

A New Spectral Clustering Algorithm and Its Simulation

Jiao Li, Xuan Wang*

School of Physics and Information Technology, Shaanxi Normal University, Xi'an Shaanxi
Email: *542437192@qq.com

Received: Nov. 7th, 2017; accepted: Nov. 20th, 2017; published: Nov. 27th, 2017

Abstract

The similarity between the samples based on Euclidean distance measurement can not reflect the probability distribution characteristics of the samples, especially the sample clustering with multi-peak distribution, and the Euclidean distance has a large deviation. In addition, the k-means algorithm widely used in the class often produces unstable clustering results due to the randomness of the initial center. In this paper, a similarity measure method which can reflect the probability distribution of the sample is proposed for the similarity measure. On this basis, the k-means algorithm in spectral clustering is improved, and a method is proposed for the initial center adaptive setting method. The experimental results show that the proposed algorithm has better clustering effect than the traditional spectral clustering in the artificial data set and the UCI real data set.

Keywords

Spectral Graph Theory, Spectral Clustering, Similarity Matrix, k-Means

一种新的谱聚类算法及其仿真

李 娇, 王 珣*

陕西师范大学物理学与信息技术学院, 陕西 西安
Email: *542437192@qq.com

收稿日期: 2017年11月7日; 录用日期: 2017年11月20日; 发布日期: 2017年11月27日

摘 要

针对传统谱聚类基于欧氏距离度量样本之间的相似性, 不能反映样本的概率分布特性, 特别是具有多峰分布的样本聚类, 欧氏距离具有较大的偏差, 另外在传统谱聚类中广泛使用的k-means算法由于初始中

*通讯作者。

心的随机性设置, 经常会产生不稳定的聚类结果。本文针对相似性度量问题提出了一种能够反映样本概率分布特性的相似性度量方法, 并在此基础上对谱聚类中的k-means算法进行了改进, 给出了一种对初始中心进行自适应设置的方法。实验结果表明, 该算法相比传统的谱聚类在人工数据集与UCI真实数据集上具有更好的聚类效果。

关键词

谱图理论, 谱聚类, 相似性矩阵, k-Means

Copyright © 2017 by authors and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

聚类分析是机器学习领域的一个重要的分支, 是人们认识和探索事物之间内在联系的有效手段[1]。其过程主要考虑特征选择、相似性度量、聚类准则、聚类算法及结果验证几个方面的内容[2]。聚类算法在数据挖掘与机器学习中具有广阔的应用前景, 近年来成为该领域的研究热点[3] [4] [5]。

传统的聚类算法主要依据样本点的距离或者密度, 根据其生成聚类方法的不同分为分割聚类与层次聚类两类。基于距离的分割聚类算法将聚类问题转化为针对全局标准的优化算法, 代表性的算法主要有K-means、CLARA 及 CLARANS 算法等; 而基于密度的分割聚类算法通过优化样本密度分布的局部目标函数实现聚类。层次聚类根据其解决问题的方式分为分裂与合并两种。上述算法都是基于假设样本空间为凸球形的, 但当样本空间不为凸或具有复杂的多峰分布时, 算法性能会显著下降。另外, 分割聚类算法大多通过基于目标的优化过程实现, 会陷入局部最优解[6] [7]。

谱聚类算法计算样本点对的局部相似性基于谱图理论挖掘样本空间的全局结构, 不需要样本空间概率分布的前提假设, 具有更广泛的适用性。因此, 受到广泛关注[8]。谱聚类通常由两个主要步骤组成, 首先是基于数据样本点的相似性度量建立关系图, 其次构建算法对图进行聚类分割[9]。现有的提升谱聚类性能的研究成果分为两种类型, 第一种类型是在样本点关系图固定的情况下, 重点提升图的分割过程; 另外一类工作是构建合适的关系图通过常用的分割算法提升聚类性能。

构建关系图的关键问题是如何定义两个样本点之间的相似性, 传统的谱聚类基于欧式距离度量不能完全反映数据聚类复杂的空间分布特性。其分割过程常用的 k-means 算法由于初始集群中心的随机性, 经常会产生不稳定的聚类结果。因此, 本文在这两方面进行了改进: 一方面本文利用 Peng Yang 等人提出的密度敏感的谱聚类方法计算相似性度量, 另一方面对 k-means 算法进行改进, 该算法不仅可以放大不同高密度区域内数据点之间的距离, 同时可以缩短同一高密度区域内数据点之间的距离, 最终有效描述数据的实际聚类分布, 而且它可以获得一个更好的聚类中心。将该算法与传统聚类算法以及常用谱聚类算法进行比较, 有效性分析和结果表明, 本文提出的算法能够获得更好的聚类效果。

2. 算法回顾

谱聚类

谱聚类算法的思想来源于谱图划分理论, 它将聚类问题看成是一个无向图的多路划分问题。该算法发展了很多不同的具体实现方法, 但是都可以归纳为以下三个主要步骤[10] [11]:

Step 1: 构建表示样本集的矩阵 Z ;

Step 2: 通过计算 Z 的前 k 个特征值与特征向量, 构建特征向量空间;

