Computer Science and Application 计算机科学与应用, 2019, 9(5), 969-984
Published Online May 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.95110

Hans汉斯

Spectral Clustering Algorithm Based on K-Threshold

Meimei Guo, Xuan Wang

Shaanxi Normal University (SUUN), Xi'an Shaanxi
Email: m15929925619@163.com, Wxuan@snnu.edu.cn

Received: May 10th, 2019; accepted: May 23rd, 2019; published: May 30th, 2019

Abstract

In the traditional NJW spectral clustering algorithm, similarity measurement between samples is determined by Euclidean distance and Gaussian kernel function. Euclidean distance is based on the whole sample set, ignoring the correlation between local samples. Moreover, the value of Gaussian kernel function is also obtained through multiple experiments, which is easy to fall into local optimization and greatly affects the clustering results. Aiming at the above problems, a new spectral clustering algorithm is proposed, which introduces the concepts of K neighborhood, shortest path and standard deviation. The parameters of similarity measurement between samples are reconstructed, and the relationship between samples is reconstructed with sample standard deviation, so as to modify the value of Gaussian kernel function and make it closer to the characteristics of samples. Through the experimental analysis, the algorithm not only inherits the convergence of spectral clustering algorithm to the whole world, and is applicable to data of various shapes, but also overcomes the limitations of similarity measurement caused by Euclidean distance and Gaussian kernel function, improves the accuracy of the algorithm, and enhances the reliability of clustering results.

Keywords

Euclidean Distance, Gaussian Kernel Function, K-Neighborhood, Shortest Path, The Standard Deviation

基于K阈值的谱聚类算法

郭梅梅，王　晅

陕西师范大学，陕西　西安
Email: m15929925619@163.com, Wxuan@snnu.edu.cn

收稿日期：2019年5月10日；录用日期：2019年5月23日；发布日期：2019年5月30日

文章引用: 郭梅梅, 王晅. 基于 K 阈值的谱聚类算法[J]. 计算机科学与应用, 2019, 9(5): 969-984.
DOI: 10.12677/csa.2019.95110

摘　要

在传统的NJW谱聚类算法中，样本间的相似性度量由欧式距离和高斯核函数公共决定；而欧式距离是基于整个样本集合的，忽略了局部样本之间的关连；且高斯核函数的值也是通过多次实验获得，容易陷入局部最优使得聚类结果大受影响。针对以上问题，提出了一种新的谱聚类算法，算法引入K邻域、最短路径以及标准差的概念；重建样本间的相似性度量的参数，并且用样本标准差重新构建样本间的联系进而修正高斯核函数的值，使其更接近样本本身的特性。通过实验分析，该算法不但继承了谱聚类算法的收敛于全局的特性，适用于各种形状的数据，而且还克服了欧式距离和高斯核函数造成的相似性度量的局限，提高了算法准确率，增强聚类结果的可信度。

关键词

欧式距离，高斯核函数，K邻域，最短路径，标准差

1. 引言

聚类分析是机器学习和模式识别领域中的一个重要分支，在人们认识和探索事物内在联系的过程中扮演着重要的角色；它能发现输入数据集的内在属性，并可以将其“分类”[1]；聚类是无监督的，它可以根据事物内在的联系将其分到不同的簇中[2]；聚类和传统的分类方法虽相似但却不同；聚类是通过挖掘数据内在联系将其分成多个簇，使得位于同一个簇中的数据具有较高的相似性，而不同簇中的数据有较大差异性[3]。传统的聚类算法有：K-means、EM 聚类法、硬聚类、模糊聚类法等[4]，但这些算法都是建立在凸球形的样本空间上，因此，当样本空间不为凸时，聚类算法就会陷入局部最优；为了克服这一缺点，谱聚类算法被提出[5]。

相比于其它聚类算法，谱聚类是利用原始数据集的亲和矩阵的特征分解矩阵对数据进行分类[6] [7] [8] [9] [10]；且它不易受数据形状影响，不易陷入局部最优，算法框架简明，容易执行。但其自身也存在一些缺点，例如：计算量较大，构造相似度矩阵时对参数较为敏感等，这些问题至今还没有得到有效的解决。

为了解决谱聚类算法存在的一些缺点，科研工作者们做出了许多的探索与研究。Donath 和 Hoffman 提出基于邻接矩阵的特征向量的图划分方法[11]；Fiedler 指出了图的二分性与图的 Laplacian 矩阵的第二特征向量之间的关系[12]，且在各个领域得到了广泛的应用。Xin-Ye Lin 和 Li-jie Guo 提出离散非负谱聚类算法[13]，利用邻域关系传播原理提出了一种新的亲和矩阵生成方法，并给出了相应的邻域关系传播算法。Yessica Nataliani 和 Miin-Shen Yang 提出基于幂高斯核函数的谱聚类算法[14]，此算法用一种改进的相关比较方法来估计高斯核相似度函数中的幂参数，利用数据点间最小距离的最大值以获得更好的聚类结果。本文基于多路规范割集准则，提出了一种新的谱聚类算法。

2. 谱聚类算法

谱聚类算法的思想源于谱图理论[15]，将每个数据样本看作是图中 G 中的顶点 V，根据样本间的相似程度，将顶点间的边 E，赋予权值 W，这样就得到一个基于样本相似度的无向加权图 $G=(V,E)$ [16]，

因此，聚类问题就转化为对图 G 的划分问题；基于图论的划分准则的优劣直接影响着聚类结果的好坏。常见的划分准则有：最小割集准则，平均割集准则，规范割集准则[17]、最小最大割集准则[18]、比例割集准则[19]、多路规范割集准则等。

谱聚类算法根据其所使用的划分准则可以分为二路谱聚类算法和多路谱聚类算法；二路谱聚类算法使用 2-way 划分准则，例如：PF 算法、SM 算法、SLH 算法、KVV 算法、Mcut 算法等；多路谱聚类算法使用 k-way 划分准则，例如：MS 算法、NJW 算法等；其中 NJW 算法是由 Ng，Jordan 等人提出的，它选取拉普拉斯矩阵的前 k 个最大特征值对应的特征向量构造新的向量空间 R，在这个新的空间内建起与原始数据的对应关系，然后进行聚类。

在传统的 NJW 谱聚类算法中，样本间的相似性可由包含欧几里德距离的高斯核函数表示，可由公式(1)表示；当高斯核函数确定时，样本间相似性随样本间距离的增加而减小。但欧几里德距离由于其自身局限并不适用于各向异性样本；马氏距离的引入很好的克服了这一点；但对应的协方差矩阵求解却为谱聚类带来巨大的计算量，使谱聚类算法更为复杂。基于以上谱聚类算法中的不足，本文提出来一种新的谱聚类算法——基于 K 阈值的谱聚类算法，简称为 NEW-SC 谱聚类。其继承了谱聚类算法的收敛于全局的特性，适用于各种形状的数据，而且克服了欧式距离造成的相似性度量局限，提高了算法准确率，增强聚类结果的可信度。

由于 NEW-SC 谱聚类算法是基于传统的谱聚类(NJW)上的算法优化，所以 NEW-SC 与 NJW 谱聚类的框架有很多相通点。谱聚类算法(NJW)的框架简单且明确，简单可以分为两部分，一是特征提取，二是利用 K-means [20]聚类特征向量空间；这也是该算法的一大优势，下面就是 NJW 算法的基本框架：

