基于低密度分割密度敏感距离的谱聚类算法

陶新民 [1] 王若彤 [1] 常 瑞 [1] 李晨曦 [1] 刘艳超 [1]

摘 要 本文提出一种基于低密度分割密度敏感距离的谱聚类算法, 该算法首先使用低密度分割密度敏感距离计算相似度矩阵, 该距离测度通过指数函数和伸缩因子实现放大不同流形体数据间的距离和缩短同一流形体数据间距离的目的, 从而有效反映数据分布的全局一致性和局部一致性特征. 另外, 算法通过增加相对密度敏感项来考虑数据的局部分布特征, 从而有效避免孤立噪声和 "桥" 噪声的影响. 文中最后给出了基于 SC (Scattering criteria) 指标的 k 近邻图 k 值选取办法和基于谱熵贡献率的特征向量选取方法. 实验部分, 讨论了参数选择对算法性能的影响并给出取值建议, 通过与其他流行谱聚类算法聚类结果的对比分析, 表明本文提出的基于低密度分割密度敏感距离的谱聚类算法聚类性能明显优于其他算法.

关键词 谱聚类, 低密度分割, 欧氏距离, 密度敏感, 鲁棒性

引用格式 陶新民, 王若彤, 常瑞, 李晨曦, 刘艳超. 基于低密度分割密度敏感距离的谱聚类算法. 自动化学报, 2020, **46**(7): 1479−1495

DOI 10.16383/j.aas.c180084

Low Density Separation Density Sensitive Distance-based Spectral Clustering Algorithm

TAO Xin-Min[1] WANG Ruo-Tong[1] CHANG Rui[1] LI Chen-Xi[1] LIU Yan-Chao[1]

Abstract This paper proposes a low density separation density sensitive distance-based spectral clustering algorithm. First, the algorithm applies the low-density separation density sensitive distance to calculate the similarity matrix. By the exponential function and flexibility factor, we can achieve increasing the distance between different manifold data and decreasing the distance between the same manifold data, which can effectively reflect the global consistency and local consistency of data distribution. In addition, by adding relative density sensitive term to take into account the local distribution characteristics of the data, isolated noise and "bridge" noise are effectively avoided. Finally, we provide the method of selecting k-value of k nearest neighbor graph based on SC (Scattering criteria) index and the method of extracting eigenvector based on spectral entropy contribution rate. In the experimental part, the effect of parameter selection on the performance of the proposed technique is discussed and some suggestions about the determination of the parameters are given. Compared with the state-of-the-art spectral clustering algorithms, the analysis results demonstrate that the proposed low density separation density sensitive distance-based spectral clustering algorithm performs well on artificial and UCI benchmark datasets.

Key words Spectral clustering, low density separation, Euclidean distance, density sensitive, robustness

Citation Tao Xin-Min, Wang Ruo-Tong, Chang Rui, Li Chen-Xi, Liu Yan-Chao. Low density separation density sensitive distance-based spectral clustering algorithm. *Acta Automatica Sinica*, 2020, **46**(7): 1479−1495

谱聚类 (Spectral clustering) 算法是一种基于谱图划分理论的子空间聚类算法[1], 与传统聚类算法相比, 因其对聚类样本空间的形状和维度没有特殊要求, 且具有收敛于全局最优解的优点, 现已被广泛应用在图像分割[2−4]、并行计算[5]、数据分类[6] 等方面. 其中, NJW (Ng-Jordan-Weiss) 算法[7] 是谱聚类算法中最核心、最常用的方法. 它通过数据的规范化拉普拉斯矩阵的特征向量作为样本集进行聚类, 相似度矩阵的确定是影响该算法性能的重要因素, 而相似度矩阵又取决于距离测度的选择[8]. 在被广泛应用的谱聚类算法中, 多采用欧氏距离和曼哈顿距离作为相似性的度量[9−10]. 除此以外, 余弦距离、皮尔森相似距离也常常被应用在谱聚类中[11]. 然而这些距离测度方法都没有考虑数据的分布特征, 如数据呈现非线性或局域流形特征时, 这些方法所表现的聚类性能并不理想[12].

收稿日期 2018-02-05 录用日期 2018-05-30
Manuscript received February 5, 2018; accepted May 30, 2018
国家自然科学基金 (31570547), 中央高校基本科研业务费专项资金 (2572017EB02, 2572017CB07), 东北林业大学双一流科研启动基金 (411112438) 资助
Supported by National Natural Science Foundation of China (31570547), Fundamental Research Funds for the Central Universities (2572017EB02, 2572017CB07), Scientific Research for Double First-class of Northeast Forestry University (411112438)
本文责任编委 曾志刚
Recommended by Associate Editor ZENG Zhi-Gang
1. 东北林业大学工程技术学院 哈尔滨 150040
1. College of Engineering & Technology, Northeast Forestry University, Harbin 150040

考虑到数据的分布特征影响, 文献 [13] 利用连接距离来测量样本点间的相似程度. 所谓连接距离, 是指从数据集全局出发, 通过寻找连接任意两点间所有路径中最大间隔距离中的最小值作为两个样本间的距离测度. 通过该距离测度能消除只考虑样本点间距离的大小而不考虑数据全局分布特征的弊端, 从而反映数据集的空间分布一致性. 但是, 该距离测度无法防止 "桥" 噪声数据的影响, 当 "桥" 噪声产生时, 原本的流形体会在该距离测度下合并, 导致无法实现正确聚类[14−16]. 为消除样本局部分布特征对聚类性能的影响, 文献 [17] 提出了一种密度敏感的谱聚类算法, 该算法使用基于密度敏感的相似度矩阵计算方法, 通过密度项放大正常样本点与孤立噪声点的距离, 同时缩小正常样本点间距离的方式消除孤立噪声的影响; 然而, 该距离测度没有考虑数据的非线性和流形分布特征对算法性能的影响. 文献 [18] 提出了一种典型的密度峰值聚类算法, 该算法通过考虑数据间的密度关系进行聚类, 满足了数据间的全部一致性. 然而该距离测度没能充分考虑数据的局部一致性特征且易受 "桥" 噪声数据的影响. 据此, Yu 等[19] 提出利用基于密度的最短几何距离谱聚类算法, 虽然算法使用基于密度的最短几何距离测度可以反映数据集分布特征的影响, 但是这种距离测度只考虑了数据集的局部空间一致性, 没能考虑全局空间一致性. 且算法的性能受 k 值的影响很大, 极端情况下当 k 取值较大时, 该距离测度等价于欧氏距离, 导致聚类效果并不理想[20−21].

为解决上述问题, 本文在综合考虑全局空间一致性和局部空间一致性的基础上, 提出一种基于低密度分割密度敏感距离的谱聚类算法. 考虑到位于同一流形体上两点间会有许多较短的边连接, 而位于不同流形体上的两点需要较长边相连, 算法使用低密度分割密度敏感距离通过指数函数和伸缩因子实现放大位于不同流形上的数据点间距离, 缩短位于同一流形上的数据点间距离的目的, 从而有效反映了数据分布的全局一致性和局部一致性特征. 为进一步防止孤立噪声以及 "桥" 噪声的影响, 该距离测度还充分考虑了数据的局部分布特征, 通过增加相对密度敏感项实现缩小正常样本点间距离、增大正常样本点与噪声样本点间距离的目的. 此外, 由于 k 最近邻参数的确定以及特征向量的选择对谱聚类算法聚类性能的影响很大, 文中给出了基于 SC (Scattering criteria) 指标的 k 近邻图 k 值选取方法和基于谱熵贡献率的特征向量选取方法. 通过实验对比可知, 本文所提出的基于低密度分割密度敏感距离的谱聚类算法在人工数据集和 UCI 数据集上均得到了较好的聚类结果, 表明了该方法的合理性与有效性.

1 NJW 谱聚类算法

谱聚类的思想来源于谱图划分, 将数据聚类问题转化为一个无向图的多路划分问题[22]. 该算法首先将数据点定义为一个无向图 $G=(P,V)$ 的顶点, 然后利用欧氏距离测度建立图中顶点之间的相似度矩阵, 并求得规范化拉普拉斯矩阵进行降维, 最后规定优化准则进行样本的聚类.

NJW 谱聚类算法的一般步骤具体如下:

首先, 定义一个无向图 $G=(P,V)$, 其中 P 为顶点集, $P=\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_n\}$, V 为顶点之间边的集合, n 为样本个数. 构建相似度矩阵 S, 则矩阵 S 的元素为

$$S_{ij}=\begin{cases}\exp\left(-\frac{d^2(\boldsymbol{p}_i,\boldsymbol{p}_j)}{2\sigma^2}\right), & i\neq j\\ 0, & i=j\end{cases} \tag{1}$$

其中, $\boldsymbol{p}_i$ 是数据样本点, $d^2(\boldsymbol{p}_i,\boldsymbol{p}_j)$ 代表样本点 $\boldsymbol{p}_i$ 和 $\boldsymbol{p}_j$ 间的欧氏距离, σ 是决定样本点之间相似程度的尺度参数.

然后构造规范化拉普拉斯矩阵 L_{sym}. 首先构造度量矩阵 B,

$$B_{ii}=\sum_{j=1}^{n}S_{ij},\quad i,j=1,2,\cdots,n \tag{2}$$

$$B=\mathrm{diag}\{B_{11},B_{22},\cdots,B_{nn}\} \tag{3}$$

对应图 G 的规范化拉普拉斯矩阵定义为

$$L_{sym}=B^{-\frac{1}{2}}SB^{-\frac{1}{2}} \tag{4}$$

最后计算 L_{sym} 矩阵的前 C 个最大特征值对应的特征向量 $\boldsymbol{l}_1,\boldsymbol{l}_2,\cdots,\boldsymbol{l}_C$, 此 C 值是需事先设定的聚类个数, 并据此构造待聚类的数据矩阵 $X=[\boldsymbol{l}_1,\boldsymbol{l}_2,\cdots,\boldsymbol{l}_C]\in\mathbf{R}^{n\times C}$, 将矩阵的行向量转变为单位向量, 得到矩阵 Y,

$$Y_{ij}=\frac{X_{ij}}{\sqrt{\sum\limits_{j}{X_{ij}}^2}},\quad i=1,\cdots,n,j=1,\cdots,C \tag{5}$$

将矩阵 Y 的每一行看作是空间中的一个点, 对其使用 K 均值算法或任意其他经典聚类算法, 得到 C 个聚类.

传统谱聚类算法对数据类型没有局限性, 可以处理不同形状和维度的数据聚类问题, 并且收敛于全局最优, 不易陷入局部最优解[21]. 然而, 当传统的谱聚类算法面对复杂的实际应用问题时, 尤其是数据呈现非线性或局域流形特征时聚类性能并不理想.

2　基于低密度分割密度敏感距离的谱聚类算法

2.1　距离选择的必要性

由上述 NJW 谱聚类算法的聚类过程可知, 相似度矩阵的确定直接影响了聚类效果, 而相似度的确定取决于距离测度的选择. 众所周知, 基于距离测度的数据间相似性测量需满足以下两个一致性关系: 1) 局部一致性, 即空间位置上相邻的数据点具有较高的相似性; 2) 全局一致性, 即位于同一流形上的数据点具有较高的相似性.

然而, 传统的谱聚类算法利用的是欧氏距离测度确定数据间的相似度. 欧氏距离由于仅考虑了数据间的局部一致性特征, 没有考虑数据间的全局一致性特征, 导致当数据呈现非线性或局域流形特征时, 无法得到合理的相似度矩阵, 进而严重影响聚类性能.

为了进一步说明该问题, 本文利用 Two-moons 人工数据集进行示例分析, 图 1 为全局一致性距离示意图. 由图 1 我们可以发现, 图中的数据点明显地呈现两个流形体分布, 根据相似性测量的全局一致性特征要求位于同一流形内的样本点应具有较高的相似性, 即 AB 的相似度应高于 AC 的相似度. 然而, 当使用欧氏距离计算数据 AB、AC 之间的距离时, AB 的相似度低于 AC 的相似度, 无法满足聚类样本的全局空间分布一致性特征. 另外, 同一流形内的样本趋于分布在一个较高的密度区域内, 而不同流形之间会存在一个较低密度的区域. 由此, 需要寻找一个能较好地体现出空间分布特征的距离测度, 使其满足数据间全局一致性特征.