Step 3: 利用 k-means 聚类算法对特征向量空间中的特征向量进行聚类。

传统的谱聚类算法的步骤如下:

输入: n 个数据点 $\{x_i\}$ 、聚类数 k , σ 为事先指定的参数。

输出: 数据的划分 $c_1, c_2, c_3, \dots, c_k$ 。

(1) 根据欧氏距离度量构造相似性矩阵 $S_{ij} = \exp\left(-\frac{dist}{2\sigma^2}\right)$;

(2) 构造拉普拉斯矩阵 $L = D - S$, 其中 D 为对角矩阵;

(3) 求拉普拉斯 L 的 k 个最大特征值对应的特征向量 $v_1, v_2, v_3, \dots, v_k$, 并且构造矩阵 $V = [v_1, v_2, \dots, v_k] \in R^{n \times k}$, 其中 v_k 为列向量;

(4) 单位化 V 的行向量, 得到矩阵 Y , 其中 $Y_{ij} = \frac{V_{ij}}{\left(\sum_j V_{ij}^2\right)^{\frac{1}{2}}}$;

(5) 将 Y 的每一行看成是 R^k 空间内的一点, 使用 K 均值将其聚为 k 类;

(6) 如果 Y 的第 i 行属于第 j 类, 则将原始数据点 x_i 也划分到第 j 类。

谱聚类算法仅与数据点的数目有关, 而与维数无关, 因而可以避免由高维特征向量造成的奇异性问题。该算法又是一种判别方法, 不针对数据的全局结构作假设。尽管它是一种极具竞争力的聚类方法, 但其目前仍处于研究初期, 算法本身存在一些亟待解决的问题。现有的算法对分析尺度的选择非常敏感, 使得尺度参数的正确选择成为算法成功的关键。在真实数据集问题中, 数据通常具有多重尺度, 而现有的谱聚类算法仍然不适合处理一些多尺度聚类问题。

在谱聚类算法中, 除了相似度矩阵的问题之外, 还涉及到了 k-means 算法的问题, 而这两个问题是直接影响聚类效果的好坏。k-means 聚类算法是目前应用最广泛的划分聚类算法之一。容易实施、简单、高效、成功的应用案例和经验是其至今流行的主要原因。但是, 该聚类算法仍然存在一些缺点, 它不仅是一个 NP 难优化问题, 也是一个贪心算法, 仅能获得局部最优, 无法获得全局最优, 且该算法的初始聚类中心是随机选取的。

3. 改进的算法

在这一部分, 我们首先针对谱聚类算法中存在的两种问题提出解决方法, 最后, 我们提出了改进的算法。

3.1. 优化的相似性度量

从上述谱聚类算法的步骤中我们可知, 相似度矩阵的选择是聚类效果性能的重要指标, 只要选择了好的相似度矩阵, 那么就会获得一个较好的聚类结果。

在创建相似性矩阵 W 的过程中, 传统的谱聚类通常基于欧氏距离通过高斯核函数计算相似度[12]。对于复杂问题, 简单的基于欧氏距离的相似性度量不能完全的反映聚类结构。因此, 本文在构造相似矩阵 W 时应用了基于密度敏感的相似性度量, 因为从数据的空间分布情况来看, 同一聚类内的数据趋向于分布在一个密度较高的区域, 而不同聚类之间存在一个数据分布相对稀疏的低密度区域, 我们根据数据密度改进相似度矩阵, 应用密度敏感的相似度矩阵[13]。

定义 1: 局部线段长度:

$$L(x_i, x_j) = \rho^{d(x_i, x_j)} - 1 \quad (1)$$

其中, $d(x_i, x_j)$ 是数据 x_i 和 x_j 之间的欧氏距离, $\rho > 1$, 称为伸缩因子。

定义 2: 将数据点集 $V = \{x\}$ 视为加权无向图 $G = \{V, E\}$ 的顶点, 边的集合表示在每对数据点之间的连接权值的集合。把 $p \in V'$ 作为长度 $l = |p| - 1$ 的路径, 其连接在点 p_1 和 p_2 之间。 p_{ij} 表示 $(p_k, p_{k+1}) \in E$ 连接数据点 $(p_k, p_{k+1}) \in E$ 所有路径的集合, 其中 $(1 \leq i, j < n)$ 。则 x_i 和 x_j 之间新的相似性距离定义为:

$$D_{ij} = \min_{p \in p_{ij}} \sum_{k=1}^{l-1} L(p_k, p_{k+1}) \quad (2)$$

其中 $L(p_k, p_{k+1})$ 表示两点之间密度可调节的线段长度。

3.2. 优化的 K-means 算法

从上述可知, 传统的谱聚类算法的最后一步是完成 k 均值算法聚类。由于 k 均值算法严格依赖于初始聚类中心, 然而, 初始聚类中心通常是随机生成的, 所以每个集群中心不是在基本算法开始时相同的, 且可能产生不稳定的聚类结果[14] [15]。

根据 k 均值算法及其相关改进方法的研究, 本文设计了一种简单有效的选择初始聚类中心的方法。该方法只是基于传统的 k 均值算法进行简单的改变, 即在随机初始化聚类中心的步骤中增加随机数, 对于每个随机初始化 k 个集群中心, 计算彼此之间的欧氏距离集群。然后, 保存并选择 k 个聚类中心的最大距离中心。该聚类中心使得聚类结果稳定并且使得聚类结果得到很大的提高。

