Computer Science and Application 计算机科学与应用, 2020, 10(2), 289-302
Published Online February 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.102030

Hans汉斯

K Neighbor Relationship for Spectral Clustering with Lager Eigengap

Wenmin Tian, Xuan Wang, Chunlin Tian

Shaanxi Normal University (SUUN), Xi'an Shaanxi
Email: sxtianwenmin@163.com

Received: Feb. 1st, 2020; accepted: Feb. 13th, 2020; published: Feb. 20th, 2020

Abstract

Recently, clustering is one of the hot method in unsupervised learning. Spectral clustering is a hot in clustering method and often shows good clustering performance. One crucial step of spectral clustering is constructing a similarity matrix. However, the traditional model cannot consider the distribution structure of the data set well, and it is difficult to truly reflect the similarity between data points. Inspired by density clustering to find the nearest neighbor relationship between data and obtain the optimized similarity matrix. In this paper, we propose a novel construction method of similarity matrix that use a graph with kneighbor relationship (KNRS), considering the kneighborhood distribution of data this model, as a result, the eigengap becomes lager. We have also made comparison with some methods on some common datasets, the experiments show the superiority of our model in the benchmark data sets.

Keywords

Spectral Clustering, Eigengap, Similarity Matrix, Euclidean Distance

基于本证间隙增大的K邻域加权谱聚类算法

田文敏，王　晅，田春霖

陕西师范大学，陕西　西安
Email: sxtianwenmin@163.com

收稿日期：2020年2月1日；录用日期：2020年2月13日；发布日期：2020年2月20日

摘　要

聚类是近些年来无监督学习的热门方法之一。而谱聚类又是聚类方法研究中的热点，并且通常表现出良

文章引用: 田文敏, 王晅, 田春霖. 基于本证间隙增大的 K 邻域加权谱聚类算法[J]. 计算机科学与应用, 2020, 10(2): 289-302. DOI: 10.12677/csa.2020.102030

好的聚类性能。谱聚类中的一个关键步骤是构建相似性矩阵。然而，一些传统的算法模型不能很好地考虑数据集的分布结构，很难真正的反映数据点之间的相似性。针对以上的问题，受密度聚类的启发，我们通过找到数据之间的最近邻关系优化相似矩阵，不仅能更好的反映数据之间真实的邻域关系，并且得到了更好的聚类效果。本文不仅提出了一种新的相似矩阵构造方法，并且证明了由本方法得到结果的拉普拉斯矩阵的本证间隙变得更大。最后，通过实验表明该方法的优越性及本证间隙的增大。

关键词

谱聚类，本证间隙，相似度矩阵，欧氏距离

1. 引言

近些年来，聚类是无监督学习中的热门方法之一。聚类是许多应用过程中的非常重要的一环，包括机器学习，计算机视觉，信号处理，模式识别等。近些年来国内外很多研究学者已经提出了许多的聚类方法，包括基于划分的聚类(K-Means，K-Medoids) [1] [2]，基于层次的聚类(CURE) [3]，基于模糊理论的聚类(FCM) [4]，基于密度的聚类(DBSCAN) [5]，基于图的聚类(Spectral Clustering) [6]，基于核的方法[7]。其中，谱聚类和基于密度的聚类是当下比较热门的方法之一，它们通常表现出良好的聚类性能。基于密度的聚类的优点是能够找到任意形状的聚类并且对噪声具有鲁棒性。密度聚类算法假设聚类结构可以通过样本分布的紧密性来确定。通常密度聚类算法从样本密度的角度考虑样本之间的连通性，并基于可连接样本不断扩展聚类以获得最终的聚类结果。基于密度的噪声应用空间聚类(DBSCAN) [5]是一种众所周知的方法，它基于一组邻域参数(ε 和 MinPts)来表征样本的数据结构。但是这两个参数 ε 和 MinPts 并不直观且难以选择。

谱聚类算法是一种基于图论的聚类算法[6] [8]。谱聚类基于数据集构造无向图，并使用非负相似矩阵表示整个无向图，从而得到一个相似度矩阵再得到拉普拉斯到矩阵，然后进行特征分解。最后通常使用 K 均值算法对结果进行后处理以获得聚类指标。谱聚类具有以下优点：(1) 易于实现；(2) 聚类结果更好，约束更少；(3) 分割非线性可分离数据的效果具有明显的优势；(4) 具有以下优点能够在任何形状的样本空间上聚类并收敛到全局最优解。基于其受欢迎程度和上述几个优点，很多研究人员提出了许多谱聚类算法。Donath W E 等人[9]首先提出了相似矩阵的特征向量。找到相似矩阵的第二个特征向量可以将数据分为两个簇[10]。Hagen 和 Kahng 在 1992 年提出的比率削减在最小化类之间的相似性时引入了类尺度的平均项，并建立了基于矩阵谱分析谱聚类算法[11]。[12]提出的归一化割，解决了一类最小割中只有一个或几个孤立点的倾斜分割问题。[13]提出了 NJW 算法(嵌入式特征空间上的 K-means 算法)。

对于给定的数据集，谱聚类基于成对相似性将数据表征为加权无向图中的节点。谱聚类的目标是将无向加权图分为两个或多个子图，以使子图的内部节点相似而子图之间的节点是不同的。然后，使用非负相似度矩阵表示整个无向图，再进行特征分解，最后通过传统聚类从特征向量中获得最终的聚类分配。总之，相似度矩阵的构建直接影响聚类效果。因此，谱聚类的相似矩阵构造问题是一个关键点[14]。通常，通过相似矩阵获得拉普拉斯矩阵，然后直接选择对应于前 k 个特征值，最后一步再进行 k-means。根据矩阵扰动理论，矩阵特征值之间的差异称为本征间隙(Eigengap) [15]，矩阵扰动理论中的 Davis-Kahan 定理

[16]表明，块对角矩阵的特征向量与矩阵对角关系相似度矩阵的子空间也由 eigengap 确定，eigengap 越大，由 k 个特征向量形成的子空间就越稳定[15]。[17]从图的划分角度出发，提出了一种定义相似度图的新方法。尽管此方法有助于推断数据集的内在结构，但是在无监督条件下获得的相似度矩阵并没有得到根本改善。[18]提出了一种局部密度自适应相似度测量方法。该方法可以放大同一类中具有公共交集的样本之间的相似度，但是不会更改没有公共交集的样本之间的相似度。