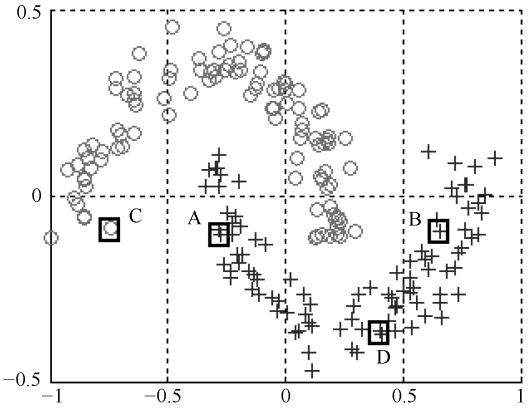


图 1　全局一致性距离示意图

Fig. 1　Global consistency distance diagram

当使用连接距离时, 可以反映这种空间全局一致性分布特征, 即使用连接任意两点 $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 所有路径中的最大距离中的最小值, 则连接距离的表达式如下:

$$D_{i,j} = \min_{\boldsymbol{p}\in P_{ij}} \max_{k<|\boldsymbol{p}|} d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1}) \tag{6}$$

其中, P_{ij} 为连接数据点 $\boldsymbol{p}_i$, $\boldsymbol{p}_j$ 的所有路径的集合, $\boldsymbol{p}$ 表示集合中的任意一条路径, $|\boldsymbol{p}|$ 表示该路径的长度, $\boldsymbol{p}_k, \boldsymbol{p}_{k+1}$ 为该路径上任意连接的两点, $d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})$ 表示两点间的欧氏距离. 通过该距离测度的计算, 能够使样本数据在同一个流形结构中的相似度高, 而在不同流形结构中的相似度低, 满足了空间分布的全局一致性要求.

然而上述的距离测度无法满足数据间局部一致性要求, 即在空间位置上相邻样本点的相似度高. 观察图 2 我们可以发现, 利用上述距离测度图中样本数据 AB 和 AD 距离的计算结果相等, 而局部一致性要求 AD 的相似度应高于 AB 的相似度, 这就违背了空间局部一致性特征. 为此在连接距离测度的基础上, 还需要考虑连接点的个数, 即长度信息的影响. 另外, 该距离测度无法防止 “桥” 数据噪声的影响. AC 数据点之间出现如图 2 所示的 “桥” 数据噪声时, 受连接距离影响导致两个流形体合并, 算法无法得出准确的聚类结果.

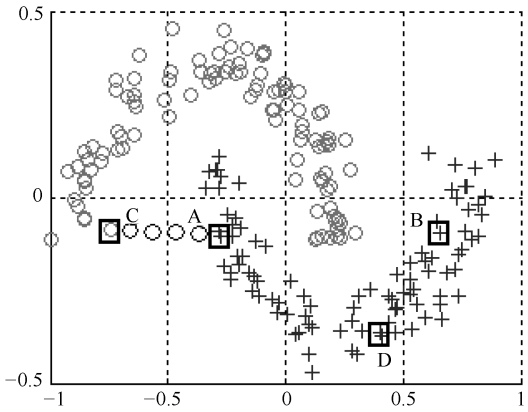


图 2　局部空间一致性距离示意图

Fig. 2　Local spatial coherence distance diagram

当选择最短几何距离测度时, 其表达式如下:

$$D_{i,j} = \min_{\boldsymbol{p}\in P_{ij}} \sum_{k=1}^{|\boldsymbol{p}|} d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1}) \tag{7}$$

由上述公式我们可以发现, 传统的最短几何距离测度, 由于考虑了路径长度的影响, 因此能够满足空间局部一致性特征. 但是当无向图 G 是全连接图时, 根据三角不等式可知, 两点的最短几何距离就等于两点的欧氏距离, 由全局一致性可知, AD 相似度应高于 AC 相似度, 但使用最短几何距离测度方法时得到的结果是 AD 相似度低于 AC 相似度, 与全局一致性特征不符.

2.2 低密度分割密度敏感距离

为提高谱聚类算法的聚类性能, 本文在原有谱聚类算法的基础上, 综合考虑聚类数据间全局一致性和局部一致性特征, 提出一种基于低密度分割密度敏感距离的谱聚类算法 (Low density separation density sensitive distance-based spectral clustering algorithm, LDSDSD-SC), 通过引入低密度分割密度敏感距离测度, 实现提升算法聚类性能的目的. 低密度分割密度敏感距离定义如下:

将样本点定义为图 $G=(P,V)$ 的顶点 P, 令 $\boldsymbol{p}\in P^l$ 表示为图上连接点 $\boldsymbol{p}_1$ 与 $\boldsymbol{p}_{|\boldsymbol{p}|}$ 的一条长度为 $l=|\boldsymbol{p}|-1$ 的路径, 其中边 $(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})\in V$. 令 P_{ij} 表示连接数据点 $\boldsymbol{p}_i$, $\boldsymbol{p}_j$ 的所有路径的集合, 其中 $\boldsymbol{p}_i,\boldsymbol{p}_j\in P$. 则 $\boldsymbol{p}_i$ 与 $\boldsymbol{p}_j$ 之间的流形距离按下式计算:

$$D_{i,j}^{\rho}=\frac{1}{\rho}\ln\left(1+\min_{\boldsymbol{p}\in P_{ij}}\sum_{k=1}^{|\boldsymbol{p}|}\left(\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{(\delta_k\delta_{k+1})}}-1\right)\right)\tag{8}$$

其中, $\min_{\boldsymbol{p}\in P_{ij}}\sum_{k=1}^{|\boldsymbol{p}|}\left(\mathrm{e}^{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})/(\delta_k\delta_{k+1})}-1\right)$ 表示的是图 G 上顶点 $\boldsymbol{p}_i$ 和 $\boldsymbol{p}_j$ 之间最短几何密度敏感距离. $d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})$ 是图 G 上顶点 $\boldsymbol{p}_i$ 到 $\boldsymbol{p}_j$ 最短路径上任意相邻两点 $\boldsymbol{p}_k$, $\boldsymbol{p}_{k+1}$ 的欧氏距离, δ_k, δ_{k+1} 为顶点 $\boldsymbol{p}_k$, $\boldsymbol{p}_{k+1}$ 的相对密度, ρ 为伸缩因子. 这里采用 $\left(\mathrm{e}^{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})/(\delta_k\delta_{k+1})}-1\right)$ 而非传统的欧氏距离, 是考虑到位于同一流形上的两点之间会有许多较短的边相连, 而位于不同流形上的两点需要较长边相连, 因此通过指数函数和伸缩因子 ρ 的值来实现放大位于不同流形体上的数据点间距离和缩短位于同一流形上的数据点间距离的目的, 从而使算法能够很好地反映数据间的全局一致性特征. 由上述公式可知, 该距离测度采用的是最短几何距离表达式, 即考虑了路径长度的影响, 因此也同时满足数据间的局部一致性特征. 另外, 该距离测度还充分考虑到了数据的局部分布特征, 通过增加相对密度敏感项, 实现缩小密集顶点间距离、增大稀疏顶点间距离的目的, 有效防止孤立噪声及 "桥" 噪声对算法聚类性能的影响.

对于数据点 $\boldsymbol{x}_i$, $\boldsymbol{x}_j$, $\boldsymbol{x}_k\in V$, $1\le i,j,k\le n$, 低密度分割密度敏感距离满足测度的 4 个性质:

1) 非负性, $D_{i,j}^{\rho}\ge 0$; 2) 自反性, $D_{i,j}^{\rho}=0$ 当且仅当 $\boldsymbol{p}_i=\boldsymbol{p}_j$; 3) 对称性, $D_{i,j}^{\rho}=D_{j,i}^{\rho}$; 4) 三角不等式, $D_{i,j}^{\rho}\le D_{i,k}^{\rho}+D_{k,j}^{\rho}$. 该性质的详细证明见附录 A.

为了从理论上说明该距离测度能够反映数据空间全局一致性和局部空间一致性特征, 下面分别给出两个有关 ρ 的极限推导:

1) 当 ρ 趋近无穷大时,

$$\lim_{\rho\to\infty}D_{i,j}^{\rho}=\lim_{\rho\to\infty}\frac{1}{\rho}\ln\left(1+\min_{\boldsymbol{p}\in P_{i,j}}\sum_{k=1}^{|\boldsymbol{p}|}\left(\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{\delta_k\delta_{k+1}}}-1\right)\right)=$$

$$\lim_{\rho\to\infty}\frac{1}{1+\min\limits_{\boldsymbol{p}\in P_{ij}}\sum\limits_{k=1}^{|\boldsymbol{p}|}\left(\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{\delta_k\delta_{k+1}}}-1\right)}\cdot\min_{\boldsymbol{p}\in P_{ij}}\sum_{k=1}^{|\boldsymbol{p}|}\left(\frac{\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{\delta_k\delta_{k+1}}}d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{\delta_k\delta_{k+1}}\right)=$$

$$\lim_{\rho\to\infty}\frac{\min\limits_{\boldsymbol{p}\in P_{ij}}\sum\limits_{k=1}^{|\boldsymbol{p}|}\left(\frac{\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{\delta_k\delta_{k+1}}}d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{\delta_k\delta_{k+1}}\right)}{\max\limits_{\frac{d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{\delta_k\delta_{k+1}}}\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{\delta_k\delta_{k+1}}}}\Big/\frac{1+\min\limits_{\boldsymbol{p}\in P_{i,j}}\sum\limits_{k=1}^{|\boldsymbol{p}|}\left(\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{\delta_k\delta_{k+1}}}-1\right)}{\max\limits_{\frac{d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{\delta_k\delta_{k+1}}}\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{\delta_k\delta_{k+1}}}}=$$

$$\min_{\boldsymbol{p}\in P_{i,j}}\max_{k<|\boldsymbol{p}|}\frac{d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{(\delta_k\delta_{k+1})}$$

当 $\rho\to\infty$ 时, $\lim\limits_{\rho\to\infty}D_{i,j}^{\rho}=D_{i,j}=\min\limits_{\boldsymbol{p}\in P_{ij}}\max\limits_{k<|\boldsymbol{p}|}d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})/(\delta_k\delta_{k+1})$, 令 $\delta_k\delta_{k+1}$ 为常量, 因此, 低密度分割密度敏感距离等价为连接距离 $D_{i,j}=\min\limits_{\boldsymbol{p}\in P_{ij}}\max\limits_{k<|\boldsymbol{p}|}d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})$.

由上述分析可知, 当 ρ 趋近无穷大时低密度分割密度敏感距离等价为连接距离. 而连接距离能体现数据间的全局一致性, 由此可知, 低密度分割密度敏感距离测度同样能体现数据间的全局一致性特征.

2) 当 ρ 趋近 0 时,

$$\lim_{\rho\to 0}D_{i,j}^{\rho}=\lim_{\rho\to 0}\frac{1}{\rho}\ln\left(1+\min_{\boldsymbol{p}\in P_{i,j}}\sum_{k=1}^{|\boldsymbol{p}|}\left(\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{\delta_k\delta_{k+1}}}-1\right)\right)=$$

$$\lim_{\rho\to 0}\frac{\min\limits_{\boldsymbol{p}\in P_{i,j}}\sum\limits_{k=1}^{|\boldsymbol{p}|}(\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{(\delta_k\delta_{k+1})}}-1)}{\rho}=$$

$$\lim_{\rho\to 0}\frac{\min\limits_{\boldsymbol{p}\in P_{i,j}}\sum\limits_{k=1}^{|\boldsymbol{p}|}\frac{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{(\delta_k\delta_{k+1})}}{\rho}=$$

$$\min_{\boldsymbol{p}\in P_{i,j}} \sum_{k=1}^{|\boldsymbol{p}|} \frac{d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})}{(\delta_k \delta_{k+1})}$$

当 $\rho \to 0$ 时, $\lim_{\rho\to\infty} D_{i,j}^{\rho} = D_{i,j}^{\rho} = \min_{\boldsymbol{p}\in P_{ij}} \sum_{k=1}^{k<|\boldsymbol{p}|} d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})/(\delta_k\delta_{k+1})$, 令 $\delta_k\delta_{k+1}$ 为常量, 因此, 低密度分割密度敏感距离等价为最短几何距离 $D_{i,j} = \min_{\boldsymbol{p}\in P_{ij}} \sum_{k=1}^{|\boldsymbol{p}|} d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})$.

由上述分析可知, 当 ρ 趋近无穷大时, 低密度分割密度敏感距离等价为最短几何距离. 而最短几何距离能体现数据间的局部一致性, 由此可知, 低密度分割密度敏感距离同样能体现数据间的局部一致性特征.