改进算法的核心如下: 随机选择 k 个数据对象作为数据集 X 的初始聚类中心; 在 k 个聚类中心之间计算欧氏距离; 然后随机地重复选择 k 个数据对象; 再次计算 k 个聚类中心之间的欧氏距离。如果这次的距离大于上一次的距离, 我们需要保存这个 k 集群中心和对应的距离, 或者对于下一个随机选择不进行修改, 直到达到设置的随机数。最终, 该算法得到了一个更好的初始聚类中心。

3.3. 改进的谱聚类算法

根据上述的两种优化方案, 可以得到改进的谱聚类算法。实验证明改进的谱聚类算法比原始的算法具有更好的性能。

改进的谱聚类算法的步骤如下:

输入: n 个数据点、聚类数 k 及伸缩因子 ρ 。

输出: 数据的划分 $c_1, c_2, c_3, \dots, c_k$ 。

(1) 根据密度敏感的相似性度量构造相似性矩阵 $W \in R^{n \times n}$,

$$\text{其中 } W_{ij} = \frac{1}{\min_{p \in p_{ij}} \left(\sum_{k=1}^{l-1} (\rho^{dist(p_k, p_{k+1})} - 1) \right)}, W_{ij} = 0;$$

(2) 构造拉普拉斯矩阵 $L = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$, 其中 D 为对角矩阵, $D_{ii} = \sum_{j=1}^n W_{ij}$;

(3) 求拉普拉斯 L 的 k 个最大特征值对应的特征向量 $v_1, v_2, v_3, \dots, v_k$, 并且构造矩阵 $V = [v_1, v_2, \dots, v_k] \in R^{n \times k}$, 其中 v_k 为列向量;

(4) 单位化 V 的行向量, 得到矩阵 Y , 其中 $Y_{ij} = \frac{V_{ij}}{\left(\sum_j V_{ij}^2 \right)^{\frac{1}{2}}}$;

- (5) 将 Y 的每一行看成是 R^k 空间内的一点, 使用改进的 K 均值算法将其聚为 k 类;
 (6) 如果 Y 的第 i 行属于第 j 类, 则将原始数据点 x_i 也划分到第 j 类。

4. 实验仿真

为了验证本文提出的改进谱聚类算法的聚类性能。将该算法应用在人工数据集[16]和 UCI 数据集的聚类问题上, 并分析实验结果。让该算法与传统的 K 均值算法(K-means), 传统的谱聚类算法(NJW-SC)和密度敏感的谱聚类算法(DNJW-SC)进行比较。结果表明, 本文提出的改进的谱聚类算法在大部分数据集集中的性能较好。

4.1. 人工数据集

本文中, 在 TwoS、2Spirals 与 circle data 三个 2-维人工数据集(图 1)中应用四种聚类算法。这是数据集都是具有“挑战性”问题的图示, 它们的类别数均为 2。