尽管谱聚类算法在过去的研究中取得了长足的进步，并且与其他聚类方法相比具有许多优点，但是常用的方法也容易出现以下两个缺点：传统聚类算法构造的相似度矩阵不能很好地考虑数据集的分布结构，很难真实地反映出数据点之间的相似度。最常用的相似度函数是高斯核函数，仅考虑数据的第 K 个最近邻居，不够全面的并且忽略其相邻点的分布。

为了克服上述两个缺点，本文提出了一种新的相似度矩阵构造方法，该方法受到密度聚类的启发，以求出数据之间的 k 个最近邻关系(KNRS)。在实验中，与广泛使用的 k 均值聚类、NJW 谱聚类(SC)、模糊聚类 FCM、SCMSA 谱聚类算法作比较，实验结果表明，我们通过新方法构建的相似矩阵不仅 eigengap 增加了，新的聚类也具有许多基本优点。主要贡献有两个方面：(1) 在我们的新方法中，考虑数据的 k 邻域分布，该模型可以方便，准确地发现数据集中包含的密度信息。可以反映数据的内在结构。(2) 关于聚类结果，eigengap 变得更大，因此矩阵分区稳定。

本文的其余部分安排如下。在第二节中，我们介绍了有关谱聚类的相关工作。在第 3 节中，我们提出了基于相似度矩阵聚类算法的新方法。在第 4 节中，对各类数据集的实验结果进行了分析。在第 5 节中总结本文。

2. 谱聚类算法及分析

2.1. 谱聚类

谱聚类是一种基于图论的聚类方法，目的是将聚类转化为图分割。基于图论的最佳划分标准：最大化位于同一子图中的样本点之间的相似度，并且最小化位于不同子图中的样本点之间的相似度。谱聚类可以大致分为以下步骤：第一步：预处理构造表示数据集的图形和相似矩阵。(计算度矩阵 D，度矩阵是对角矩阵，其元素是点的度数。)第二步：谱表达，根据相似矩阵，得到相应的正则化或不规则拉普拉斯矩阵，计算 L 的第一个特征向量。通过拉普拉斯矩阵的特征分解，得到特征值及其对应的特征向量。第三步：选择特征向量以形成特征空间，并通过聚类算法(例如 K-means 算法)将特征空间离散化(可以在最后阶段使用除 k-means 之外的简单算法，例如 simple linkage，k-lines，elongated k-means，混合模型等)最后获得聚类结果。

从图的划分角度来看，谱聚类根据所使用的划分标准将算法分为两类，即 2-way 谱聚类和 k-way 谱聚类。2-way 谱聚类：Perona 和 Freeman 提出的 PF [19]，它是最简单的迭代 2 路分区算法，提出与相似矩阵 W 的第一特征向量聚类。Shiand Malik 在 2000 年提出了 SM 方法[12] SM 算法对获得的拉普拉斯矩阵进行特征分解，并采用第二个小的特征向量(即 Fiedler 向量)进行聚类。在 Fiedler 向量中大于某个值的数据点属于一个类别，小于某个值的数据点属于另一个类别。Scott 等人提出了 SLH 重定位方法[20]。Ding 等人提出了 Mcut 方法[21]，Mcut 算法可以产生更加均衡的分区结果，尤其是当类之间的重叠很大时，效果更加明显。近年来，已经发现使用更多特征向量并直接计算 k-way 分割会得到更好的聚类结果[22] [23]。NJW 算法通常被称为最经典的多向谱聚类算法，由 Ng 等人[13]提出，该算法直接选择与 Laplacian 矩阵的前 k 个特征值相对应的特征向量进行正则化。MS 方法是由 Meila 等人[24]提出的，Meila 将相似性解释为马尔可夫链中的随机游动，并分析了该随机游动的概率传递矩阵。

2.2. 相似度矩阵

给定数据集$\{x_i\}_{i=1}^n$，$x_i \in R^d$ 其中 d 是每个点的维数。我们首先将这些数据点映射到无向图$G=(V,E)$，可以得到一个$n\times n$矩阵 W 表示整个无向图，其元素 W_{ij} 表示边缘权重且$W_{IJ}=W_{JI}$，表示两个顶点之间的相似性。构造相似矩阵的大多数方法都是数据集的稀疏表示，并且具有计算优势。以下是一些构建方法：ε-邻图，k 最近邻图和完全连通图。一般通常使用全连接图的方法但这种构造仅在相似度函数对局部邻域建模时才有用。选择相似度函数会极大影响之后谱聚类的结果。点的特征在于它们的成对相似性，即与每对点(x_i,x_j)相关联的是相似度值 $f\left(d\left(x_i,x_j\right);\theta\right)$。相似性矩阵$W\in R_{n\times n}$由项 w_{ij} 组成，这些项可以是：余弦型，模糊型，高斯类型。在大多数情况下选择高斯核函数，而其他相似性(例如余弦相似性)用于特定应用。最初提出相似矩阵的特征向量用于划分数据是由 Donath 等人提出[9]。Fiedler 发现相似矩阵的第二特征向量可以将数据分为两个簇[10]，Hagen 和 Kahng 在 1992 年提出的“比率削减”在最小化类之间的相似性时引入了类尺度平均项，并建立了基于矩阵谱分析[11]，之后，学者们也提出了许多构建相似度矩阵的方法。R. Liu 等。提出的 FLR-MRW 方法假定正则矩阵通过对正则矩阵的核规范项进行正则化来假设其亲和度较低[25]。BD-LRR 关注子空间聚类的相似度矩阵的构建[14]。提出了局部密度自适应相似度测量的频谱聚类 SC-DA 方法[18]。[26]提出了一种基于邻域传播在谱聚类中构造亲和度矩阵的方法。[27]提出了一种基于模糊的学习算法，用于 AFSSC 谱聚类。

2.3. 本征间隙

理想情况下，能否在相似度矩阵中反映相互分离的 k 个数据集的间隔，应该使该阵对角线上分布 k 个全 1 分块矩阵，其余位置都为 0。实际上，构造这样的亲和力矩阵是困难的。通过矩阵扰动理论已知，实际相似度矩阵对角线上的块矩阵元素不全为 1，块对角矩阵外的元素也不全为 0。而分别是全 1 矩阵加上一个负的扰动量和一个正扰动量。根据矩阵摄动理论,由相似度矩阵得到的拉普拉斯矩阵的第 k 个和第 $k+1$ 个特征值之间的差称之为本征间隙(Eigengap)，本征间隙越大，选取的 k 个特征向量所构成的子空间就越稳定[15]。