2.3　相对密度敏感项计算

本文相对密度敏感项 δ_i 的估计方法采用的是 Parzen-window 算法[23], 定义如下:

$$\delta_i = \exp\left\{\omega \frac{Par(\boldsymbol{p}_i)}{\zeta}\right\}, \quad \forall i = 1, 2, \cdots, n \tag{9}$$

$$Par(\boldsymbol{p}_i) = \frac{1}{|\Psi(\boldsymbol{p}_i)|} \sum_{x_j \in |\Psi(\boldsymbol{p}_i)|} \left(\frac{1}{\sqrt{(2\pi)^D s}}\right) \exp\left(-\frac{1}{2s} d^2(\boldsymbol{p}_i, \boldsymbol{p}_j)\right) \tag{10}$$

$$\zeta = \frac{1}{|\Psi(\boldsymbol{p}_i)|} \sum_{i=1}^{n} Par(\boldsymbol{p}_i) \tag{11}$$

其中, D 为输入数据的维度, ω 为权重, 缺省设置为 1, s 是 Parzen-window 的平滑参数, 缺省设置为 3. 由上述公式不难看出, 训练样本点 $\boldsymbol{p}_i$ 获取的互近邻相对密度值 δ_i 越大, 则表明 $\boldsymbol{p}_i$ 所处的区域越密集.

$$\Psi(\boldsymbol{p}_i) = \{\boldsymbol{p}_j | \boldsymbol{p}_j \in K(\boldsymbol{p}_i) \text{ and } \boldsymbol{p}_i \in K(\boldsymbol{p}_j)\} \tag{12}$$

$$K(\boldsymbol{p}_i) = \{\boldsymbol{p}_j | d(\boldsymbol{p}_i, \boldsymbol{p}_j) \le d_i^k\} \tag{13}$$

式 (13) 中 $d(\boldsymbol{p}_i, \boldsymbol{p}_j)$ 代表 $\boldsymbol{p}_i$ 和 $\boldsymbol{p}_j$ 之间的欧氏距离, d_i^k 代表 $\boldsymbol{p}_i$ 样本与其他样本间的第 k 个最小欧氏距离. $K(\boldsymbol{p}_i)$ 代表 $\boldsymbol{p}_i$ 的 k 最近邻. $\Psi(\boldsymbol{p}_i)$ 代表 $\boldsymbol{p}_i$ 的 k 互近邻.

观察式 (8) 低密度分割密度敏感距离公式我们可以得到, 在互近邻密度 δ_i 较小的条件下, 两点间的距离通过指数函数的放大作用进一步增大, 从而使经过相对密度小的样本点间的路径距离变大; 而在互近邻密度 δ_i 较大的条件下, 两点间的距离受指数函数的放大作用影响较小, 从而使经过相对密度大的样本点间的路径距离变小. 综上所述, 低密度分割密度敏感距离测度通过样本点密度项来调整两点间的欧氏距离 $d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})$, 有效降低了孤立噪声和 "桥" 噪声对算法聚类性能的影响.

2.4　参数确定

由于谱聚类算法中最近邻个数 k 参数的确定和拉普拉斯矩阵特征向量的选择对算法聚类性能的影响很大, 为此, 本文进一步给出了基于 SC 指标的 k 近邻图 k 值选取方法和基于谱熵贡献率的特征向量选取方法.

1) 最近邻个数 k 参数的确定

基于 SC 指标的 k 近邻图 k 值选取方法描述如下:

$$W_j = \sum_{\boldsymbol{p}_i \in \boldsymbol{c}_j} \frac{1}{n_j} (\boldsymbol{p}_i - \boldsymbol{m}_j)(\boldsymbol{p}_i - \boldsymbol{m}_j)^{\mathrm{T}} \tag{14}$$

$$W_C = \sum_{j=1}^{C} W_j \tag{15}$$

$$W_B = \sum_{j=1}^{C} (\boldsymbol{m}_j - \boldsymbol{m})(\boldsymbol{m}_j - \boldsymbol{m})^{\mathrm{T}} \tag{16}$$

$\boldsymbol{m}$ 是全体样本的均值, $\boldsymbol{m}_j$, n_j 是属于第 j 类簇 $\boldsymbol{c}_j$ 所有样本的均值和个数.

$$SC = \mathrm{tr}(W_C^{-1} W_B) \tag{17}$$

其中, tr 为求解矩阵的迹函数, 为了防止 W_C 矩阵是奇异的, 这里通常采用 $SC = \mathrm{tr}(W_B)/\mathrm{tr}(W_C)$, SC 值越大聚类效果越好.

2) 特征向量的选择

本文基于谱熵贡献率的特征向量选取方法是根据核熵成分分析[24] 的思想, 计算各特征向量的贡献率, 从而选取前 C 个贡献率高的特征向量. 具体方法如下:

将本文算法得到的拉普拉斯矩阵 L_{sym} 进行分解得到 $L_{sym} = LML^{\mathrm{T}}$, 其中 $M = \mathrm{diag}\{\lambda_1, \cdots, \lambda_n\}$, $E = (\boldsymbol{l}_1, \cdots, \boldsymbol{l}_n)$. λ_i, $\boldsymbol{l}_i$ 分别为拉普拉斯矩阵 L_{sym} 的特征值和特征向量. 计算特征向量贡献率为:

$$\overline{V}(\boldsymbol{l}_i) = (\sqrt{\lambda_i} \boldsymbol{l}_i^{\mathrm{T}} \mathbf{1})^2, \quad i = 1, \cdots, n \tag{18}$$

由式 (18) 可以看出, 拉普拉斯矩阵 L_{sym} 的每个特征值及特征向量对熵估计的贡献不同. 本文选取前 C 个对应贡献率 $\overline{V}(\boldsymbol{l}_i)$ 值最大的特征向量构造待聚类的数据矩阵 $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_C] \in \mathbf{R}^{n\times C}$.

2.5　算法流程

输入.　n 个样本点 $\{\boldsymbol{x}_i\}_{i=1}^n$, 聚类个数 C, 伸缩因子 ρ, 输入数据的维度 D, 权重 ω, 缺省设置为 1, Parzen-window 的平滑参数 s, 缺省设置为 3.

输出. 样本点的划分 $c_1, c_2, \cdots, c_C$

步骤 1. 构造 k 互近邻无向图 $G = (P, V)$, P

是样本顶点, V 是 G 的边集合, 其中 k 值的确定通过基于 SC 指标进行最优选取;

步骤 2. 根据 G 计算每个顶点 $\boldsymbol{p}_i$ 的密度信息 δ_i, $i = 1, \cdots, n$;

步骤 3. 根据样本相似性度量方法构造样本相似度矩阵 $S \in \mathbf{R}^{n\times n}$, 其中矩阵中的任意元素表示为:

$$D_{i,j}^{\rho} = \frac{1}{\rho}\ln\left(1+\min_{\boldsymbol{p}\in P_{ij}}\sum_{k=1}^{|\boldsymbol{p}|}(\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k,\boldsymbol{p}_{k+1})}{(\delta_k\rho_{k+1})}})-1\right) \tag{19}$$

$$S_{ij} = \exp\left(-\frac{D_{i,j}^{\rho}}{2\sigma^2}\right) \tag{20}$$

这里设置对角线上的元素 $S_{ii} = 0$, 其中 $1 \le i, j \le n$;

步骤 4. 构造拉普拉斯矩阵 $L_{sym} = B^{-\frac{1}{2}}SB^{-\frac{1}{2}}$, 其中 B 为对角矩阵 $B_{ii} = \sum_{j\in\Psi(\boldsymbol{p}_i)} S_{ij}$;

步骤 5. 得到所有的特征向量 $L_{sym} = \{\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_n\}$, 并根据其谱熵贡献率进行排序得到 $\{Rv_1, Rv_2, \cdots, Rv_n\}$, 选取前 C 个谱熵贡献率最大的特征向量构成谱空间, 生成待聚类的数据矩阵 $X = \{\boldsymbol{l}_{Rv_1}, \boldsymbol{l}_{Rv_2}, \cdots, \boldsymbol{l}_{Rv_n}\}$;

步骤 6. 单位化 X 的行向量, 得到矩阵 $Y \in \mathbf{R}^{n\times C}$, 其中, $Y_{ij} = \frac{X_{ij}}{\sqrt{\sum\limits_j X_{ij}^2}}, i = 1, \cdots, n, j = 1, \cdots, C$

步骤 7. 将 Y 的每一行看成是空间 $\mathbf{R}^C$ 内的一点, 使用最大熵聚类算法将其聚为 C 类.

2.6 算法稳定性

本文根据 Davis-Kahan 定理[25] 来评价算法的稳定性. 定理内容如下:

Davis-Kahan 定理: 设 $A, H \in \mathbf{R}^{n\times n}$ 是对称矩阵, 令 $\|\cdot\|$ 是 Frobenius 矩阵范数. 扰动后的矩阵为 $\widetilde{A} = A + H$, H 为扰动阵. 令区间 $Z_1 \subset \mathbf{R}$, $\sigma_{Z_1}(A)$ 为包含在 Z_1 中所有 A 矩阵特征值的集合, V_1 是 $\sigma_{Z_1}(A)$ 对应的特征向量子空间, $\sigma_{Z_1}(\widetilde{A})$ 和 $\widetilde{V}_1$ 分别是包含在 Z_1 中的所有 $\widetilde{A}$ 矩阵特征值的集合以及 $\sigma_{Z_1}(\widetilde{A})$ 对应的特征向量子空间, 则定义空间 Z_1 和除 Z_1 之外的 A 矩阵的互补子空间之间的距离为:

$$\delta = \min\{|\lambda - s|,\ \lambda \text{ 是 } A \text{ 矩阵的特征值},\ \lambda \notin Z_1, s \in Z_1\} \tag{21}$$

则两个子空间 V_1 和 $\widetilde{V}_1$ 的距离 $d(V_1, \widetilde{V}_1)$ 的上界为:

$$d(V_1, \widetilde{V}_1) \le \frac{\|H\|}{\delta} \tag{22}$$

根据上述定理, 令理想状态下的拉普拉斯图矩阵为 L_{sym}, 经不同算法得到的拉普拉斯矩阵为 $\widetilde{L}_{sym}$, 这里选择设定 Z_1 包括 L_{sym} 和 $\widetilde{L}_{sym}$ 的前 C 个特征值, 则 $Z_1 = [\overline{V}(l_{Rv_1}), \overline{V}(l_{Rv_C})]$, $\delta = |\overline{V}(l_{Rv_C}) - \overline{V}(l_{Rv_{C+1}})|$. 根据上述 Davis-Kahan 定理可知, 特征子空间的估计近似程度取决于 $\|H\| = \|\widetilde{L}_{sym} - L_{sym}\|$ 和 $|\overline{V}(l_{Rv_C}) - \overline{V}(l_{Rv_{C+1}})|$ 的大小, 在 $|\overline{V}(l_{Rv_C}) - \overline{V}(l_{Rv_{C+1}})|$ 固定的前提下, $\|H\| = \|\widetilde{L}_{sym} - L_{sym}\|$ 越小, 子空间 V_1 和 $\widetilde{V}_1$ 的距离越小, 则特征子空间的估计近似程度越大, 谱聚类的效果越好.