在(a)图中, 正确的分类标准应该是右上角的倒“S”是一类, 左下角的正“S”是另一类。

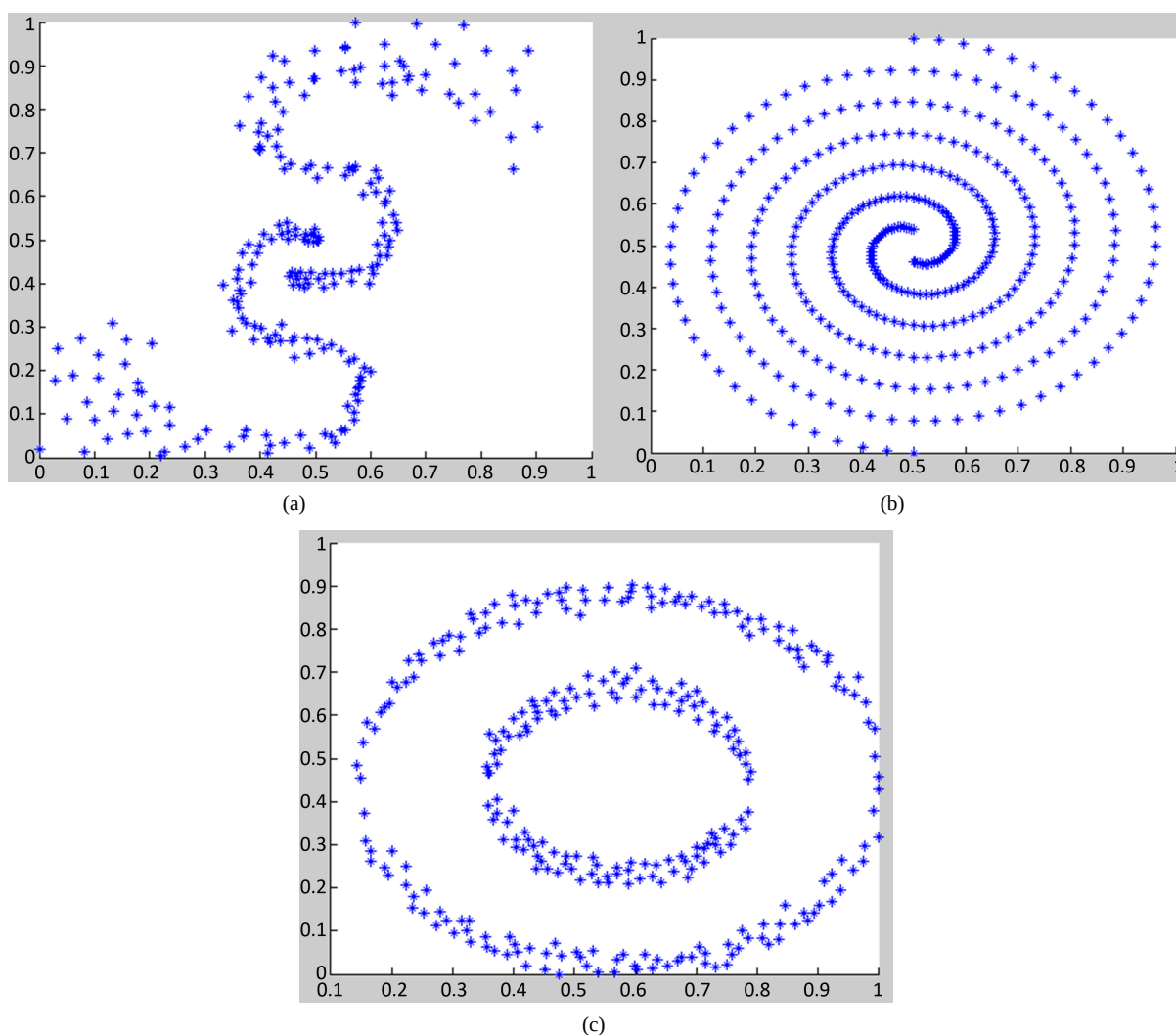


Figure 1. The original datasets

图 1. 原始数据集

在(b)图中, 正确的分类标准应该是以靠左边的点为起点画出的螺旋是一类, 靠右边点为起点画出的螺旋是另一类。

在(c)图中, 正确的分类标准应该是内椭圆是一类, 而外椭圆是另外一类。

如图 2~图 4 四种聚类算法应用在 3 个不同的人工数据集上得到的聚类结果。其中(a) 传统的 K 均值算法, (b) 传统的谱聚类算法, (c) 密度敏感的谱聚类算法, (d) 改进的谱聚类算法。

4.2. UCI 数据集

UCI 数据集是一个用于机器学习的常用标准测试集, 是 University of California Irvine 提出的真实数据集。下面选用该测试集中 4 个真实数据集进一步验证本文提出算法的优越性, 具体数据特征如表 1 所示。

为了评估聚类的性能, 本文采用了 Rand Index(RI)和 F-measure 两种聚类评价指标来判断。

第一种: 采用 Rand Index(RI)评估聚类结果的正确率, 实际上这是一种用排列组合原理来对聚类进行评价的手段[17], 记为 RI, 公式如下:

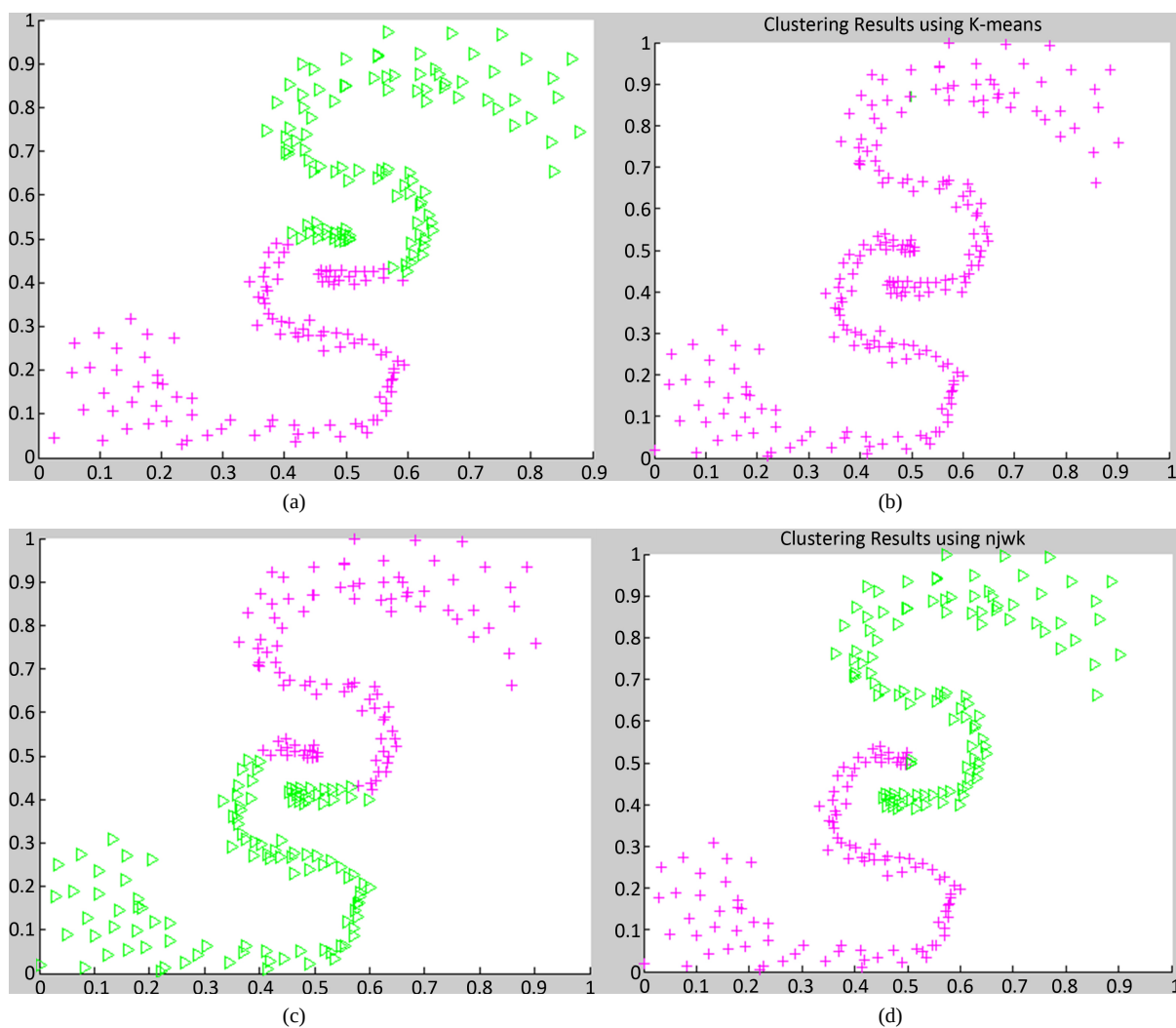


Figure 2. Contrast of four algorithms on synthetic datasets TwoS

图 2. 人工数据集 TwoS 的 4 种算法对比

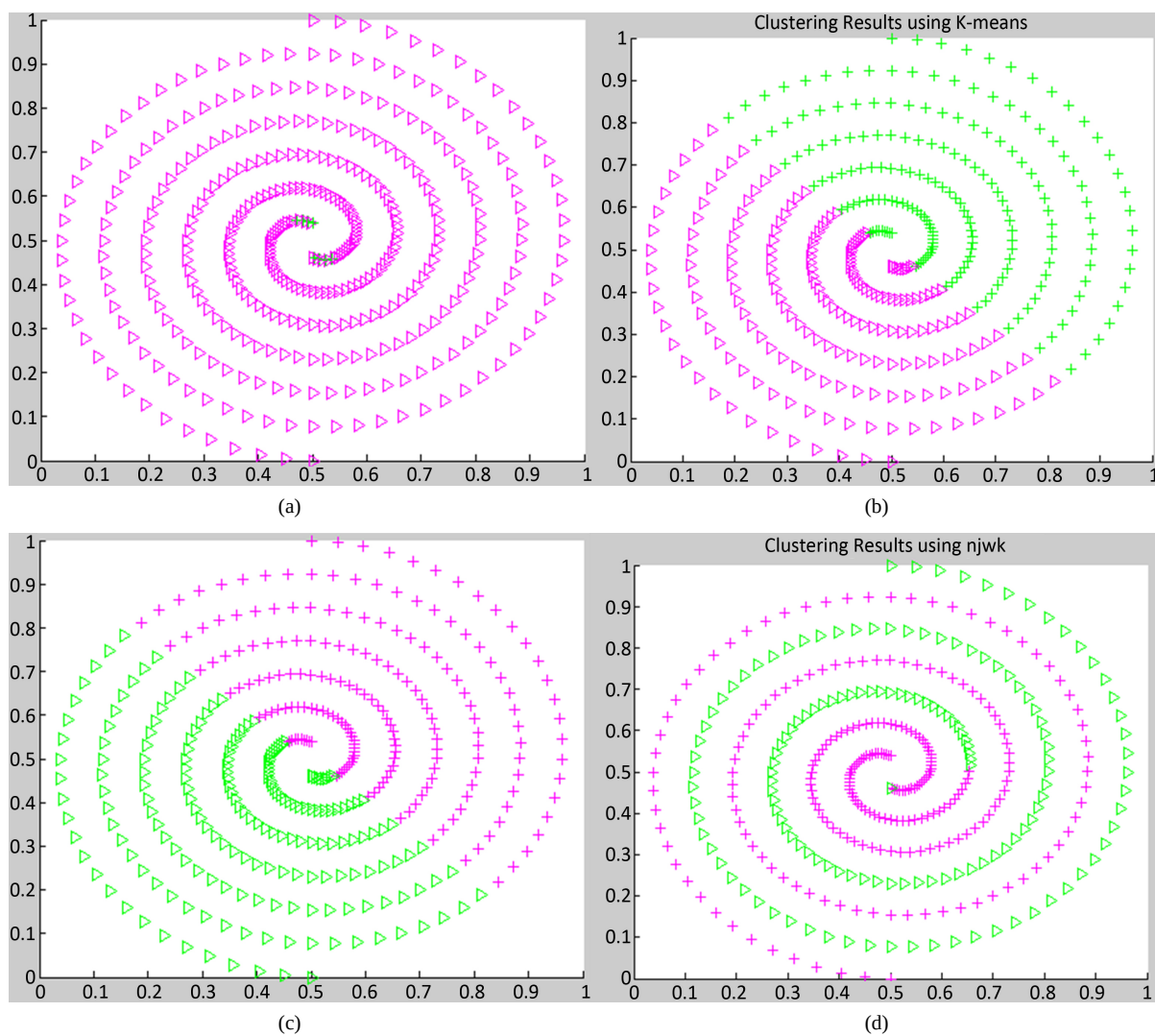
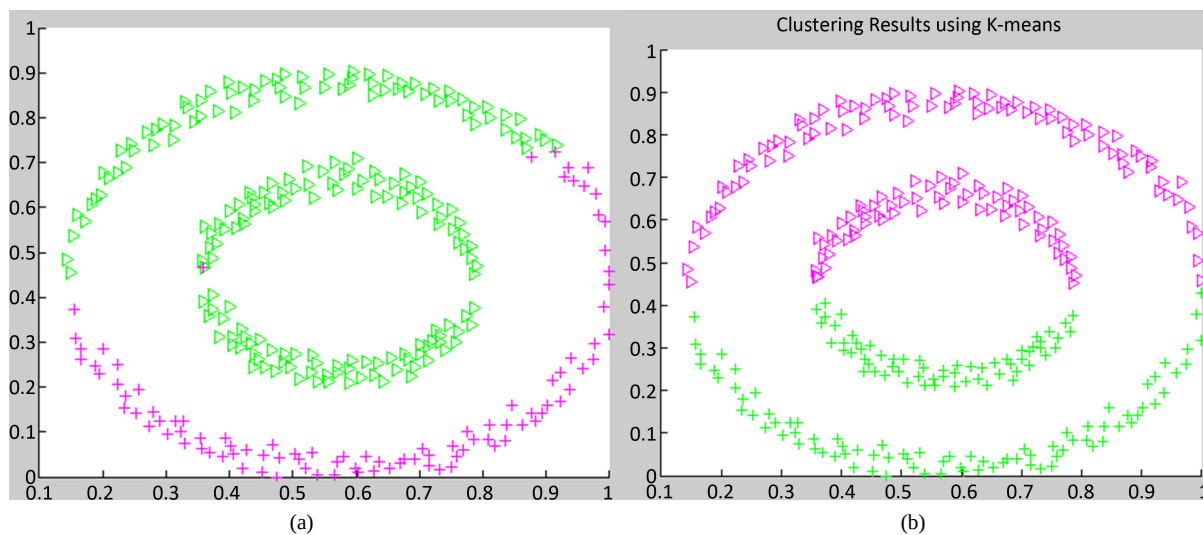