在线性代数中，线性算子的特征是两个连续特征值之间的差异，其中特征值以升序排序。在谱聚类中，eigengap 通常称为谱间隙：(以钱德勒·戴维斯(Chandler Davis)和威廉·卡汉(William Kahan)的名字命名的戴维斯·卡汉定理，用 eigengap 表示算子的本征空间在扰动下如何变化。块对角相似度矩阵矩阵和实感矩阵也由 eigengap 确定，eigengap 越大，由 k 个特征向量稳定形成的子空间就越多[15] [28]。[29]基于 eigengap 和正交特征向量提出了自适应谱聚类。

3. KN-SC 算法及证明

为了清楚地描述我们提出的新方法。我们首先在第 3.1 节中定义基本概念，然后在第 3.2 节中描述建议的 KNRS 方法。

3.1. 概念定义

给定数据集，$D=\{X_1,X_2,\cdots,X_n\}$，d 是每个点的维数。

定义 1：k 近邻图矩阵 A。

点之间的欧几里得距离定义为：

$$d(i,j)=\sqrt{\left(x_{i1}-x_{j1}\right)^2+\left(x_{i2}-x_{j2}\right)^2+\cdots+\left(x_{in}-x_{jn}\right)^2} \tag{1}$$

$$A(i,j)=\begin{cases}1; & \text{if } j\text{th point distance} \le k\text{th points distance}\\ 0; & \text{if } j\text{th point distance} > k\text{th points distance}\end{cases} \tag{2}$$

k 是 KNRS 的参数，k 用于找出每个点的 k 个最近邻。

定义 2：k 近邻关系 ρ 。

$$\rho = \exp\left(-x + \log\left(1-\alpha\right)\right) \tag{3}$$

式(3)是归一化的。是 KNRS 的参数通常用于调整，可选参数的范围易于确定，用户可以轻松快捷地选择最佳参数。(根据重复的实验验证和结果比较，值通常为 0.8，0.9)。x 是矩阵 A 的每一行的总和除以 k。

$$x = \left(\sum_{i=1}^{n} A\right) \Big/ k \tag{4}$$

定义 3：eigengap δ。

定理 1：根据 Davis-Kahan 定理，$H \in R^{n\times n}$ 为对称矩阵，而 $\|\cdot\|$ 为 frobenius 范数或矩阵的 2 范数。将 $\tilde{A} = A + H$ 视作 A 加扰动。令 $S_1 \subset R$ 为间隔。用 $\sigma_{s_1}(A)$ 表示包含在其中的 A 个特征值的集合，用 V_1 表示与所有这些特征值相对应的本征空间，将 S_1 之外的 A 的谱与 S_1 之间的距离定义为：

$$\delta = \min\left\{\left|\lambda - s\right|; \lambda \text{ eigenvalue of } A, \lambda \notin S_1, s \in S_1\right\} \tag{5}$$

V_1 和 $\tilde{V}_1$ 之间的距离 $d\left(V_1, \tilde{V}_1\right)$ 的范围为：

$$d\left(V_1, \tilde{V}_1\right) \le \frac{\|H\|}{\delta} \tag{6}$$

这表明，d 越小越好，越接近理想分割，δ 越大 d 值便越小即效果越好。在本文中我们将 δ 定义为 eigengap。

3.2. KNRS 算法

KNRS 方法包括五个阶段，大致流程如图 1 所示。

第一步：为所有的数据点构造完整的连接图(每个样本的邻居数)，我们使用到完整连接图的欧几里德距离并获得连接矩阵。按距离 d 排序，根据给定 k 的数量选择最近的 k 个点。根据等式(2)，最后我们得到一个最近邻图矩阵 A。

第二步：找到 k 个邻居关系 ρ 。根据等式(3)，我们得到 k 个邻居关系 ρ 。由于相似矩阵具有对称性，因此将 ρ 转换为对称矩阵可得到最终 ρ 。通过 ρ 的模型可以方便准确地发现数据集中包含的密度信息。可以反映数据的内在结构，考虑到邻居对他们密度的影响。生成的新亲和度矩阵可以增加应该在同一类中的点对的相似度，并使近点越来越近。使改进的算法对离群点不那么敏感。

第三步：得到 KNRS 相似度矩阵 W。基于完全连接图和 k 邻居关系 ρ 计算的相似度矩阵，更新相似度矩阵。(此类相似性函数的一个示例是高斯相似度函数，其中参数控制邻域的宽度)。

根据上述步骤获得的更新的相似度矩阵，KNRS 相似度矩阵 W 不仅考虑了其邻居的分布，而且还充分考虑了数据集的分布结构，真正体现了数据点之间的相似度。

第四步，根据相似度矩阵，得到相应的正则化或不规则化的拉普拉斯矩阵，通过对拉普拉斯矩阵进行特征分解，得到特征值及其对应的特征向量。计算 D 矩阵，度矩阵 D 是对角矩阵，其元素 $D_{ii}\left(D_{ii} = \sum_{j=1}^{n} W_{ij}\right)$ 是点 x_i 的度。拉普拉斯矩阵[30] [31] [32]通常分为两类[6]。非正则化拉普拉斯矩阵和正则化的拉普拉斯矩阵。