输入：样本集 $X=\{X_1,X_2,X_3,\cdots,X_n\}$，$X_1=[X_{11},X_{12},\cdots,X_{1m}]$，$n$ 为样本数，m 为样本维度，聚类数为 C；

1) 由输入样本集合 X 计算样本间欧氏距离 D_d；$D=\{d_{ij}\}$；d_{ij} 为样本 X_i 与 X_j 之间的欧氏距离，由公式(2)可得；

2) 相似度矩阵 $S=\{S_{ij}\}$，可以根据给定的公式(1)计算得到；D 为度矩阵，是由相似矩阵 S 得到的对角阵，同时由公式(3)得到拉普拉斯矩阵 L；

3) 选出 L 矩阵中的前 C 个最大特征值所对应的特征向量，构成新的特征向量空间 Z；

4) 用 K-means 算法对特征向量空间 Z 进行聚类，得到聚类结果 Y。

输出：Y

$$d_{ij}=\sqrt{\left(x_{i1}-x_{j1}\right)^2+\left(x_{i2}-x_{j2}\right)^2+\cdots+\left(x_{im}-x_{jm}\right)^2} \tag{1}$$

$$S_{ij}=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2\sigma^2}\right) \tag{2}$$

$$L=D^{-\frac{1}{2}}SD^{-\frac{1}{2}} \tag{3}$$

3. NEW-SC 谱聚类算法

谱聚类算法的核心思想是将输入数据转换为新的特征空间，然后在新的特征空间中应用 K-means 聚类方法[21]；而 NEW-SC 算法主要是在 NJW 的算法框架上的算法改进；在新方法中，主要针对相似度量函数进行优化，通过对函数中的距离度量、高斯核函数以及加权项三个方面的分析研究，优化提升，使得我们的算法较原算法更具有优越性。

欧式距离只考虑两两样本之间的相关性，不但忽略了与其他样本之间的整体关连，而且欧式距离对

于多维样本的不同维度“一视同仁”，不利于样本特征的提取；而马氏距离引入很好的克服这一缺点，它的协方差矩阵给样本的特征提供权值，使得样本特征更加鲜明；却为聚类算法引入巨大的计算量。K近邻的引入不但可以克服马氏距离计算量大的问题，还可以克服欧式距离点与点的关联模式，建立邻域相关性。

即使引入了K近邻，但也需要距离度量来建立样本与样本集合以及与局部样本之间的联系；先由欧式距离建立联系，在由K近邻选定每个样本前K的近邻值，但是对于不同的样本周围的样本稀疏程度不一，而基于K值的限定也得保留K个有效值，就使得每个样本的近邻域范围小各不相同，即：稀疏样本集的近邻域范围大而稠密样本集的近邻域范围却小，范围的大小仅仅与K值又关。而K阈值的提出便能很好的解决这一问题；在K近邻的基础上选定所需要的近邻的样本值，针对每个样本取其K近邻样本集合中的标准差，标准差是一组数据平均值分散程度的一种度量，该值就是每个样本的K阈值；使得邻域范围不但与K值相关，还与样本自身属性紧密相关。

对于K近邻，每一个样本都是统一的，没有考虑样本间的个体差异性；通过K阈值筛选后，每个样本根据自身属性保留了各自K范围内的样本间的相关性，消除了由统一定量而忽略对局部的影响。在K阈值范围内的样本联系中，引入最短路径建立连通图；全图的连接路径都是在已定K阈值路径基础上生成的，所以一些样本间的联系是在其余某些样本间关系形成的，这就使得样本间，样本与集合间的关系都更为紧密。

高斯核函数不确定性导致了聚类结果的不确定性；即：不同的高斯核取值使得实验结果区别很大，在以往的实验中高斯核函数的取值有 0.1、0.5、1、1.5 等；一般是经过大量实验，由最佳的实验结果来决定取值范围，无法说明核函数取值与样本之间的关联。而且这种取值方法耗时，得到的值也可能是局部最优。也有学者提出自动选择高斯核值 σ，其实就是对一些 σ 值重复运行它们的聚类算法，从中选择聚类失真最少的 σ 值[21]。

我们提出用样本标准差给高斯核函数取值，不但避免了大量的重复实验，而且还使影响实验结果的因素更多与样本本身相关。在之前取K阈值时，对每个样本都定义了相应的“邻域”，此时的标准差的取值就来自于“邻域”的样本集合；将样本与“邻域”样本集合之间的关系也加入到影响聚类结果的因素之中，减少了聚类结果的随机性。

相似性度量更多的依赖距离度量函数和高斯函数，但在此基础上如果变更相应的权值函数，也会对实验结果产生明显影响。所以新的权重参量引入就很关键；新的权重参数是基于 K 近邻算法提出的。K近邻算法确定了每个样本的K阈值，取各个阈值范围内的样本的度 N_i，再取K阈值下样本集的度 N，因此，权值为：$Q=N/N_i$。对于一个具体的样本集合，当K阈值确定以后，N 值就是一个固定值，而局部度 N_i 也会因此确定；对于一个较为“稀疏”的局部，Q 值较大；而较为“稠密”的局部，Q 值较小；将其引入相似性度量就越大，使得局部的相关性更强。

通过以上三点，对谱聚类算法的优化提升，使得我们的算法较原算法更具有优越性，使信息量中包含的相似性更加全面，也提高了聚类算法的性能。其中，最短路径的应用也是关键，它能使我们的流行数据算法具有良好的聚类效果。下面是我们基于以上改进后提出的新算法(NEW-SC)的基本思想，其基本的步骤如下：

输入：样本集 $X=\{X_1,X_2,X_3,\cdots,X_n\}$，$X_1=[X_{11},X_{12},\cdots,X_{1m}]$，$n$ 为样本数，m 为样本维度，聚类数为 C；K 为K近邻值；

1) 由输入样本集合 X 计算样本间欧氏距离 D；$D=\{d_{ij}\}$；d_{ij} 为样本 X_i 与 X_j 之间的欧氏距离，由公式(2)可得；

2) 根据给定的 K 值，对每个样本的K近邻样本集合取标准差 B_i；

3) 对每个样本 X_i 以其对应的标准差 B_i 作为阈值，并保留阈值内的样本集合；此时，每个样本 X_i 中保留的样本数量与样本集本身性质和 K 值有关；

4) 每个样本仅在其阈值范围内保留了其与其他样本的距离度量，与阈值之外设为样本间距离是间接获得的，即：利用最短距离方法来保持样本集可以成为一个完整的连通图；

5) 用 3 中得到的标准差 B_i，代替相似性度量中的高斯核函数，消除核函数的取值难问题；

6) 于样本 X_i 在阈值 K 中的近邻样本数目为 N_i，D_i 是阈值 K 中的近邻样本集到 X_i 的距离均值；而于所有样本集 X，在其阈值 K 中的近邻样本数目为 N，其中 $N = N_1 + N_2 + \cdots + N_n$；$D$ 是阈值 K 中的近邻样本集到分别到其对应的 X_i 的距离均值；权重值是 $Q = D/D_i$ 是对局部距离和全局距离的统一调整参数，通过每次建立局部与整体的关系，将所有局部距离建立起相关性，使得相似性度量更具有全局性；