Figure 3. Contrast of four algorithms on synthetic datasets 2Spiral

图 3. 人工数据集 2Spiral 的 4 种算法对比



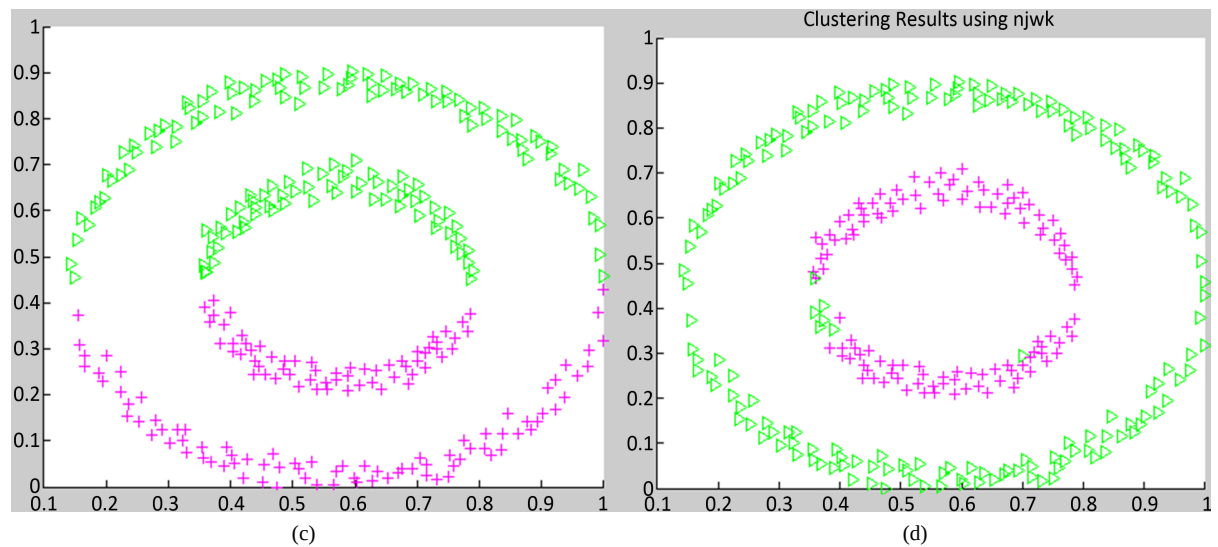


Figure 4. Contrast of four algorithms on synthetic datasets circle data

图 4. 人工数据集 circle data 的 4 种算法对比

Table 1. Experiment dataset and data characteristics

表 1. 实验数据集及数据特征

数据集	样本个数	维数	类别数
Iris	150	4	3
Soybean	47	35	4
Wine	178	13	3
glass	214	9	6

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

其中, TP 是指被聚在一类的文档被正确分类了, TN 是指不应该被聚在一类的文档被正确分开了, FP 是指不应该放在一类的文档被错误的放在了一类, FN 是指不应该分开的文档被错误的分开了。