非正则化拉普拉斯矩阵定义为：

$$L = D - W \tag{7}$$

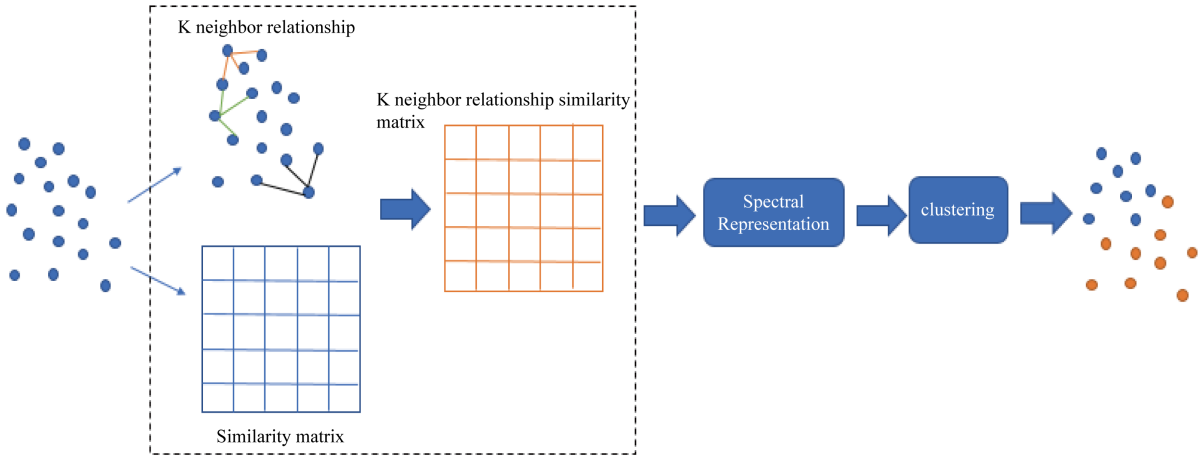

Figure 1. KNRS algorithm flowchart
图 1. KNRS 算法流程图

正则化拉普拉斯矩阵分为两种矩阵(L_{sym} 和 L_{rw})：对称正则化的拉普拉斯矩阵和非对称正则化的拉普拉斯矩阵。

定理 2： G 是具有非负权重的无向图，具有 k 个连通子图 $A_1, A_2, \cdots, A_k$。L_{sym} 和 L_{rw} 的特征值 0 的数量等于图中连接子图的数量。对于 L_{rw} 来说，特征向量 $I_1, I_2, \cdots, I_k$ 划分 k 个特征空间，对于 L_{sym} 来说，可通过特征向量 $D^{\frac{1}{2}}I_{A_i}$ 获得 k 个特征子空间。两个矩阵彼此密切相关，并定义为：

$$L_{sym} = D^{-1/2}LD^{1/2} = I - D^{\frac{-1}{2}}WD^{\frac{1}{2}} \tag{8}$$

$$L_{rw} = D^{-1}L = I - D^{-1}W\ . \tag{9}$$

命题 1：方程(8)正则化拉普拉斯矩阵的性质。

对于每一个 $f \in R_n$，我们都有等式(10)。L_{sym} 和 L_{rw} 是正半定值，并且具有非负实值特征值 $0 = \lambda_1 \le \cdots \le \lambda_n$，最小特征值是 0。

$$fLf = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}}\right)^2 \tag{10}$$

由于上述定理 2 和命题 1：

证明 1：证明方程式

$$\begin{aligned} f^{\mathrm{T}}Lf &= f^{\mathrm{T}}\left(I - D^{\frac{-1}{2}}WD^{\frac{-1}{2}}\right)f \\ &= \sum_{i=1}^{n} f_i^2 - \sum_{i,j=1}^{n} f_i f_j \frac{w_{ij}}{\sqrt{d_i d_j}} \\ &= \frac{1}{2}\left(\sum_{i=1}^{n} f_i^2 + \sum_{j=1}^{n} f_j^2 - 2\sum_{i,j=1}^{n} f_i f_j \frac{w_{ij}}{\sqrt{d_i d_j}}\right) \\ &= \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}}\right)^2 \end{aligned} \tag{11}$$

最后，得到拉普拉斯矩阵 L，然后选择特征向量形成特征向量空间，并在特征向量空间中采用聚类算法如 k 均值算法。算法流总结为：

1. 第一步，构建完整的连接图并选择最近的 k 个点，获得 k 个最近的邻居矩阵 W。
2. 第二步，找到 k 邻居关系。
3. 第三步，获得新的相似矩阵 W。
4. 第四步，获得拉普拉斯矩阵 L。
5. 第五步，使用例如 K-均值算法得到最终的聚类结果。

3.3. KNRS 算法相关证明

由于(等式(8))，得到 $L = D^{-\frac{1}{2}}(D-W)D^{\frac{1}{2}}$，eigengap 取决于 L 的特征值，L 的特征值基本上由 $D-W$ 确定。令 $f(\cdot)$ 为计算特征值的函数。

引理 1：$f(D-W)$ 下限的渐近分析：

$$f(D-W) \geq f(e_{\min} I - W) \sim e_{\min} - \lambda \geq e_{\min} - \lambda_{\max} \geq e_{\min} - \min\left(\|W\|_{\infty}, \|W\|_{1}\right) \tag{12}$$

矩阵 A 的每个特征值的模数(绝对值)不超过矩阵 A 的范数 $\|A\|$。$\|\lambda_i\| < \|A\|$。“~”代表渐近分析。

$$\max_{1\leq i\leq n} |\lambda_i| \leq \|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \tag{13}$$

$$\max_{1\leq i\leq n} |\lambda_i| \leq \|A\|_{\infty} = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \tag{14}$$

引理 2：$f(D-W)$ 上限的渐近分析：

$$f(D-W) \leq f(e_{\max} I - W) \sim e_{\max} - \lambda \leq e_{\max} - \lambda_{\min} \leq \min\left(\|W\|_{\infty}, \|W\|_{1}\right) - \lambda_{\min} \tag{15}$$

设 $sup\,(L)$的上限和 $sub\,(L)$下限，则差为Δ ($\Delta = \|\text{eigengap}\|$)。根据引理 1 和引理 2：

$$\begin{aligned} \Delta &= sup(L) - sub(L) \\ &= \left(\min\left(\|W\|_{\infty}, \|W\|_{1}\right) - \lambda_{\min}\right) - \left(e_{\min} - \min\left(\|W\|_{\infty}, \|W\|_{1}\right)\right) \\ &= -\lambda_{\min} - e_{\min} + 2\min\left(\|W\|_{\infty}, \|W\|_{1}\right) \\ &= \min\left(\|W\|_{\infty}, \|W\|_{1}\right) - \lambda_{\min} \ (\text{as sym. matrix } W) \\ &\leq \lambda_{\max} - \lambda_{\min} \leq \lambda_{\max} \end{aligned} \tag{16}$$

引理 3：eigengap 的边界取决于 $\lambda_{\max}$。根据戴维斯–卡汉(Davis-Kahan)定理 1 和方程式(7)，(这里我们将 KNRS 度矩阵和 KNRS 相似度矩阵分别称为 D'、W')，然后，我们对原始 L 和 A 进行运算，效果取决于 H'和 eigengap。

定理 3：由矩阵的谱半径(即矩阵的特征值)不超过矩阵的任何范数可得：

$$\begin{aligned} \Delta &= sup(L) - sub(L) \\ &\leq \lambda_{\max} \\ &\leq \|L + H'\| \\ &\leq \|L\| + \|H'\| \end{aligned} \tag{17}$$