为了显式地说明上述定理, 我们以 4 个具有流形结构的人工数据集为例进行说明, 其中 Ring 数据集包括 2 类, Eyes 数据集包括 3 类, Four-lines 数据集包括 4 类和 Three-guasses 数据集包括 3 类. 其中本文算法的相关参数根据实验经验设置如下: 最近邻个数 $k = 15$, 伸缩因子 $\rho = 100$, 相似度尺度参数 $\sigma_2 = 5$. 由于拉普拉斯图矩阵完全取决于相似度矩阵, 因此图 3～6 显示基于传统欧氏距离、连接距离、密度敏感、基于密度的最短几何距离和低密度分割密度敏感距离的相似性度量计算出的相似度矩阵. 对矩阵按照样本聚类重新排序后, 可明显看出, 低密度分割密度敏感距离的相似性度量得到的相似度矩阵效果最好, 其中图中白色块状表示数值为 0, 黑色块状表示数值为 1, 表明相似度只有 0 和 1, 区分效果明显. 该示例说明因本文距离测度能同时满足数据间的全局一致性和局部一致

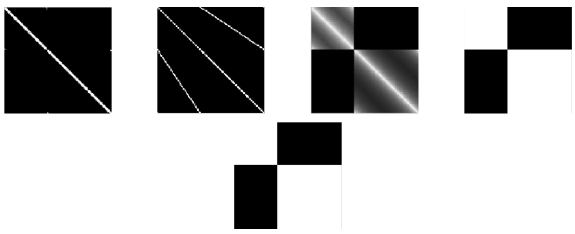


图 3 Ring 数据集相似度矩阵

Fig. 3 Ring dataset similarity matrix

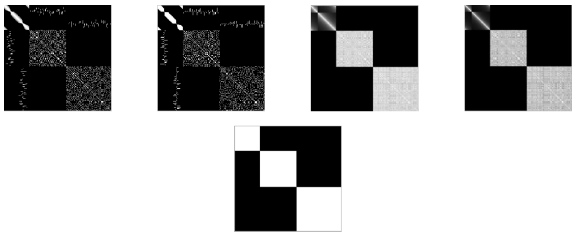


图 4 Cirlcetwopoints (Eyes) 数据集相似度矩阵

Fig. 4 Cirlcetwopoints (Eyes) dataset similarity matrix

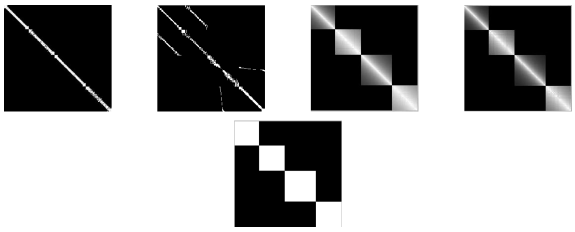


图 5　Four-lines 数据集相似度矩阵

Fig. 5　Four-lines dataset similarity matrix

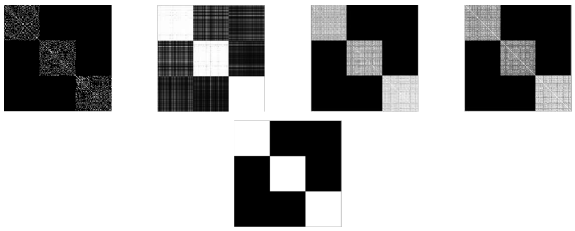


图 6　Three-Guasses 数据集相似度矩阵

Fig. 6　Three-Guasses dataset similarity matrix

性, 计算出来样本相似矩阵与理想相似矩阵相一致, 从而使算法更加稳定, 聚类效果更好.

2.7　算法复杂度

因本文算法采用了 Dijkstra 最短路径算法[26] 求解最短路径, 该算法的计算复杂度为 $\mathrm{O}(n^2)$, n 为样本个数. 同时该算法涉及到谱聚类算法的拉普拉斯矩阵特征值分解, 该问题的计算复杂度为 $\mathrm{O}(n^2)$. 因此可以说本文算法的计算复杂度为 $\mathrm{O}(n^2)$, 与原有谱聚类算法同一个数量级.

3　实验分析

本文通过 4 个实验来测试本文算法的聚类性能. 首先将本文提出的算法应用到 8 个人工数据集中并与传统的 NJW 谱聚类算法[7] (NJW-SC)、密度敏感的谱聚类算法[17] (DS-SC)、密度峰值聚类算法[18] (DP-SC) 和基于密度的最短几何距离谱聚类算法[18] (DSGD-SC) 的聚类结果进行比较分析. 其次, 对本文算法中伸缩因子 ρ 和 k 近邻参数的选择进行讨论分析, 并给出取值建议, 接着通过 8 个 UCI 数据集验证参数选择的正确性. 最后, 对 5 种算法聚类问题的鲁棒性进行分析. 实验环境: Windows 7 操作系统, CPU: Intel i7, 3.4 GHz 处理器, 仿真软件为 Matlab 2010b.

3.1　实验评价指标

为了进行聚类效果的对比分析, 实验中使用两种评价指标, 标准互信息 (Normalized mutual information) 和兰德调整指数 (Adjusted rand index).

互信息 (Mutual information) 是用来衡量两个数据分布的吻合程度. 假设 U 与 V 是对 N 个样本标签的分配情况, 则两种分布的熵 (熵表示的是不确定程度) 分别为:

$$H(U)=\sum_{i=1}^{|U|}P(i)\log\left(P(i)\right) \tag{23}$$

$$H(V)=\sum_{j=1}^{|V|}P'(j)\log\left(P'(j)\right) \tag{24}$$

其中, $P(i)=|U_i|/N$, $P'(j)=|V_j|/N$, U 与 V 的互信息定义为:

$$MI(U,V)=\sum_{i=1}^{|U|}\sum_{j=1}^{|V|}P(i,j)\log\left(\frac{P(i,j)}{P(i)P'(j)}\right) \tag{25}$$

其中, $P(i,j)=|U_i\cap V_j|/N$, 标准化后的互信息 (Normalized mutual information, NMI) 其值越大聚类效果越好, 计算公式为:

$$NMI(U,V)=\frac{MI(U,V)}{\sqrt{H(U)}\sqrt{H(V)}} \tag{26}$$

调整蓝德指数 (Adjusted rand index, ARI) 是为了定量地描述算法的聚类性能, 它将类别的划分看成是样本之间的关系. 每对样本要么划分为同类, 要么划分为不同类, 通过统计正确决策对数来评价聚类算法的性能, 对于一个有 N 个样本的数据集而言, RI 指标的计算公式如下:

$$R(U,V)=\frac{\sum\limits_{lk}\binom{N_{lk}}{2}-\frac{\left[\sum\limits_{l}\binom{N_{l\cdot}}{2}\cdot\sum\limits_{k}\binom{N_{\cdot k}}{2}\right]}{\binom{N}{2}}}{\frac{1}{2}\left[\sum\limits_{l}\binom{N_{l}}{2}+\sum\limits_{k}\binom{N_{\cdot k}}{2}\right]-\frac{\left[\sum\limits_{l}\binom{N_{l}}{2}\sum\limits_{k}\binom{N_{\cdot k}}{2}\right]}{\binom{N}{2}}} \tag{27}$$

其中, N_{lk} 表示被划分到类别 l 和 k 的样本的个数, $(U,V)\in[0,1]$, 其数值越大, 说明聚类划分的正确率越高.

3.2　仿真数据实验分析

实验选取了 8 个人工合成数据集, Spiral 数据集、Three-circles 数据集、Two-moons 数据集、Ring 数据集、Eyes 数据集、Three-Gausses 数据集、Four-lines 数据集和 Square 数据集, 并对 8 个数据集分别使用不同的谱聚类算法进行聚类. 实验参数设置为, 伸缩因子 $\rho=50$, 最近邻个数 $k=15$. 依据以往实验经验设置相对密度的参数项为: 权重 $\omega=1$, 平滑参数 $s=3$. 其中, 密度敏感的谱聚类算

法利用的是全距离公式进行距离的测度, 其伸缩因子设置为 $\rho = 3$.

首先, 实验将本文提出的基于低密度分割密度敏感距离的谱聚类算法与其他 4 种谱聚类算法的聚类结果进行比较, 分别计算出评价指标 NMI 和 RI 值, 如表 1 所示.

由表 1 可以看出, DS-SC 算法、DP-SC 算法以及 DSGD-SC 算法在人工数据集上的聚类效果不如传统的 NJW 谱聚类算法, 这主要是由于上述三种算法所采用的距离测度, 没能很好地体现样本数据间的全局一致性和局部一致性特征, 导致相似度矩阵不符合真实的数据分布情况, 因此严重影响了聚类性能. 观察传统的 NJW 谱聚类算法的聚类结果可知, 除 Spiral、Two-moons、Eyes 和 Square 数据集外, 其余数据集聚类结果的性能评价指标 NMI 和 RI 均达到了最佳. 这说明虽然传统的 NJW 谱聚类采用的是欧氏距离测度, 但由于受到 $k = 15$ 最近邻图的影响, 使其部分兼顾了数据间的全局一致性特征, 因此聚类效果优于其余算法. 相比较而言, 本文提出的基于低密度分割密度敏感距离的谱聚类算法, 除 Square 数据集外其余数据集聚类结果的性能评价指标 NMI 和 RI 均优于其他算法. 这是由于本文算法所采用的距离测度能同时满足数据间全局一致性和局部一致性要求, 使得到的相似度矩阵更加符合数据的流形结构分布特点, 进而大大提高了算法的聚类性能.

为了更直观地说明本文算法对仿真数据的聚类效果, 实验中对 Spiral 数据集、Three-circles 数据集、Ring 数据集、Eyes 数据集 4 个人工合成数据集使用 5 种不同的谱聚类算法进行聚类分析. 同时为了显示地说明 "桥" 噪声对算法性能的影响, 在 Eyes 数据集中增加了 15 个 "桥" 噪声, 并使用基于谱熵贡献率的特征向量选取方法选取特征向量, 图 7 ~ 14 分别为各种谱聚类算法的聚类结果和本文算法通过基于谱熵贡献率选取的各个数据集特征向量分布情况.

由各数据集的聚类结果图可以看出, 密度敏感的谱聚类算法和基于密度的最短几何距离谱聚类算法未能得到准确的聚类结果; 密度峰值聚类算法虽然能够在 Spiral 数据集、Three-circles 数据集和 Ring 数据集上得到清晰的聚类结果, 但是在加入了 "桥" 噪声的 Eyes 数据集上的聚类结果不如基于低密度分割密度敏感距离的谱聚类算法, 可见该算法虽能对流形结构数据有较好的聚类但易受 "桥" 噪声数据影响; 传统的 NJW 谱聚类算法除了 Spiral 数据集和 Eyes 数据集, 对其余数据集均得到较好的聚类结果, 这与上述实验结果相一致. 相比较而言, 本文算法在 4 个数据集上均呈现出理想清晰的聚类结果. 这说明在使用距离测度对样本数据进行相似性度量计算时, 仅考虑了度量关系中任何一种一致性关系要求, 都容易导致算法对 k 最近邻图过于敏感, 当设置不当时两个流形体间因受连接 "桥" 噪声

表 1 不同谱聚类算法对 8 种人工合成数据集关于 NMI 和 RI 的统计结果

Table 1 Statistics of different clustering algorithms on eight synthetic datasets in terms of NMI and RI

数据集	NJW-SC	DS-SC	DP-SC	DSGD-SC	LDSDSD-SC
Spiral	0.5635 ± 0	0.0290 ± 0	$\mathbf{1 \pm 0}$	0.0138 ± 0	$\mathbf{1 \pm 0}$
	0.6720 ± 0	0.0390 ± 0	$\mathbf{1 \pm 0}$	$8.0080 \times 10^{-6} \pm 0$	$\mathbf{1 \pm 0}$
Three-circles	$\mathbf{1 \pm 0}$	0.0456 ± 0	$\mathbf{1 \pm 0}$	0.0495 ± 0	$\mathbf{1 \pm 0}$
	$\mathbf{1 \pm 0}$	0.0195 ± 0	$\mathbf{1 \pm 0}$	0.0069 ± 0	$\mathbf{1 \pm 0}$
Two-moons	0.9595 ± 0	0.1316 ± 0	$\mathbf{1 \pm 0}$	0.0334 ± 0	$\mathbf{1 \pm 0}$
	0.9799 ± 0	0.1722 ± 0	$\mathbf{1 \pm 0}$	0.0013 ± 0	$\mathbf{1 \pm 0}$
Ring	$\mathbf{1 \pm 0}$	0.0277 ± 0	$\mathbf{1 \pm 0}$	0.0277 ± 0	$\mathbf{1 \pm 0}$
	$\mathbf{1 \pm 0}$	0.0040 ± 0	$\mathbf{1 \pm 0}$	0.0040 ± 0	$\mathbf{1 \pm 0}$
Eyes	0.4998 ± 0	0.5913 ± 0	0.5117 ± 0	0.0508 ± 0	$\mathbf{1 \pm 0}$
	0.4642 ± 0	0.6210 ± 0	0.4366 ± 0	0.0060 ± 0	$\mathbf{1 \pm 0}$
Three-Gausses	$\mathbf{1 \pm 0}$	$\mathbf{1 \pm 0}$	$\mathbf{1 \pm 0}$	0.0332 ± 0	$\mathbf{1 \pm 0}$
	$\mathbf{1 \pm 0}$	$\mathbf{1 \pm 0}$	$\mathbf{1 \pm 0}$	$4.4894 \times 10^{-5} \pm 0$	$\mathbf{1 \pm 0}$
Four-lines	$\mathbf{1 \pm 0}$	0.7677 ± 0	0.8955 ± 0	0.0293 ± 0	$\mathbf{1 \pm 0}$
	$\mathbf{1 \pm 0}$	0.6729 ± 0	0.7725 ± 0	0.0005 ± 0	$\mathbf{1 \pm 0}$
Square	$\mathbf{0.7978 \pm 0}$	0.7830 ± 0	0.2335 ± 0.0017	0.2284 ± 0.1978	0.7636 ± 0
	0.8348 ± 0	0.8227 ± 0	$\mathbf{0.8817 \pm 0.1113}$	0.2054 ± 0.1781	0.8038 ± 0