7) 把上述各变量带入为相似度计算公式(4)得相似性矩阵 S；

8) 将 S 矩阵带回谱聚类算法(NJW)的第 2 步后继续执行后续实验步骤；

输出：Y。

$$S_{ij} = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2 * B_i B_j} Q\right) \tag{4}$$

4. 实验

4.1. 实验数据集

在本节中，我们将提出的新谱聚类算法(NEW-SC)在人工数据和 UCI 数据中进行多次实验，且将实验结果与 k-means、SC、SC-ND、PGSC 等算法进行比较。实验中使用的 UCI 数据集有：Iris、Wine、Seeds、Heart、Fertility-Diagnose、Four-gauss、Haberman、Bupa 等；使用的人工数据集有：Two-cluster、Two moons、Spiral、Three-circles、Checker-board、Three-cluster。我们所选的人工数据集主要是二维的，所以能很好的以二维图片的形式展示；如图 1 所示：在本实验中我们使用的 UCI 数据皆为多维的，数据的基本信息包括数据样本的数量、样本维数和聚类数以图表的形式进行说明，如表 1 所示：

Table 1. UCI data set
表 1. UCI 数据集信息

数据集	Sample size	Dimension	Cluster number
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Heart	270	13	2
Fertility-Diagnose	100	9	2
Four-gauss	100	12	4
Haberman	306	3	2
Bupa	345	6	2

4.2. 性能测量

对于谱聚类算法的性能，我们通常用聚类准确率来度量(即：聚类精度 ACC)，其的定义式如(3~6)示；在公式中，C_i 表示已知的真实类标签，C_i' 是通过聚类算法得到的样本集的聚类标签；$\delta(\cdot)$是 δ 函数，它的

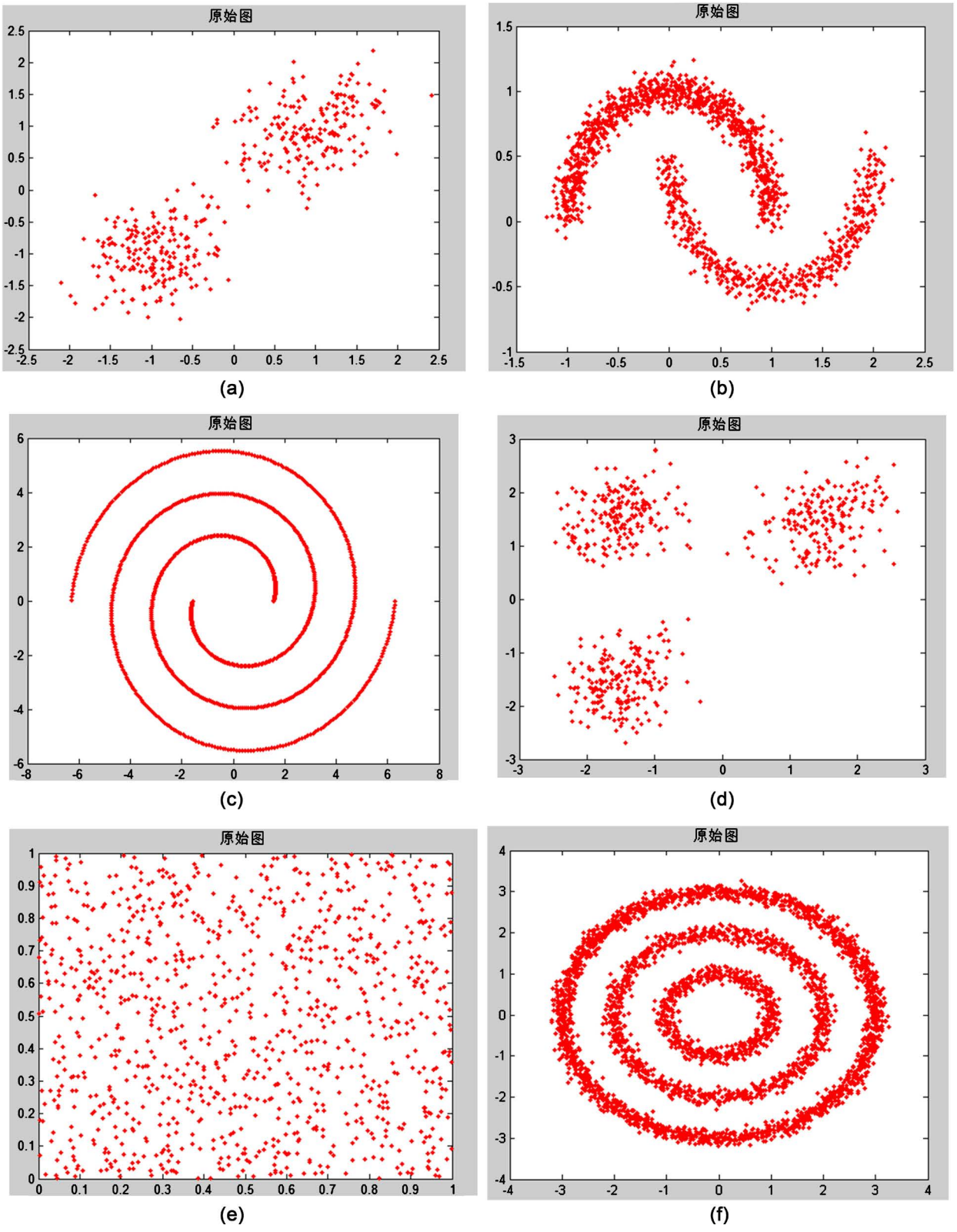

Figure 1. Artificial data set information graph: (a) Two-cluster; (b) Two moons; (c) Spiral; (d) Three-circles; (e) Checker-board; (f) Three-cluster

图 1. 人工数据集信息图：(a) Two-cluster；(b) Two moons；(c) Spiral；(d) Three-circles；(e) Checker-board；(f) Three-cluster

值是 1 或 0；如果 $\delta(a,b)$ 函数中的 $a=b$，则 $\delta(a,b)=1$，否则 $\delta(a,b)=0$。$map(\cdot)$是最佳映射函数，如果 $map(\cdot)$函数与真类标号匹配，用 Kuhn-Munkres 算法求解最佳映射。*ACC* 的取值范围从 0 到 1，其中 *ACC* 值较大时，表示聚类的识别率比较高，算法性能更好。其中 *ACC* 值较小时，表示聚类的识别率低，算法不成熟。

$$ACC=\frac{\sum_{i=1}^{n}\delta\left(C_i,map\left(C_i'\right)\right)}{n} \tag{6}$$

4.3. 人工数据集实验结果

使用 k-means、SC、SC-ND [13]、PGSC [14]以及 NEW-SC 算法对人工数据进行聚类，k-means 算法聚类结果见图 2 所示，从实验结果可以清新的看出，K-means 聚类算法虽然有其自身很好的一些性能，对于简单数据可以获得良好的聚类结果，其结论结果可以达到 1，但对于流形数据的聚类结果却不是很理想，不能很好识别出数据的形状，其结果与聚类真实意图差别较大。