与以前 $\Delta \leq \|L\|$相比，eigengap 变大。

4. 实验

在本节中，我们显示了所提出的模型 KRNA 与一些模型(例如 K 均值，广泛使用的 NJW 谱聚类(SC)、模糊聚类 FCM，SCMSA 算法的结果。显示并列出了实验结果。表 1 有 6 个数据集，其中 4 个来自 UCI [33]，两个来自人工数据集和 Mnist 数据集。

4.1. 数据集

Iris 数据集包含 3 个类别，每个类别有 50 个实例，具有 4 个属性，其中每个类别都表示一种 Iris 植物。Wine 数据集包含三种葡萄酒中 178 种样品中每种的 13 种成分的数量。Soybean 数据集有 47 个样本，35 个属性和 7 个类别。Seeds 数据集具有 210 个样本，7 个属性和 3 个类别。Two circles 的人工数据集在两个圆圈中具有 420 个样本，在两个类别中具有三个属性。Boat 数据集包含 100 个样本，2 个属性和 3 个类别。这两个人工数据集包含两种类型的样本，即中等密度密集区域和周围稀疏区域。MNIST 手写数字数据库的训练集为 60,000 个示例，测试集为 10,000 个示例，784 个属性和 10 个类别用来验证算法不仅对密度均匀的数据集聚类有效，而且对于密度分布不均匀的数据集聚类也有效。标准数据集用于进一步检查算法的有效性和泛化能力。

Table 1. Benchmark datasets
表 1. 实验数据集

数据集	样本数	属性个数	类别
Iris	150	4	3
Wine	178	13	3
Soybean	47	35	7
Seeds	210	7	3
Two circles	420	3	2
Boat	100	2	3
MNIST	60000	784	10

4.2. 聚类评价指标

4.2.1. NMI

本文使用称为标准化互信息(NMI)的外部索引。NMI 是使用最广泛的外部有效性指标之一[34] [35]。它用于度量通过聚类方法获得的聚类标签与基础类标签之间的相似度。两个随机向量 X 和 Y 之间的互信息为：

$$NMI = \frac{I(H,Y)}{\sqrt{H(X)H(Y)}} \tag{18}$$

其中 $I(X,Y)$ 是 X 和 Y 之间的互信息，$H(X)$ 和 $H(Y)$ 分别是 X 和 Y 的熵。具体可以表示为：

$$NMI = \frac{\sum_{k=1}^{C}\sum_{m=1}^{C} n_{k,m}\log\left(\frac{n\times n_{k,m}}{n_k\hat{n}_m}\right)}{\sqrt{\left(\sum_{k=1}^{C} n_k\log\frac{n_k}{n}\right)\left(\sum_{m=1}^{C}\hat{n}_m\log\frac{\hat{n}_m}{n}\right)}} \tag{19}$$

其中 n_k 表示 C_k 类中的元素数，$\hat{n}_m$ 表示属于第 m 类的元素数，$n_{k,m}$ 表示 C 类与第 m 类之间的相交数。上述公式计算的结果越大，聚类效果越好。

4.3. 实验总结分析

4.3.1. 章节标题

在实验中，我们将模型与其他模型的性能进行了比较。表 2 显示了实验的各个聚类方法 NMI 在 6 个数据集(Iris，Wine，Soybean，Seeds，Two circles，Boat)上的结果。关于 NMI 评估指数，在大多数数据集上，我们的 KNRS 优于 *k*-means，FCM，SC 和 SC-MSA。例如，在 Wine 数据集中，实验结果表明，我们的 KNRS 模型在基准数据集上实现了最先进的性能。尽管密度稀疏空间和稠密空间如 Boat 数据集，对谱聚类方法提出了一些挑战，但 KNRS 方法的性能还是分离的很准确。仅在“种子”数据集中，比 SCMSA 稍差一点。表 2 是对 6 个标准数据集的聚类结果。图 2 是由 KNRS 和实际获得的结果，和与 SC 聚类比较的真是结果。例如在 Iris 和 Two circles 数据集上，左边是 KNRS 的结果，右边是真实的标记效果。