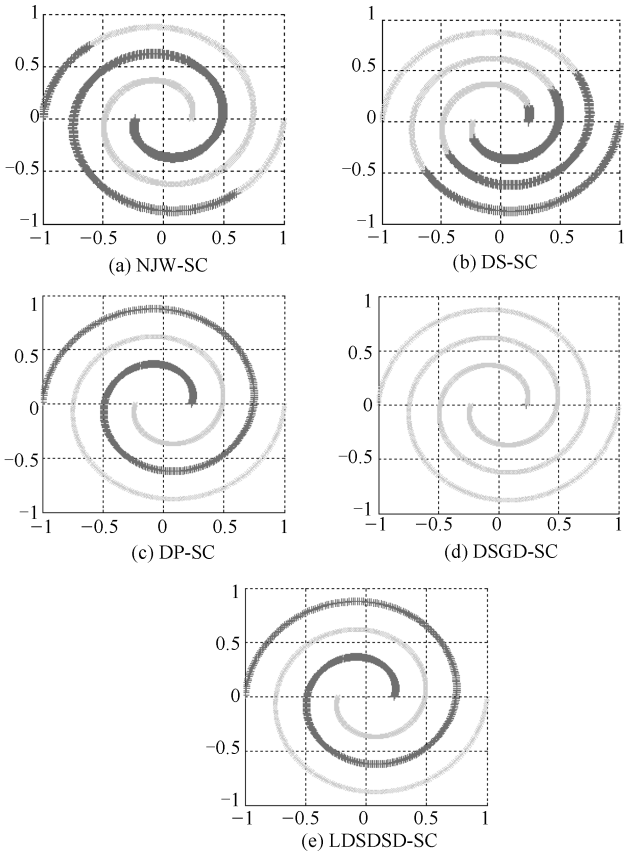


图 7　Spiral 数据集 5 种谱聚类算法的聚类结果图

Fig. 7　Clustering results of five different spectral clustering algorithms on Spiral dataset

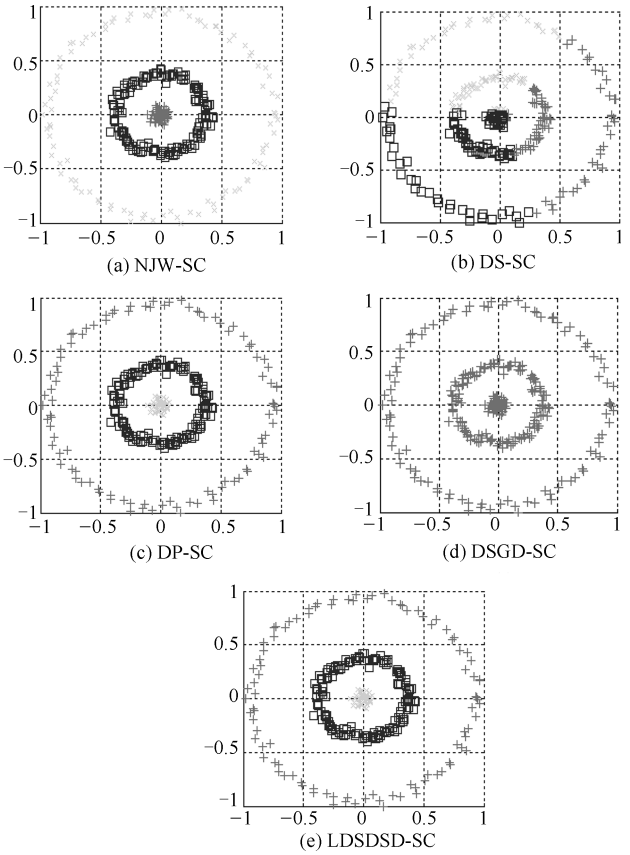


图 9　Three-circles 数据集 5 种谱聚类算法的聚类结果图

Fig. 9　Clustering results of five different spectral clustering algorithms on Three-circles dataset

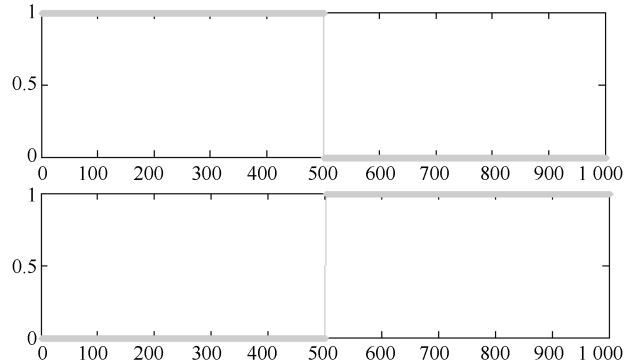


图 8　Spiral 数据集前两个特征向量分布情况

Fig. 8　The first two eigenvectors distribution on Spiral dataset

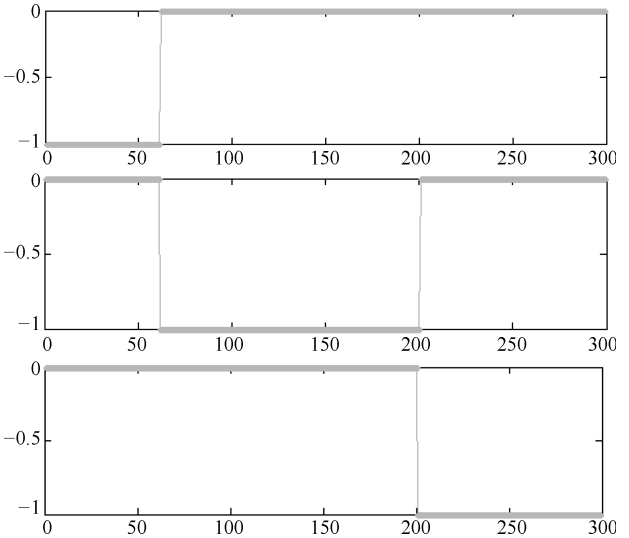


图 10　Three-circles 数据集前三个特征向量分布情况

Fig. 10　The first three eigenvectors distribution on Three-circles dataset

影响易发生合并现象, 如图 7 (b) 和 (d), 图 9 (b) 和 (d), 图 11 (b) 和 (d), 图 13 (b) 和 (d) 所示, 从而导致聚类结果不准确.

除此之外, 图 8、10、12、14 为本文算法利用基于谱熵贡献率方法选取的 4 种数据集特征向量分布情况. 由特征向量分布图可以看出, 各特征向量分布清晰, 且每个数据集选取出来的特征向量均为正交关系, 数据区分度明显, 进而使算法的聚类性能大大提高. 该实验结果进一步说明本文算法采用的距离测度计算出来的相似度矩阵更加合理, 同时基于谱熵贡献率的特征向量选取方法选取出的特征向量能

最大程度地保留了原始数据的空间分布特征, 从而大大提高了算法的聚类精度.

综合以上实验结果, 可以说明本文提出的基于低密度分割密度敏感距离的谱聚类算法性能稳定, 对于非线性或具有流形结构特征的数据聚类能够获得理想的聚类结果.

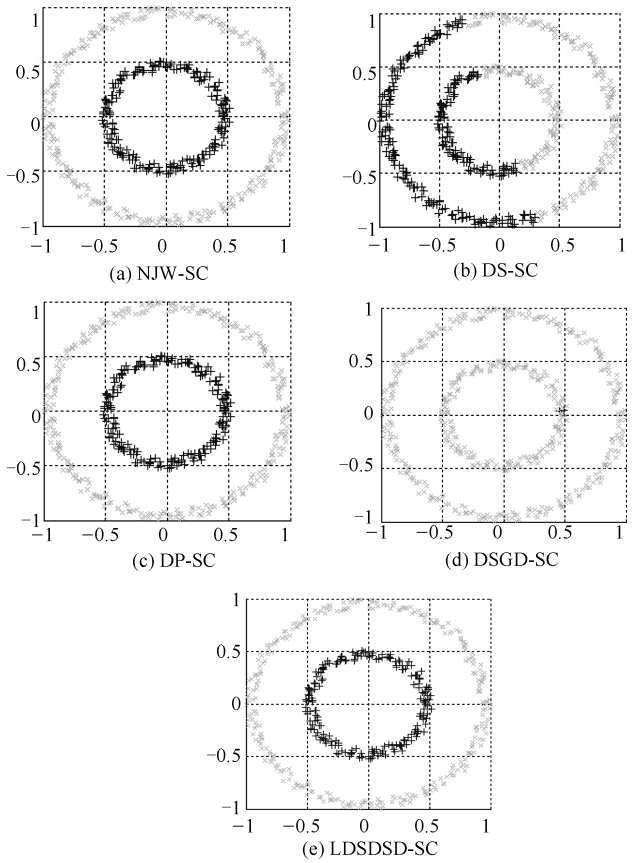


图 11 Ring 数据集 5 种谱聚类算法的聚类结果图

Fig. 11 Clustering results of five different spectral clustering algorithms on Ring dataset

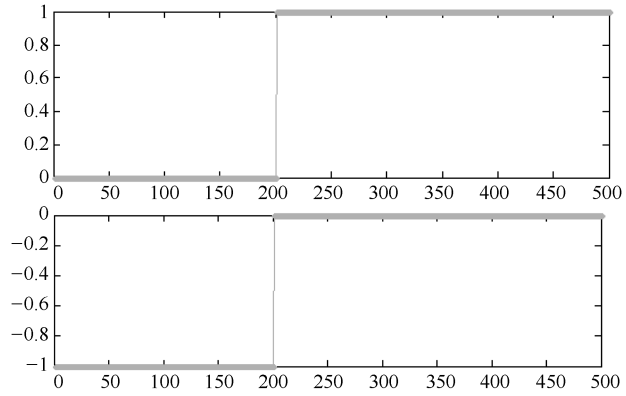


图 12 Ring 数据集前两个特征向量分布情况

Fig. 12 The first two eigenvectors distribution on Ring dataset

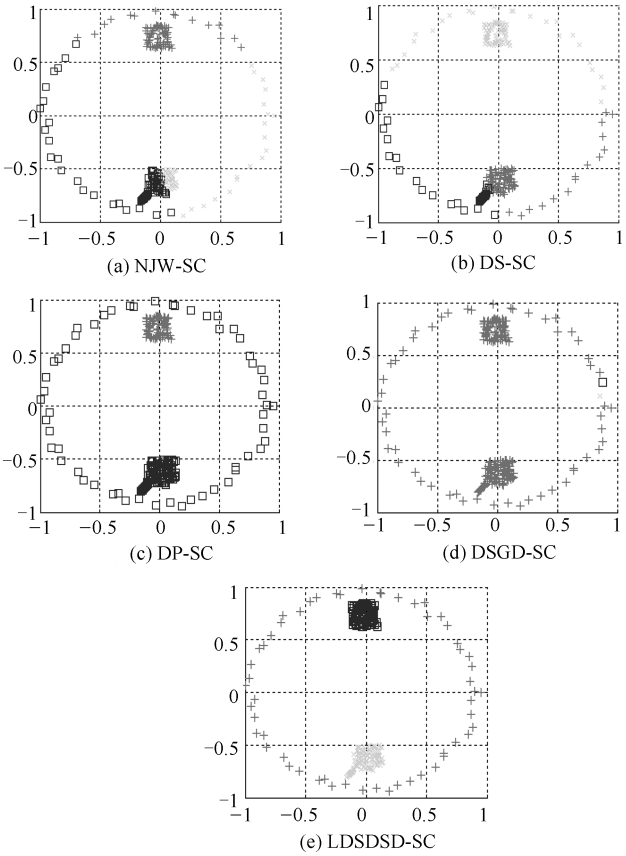


图 13 Eyes 数据集 5 种谱聚类算法的聚类结果图

Fig. 13 Clustering results of five different spectral clustering algorithms on Eyes dataset

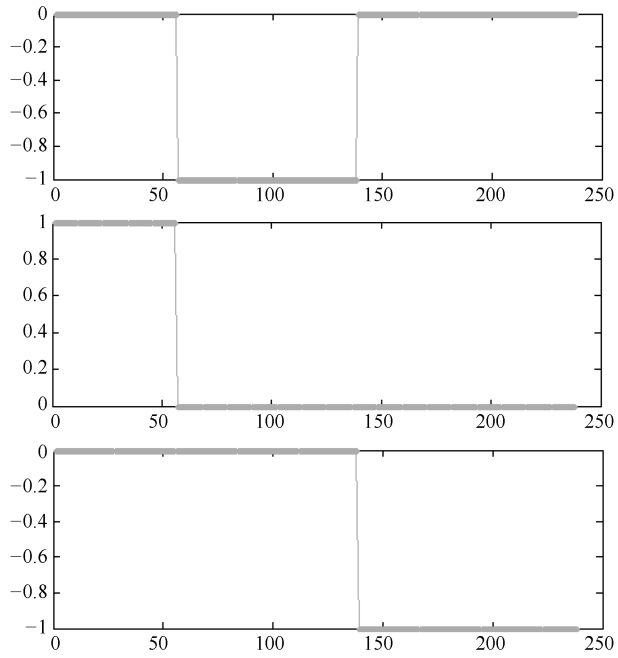


图 14 Eyes 数据集前三个特征向量分布情况

Fig. 14 The first three eigenvectors distribution on Eyes dataset

3.3　参数选择实验

为了讨论伸缩因子 ρ 的取值对本文算法聚类性能的影响, 实验利用 8 个人工合成数据集来分析本文提出的算法在不同 ρ 值下聚类结果, 为消除随机影响, 针对不同 ρ 值分别进行了 20 次实验并计算 NMI 和 RI 指标值的均值和方差, 结果如图 15 和图 16 所示.

