图像经过了大小归一化，得到了 1 616 张 256 维的灰度图像。它包含 7 291 个训练

实例和 2 007 个测试实例 (总共 9 298 个)。为了展示该方法的可伸缩性, 考虑了不同数量的集群。具体来说, 数字子集{0,8}、{4,9}、{0,5,8}、{3,5,8}、{1,2,3,4}、{0,2,4,6,7}和整个集合{0,1,⋯,9}用于测试本文提出的算法。这些子集的详细信息如表 4 所示。分别在每个子集上进行实验, 并使用 CE 和 NMI 来测量性能。

表 4　USPS 实验数据集的特性

Table 4　Characteristics of the USPS experimental datasets

数据集	数据总数	类数	维数
{0,8}	2 261	2	256
{4,9}	1 673	2	256
{0,5,8}	2 977	3	256
{3,5,8}	2 248	3	256
{1,2,3,4}	3 837	4	256
{0,2,4,6,7}	4 960	5	256
{0,1,⋯,9}	9 298	10	256

从图 1 可以看出, 在 CE 方面, 所提算法在所有的情况下都优于 STSC 和 SC, 尤其在{0,8}、{0,5,8}、{3,5,8}、{0,2,4, 6,7}、{0,1,⋯,9}数据集上 CE 均改善了 5% 以上, 甚至在{3,5,8}上 CE 相较于其他对比算法, 所提算法改进了 10% 以上。总体而言与 SC 和 STSC 相比, 可以从图 1 中看出所提出的方法均得到明显的改善。

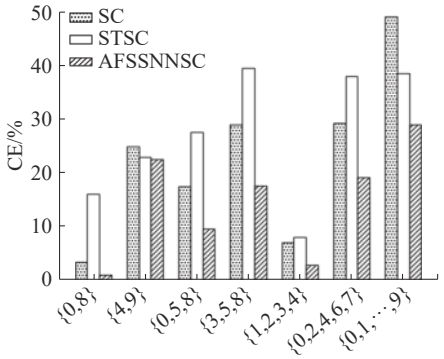

图 1　USPS 数据集上 CE 的性能比较

Fig. 1　Performance comparison of CE on the USPS datasets

图 2 显示了基于 NMI 的 USPS 数据集的结果。从图 2 中可以看出, 所提出的方法在所有情况下都比 SC 和 STSC 有优势。在{0,8}、{1,2,3,4}、{0,1,⋯,9}上相较于其他对比算法, 所提算法的 NMI 都提高了 10% 以上, 特别是对于具有挑战性的情况{3,5,8}和多类情况{1,2,3,4}、{0,2,4,6,7}、{0,1,⋯,9}, 所提出的算法都具有一定的优越性。

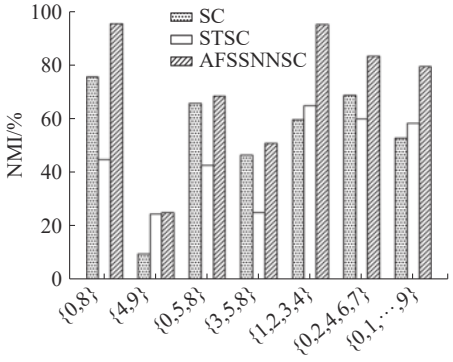

图 2　USPS 数据集上 NMI 的性能比较

Fig. 2　Performance comparison of NMI on the USPS datasets

4　结束语

本文提出了一种新的无监督广义数据亲和图的构造方法, 该方法具有更强的鲁棒性和更有意义的数据亲和图, 以提高谱聚类精度。采用模糊理论定义数据相似度, 利用模糊隶属度函数导出亲和图。此外, 亲和图不是盲目地相信所有可用变量, 而是通过捕捉和通过对每个样本的模糊描述, 确定了特征子空间中组合分布的微妙两两相似关系。同时采用共享近邻的方法发现密集区域样本点分布的结构和密度信息, 并且根据每个点所处领域的稠密程度自动调节参数 σ, 从而生成更强大的亲和矩阵, 进一步提高聚类准确率, 证明了该方法对不同类型数据集的有效性。实验结果表明, 该方法与其他先进的方法相比具有一定的优越性。数据大小的多样性在一定程度上体现了该方法对于大数据集的可扩展性。在未来将通过系统地将所提出的算法与一些采样或量化策略相结合来处理一般的可伸缩性问题。

参考文献:

[1] XU Dongkuan, TIAN Yingjie. A comprehensive survey of clustering algorithms[J]. Annals of data science, 2015, 2(2): 165–193.

[2] LIU Hanqiang, ZHAO Feng, JIAO Licheng. Fuzzy spectral clustering with robust spatial information for image segmentation[J]. Applied soft computing, 2012, 12(11): 3636–3647.

[3] TUNG F, WONG A, CLAUSI D A. Enabling scalable spectral clustering for image segmentation[J]. Pattern recognition, 2010, 43(12): 4069–4076.

[4] ZENG Shan, HUANG Rui, KANG Zhen, et al. Image segmentation using spectral clustering of Gaussian mixture models[J]. Neurocomputing, 2014, 144: 346–356.