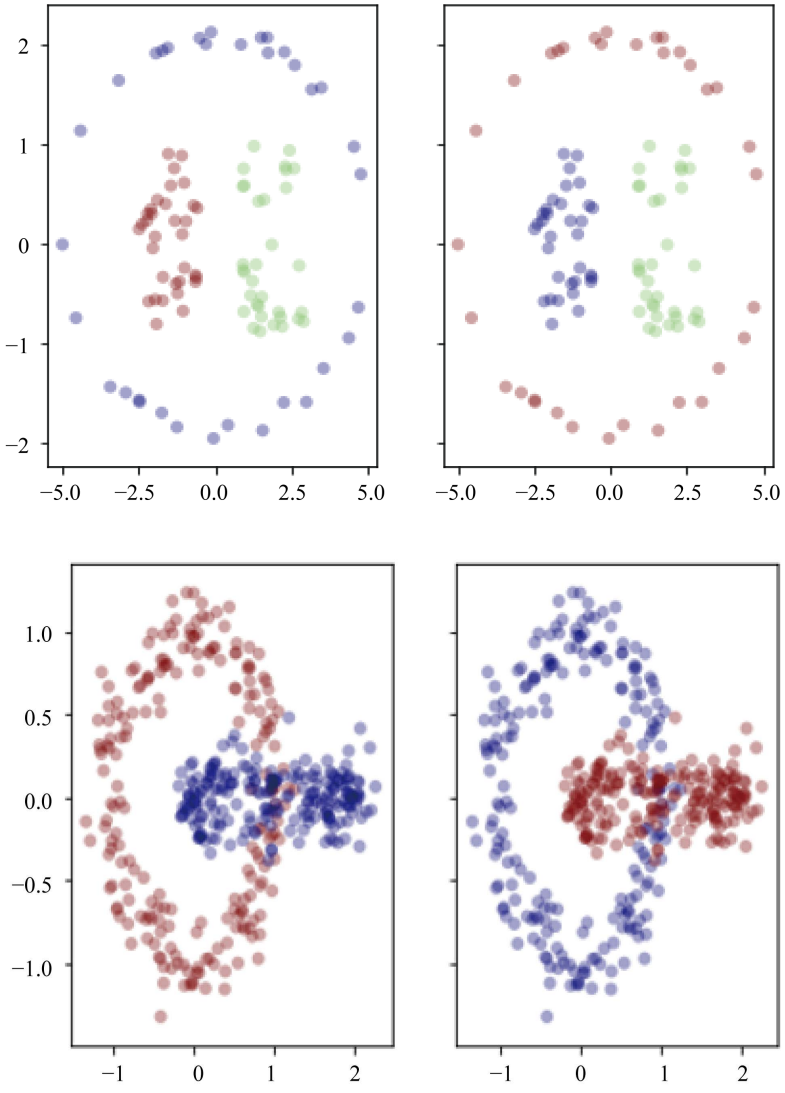

Figure 2. KNRS clustering results on benchmark datasets
图 2. KNRS 聚类在实验数据上的结果

Table 2. Results on datasets by NMI
表 2. 实验数据集 NMI 结果

数据集	*K*-means	FCM	SC	SCMSA	KNRS
Iris	0.7371	0.6551	0.7419	0.7016	**0.7421**
Wine	0.4286	0.4168	0.4149	0.4292	**0.9106**
Soybean	0.6758	0.7450	0.6942	0.7204	**0.8489**
Seeds	0.7025	0.6949	0.6655	**0.8984**	0.6949
Two circles	0.3405	0.6381	0.5405	0.8863	**1**
Boat	0.4500	0.5000	0.5405	0.8863	**1**

MNIST 是手写数字数据库如图 3 所示，对于那些想在真实数据上尝试学习和模式识别方法同时又在预处理和格式化上花费最少精力的人们来说，是一个很好的数据库。本文从 MNIST 的训练集中选择了 500 个图像中的 150 个属性进行实验。NMI 对 MNIST 数据集的聚类结果总结在图 4 中。根据图 4，我们可以看到 KNRS 方法优于其他比较方法。

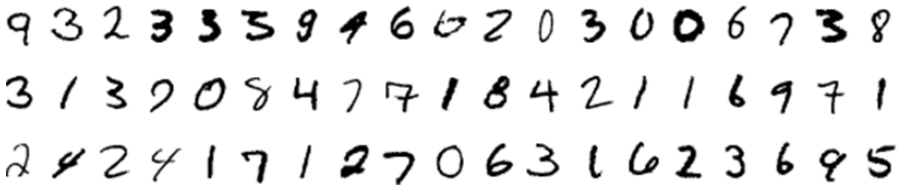

Figure 3. MNIST dataset
图 3. MNIST 数据集

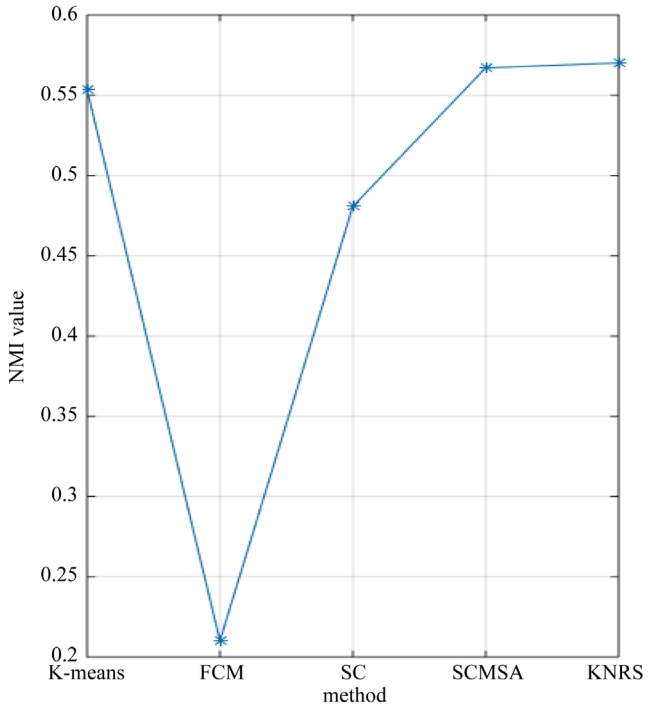

Figure 4. Results on MNIST datasets by NMI
图 4. 在数据集 MNIST 上的 NMI 结果

关于 eigengap 值的实验结果显示在图 5 中。SC-eigengap 表示通过 SC 方法获得的 eigengap 值，KNRS-eigengap 表示通过 KNRS 方法获得的 eigengap 值，large-eigengap 表示增大的 eigengap 的值，由实验可知与 SC 方法相比，KNRS 方法增加了 eigengap。从图 2 和图 5 可以看出，KNRS 的性能优于 SC，KNRS 可以使 eigengap 更大。图 6 展示出了使用 SC 和 KNRS 方法获得的 eigengap 值和 NMI。在图中，我们可以清楚地看到 eigengap 和 NMI 的值增加，这意味着随着 eigengap 的增大，聚类效果确实会越来越好。

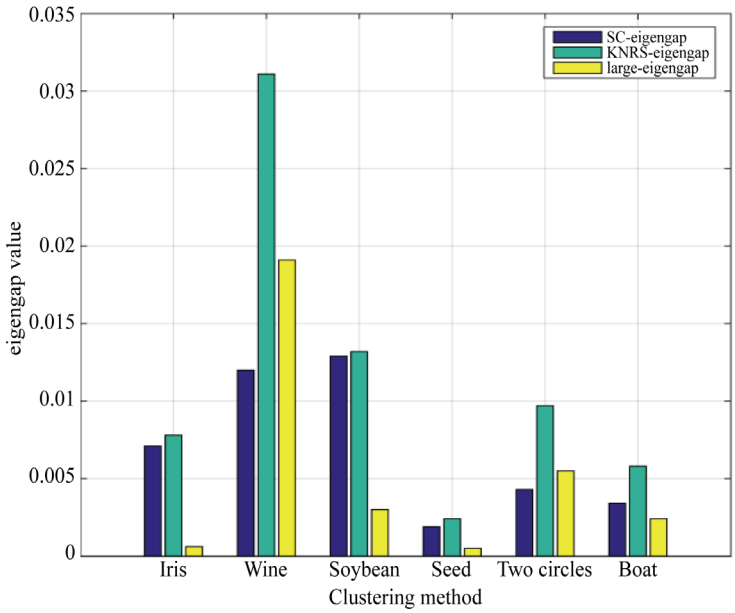

Figure 5. Results of increased eigengap of dataset in KNRS and SC methods
图 5. KNRS 和 SC 方法中 6 组数据集中本征间隙增大值结果