图 5 表示的是 K-means 聚类算法、传统聚类算法、密度敏感的谱聚类和本文提出的算法对 4 个 UCI 数据集进行聚类所得结果的 RI 值。

第二种: 采用基于人工标注的评价指标 F-measure 对聚类质量进行评价, 对于聚类结果簇 i 和簇 j , 计算准确率(precision) P 、召回率(recall) R 以及 P 、 R 的综合评价指标 F-measure (记为 F):

$$P(i, j) = \frac{n_{ij}}{n_j} \quad (4)$$

$$R(i, j) = \frac{n_{ij}}{n_i} \quad (5)$$

$$F(i, j) = \frac{2 \times P(i, j) \times R(i, j)}{P(i, j) + R(i, j)} \quad (6)$$

式中, n_i 为聚类算法识别出的簇 i 所包含的数据样本数目, n_j 为簇 j 所包含的数据样本数目, n_{ij} 表示属于簇 j 却被错划为簇 i 的数据样本数目。

对数据集进行聚类后最终所得到的结果, 通过取它对所有簇的最大 F-measure 的加权平均来评价其聚类质量, 该指标为聚类结果的总体正确率 F-measure:

$$F = \sum_i \frac{n_i}{n} \max_j \{F(i, j)\} \quad (7)$$

$F \in [0,1]$, F 值越高表明聚类效果越好。图 6 表示的是 k-means 聚类算法、传统聚类算法、密度敏感

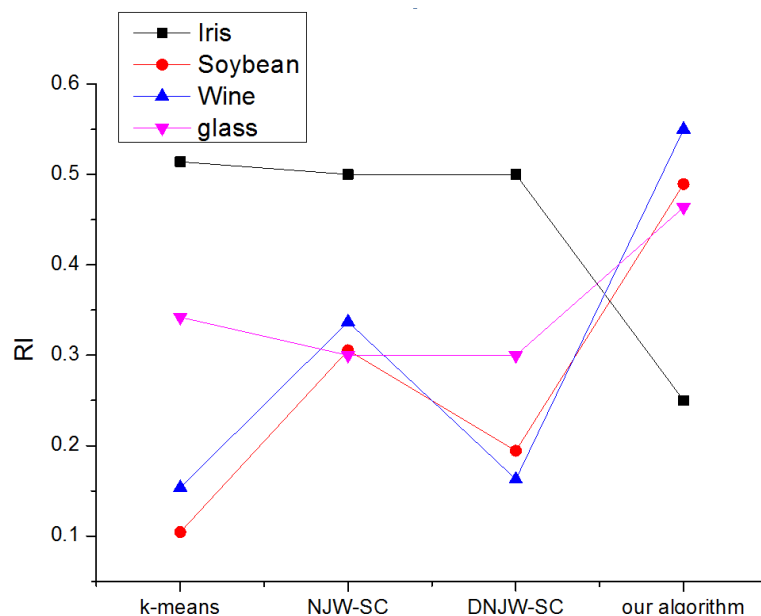


Figure 5. Comparison of the RI indicator figures of the four sets of data clustered by the four algorithms

图 5. 四种算法对 4 组数据聚类后的 RI 性能指标比较

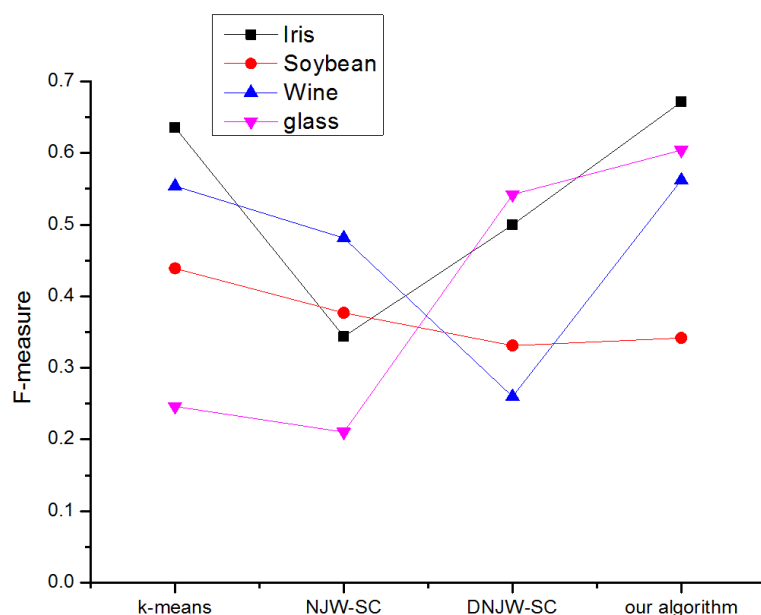


Figure 6. Comparison of the F-measure indicator figures of the four sets of data clustered by the four algorithms