由图 15 和图 16, 我们可以得到当 ρ 取 [50, 100] 附近时 NMI 指标和 RI 指标均达到了最大值, 聚类结果最为理想. 该实验结果与本文理论分析的结果一致, ρ 取过小, 距离测度容易趋于最短几何距离, 导致相似度计算不能兼顾数据间的全局一致性特征; 相反 ρ 取过大, 距离测度容易趋于连接距离, 导致相似度计算不能兼顾数据间的局部一致性特征. 因此, 本文建议伸缩因子的取值范围为 [50, 100].

为验证基于 SC 指标的 k 近邻图 k 值选取方法的有效性, 实验数据采用的是上一实验中 $k = 15$ 时本文算法聚类效果表现不佳的 Square 数据集. 对该数据集在不同 k 值下使用本文算法聚类后的 SC 指标进行统计分析, 结果如图 17 所示. 为了便于比较分析, 实验中同时给出了不同 k 值本文算法的聚类结果, 通过计算 NMI 和 RI 评价指标之和来表示, 结果如图 18 所示. 通过两个图的对比分析可以看出, 当 k 取 10 时, SC 指标达到最大, 为 11.8689, 同时本文算法的聚类性能评价指标之和也取得最大, 其中 NMI = 0.7844, RI = 0.8231, 这表明在基于 SC 指标的 k 近邻图 k 值选取方法选择的 k 值下, 本文算法对 Square 数据集聚类结果是比较理想的, 聚类结果清晰, 区分度明显, 如图 19 所示, 该实验结果验证了基于 SC 指标的 k 近邻图 k 值选取方法的可行性和有效性.

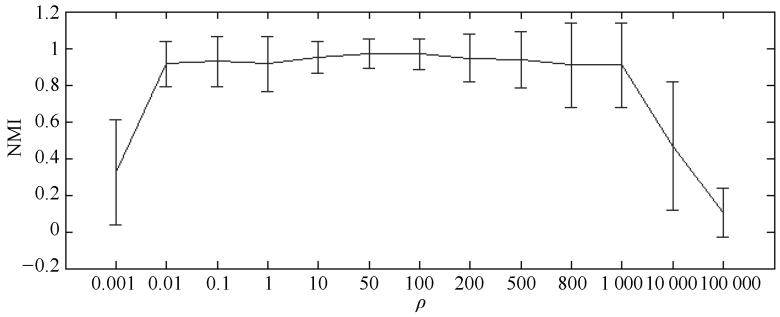


图 15　不同 ρ 值下 8 个人工数据集关于 NMI 的误差棒图

Fig. 15　NMI performance metrics on 8 synthetic datasets with different ρ values

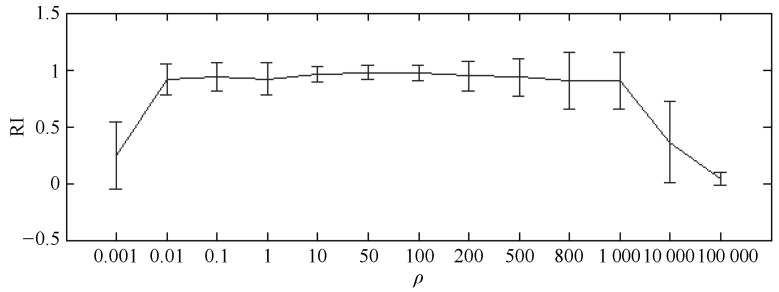


图 16　不同 ρ 值下 8 个人工数据集关于 RI 的误差棒图

Fig. 16　RI performance metrics on 8 synthetic datasets with different ρ values

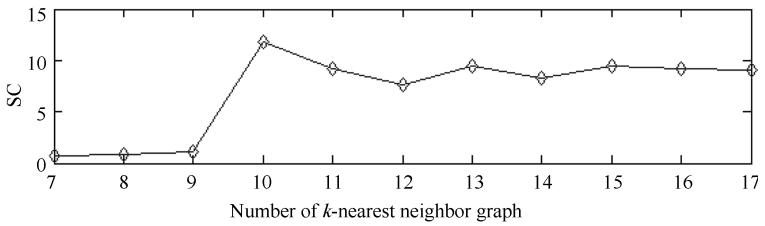


图 17　Square 数据集的 SC 性能指标随 k 值变化图

Fig. 17　SC performance metric obtained by the proposed approach on Square dataset with different k values

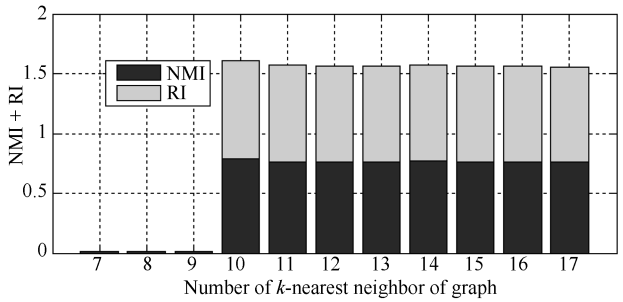


图 18 Square 数据集的 NMI + RI 性能指标随 k 值变化图

Fig. 18 NMI + RI performance metrics on Square dataset with different k values

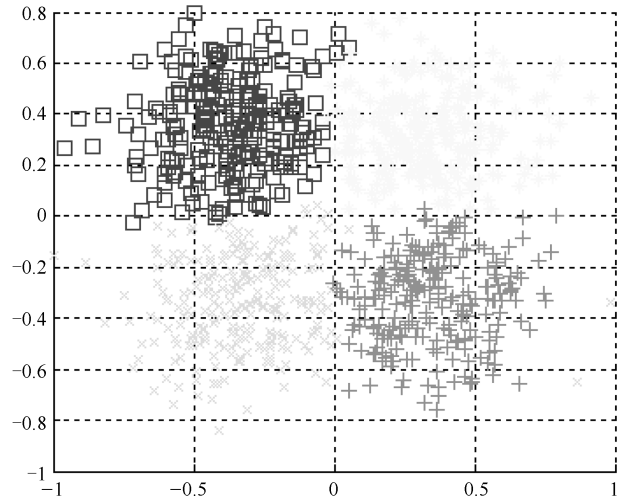


图 19 $k = 10$ 时 Square 数据集聚类的结果图

Fig. 19 The clustering result on Square dataset when $k = 10$

3.4 UCI 数据实验分析

为验证本文算法针对不同复杂结构数据的聚类性能, 实验数据采用来源于国际机器学习标准数据库 UCI 中的 8 组不同数据集, 分别为 WINE、IRIS、BREAST CANCER、GLASS、CAR、VEHICLE、LETTER(A, B) 和 LETTER(C, D), 数据的特征信息见表 2. 实验参数设置为: 伸缩因子 $\rho = 50$, 相对密度的参数设置为权重 $\omega = 1$, 平滑参数 $s = 3$, 实验评价指标为 NMI 和 RI, 本文算法采用基于谱熵贡献率进行特征向量选取.

表 2 实验数据集描述

Table 2 Description of experimental datasets

数据集	属性	样本个数	聚类数
WINE	13	178	3
IRIS	4	150	3
BREAST CANCER	9	683	2
GLASS	10	214	6
CAR	12	84	8
VEHICLE	18	846	4
LETTER (A, B)	16	1 555	2
LETTER (C, D)	16	1 541	2

实验首先根据基于 SC 指标的 k 近邻图 k 值选取方法对 8 个 UCI 数据集选取 k 值; 其次根据选择的 k 值, 对 8 个 UCI 数据集使用不同的谱聚类算法进行聚类, 并根据 5 种谱聚类算法的聚类结果计算评价指标 NMI 和 RI 的值, 实验结果如表 3 所示.

表 3 不同谱聚类算法对 8 种 UCI 数据集关于 NMI 和 RI 的统计结果

Table 3 Statistics of different clustering algorithms on eight UCI datasets in terms of NMI and RI

数据集	NJW-SC	DS-SC	DP-SC	DSGD-SC	LDSDSD-SC
WINE	0.8758 ± 0	0.8363 ± 0	0.8335 ± 0.1361	0.8399 ± 0.0640	$\mathbf{0.8781 \pm 0}$ $\mathbf{(k = 20)}$
	0.8974 ± 0	0.8515 ± 0	0.8116 ± 0.2470	0.8632 ± 0.0660	$\mathbf{0.8991 \pm 0}$ $\mathbf{(k = 20)}$
IRIS	0.6818 ± 0	0.7549 ± 0	0.7853 ± 0.0115	0.7552 ± 0.3187	$\mathbf{0.8571 \pm 0 (k = 13)}$
	0.6525 ± 0	0.7559 ± 0	0.7276 ± 0.1237	0.7251 ± 0.2611	$\mathbf{0.8682 \pm 0}$ $\mathbf{(k = 13)}$
BREAST CANCER	0.7903 ± 0	$\mathbf{0.8143 \pm 0}$	0.7467 ± 0.2135	0.7478 ± 0	0.7921 ± 0 $(k = 20)$
	0.8796 ± 0	0.8909 ± 0	0.8755 ± 0.1306	$\mathbf{0.9487 \pm 0}$	0.8797 ± 0 $(k = 20)$
GLASS	0.3065 ± 0	0.2692 ± 0	0.1913 ± 0.0157	0.1778 ± 0.1488	$\mathbf{0.7411 \pm 0}$ $\mathbf{(k = 30)}$
	0.1923 ± 0	0.1490 ± 0	0.1515 ± 0.0175	0.1031 ± 0.1196	$\mathbf{0.6644 \pm 0}$ $\mathbf{(k = 30)}$
CAR	0.6424 ± 0	0.7604 ± 0	0.6457 ± 0.0103	0.7126 ± 0	$\mathbf{0.7711 \pm 0}$ $\mathbf{(k = 7)}$
	0.3797 ± 0	0.5583 ± 0	$\mathbf{0.5773 \pm 0.0141}$	0.5571 ± 0	0.5697 ± 0 $(k = 7)$
VEHICLE	0.1280 ± 0	0.2032 ± 0	0.1125 ± 0.0377	0.0462 ± 0.0220	$\mathbf{0.2114 \pm 0}$ $\mathbf{(k = 15)}$
	0.1006 ± 0	0.1563 ± 0	$\mathbf{0.3759 \pm 0.0435}$	0.0333 ± 0.0167	0.1783 ± 0 $(k = 15)$
LETTER (A, B)	0.2593 ± 0	0.7216 ± 0	0.3103 ± 0.0016	0.7168 ± 0	$\mathbf{0.7396 \pm 0}$ $\mathbf{(k = 25)}$
	0.3367 ± 0	0.7658 ± 0	0.5667 ± 0.01005	0.5975 ± 0	$\mathbf{0.7863 \pm 0}$ $\mathbf{(k = 25)}$
LETTER (C, D)	0.0724 ± 0	$\mathbf{0.6876 \pm 0}$	0.5517 ± 0.1953	0.6131 ± 0	0.6535 ± 0 $(k = 30)$
	0.0984 ± 0	$\mathbf{0.7280 \pm 0}$	0.6505 ± 0.0167	0.6578 ± 0	0.6864 ± 0 $(k = 30)$

由表 3 可以看出, 本文提出的 LDSDSD-SC 算法聚类, 除了在 CAR、VEHICLE 数据集的 RI 指标小于 DP-SC 算法, 在 BREAST CANCER 和 LETTER (C, D) 数据集的指标稍差, 其余聚类结果的评价指标 NMI 和 RI 指标值均达到最佳状态. 实验结果表明, 本文算法通过采用同时满足数据间全局一致性和局部一致性要求的距离测度进行相似度的计算, 使相似度矩阵更加符合实际数据分布情况. 另外, 算法中通过基于谱熵贡献率进行特征向量选取, 使得到的特征向量最大程度地保留了原始数据的分布特征, 从而使本文算法能适应各种复杂数据结构的分布, 聚类性能大大提高.