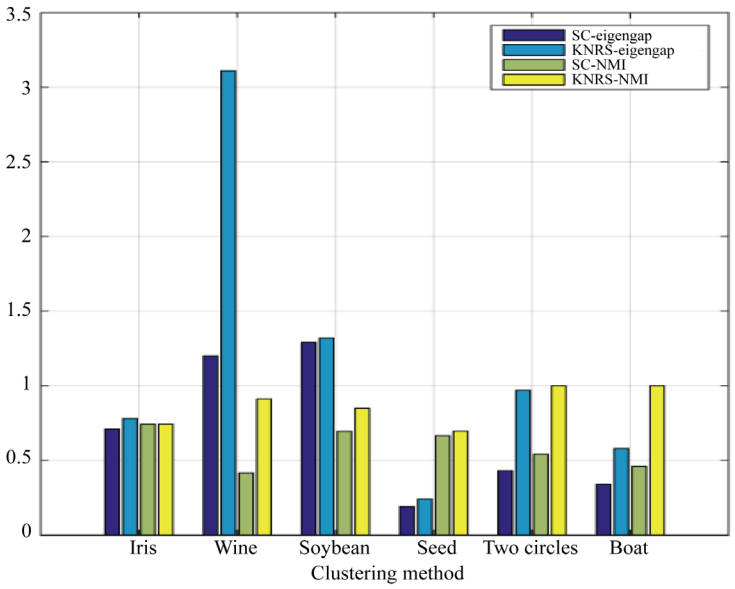

Figure 6. Comparison of the effect of increasing the eigenspace on NMI on 6 sets of datasets
图 6. 6 组数据集上增大本征间隙对 NMI 影响的比较

谱聚类的一个关键步骤是为参数选择合适的值。为了观察参数选择对聚类结果的影响，我们在人工数据集 Boat 上测试了 KNRS 的性能。图 7 描绘了参数 δ 从(0.1，0.2，0.4，0.8，1)和参数 k 从(1，3，5，7，9，10，13，15，16，17，18，19，20，21)以及在 Boat 上的聚类结果的相应精度值。如图 7 所示，随着 k 的增长，聚类效应变得更好，并且在达到最佳效应后趋于稳定。这表明我们的算法对参数是稳定的。结果如图 7 所示。从这个图可以看出，KNRS 对参数不敏感。

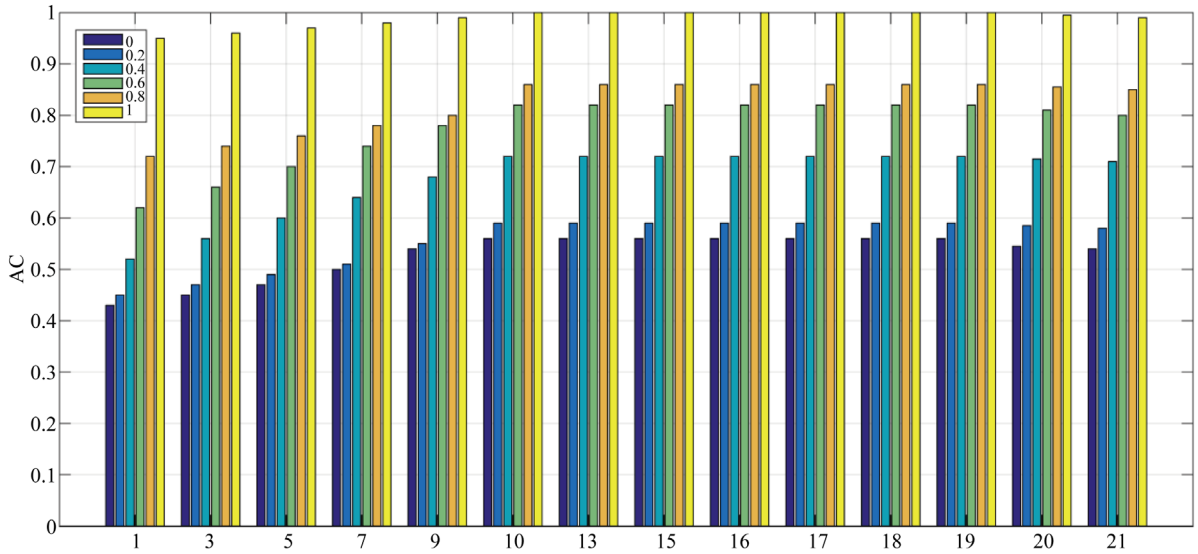

Figure 7. Clustering results of KNRS method for different δ and k. (δ = 0.1, 0.2, 0.4, 0.8, 1); number of neighbors k = 1, 3, 5, 7, 9, 10, 13, 15, 16, 17, 18, 19, 20, 21

图 7. 在不同的 δ = 0.1, 0.2, 0.4, 0.8, 1 与 k = 1, 3, 5, 7, 9, 10, 13, 15, 16, 17, 18, 19, 20, 21 值下 KNRS 聚类结果

5. 总结

在本文中，我们提出了一种新的基于 k 邻域关系构建相似矩阵方法而非传统的用 k 近邻图或者全连接图构建，经证明确实可以得到具有较大 eigengap 的拉普拉斯矩阵，从而提升谱聚类效果。我们的新方法中，考虑数据的 k 邻域分布，该模型可以方便，准确地发现数据集中包含的密度信息。可以反映数据的内在结构。关于聚类结果，eigengap 变得更大，因此矩阵分区稳定。无论密度稀疏空间还是稠密空间，KNRS 都具有良好的性能，KNRS 方法的性能优于其他方法，并且可以更准确地聚类，和提高 eigengap。由实验可以看出我们提出的 KNRS 方法，在七个基准数据集上的实验表明该模型表现良好。我们观察到，所提出的方法明显优于其他一些最新方法。将来，我们计划在其他条件下应用 KNRS，并持续完善这个模型。

致　谢

时光飞逝，一转眼三年就过去了，在这三年里我学到了很多也成长了很多。上学的时候总是想着什么时候可以工作，但是越来越临近毕业季的时候却越发的留恋。回顾读研究生这短短三年时光，感慨颇多，收获颇丰。很感恩和感激每一个帮助过我的人，感谢我的研究生导师王老师，他的和蔼可亲对学术有很高的追求，循循善诱，勤勤恳恳又很理解学生，非常的让人念念不忘。也感谢我的师兄师姐和师弟师妹们，谢谢大家对我的帮助和支持。最后也要感谢我的父母和一直支持的我的亲朋们。我想以后我要用我学到的知识努力的回报社会，做一个有理想有抱负的人，为祖国的繁荣昌盛添砖添瓦。

参考文献

[1] Jain, A.K. (2008) Data Clustering: 50 Years beyond K-Means.