基于 K-means 的谱聚类算法在特征提取上得到了优化，但在人工数据集中的实验结果相较于 K 均值聚类算法并未得到很好的提高，如图 3 所示；由人工数据集在光谱聚类 SC(NJW)算法下获得的实验结果图像，可以看出，光谱聚类与 K-means 相似，对于简单数据可以获得良好的聚类结果，但对于流形数据的聚类结果与不是很理想。也可以得出：虽然谱聚类在 K 均值的基础上对数据进行了特征提取，但是如果选择的特征不合适或是在提取特征的度量方式不正确时，都不会对聚类算法产生好的影响。

如图 4 所示，由人工数据集在 NP-SC 算法下获得的实验结果图像，从实验结果可以看出，NP-SC 算法与 K-means 有相同之处，即：对于简单数据可以获得良好的聚类结果；然而 SC 算法与 K-means 对于流形数据的聚类结果不是很理想，但是，NP-SC 算法对流形数据却有良好的聚类结果。在 NP-SC 算法中加入了数据之间的近邻的概念，所以能发觉数据周围细微的变化，使得聚类结果更符合预期。

由人工数据集在 PGK-SC 算法下获得的实验结果图像如图 5 所示，从实验结果可以看出，PGK-SC 算法即与 SC 算法与 K-means 算法相似，对于简单数据可以获得良好的聚类结果；又与 NP-SC 算法有相同，对流形数据也有良好的聚类结果。PGK-SC 算法中充分的考虑了高斯核函数的影响，比 SC 更能适应样本数据类型的变化。

由人工数据集在我们提出的新算法 NEW-SC 下获得的实验结果图像可以得出，从实验结果可以看出，NEW-SC 算法与 SC 算法与 K-means 算法相同，对简单数据可以获得良好的聚类结果，而且又 PGK-SC 算法和 NP-SC 算法相似，对流形数据也有良好的聚类结果，实验结果如图 6 所示：NEW-SC 算法中加入了数据之间的近邻的概念，所以能发觉数据周围细微的变化；还充分的考虑了高斯核函数的影响使得聚类结果更符合预期。

以上实验是将我们提出的聚类方法(NEW-SC)与一些现有的聚类方法进行了比较，在上述实验中，将谱聚类算法 SC(NJW)、NP-SC 聚类算法中涉及到高斯核的值的相关参数都设置为 1，以便使实验有更多的初始值，其对比实验更具有可信度。对人工数据集中的数据在 K-均值、谱聚类 SC(NJW)、NP-SC、PGK-SC 和新 NEW-SC 聚类方法下的实验结果的准确率进行了比较，如表 2 所示就是具体的实验结果数据。

在图 7 中的柱状图中，可以清晰的看到人工数据集中 Two-cluster、Two moons、Spiral、Three-circles、Checker-board、Three-cluster 等数据在 k-means、SC、SC-ND、PGSC 以及我们的新算法(NEW-SC)中的实验结果。在五组实验中，k-means、SC 算法对整体的数据集的聚类结果整体偏低。SC-ND、PGSC、算法对数据集的聚类结果整体较好，其中我们提出的新算法(NEW-SC)既能实现对离散数据的聚类，也能像 SC-ND、PGSC、算法一样，对流行数据有很好的聚类效果。

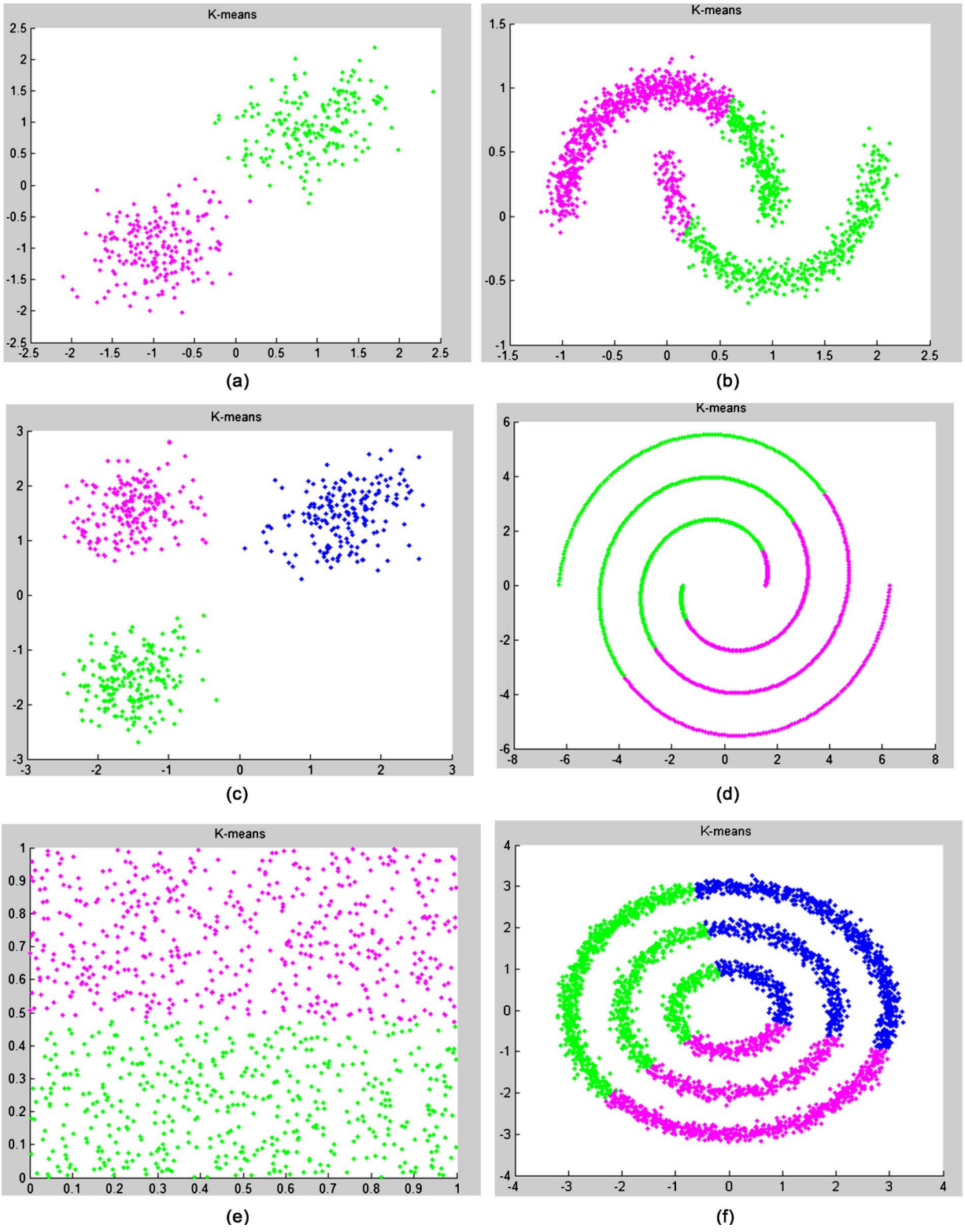

Figure 2. K-Mean clustering result graph of artificial data set: (a) Two-cluster; (b) Two moons; (c) Spiral; (d) Three-circles; (e) Checker-board; (f) Three-cluster

图 2. 人工数据集的 K 均值聚类结果图：(a) Two-cluster；(b) Two moons；(c) Spiral；(d) Three-circles；(e) Checker-board；(f) Three-cluster