为了进一步验证基于 SC 指标的 k 近邻图 k 值选取方法在不同结构数据下的有效性, 图 20、21 和图 22、23 分别为 IRIS 以及 WINE 两个 UCI 数据集的 SC 指标随着 k 近邻图中 k 值的变化情况图, 对于 IRIS 数据集而言当 $k = 13$ 时, SC 性能指标最优, 如图 20 所示. 通过图 21 的 NMI 和 RI 性能指标和的结果可以看出, IRIS 数据集在 $k = 13$ 的聚类性能指标最大. 对于 WINE 数据集而言当 $k = 20$ 时 SC 性能指标最优, 如图 22 所示, 同样通过图 23 的 NMI 和 RI 性能指标和的结果可以看出 WINE 数据集在 $k = 20$ 时的聚类性能指标最大. 该实验结果也进一步验证了基于 SC 指标的 k 近邻图 k 值选取的正确性, 在使用基于 SC 指标的 k 近邻图 k 值选取方法时, 可以选出使算法性能达到最佳的 k 值.

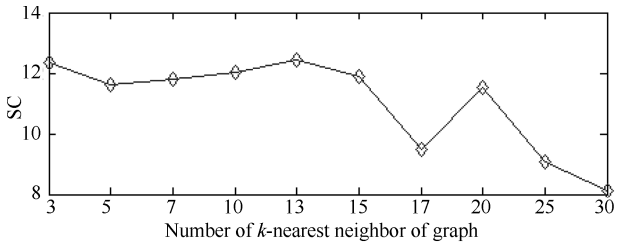


图 20　IRIS 数据集的 SC 性能指标随 k 值变化图

Fig. 20　SC metrics on IRIS dataset with different k values

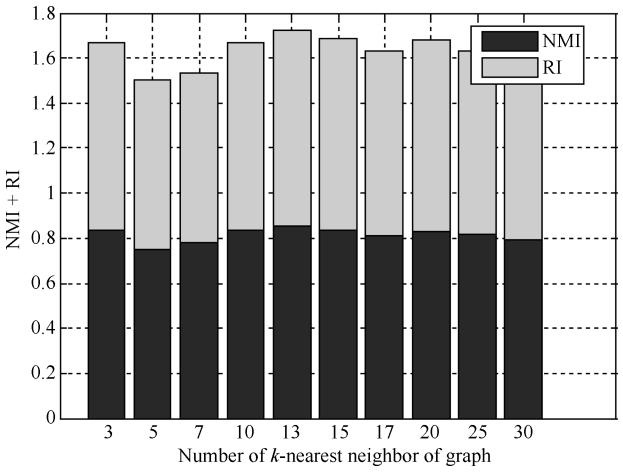


图 21　IRIS 数据集的 NMI+RI 性能指标随 k 值变化图

Fig. 21　NMI + RI performance metrics on IRIS dataset with different k values

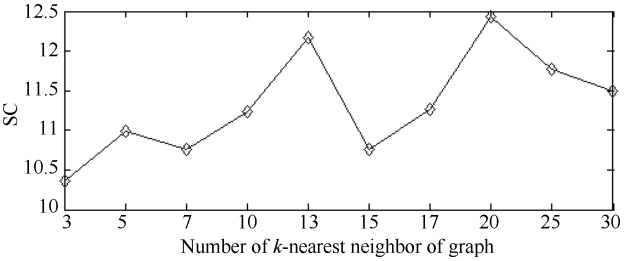


图 22　WINE 数据集的 SC 性能指标随 k 值变化图

Fig. 22　SC metrics on WINE dataset with different k values

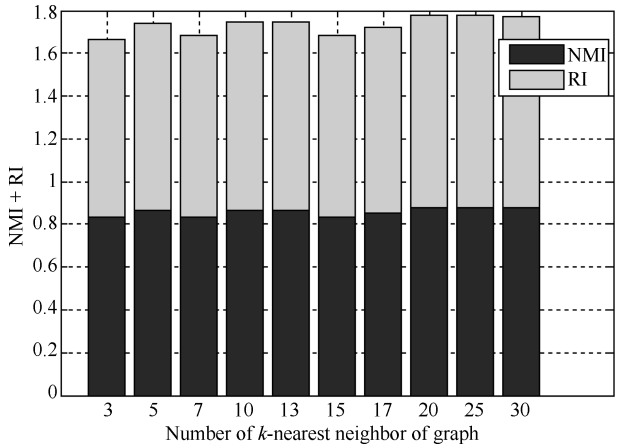


图 23　WINE 数据集的 NMI+RI 性能指标随 k 值变化图

Fig. 23　NMI + RI performance metrics on WINE dataset with different k values

3.5　算法鲁棒性

为验证本文提出的基于低密度分割密度敏感距离的谱聚类算法的鲁棒性, 本文借鉴文献 [27] 中的鲁棒性分析方法对 5 种算法在求解以上 8 个人工数据集以及 8 个 UCI 数据集聚类问题时的鲁棒性进行比较. 具体地, 算法 m (m 是指 5 种算法中的某一个算法) 在某一特定人工数据集或 UCI 数据集上的相对性能用如下两个比值来衡量:

$$b_m = \frac{NMI_m}{\max\limits_m NMI_m} \tag{28}$$

$$c_m = \frac{CR_m}{\max\limits_m CR_m} \tag{29}$$

因此, 在某个人工数据集或 UCI 数据集上表现最好的算法 m^* (m^* 是指性能最好的算法) 的相对

性能 $b_{m^*}=1$ 或 $c_{m^*}=1$, 而其他算法的 NMI 指标相对性能 $b_m<1$ 或者 RI 指标相对性能 $c_m<1$, b_m 或 c_m 值越大, 则算法 m 值越大, 则算法 m 在所有数据集上的 b_m 或 c_m 值的总和可以用来客观评价算法的鲁棒性, 总和越大鲁棒性越好.

根据上述分析, 实验对仿真数据集和 UCI 数据集聚类问题中 5 种算法的 b_m 和 c_m 值进行计算并比较, 图 24 和图 25 为 5 种算法的 NMI 性能指标和 RI 性能指标鲁棒性比较结果柱状图, 每一个算法对应的柱状图顶部所标数值为对应算法在所有 8 个人工数据集以及 8 个 UCI 数据集聚类问题上的 b_m 值的总和. 从图 24 中可以看出, 本文提出的 LDSDSD-SC 算法获得了最高的总和值, 达到了 15.8802, 密度峰值聚类算法次之, 达到了 12.3318.

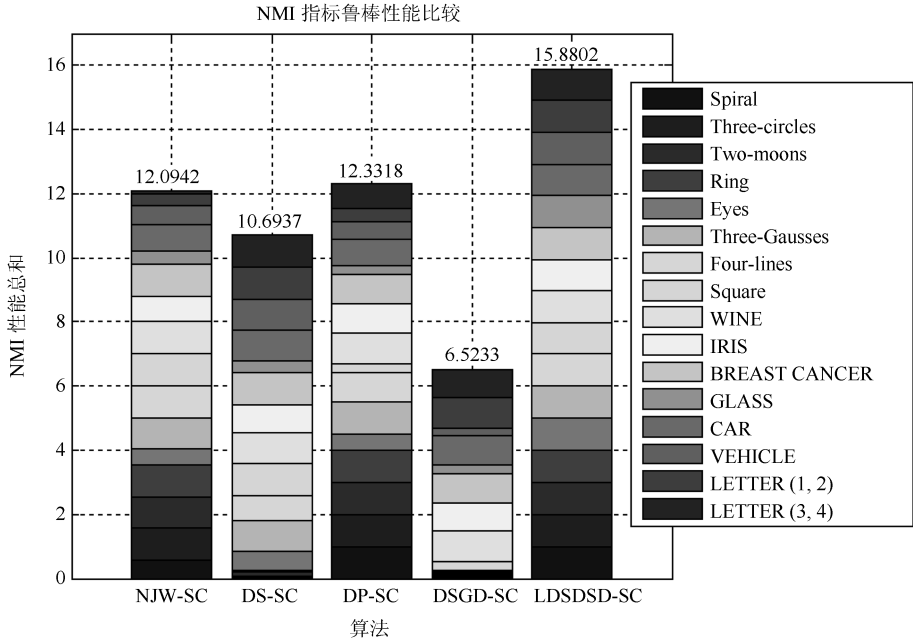


图 24 5 种算法的 NMI 性能指标鲁棒性比较结果

Fig. 24 Comparison results of NMI performance index of five algorithms robustness

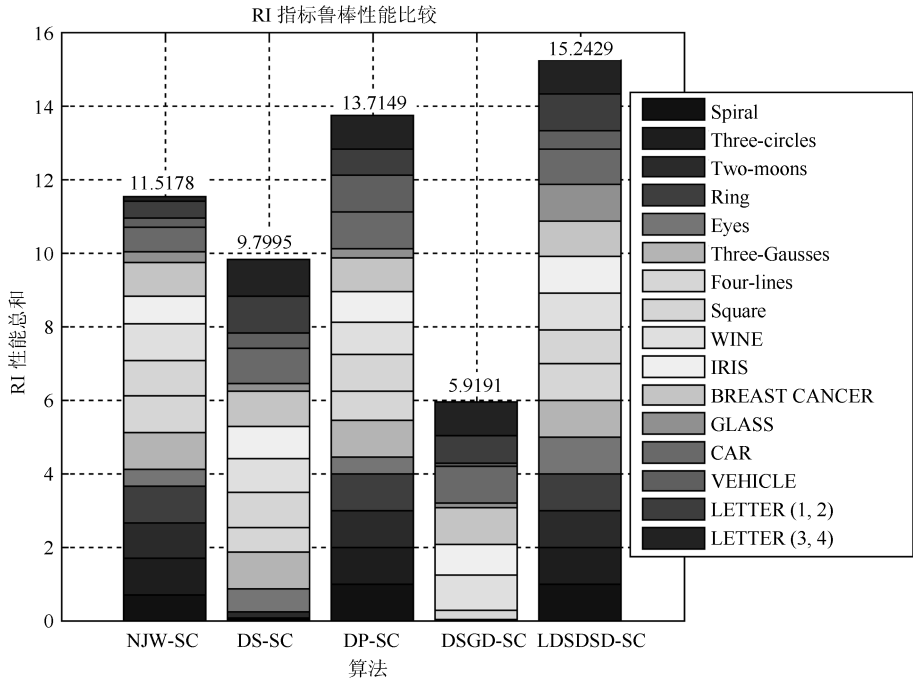


图 25 5 种算法的 RI 性能指标鲁棒性比较结果

Fig. 25 Comparison results of RI performance index of five algorithms robustness

其中, LDSDSD-SC 算法的 b_m 值在测试的 8 个人工数据集聚类问题中的 Spiral、Three-circles、Two-moons、Ring、Eyes、Three-Gausses 数据集上以及 8 个 UCI 数据集聚类问题中的 WINE、IRIS、BREAST CANCER、GLASS、CAR、VEHECLE、LETTER (1, 2)、LETTER (3, 4) 数据集上均为 1. 同样地, 从图 25 可以看出, 本文提出的 LDSDSD-SC 算法获得了最高的总和值, 达到了 15.8329, 密度峰值聚类算法次之, 达到了 13.7149. 这充分说明本文提出的基于低密度分割密度敏感距离的谱聚类算法对不同流型结构不同维度数据集的聚类问题均表现出很好的性能, 在所有比较的 5 种算法中具有最好的鲁棒性.

综上所述, 由于本文提出的基于低密度分割密度敏感距离的谱聚类算法使用低密度分割密度敏感距离能够综合考虑连接距离以及最短几何距离两个因素的影响. 同时, 算法考虑数据的局部密度特征增加相对密度敏感项, 从而有效防止孤立噪声以及 "桥" 噪声的影响, 得到较为理想的聚类结果, 提高了算法的鲁棒性. 另外, 算法通过基于 SC 指标的进行 k 近邻图中 k 值选取办法确定最优 k 近邻个数以及基于谱熵贡献率的特征向量选取方法, 也使得本文算法的聚类性能大大提高.

4 结论

本文提出一种基于低密度分割密度敏感距离的谱聚类算法, 通过实验分析得到如下结论:

1) 为解决数据分布特征对现有谱聚类算法聚类性能的影响, 本文引入低密度分割密度敏感距离进行数据间相似性测量, 并通过理论推导和实验证明本文算法所采用的距离测度能同时满足数据间的全局一致性和局部一致性特征要求, 使获得的相似度矩阵更符合实际数据的分布情况, 同时通过增加相对密度敏感项来考虑样本局部分布特征的影响, 大大提高了算法聚类性能.