[2] Park, H.S. and Jun, C.H. (2009) A Simple and Fast Algorithm for k-Medoids Clustering. *Expert Systems with Applications*, **36**, 3336-3341. https://doi.org/10.1016/j.eswa.2008.01.039

[3] Guha, S., Rastogi, R. and Shim, K. (1998) Cure: An Efficient Clustering Algorithm for Large Databases. *Information Systems*, **26**, 35-58. https://doi.org/10.1016/S0306-4379(01)00008-4

[4] Bezdek, J.C., Ehrlich, R. and Full, W. (1984) FCM: The Fuzzy C-Means Clustering Algorithm. *Computers & Geosciences*, **10**, 191-203. https://doi.org/10.1016/0098-3004(84)90020-7

[5] Ester, M., Kriegel, H.P., Sander, J., *et al*. (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of International Conference on Knowledge Discovery & Data Mining*, Vol. 96, 226-231.

[6] Luxburg, U.V. (2007) A Tutorial on Spectral Clustering. *Statistics & Computing*, **17**, 395-416. https://doi.org/10.1007/s11222-007-9033-z

[7] Schlkopf, B., Smola, A. and Mller, K. (1998) Nonlinear Component Analysis as a Kernel Eigen Value Problem. *Neural Computation*, **10**, 1299-1319.

[8] Xu, D. and Tian, Y. (2015) A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, **2**, 165-193. https://doi.org/10.1007/s40745-015-0040-1

[9] Donath, W.E. and Hoffman, A.J. (1973) Lower Bounds for the Partitioning of Graphs. *IBM Journal of Research & Development*, **17**, 420-425. https://doi.org/10.1147/rd.175.0420

[10] Fiedler, M. (1976) Algebraic Connectivity of Graphs. *Czechoslovak Mathematical Journal*, **23**, 298-305.

[11] Hagen, L. and Kahng, A.B. (2002) New Spectral Methods for Ratio Cut Partitioning and Clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, **11**, 1074-1085. https://doi.org/10.1109/43.159993

[12] Shi, J. and Malik, J. (2000) Normalized Cuts and Image Segmentation. Departmental Papers (CIS), 107.

[13] Ng, A.Y., Jordan, M.I. and Weiss, Y. (2002) On Spectral Clustering: Analysis and an Algorithm. In: *Advances in Neural Information Processing Systems*, Springer, The Netherlands, 849-856.

[14] Feng, J., Lin, Z., Xu, H. and Yan, S. (2014) Robust Subspace Segmentation with Block-Diagonal Prior. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 23-28 June 2014, 3818-3825. https://doi.org/10.1109/CVPR.2014.482

[15] Xia, T., Cao, J., Zhang, Y.-D. and Li, J.-T. (2009) On Defining Affinity Graph for Spectral Clustering through Ranking on Manifolds. https://doi.org/10.1016/j.neucom.2009.03.012

[16] Zhang, X., Li, J. and Yu, H. (2011) Local Density Adaptive Similarity Measurement for Spectral Clustering. *Pattern Recognition Letters*, **32**, 352-358. https://doi.org/10.1016/j.patrec.2010.09.014

[17] 孙继广. 矩阵扰动分析[M]. 北京: 科学出版社, 2001: 146-160.

[18] Davis, C. and Kahan, W.M. (1970) The Rotation of Eigenvectors by a Perturbation. *SIAM Journal on Numerical Analysis*, **7**, 1-46. https://doi.org/10.1137/0707001

[19] Perona, P. and Freeman, W. (1998) A Factorization Approach to Grouping. In: *European Conference on Computer Vision*, Springer, Netherlands, 655-670. https://doi.org/10.1007/BFb0055696

[20] Scott, G.L. and Longuet-Higgins, H.C. (1990) Feature Grouping by 'Relocalisation' of Eigenvectors of the Proximity Matrix. BMVC, 1-6. https://doi.org/10.5244/C.4.20

[21] Ding, C.H., He, X., Zha, H., Gu, M. and Simon, H.D. (2001) A Min-Max Cut Algorithm for Graphpartitioning and Data Clustering. *Proceedings* 2001 *IEEE International Conference on Data Mining*, San Jose, CA, 29 November-2 December 2001, 107-114.

[22] 王丽. 图论在算法设计中的应用[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2010.

[23] Malik, J., Belongie, S., Leung, T. and Shi, J. (2001) Contour and Texture Analysis for Image Segmentation. *International Journal of Computer Vision*, **43**, 7-27.

[24] Meila, M. and Shi, J. (2001) Learning Segmentation by Random Walks. In: *Advances in Neural Information Processing Systems*, Springer, The Netherlands, 873-879.

[25] Liu, R., Lin, Z., and Su, Z. (2014) Learning Markov Random Walks for Robust Subspace Clustering and Estimation. *Neural Networks*, **59**, 1-15. https://doi.org/10.1016/j.neunet.2014.06.005

[26] Li, X.-Y. and Guo, L.-J. (2012) Constructing Affinity Matrix in Spectral Clustering Based on Neighbor Propagation.

Neurocomputing, **97**, 125-130. https://doi.org/10.1016/j.neucom.2012.06.023

[27] Li, Q., Ren, Y., Li, L. and Liu, W. (2016) Fuzzy Based Affinity Learning for Spectral Clustering. *Pattern Recognition*, **60**, 531-542. https://doi.org/10.1016/j.patcog.2016.06.011

[28] 田铮, 李小斌, 句彦伟. 谱聚类的扰动分析[J]. 中国科学, 2007, 37(4): 527-543.

[29] 孔万增, 孙志海, 杨灿. 基于本征间隙与正交特征向量的自动谱聚类[J]. 电子学报, 2010, 38(8): 1980-1985.

[30] Bhatia, R. (2013) Matrix Analysis. Springer Science & Business Media, New York.

[31] He, X., Cai, D. and Niyogi, P. (2006) Laplacian Score for Feature Selection. In: *Advances in Neural Information Processing Systems*, Springer, The Netherlands, 507-514.

[32] Xiang, T. and Gong, S. (2008) Spectral Clustering with Eigenvector Selection. *Pattern Recognition*, **41**, 1012-1029. https://doi.org/10.1016/j.patcog.2007.07.023

[33] Asuncion, A. and Newman, D. (2007) UCI Machine Learning Repository.

[34] Strehl, A. and Ghosh, J. (2002) Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, **3**, 583-617.

[35] Jin, R., Kang, F. and Ding, C.H. (2006) A Probabilistic Approach for Optimizing Spectral Clustering. In: *Advances in Neural Information Processing Systems*, Springer, The Netherlands, 571-578.