[5] JIANG J Q, DRESS A W M, YANG Genke. A spectral clustering-based framework for detecting community struc-

tures in complex networks[J]. Applied mathematics letters, 2009, 22(9): 1479–1482.

[6] FORESTIER G, WEMMERT C. Semi-supervised learning using multiple clusterings with limited labeled data[J]. Information sciences, 2016, 361−362: 48–65.

[7] 赵晓晓, 周治平. 结合稀疏表示与约束传递的半监督谱聚类算法 [J]. 智能系统学报, 2018, 13(5): 855–863.
ZHAO Xiaoxiao, ZHOU Zhiping. A semi-supervised spectral clustering algorithm combined with sparse representation and constraint propagation[J]. CAAI transactions on intelligent systems, 2018, 13(5): 855–863.

[8] 林大华, 杨利锋, 邓振云, 等. 稀疏样本自表达子空间聚类算法 [J]. 智能系统学报, 2016, 11(5): 696–702.
LIN Dahua, YANG Lifeng, DENG Zhenyun, et al. Sparse sample self-representation for subspace clustering[J]. CAAI transactions on intelligent systems, 2016, 11(5): 696–702.

[9] CHANG Yanshuo, NIE Feiping, LI Zhihui, et al. Refined spectral clustering via embedded label propagation[J]. Neural computation, 2017, 29(12): 3381–3396.

[10] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm[C]//Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Vancouver, Canada, 2001: 849−856.

[11] YE Xiucai, SAKURAI T. Robust similarity measure for spectral clustering based on shared neighbors[J]. ETRI journal, 2016, 38(3): 540–550.

[12] JIA Hongjie, DING Shifei, DU Mingjing. Self-tuning p -spectral clustering based on shared nearest neighbors[J]. Cognitive computation, 2015, 7(5): 622–632.

[13] 王雅琳, 陈斌, 王晓丽, 等. 基于密度调整的改进自适应谱聚类算法 [J]. 控制与决策, 2014, 29(9): 1683–1687.
WANG Yalin, CHEN Bin, WANG Xiaoli, et al. Improved adaptive spectral clustering algorithm based on density adjustment[J]. Control and decision, 2014, 29(9): 1683–1687.

[14] SHI Jianbo, MALIK J. Normalized cuts and image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2000, 22(8): 888–905.

[15] LIU Xiaodong. The fuzzy theory based on AFS algebras and AFS structure[J]. Journal of mathematical analysis and applications, 1998, 217(2): 459–478.

[16] LIU Xiaodong, PEDRYCZ W, ZHANG Qingling. Axiomatics fuzzy sets logic[C]//Proceedings of the12th IEEE International Conference on Fuzzy Systems. St Louis, USA, 2003: 55-60.

[17] LIU Xiaodong, PEDRYCZ W. Axiomatic fuzzy set theory and its applications[M]. Berlin, Heidelberg: Springer, 2009.

[18] LIU Xiaodong, PEDRYCZ W, CHAI Tianyou, et al. The development of fuzzy rough sets with the use of structures and algebras of axiomatic fuzzy sets[J]. IEEE transactions on knowledge and data engineering, 2009, 21(3): 443–462.

[19] LIU Xiaodong, 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.

[20] LIU Xiaodong, WANG Xianchang, PEDRYCZ W. Fuzzy clustering with semantic interpretation[J]. Applied soft computing, 2015, 26: 21–30.

[21] LIU Xiaodong, WANG Wei, CHAI T. The fuzzy clustering analysis based on AFS theory[J]. IEEE transactions on systems, man, and cybernetics, part B, 2005, 35(5): 1013–1027.

[22] ZELNIK-Manor L, PERONA P. Self-tuning spectral clustering[C]//Proceedings of the 17th International Conference on Neural Information Processing Systems. Pasadena, USA, 2004: 1601−1608.

[23] YAN Donghui, HUANG Ling, JORDAN M I. Fast approximate spectral clustering[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 907−916.

[24] LI Mu, KWOK J T, LU Baoliang. Making large-scale nyström approximation possible[C]//Proceedings of the 27th International Conference on International Conference on Machine Learning. Haifa, Israel, 2010: 631−638.

[25] CAI Deng, CHEN Xinlei. Large scale spectral clustering via landmark-based sparse representation[J]. IEEE transactions on cybernetics, 2015, 45(8): 1669–1680.

[26] SCHÖLKOPF B, PLATT J, HOFMANN T. A local learning approach for clustering[C]//Proceedings of the 19th International Conference on Neural Information Processing Systems. Doha, Qatar, 2007: 1529−1536.

[27] STREHL A, GHOSH J. Cluster ensembles: a knowledge reuse framework for combining partitionings[C]//Proceedings of the 18th National Conference on Artificial Intelligence. Alberta, Canada, 2003: 583–617.

作者简介：

储德润，男，1994 年生，硕士研究生，主要研究方向为数据挖掘。

周治平，男，1962 年生，教授，博士，主要研究方向为智能检测、网络安全。发表学术论文 20 余篇。