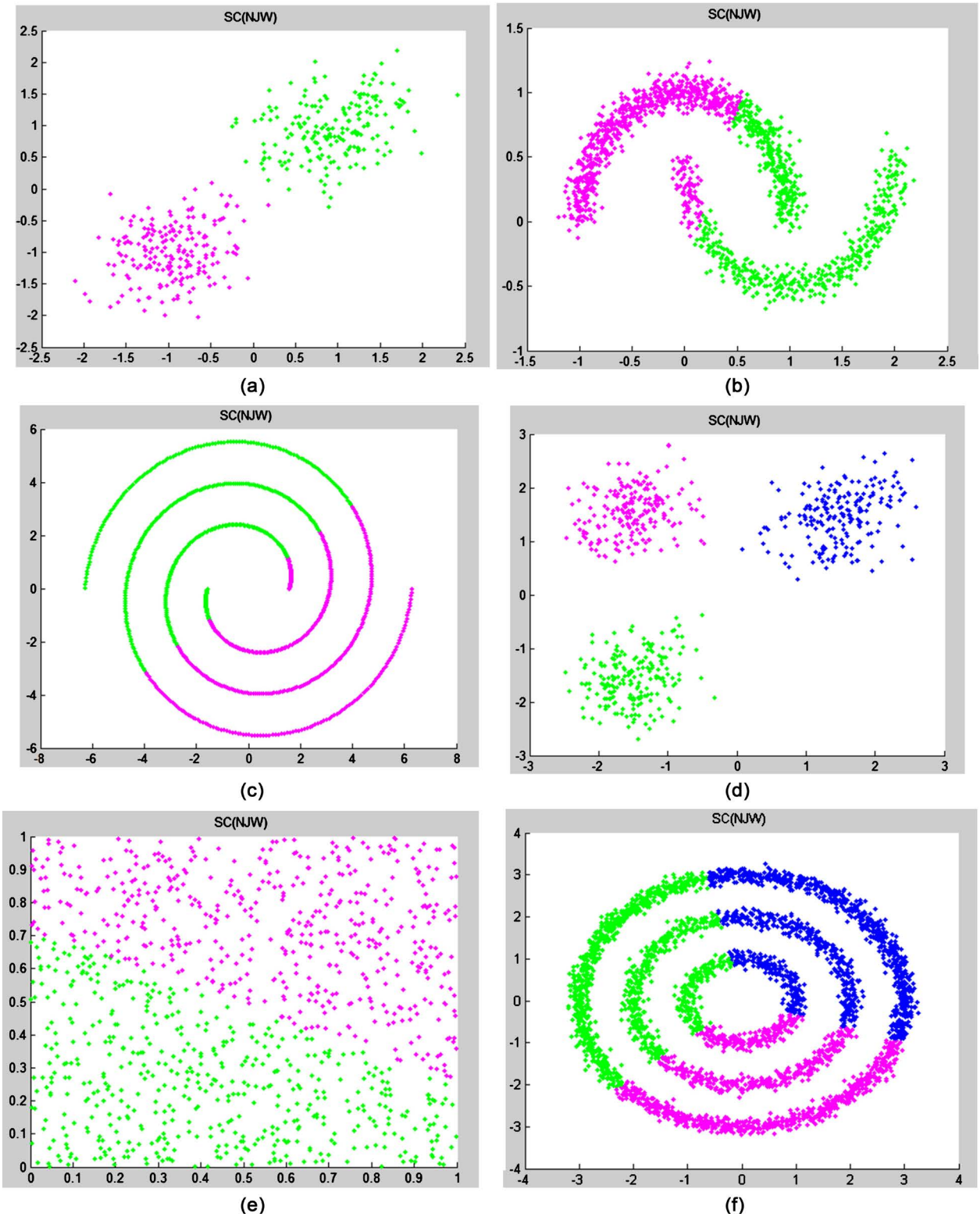

Figure 3. Spectral clustering (NJW) result graph of artificial data set: (a) Two-cluster; (b) Two moons; (c) Spiral; (d) Three-circles; (e) Checker-board; (f) Three-cluster

图 3. 人工数据集的谱聚类 SC 结果图：(a) Two-cluster；(b) Two moons；(c) Spiral；(d) Three-circles；(e) Checker-board；(f) Three-cluster

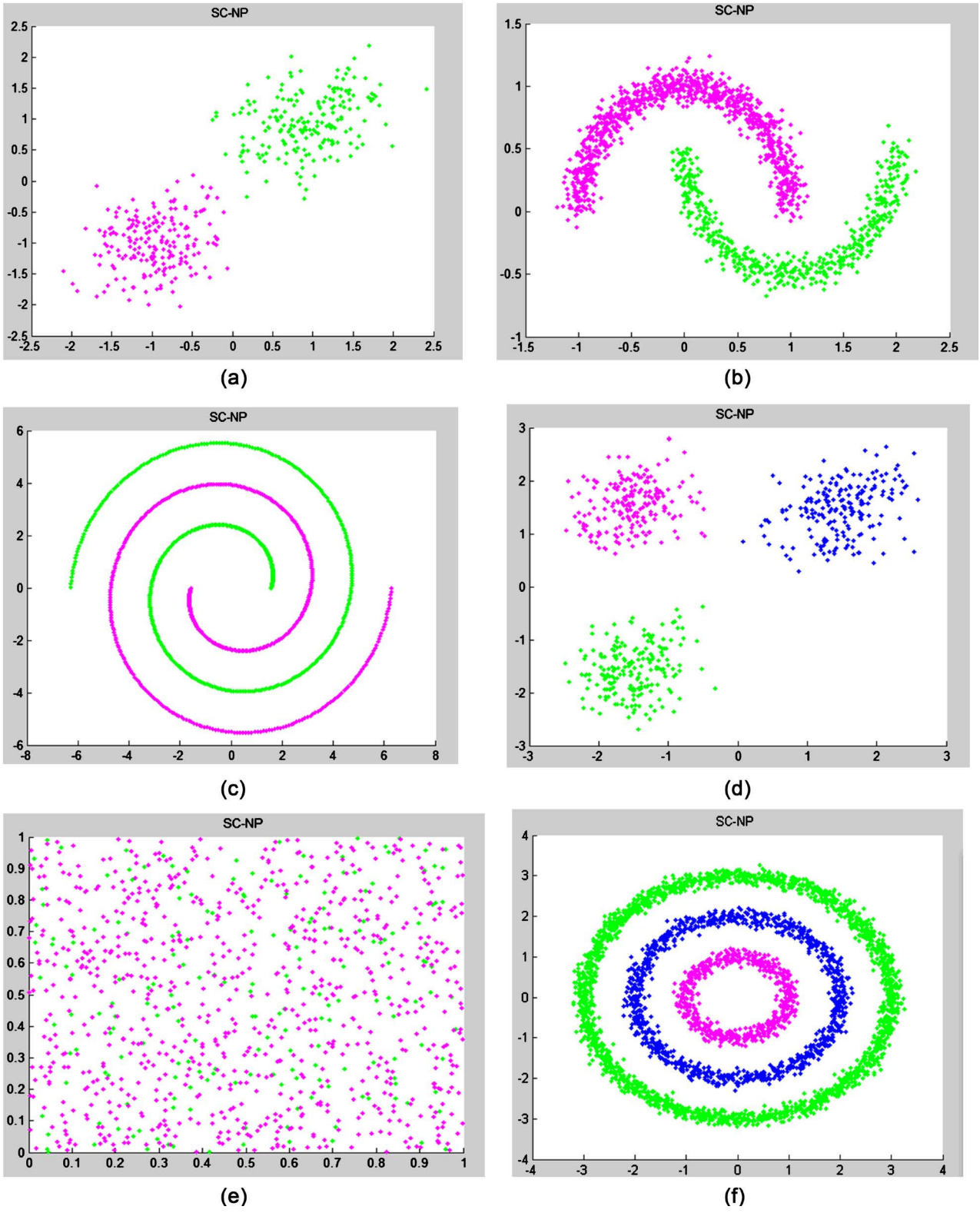

Figure 4. NP-SC clustering result graph of artificial data set: (a) Two-cluster; (b) Two moons; (c) Spiral; (d) Three-circles; (e) Checker-board; (f) Three-cluster