2) 由于最近邻个数的确定以及特征向量的选择对谱聚类算法的聚类性能影响很大, 本文进一步给出了基于 SC 指标的 k 近邻图 k 值选取方法和基于谱熵贡献率的特征向量选取方法. 实验部分通过对仿真和 UCI 数据集在不同参数设置下聚类性能的对比分析, 结果验证了上述两种方法的正确性和有效性.

3) 实验最后对不同谱聚类算法的鲁棒性进行了比较, 结果表明本文 LDSDSD-SC 算法在所有比较的 5 种算法中鲁棒性能最好, 这说明本文算法采用的低密度分割密度敏感距离不仅能同时满足聚类的全局一致性和局部一致性的特征, 且能有效防止孤立噪声和 "桥" 噪声对聚类性能的影响, 使得算法的鲁棒性显著提高. 需要说明的是, 如何选取特征向量的个数以及在聚类个数未知的情况下如何确定聚类个数都将是本课题下一阶段研究的重点.

附录 A

低密度分割密度敏感距离满足测度的 4 个性质, 证明如下:

1) 自反性

当且仅当 $\boldsymbol{x}_i = \boldsymbol{x}_j$;

证明.

当 $\boldsymbol{x}_i = \boldsymbol{x}_j$, 因 $d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})$ 代表图 G 上顶点 $\boldsymbol{x}_i$ 到 $\boldsymbol{x}_j$ 最短路径上任意相邻两点 $\boldsymbol{p}_k$, $\boldsymbol{p}_{k+1}$ 的欧氏距离. $\boldsymbol{x}_i = \boldsymbol{x}_j$, 则 $\boldsymbol{x}_i$ 到 $\boldsymbol{x}_j$ 最短路径为零, 因此 $d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1}) = 0, k = 1, \cdots, |\boldsymbol{p}|$.

$$D^{\rho}_{i,j} = \frac{1}{\rho} \ln \left(1 + \sum_{k=1}^{|\boldsymbol{p}|} \left(\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})}{\delta_k \delta_{k+1}}} - 1 \right) \right) = \frac{1}{\rho} \ln(1+0) = 0$$

□

2) 非负性

证明.

$$\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})}{\delta_k \delta_{k+1}}} \geq 1, \left(\mathrm{e}^{\frac{\rho d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})}{\delta_k \delta_{k+1}}} - 1 \right) \geq 0,$$

则:

$$1 + \sum_{k=1}^{|\boldsymbol{p}|} \left(\mathrm{e}^{\rho d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})/(\delta_k \delta_{k+1})} - 1 \right) \geq 1$$

$$\frac{1}{\rho} \ln \left(1 + \sum_{k=1}^{|\boldsymbol{p}|} (\mathrm{e}^{\rho d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})/(\delta_k \delta_{k+1})} - 1) \right) \geq 0$$

□

3) 对称性

证明.

因为

$$\frac{d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})}{\delta_k \delta_{k+1}} = \frac{d(\boldsymbol{p}_{k+1}, \boldsymbol{p}_k)}{\delta_{k+1} \delta_k}$$

所以

$$D^{\rho}_{i,j} = D^{\rho}_{j,i}$$

□

4) 三角不等式

证明.

假设

$$D^{\rho}_{i,j} > D^{\rho}_{i,m} + D^{\rho}_{m,j}$$

则根据定义, $\min_{\boldsymbol{p} \in P_{ij}} \sum_{k=1}^{|\boldsymbol{p}|} \left(\mathrm{e}^{\rho d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})/(\delta_k \delta_{k+1})} - 1 \right)$ 就至少应该取 $\min_{\boldsymbol{p} \in P_{im}} \sum_{k=1}^{|\boldsymbol{p}|} \left(\mathrm{e}^{\rho d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})/(\delta_k \delta_{k+1})} - 1 \right) + \min_{\boldsymbol{p} \in P_{mj}} \sum_{k=1}^{|\boldsymbol{p}|} \left(\mathrm{e}^{\rho d(\boldsymbol{p}_k, \boldsymbol{p}_{k+1})/(\delta_k \delta_{k+1})} - 1 \right)$, 因此假设不成立, 由此证明 $D^{\rho}_{i,j} < D^{\rho}_{i,m} + D^{\rho}_{m,j}$ □

References

1 Qin F P, Zhang A W, Wang S M, Meng X G, Hu S X, Sun W D. Hyperspectral band selection based on spectral clustering and inter-class separability factor. *Spectroscopy and Spectral Analysis*, 2015, **35**(5): 1357−1364

2 Goyal S, Kumar S, Zaveri M A, Shukla A K. Fuzzy similarity measure based spectral clustering framework for noisy image segmentation. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 2017, **25**(4): 649−673

3 Choy S K, Lam S Y, Yu K W, Lee W Y, Leung K T. Fuzzy model-based clustering and its application in image segmentation. *Pattern Recognition*, 2017, **68**: 141−157

4 Akbarizadeh G, Rahmani M. Efficient combination of texture and color features in a new spectral clustering method for PolSAR image segmentation. *National Academy Science Letters*, 2017, **40**(2): 117−120

5 Jang M, Song Y, Chang J W. A parallel computation of skyline using multiple regression analysis-based filtering on MapReduce. *Distributed & Parallel Databases*, 2017, **35**(3−4): 383−409

6 Mao Guo-Jun, Hu Dian-Jun, Xie Song-Yan. Models and algorithms for classifying big data based on distributed data streams. *Chinese Journal of Computers*, 2017, **40**(1): 161−174
(毛国君, 胡殿军, 谢松燕. 基于分布式数据流的大数据分类模型和算法. 计算机学报, 2017, **40**(1): 161−174)

7 Yang Y, Shen F M, Huang Z, Shen H T, Li X L. Discrete nonnegative spectral clustering. *IEEE Transactions on Knowledge & Data Engineering*, 2017, **29**(9): 1834−1845

8 Li Y G, Zhang S C, Cheng D B, He W, Wen G Q, Xie Q. Spectral clustering based on hypergraph and self-re-presentation. *Multimedia Tools & Applications*, 2017, **76**(16): 17559−17576

9 Hosseini M, Azar F T. A new eigenvector selection strategy applied to develop spectral clustering. *Multidimensional Systems & Signal Processing*, 2017, **28**(4): 1227−1248

10 Tan P N, Steinbach M, Kumar V. *Introduction to Data Mining*. Boston, MA: Addison Wesley, 2005.

11 Liu W F, Ye M, Wei J H, Hu X X. Compressed constrained spectral clustering framework for large-scale data sets. *Knowledge-Based Systems*, 2017, **135**: 77−88

12 Zhang R, Nie F P, Li X L. Self-weighted spectral clustering with parameter-free constraint. *Neurocomputing*, 2017, **241**: 164−170

13 Li X Y, Guo L J. Constructing affinity matrix in spectral clustering based on neighbor propagation. *Neurocomputing*, 2012, **97**: 125−130

14 Zhan Q, Mao Y. Improved spectral clustering based on Nyström method. *Multimedia Tools & Applications*, 2017, **76**(19): 20149−20165

15 Chen J S, Li Z Q, Huang B. Linear spectral clustering superpixel. *IEEE Transactions on Image Processing*, 2017, **26**(7): 3317−3330

16 Wang Ling, Bo Lie-Feng, Jiao Li-Cheng. Density-sensitive spectral clustering. *Acta Electronica Sinica*, 2007, **35**(8): 1577−1581
(王玲, 薄列峰, 焦李成. 密度敏感的谱聚类. 电子学报, 2007, **35**(8): 1577−1581)

17 Wu Jian, Cui Zhi-Ming, Shi Yu-Jie, Sheng Sheng-Li, Gong Sheng-Rong. Local density-based similarity matrix construction for spectral clustering. *Journal on Communications*, 2013, **34**(3): 14−22
(吴健, 崔志明, 时玉杰, 盛胜利, 龚声蓉. 基于局部密度构造相似矩阵的谱聚类算法. 通信学报, 2013, **34**(3): 14−22)

18 Rodriguez A, Laio A. Clustering by fast search and find of density peaks. *Science*, 2014, **344**(6191): 1492−1496

19 Yu J, Kim S B. Density-based geodesic distance for identifying the noisy and nonlinear clusters. *Information Sciences*, 2016, **360**: 231−243

20 Qian P J, Jiang Y Z, Wang S T, Su K H, Wang J, Hu L Z, et al. Affinity and penalty jointly constrained spectral clustering with all-compatibility, flexibility, and robustness. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, **28**(5): 1123−1138

21 Wang B J, Zhang L, Wu C L, Li F Z, Zhang Z. Spectral clustering based on similarity and dissimilarity criterion. *Pattern Analysis & Applications*, 2017, **20**(2): 495−506

22 Zhou Lin, Ping Xi-Jian, Xu Sen, Zhang Tao. Cluster ensemble based on spectral clustering. *Acta Automatica Sinica*, 2012, **38**(8): 1335−1342
(周林, 平西建, 徐森, 张涛. 基于谱聚类的聚类集成算法. 自动化学报, 2012, **38**(8): 1335−1342)

23 Cheng Hao-Xiang, Wang-Jian. Support vector data description based on fast clustering analysis. *Control and Decision*, 2016, **31**(3): 551−554
(程昊翔, 王坚. 基于快速聚类分析的支持向量数据描述算法. 控制与决策, 2016, **31**(3): 551−554)

24 Zhou H D, Shi T L, Liao G L, Xuan J P, Su L, He Z Z, et al. Using supervised Kernel entropy component analysis for fault diagnosis of rolling bearings. *Journal of Vibration & Control*, 2017, **23**(13): 2167−2178

25 Yu Y, Wang T, Samworth R J. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 2015, **102**(2): 315−323

26 Wu T F, Tsai P S, Hu N T, Chen J Y. Combining turning point detection and Dijkstra′s algorithm to search the shortest path. *Advances in Mechanical Engineering*, 2017, **9**(2), doi: 10.1177/1687814016683353

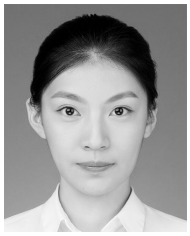
27 Tao Xin-Min, Xu Jing, Fu Qiang, Liu Xing-Li. Kernel fuzzy C-means algorithm based on distribution density and its application in fault diagnosis. *Journal of Vibration and Shock*, 2009, **28**(8): 61−64
(陶新民, 徐晶, 付强, 刘兴丽. 基于样本密度 KFCM 新算法及其在故障诊断的应用. 振动与冲击, 2009, **28**(8): 61−64)

陶新民 东北林业大学工程技术学院教授. 主要研究方向为智能信号处理, 物联网技术, 故障诊断, 软计算方法, 模式识别, 网络安全. 本文通信作者.
E-mail: taoxinmin@nefu.edu.cn

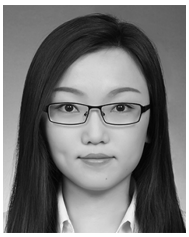
(**TAO Xin-Min** Professor at the College of Engineering & Technology, Northeast Forestry University. His research interest covers intelligent signal processing, internet

of things, fault diagnosis, soft computing, pattern recognition, and network security. Corresponding author of this paper.)

王若彤 东北林业大学工程技术学院硕士研究生. 主要研究方向为物联网技术, 人工智能, 模式识别.
E-mail: celia_wangrt@163.com
(**WANG Ruo-Tong** Master student at the College of Engineering & Technology, Northeast Forestry University. Her research interest covers internet of things, artificial intelligence, and pattern recognition.)



常 瑞 东北林业大学工程技术学院硕士研究生. 主要研究方向为物联网技术, 人工智能, 模式识别.
E-mail: m15765549429@163.com
(**CHANG Rui** Master student at the College of Engineering & Technology, Northeast Forestry University. Her research interest covers internet of things, artificial intelligence, and pattern recognition.)



李晨曦 东北林业大学工程技术学院硕士研究生. 主要研究方向为物联网技术, 人工智能, 模式识别.
E-mail: chenxili0613@163.com
(**LI Chen-Xi** Master student at the College of Engineering & Technology, Northeast Forestry University. Her research interest covers internet of things, artificial intelligence, and pattern recognition.)

刘艳超 东北林业大学工程技术学院硕士研究生. 主要研究方向为物联网技术, 智慧物流, 数字通信技术.
E-mail: 15776802213@163.com
(**LIU Yan-Chao** Master student at the College of Engineering & Technology, Northeast Forestry University. Her research interest covers internet of things, intelligence logistics, and digital communication technology.)