图 4. 人工数据集的 NP-SC 谱聚类结果图：(a) Two-cluster；(b) Two moons；(c) Spiral；(d) Three-circles；(e) Checker-board；(f) Three-cluster

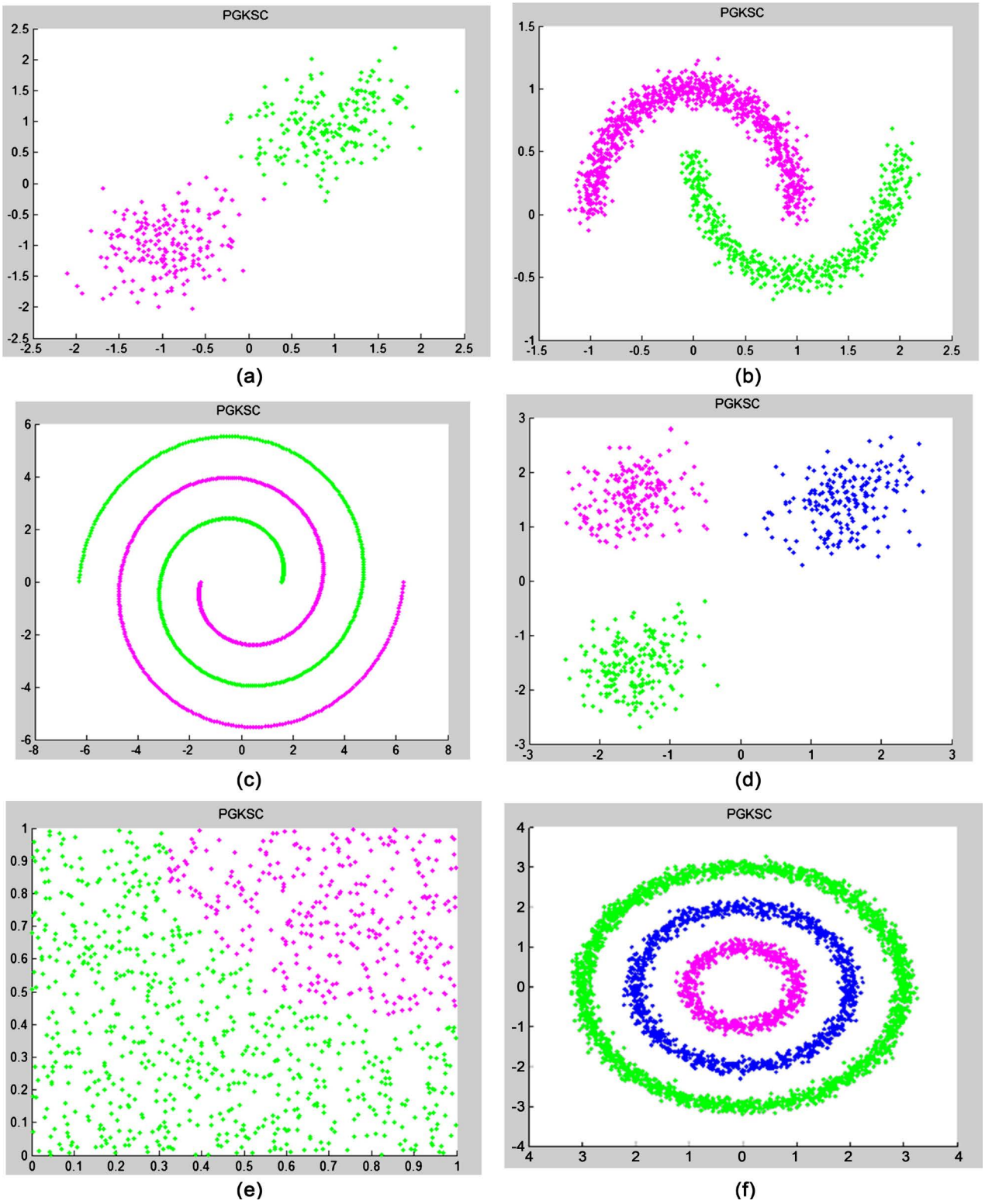

Figure 5. PGK-SC clustering result graph of artificial data set: (a) Two-cluster; (b) Two moons; (c) Spiral; (d) Three-circles; (e) Checker-board; (f) Three-cluster

图 5. 人工数据集的 PGK-SC 聚类结果图：(a) Two-cluster；(b) Two moons；(c) Spiral；(d) Three-circles；(e) Checker-board；(f) Three-cluster

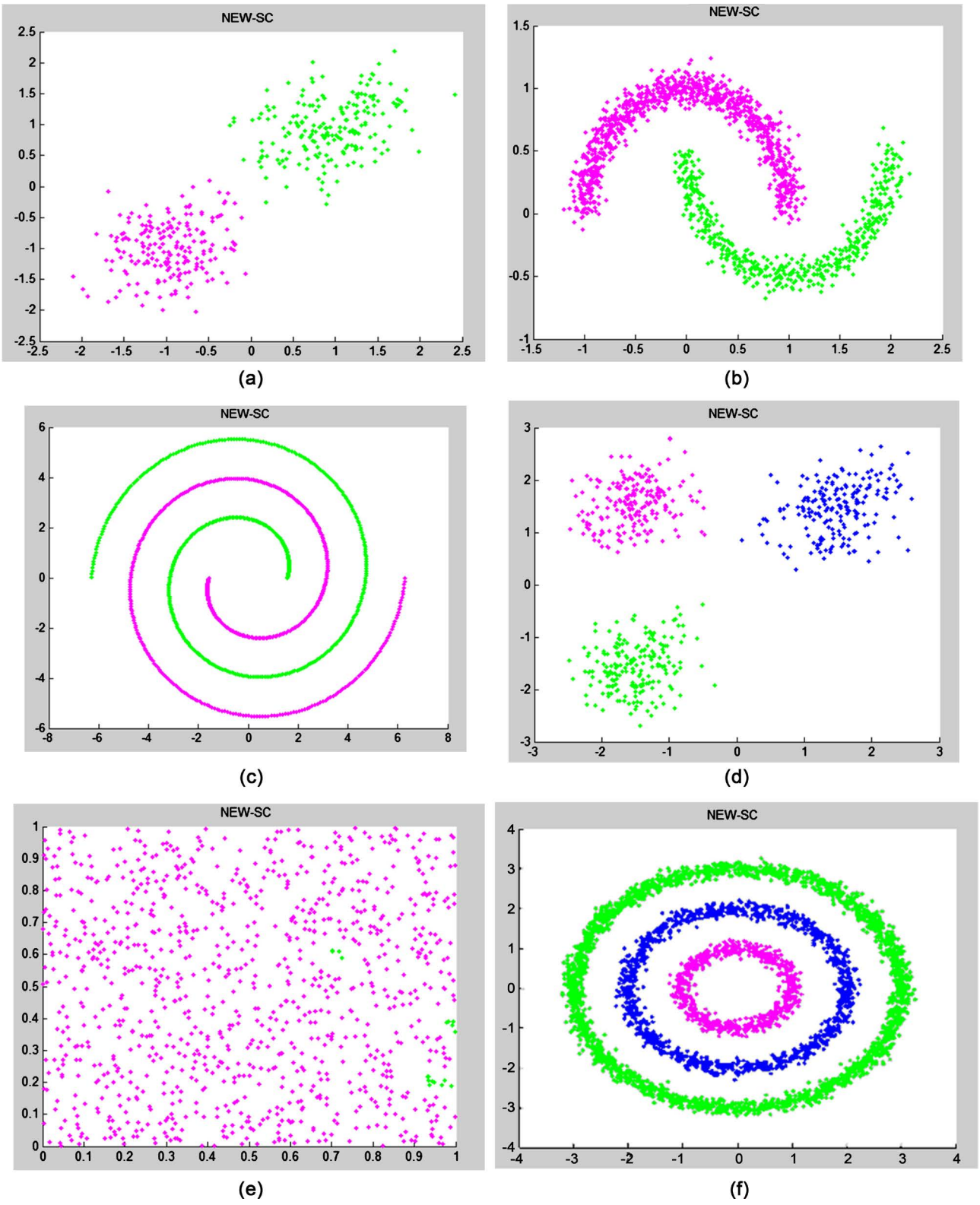

Figure 6. NEW-SC clustering result graph of artificial data set: (a) Two-cluster; (b) Two moons; (c) Spiral; (d) Three-circles; (e) Checker-board; (f) Three-cluster

图 6. 人工数据集的 NEW-SC 聚类结果图：(a) Two-cluster；(b) Two moons；(c)Spiral；(d) Three-circles；(e) Checker-board；(f) Three-cluster

Table 2. Accuracy rates of artificial data sets (ACC)
表 2. 人工数据集的准确率(ACC)

数据集	K-means	SC(NJW)	NP-SC	PGK-SC	NEW-SC
Checker-board	0.5710	0.5710	0.5740	0.5790	0.5980
Spiral	0.5890	0.6610	1	1	1
Two-cluster	0.9975	0.9975	1	1	1
Two moons	0.7337	0.7250	1	1	1
Three-cluster	1	1	1	1	1
Three-circles	0.3372	0.3369	1	1	1

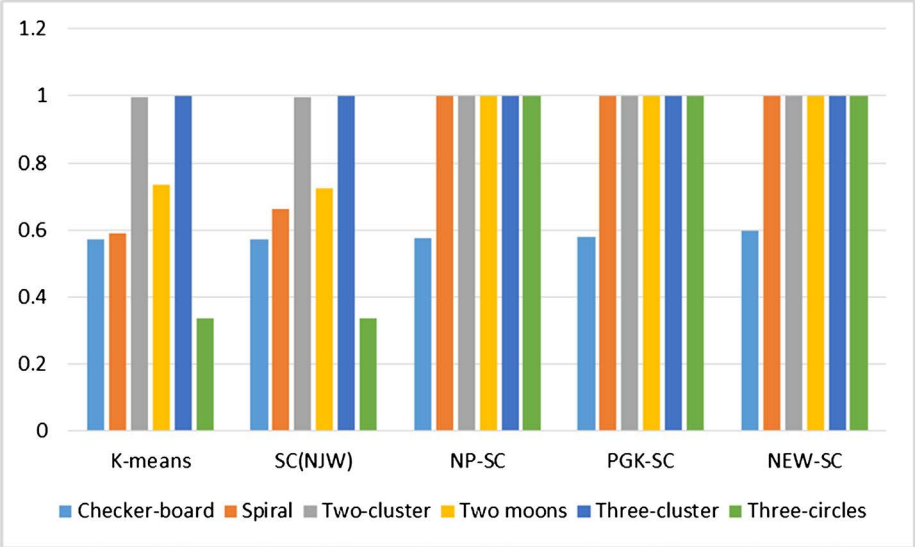

Figure 7. Accuracy histogram of artificial data sets
图 7. 人工数据集的准确率柱状图

4.4. UCI 数据集的实验结果

在本实验中我们使用的 UCI 数据有：Iris、Wine、Seeds、Heart、Fertility-Diagnose、Four-gauss、Haberman、Bupa；表 2 是上述数据集的基本信息，包括数据样本的数量、样本维数和聚类数。表 3 是上述所有数据集在 K 均值和 SC、NP-SC、PGK-SC、NEW-SC 聚类后得到的算法准确率(ACC)。

Table 3. Accuracy rates of UCI data set (ACC)
表 3. 算法准确率(ACC)

数据集	K-means	SC(NJW)	NP-SC	PGK-SC	NEW-SC
Iris	0.8933	0.9000	0.6800	0.9067	0.9267
Wine	0.7022	0.6292	0.4270	0.6573	0.7416
Seeds	0.8905	0.8905	0.3714	0.8857	0.9095
Heart	0.5926	0.5444	0.5519	0.5556	0.6259
Fertility-Diagnose	0.6700	0.6600	0.6200	0.7000	0.8900
Four-gauss	0.9800	0.9800	0.8800	0.9800	0.9900
Haberman	0.4739	0.7353	0.7386	0.7320	0.7386
Bupa	0.5536	0.5623	0.5652	0.5652	0.5826

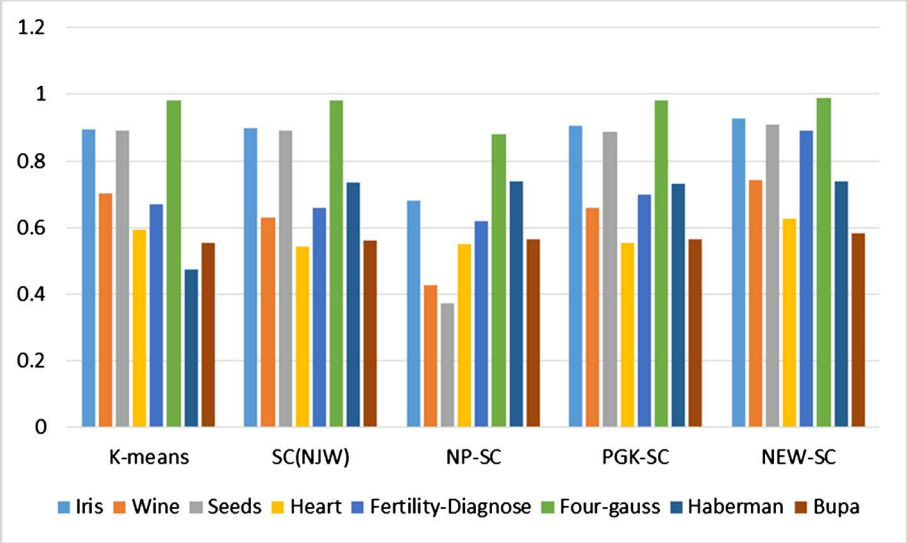

Figure 8. Accuracy histogram of UCI data sets
图 8. UCI 数据集的准确率柱状图

在图 8 中的柱状图中，可以清晰的看到 UCI 数据有：Iris、Wine、Seeds、Heart、Fertility-Diagnose、Four-gauss、Haberman、Bupa；等数据在 k-means、SC、SC-ND、PGSC 以及我们的新算法(NEW-SC)中的实验结果。在五组实验中，k-means、SC 算法对整体的数据集的聚类结果整体偏低。SC-ND、PGSC、NEW-SC 算法对数据集的聚类结果整体较好，其中我们提出的新算法的性能更好。

实验对比分析，由人工数据和 UCI 上的数据实验，可以得知，NEW-SC 算法不但在离散的数据样本上很好的聚类效果，而且还克服了传统算法的一些弊端，使得可以在流行数据上有较好的聚类效果，通过在 UCI 上的实验可知，NEW-SC 算法不但在低维度的数据上有较好的聚类性能，而且在较高维度的数据集中的实验结果也不错。

5. 实验结论

本文提出了一种新的 SC 聚类算法，首先考虑欧几里得距离矩阵在计算过程中的相似局限性，引入 K 邻域的概念，然后根据 K 邻域中的基本概念对标准差进行聚类，最后根据标准差进行聚类。利用最短路径算法求出样本间的最短路径，从而克服了欧几里德距离的诸多缺点，提高了同一聚类中样本在不同聚类之间的相似性和样本对异性之间的相似性，提高了聚类的精度。结果。此外，我们还认为在问题处理过程中很难确定高斯核相似矩阵的值，因为不同的值会严重影响聚类性能，在此基础上提出了每个样本的相关标准来代替高斯核，使得两个样本之间的相似性不仅与他们相关，还与数字本身、与周围样本和样本的距离有关。通过对人工数据集的实验，可以看出我们的算法不仅能够实现一些其他的算法，精确的方式使得流形数据集合为 1，而且在其他形状数据集合中也达到了较高的精度；实验用 UCI 数据集合更证明了这一点。

致　谢

光阴似箭，日月如梭，自从 2016 年 9 月入学至今，三年的研究生生涯即将结束。回顾在陕西师范大学所经历的点点滴滴，除了满满的收获之外，带给我更多的是对人生的思考与感悟，我相信三年的研究生经历将成为我人生中浓墨重彩的一部分，将帮助我在以后的道理上勇敢前行。在这里，我要向三年来给予我指导、帮助和鼓励的人表达诚挚的感谢！

感谢我的父母，感谢父母对我无私的付出和养育之恩，感谢父母对我学业的鼓励与支持，感谢父母教会我做人做事的道理，是他们无私的辛苦与付出才换来我如今的成绩，谁言寸草心，报得三春晖，我将努力提高自己，不负父母所望，为家庭、社会做出力所能及的贡献。

感谢我的导师王晅教授。王老师是一位品德高尚、认真负责、学识渊博、勇于创新、关心学生的导师，他具有严谨的学术态度、踏实认真的钻研精神以及丰富的工程实践经验。三年来，刘老师给予我无数次的耐心指导，无数次的关心与帮助。他踏实刻苦、忘我的工作态度是我学习的榜样，和蔼可亲、平易近人的待人态度深深影响这我。在这里，衷心感谢我的导师王老师。

感谢李娇师姐、叶爽师姐、感谢我的师弟师妹们，是你们的鼓励与陪伴，才能使我顺利完成今天的学业。与你们在一起的点点滴滴将成为我人生中重要的一笔财富，衷心感谢你们。

参考文献

[1] Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) Data Clustering: A Review. *ACM Computing Surveys*, **31**, 264-323. https://doi.org/10.1145/331499.331504

[2] 王千, 王成, 冯振元, 等. K-means 聚类算法研究综述[J]. 电子设计工程, 2012(4).

[3] 孙伟. 谱聚类算法的改进研究[D]: [硕士学位论文]. 兰州: 兰州商学院, 2012.

[4] Cai, D., He, X., Han, J. and Member, S. (2005) Document Clustering Using Locality Preserving Indexing. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 1624-1637. https://doi.org/10.1109/TKDE.2005.198

[5] 蔡晓研, 戴冠中, 杨黎斌. 谱聚类算法综述[J]. 计算机科学, 2008, 35(7): 14-18.

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

[7] Filipponea, M., Camastrab, F., Masullia, F. and Rovetta, S. (2008) A Survey of Kernel and Spectral Methods for Clustering. *Pattern Recognition*, **41**, 176-190. https://doi.org/10.1016/j.patcog.2007.05.018

[8] Nascimento, M.C.V. and de Carvalho, A.C.P.L.F. (2011) Spectral Methods for Graph Clustering—A Survey. *European Journal of Operational Research*, **211**, 221-231. https://doi.org/10.1016/j.ejor.2010.08.012

[9] Chen, W. and Feng, G. (2012) Spectral Clustering: A Semi-Supervised Approach. *Neurocomputing*, **77**, 229-242. https://doi.org/10.1016/j.neucom.2011.09.002

[10] Ozertem, U., Erdogmus, D. and Jenssen, R. (2008) Mean Shift Spectral Clustering. *Pattern Recognition*, **41**, 1924-1938. https://doi.org/10.1016/j.patcog.2007.09.009

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

[12] Fiedler, M. (1973) Algebraic Connectivity of Graphs. *Czechoslovak Mathematical Journal*, **23**, 298-305.

[13] 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

[14] Nataliani, Y. and Yang, M.-S. (2017) Powered Gaussian Kernel Spectral Clustering. *Neural Computing and Applications*, **31**, 557-572. https://doi.org/10.1007/s00521-017-3036-2

[15] Gong, Y.C. and Chen, C. (2008) Locality Spectral Clutering. *Proceeding of the* 21*st Australasian Joint Conference on Artificial Intelligence*: *Advance in Artificial Intelligence*, Springer-Verlag, 348-354. https://doi.org/10.1007/978-3-540-89378-3_34

[16] Shi, J. and Malik, J. (2000) Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 888-905. https://doi.org/10.1109/34.868688

[17] Dhillon, I.S., Guan, Y. and Kulis, B. (2004) Kernel K-Means: Spectral Clustering and Normalized Cuts. In: *Proceedings of the* 10*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, 551-556. https://doi.org/10.1145/1014052.1014118

[18] Ding, C., He, X., Zha, H.-Y., Gu, M. and Simon, H.D. (2001) A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering. *Proceedings of the* 2001 *IEEE International Conference on Data Mining*, San Jose, CA, 29 November-2 December 2001, 107-114. https://doi.org/10.1109/ICDM.2001.989507

[19] Hagen, L. and Kahng, A.B. (1992) 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

[20] McQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of* 5-*th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281-297.

[21] Ng, A.Y., Jordan, M.I. and Weiss, Y. (2002) On Spectral Clustering: Analysis and an Algorithm. In: Dietterich, T.G., Becker, S. and Ghahramani, Z., Eds., *Proceedings of the* 14*th Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, 849-856.