图 6. 四种算法对 4 组数据聚类后的 F-measure 性能指标比较

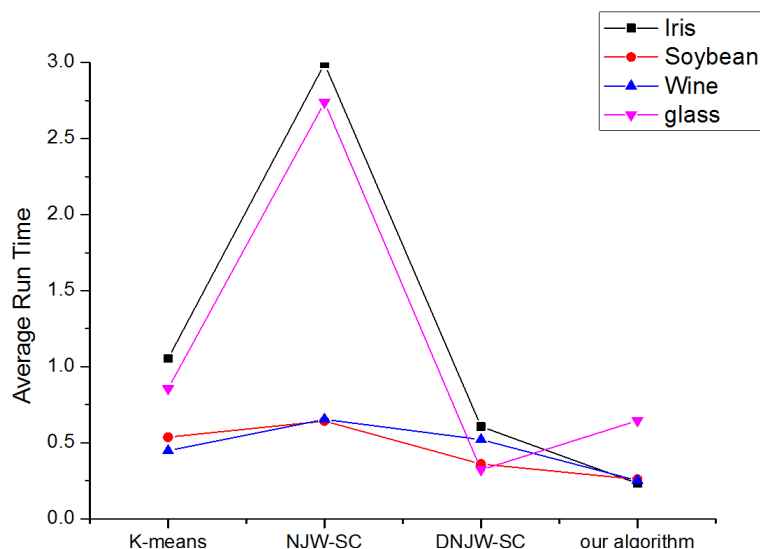


Figure 7. Comparison of average run time of the four sets of data clustered by the four algorithms

图 7. 四种算法对 4 组数据聚类后的平均运行时间比较

的谱聚类和本文提出的算法对 4 个 UCI 数据集进行聚类所得结果的 F-measure 测量值[18] [19] [20]。

由于聚类算法选取的不同, 将会导致平均运行时间不同, 本文中也得到了四种不同算法的平均运行时间, 平均运行时间越短, 说明该算法更节省时间, 效率更高。运行结果如图 7 所示。

5. 结论

本文研究了影响谱聚类算法聚类效果的两个因素, 提出了一个新颖的 k 均值算法, 得到了改进 k 均值的谱聚类算法。该算法一方面不仅能够处理多尺度聚类的问题, 而且对参数选择相对不敏感, 另一方面提出的 K 均值算法得到的聚类中心, 该聚类中心使得聚类结果稳定并且得到很大的提高。然而, 本方法仍需进一步改进, 以使得聚类效果得到最佳。另外, 如何将本文提到的算法应用到图像处理领域将是本课题下一阶段的研究重点。

参考文献 (References)

- [1] Jain, A., Murty, M. and Flynn, P. (1999) Data Clustering: A Review. *ACM Computing Surveys*, **31**, 264-323. <https://doi.org/10.1145/331499.331504>
- [2] 王珣, 马建峰. 数字图像分析与模式识别[M]. 北京: 科学出版社, 2011: 228-229.
- [3] Huang, Z. (1998) Extensions to the k-Means Algorithm for Clustering Large Datasets with Categorical Values. *Data Mining and Knowledge Discovery*, 283-304.
- [4] Wu, X. and Zhang, S. (2003) Synthesizing High-Frequency Rules from Different Data Sources. *IEEE Transactions on Knowledge and Data Engineering*, **15**, 353-367.
- [5] Wu, X., Zhang, C. and Zhang, S. (2005) Database Classification for Multi-Database Mining. *Information Systems*, **30**, 71-88. <https://doi.org/10.1016/j.is.2003.10.001>
- [6] Luxburg, U.V. (2007) A Tutorial on Spectral. *Statistics and Computing*, **17**, 395-416. <https://doi.org/10.1007/s11222-007-9033-z>
- [7] Yan, D. (2009) Fast Approximate Spectral Clustering. UC Berkeley, 907-915. <https://doi.org/10.1145/1557019.1557118>
- [8] 蔡晓妍, 戴冠中, 等. 谱聚类算法综述[J]. 计算机科学, 2008, 35(7): 14-18.
- [9] Wu, Z. and Leahy, R. (1993) An Optimal Graph Theoretic Approach to Data Clustering: Theory and Its Application to Image Segmentation. *IEEE Transactions on PAMI*, **15**, 1101-1113.

- [10] Chang, H. and Yeung, D.Y. (2011) Robust Path-Based Spectral Clustering. *Applied and Computational Harmonic Analysis*, **30**, 319-336.
- [11] Chi, Y. and Song, X.D. (2009) On Evolutionary Spectral Clustering. *ACM Transactions on Knowledge Discovery from Data*, **3**, Article No. 17. <https://doi.org/10.1145/1631162.1631165>
- [12] Tulin, I. (2015) A Parameter-Free Similarity Graph for Spectral Clustering. *Expert Systems with Applications*, **42**, 9489-9498. <https://doi.org/10.1016/j.eswa.2015.07.074>
- [13] Mario, B. (2015) A Density-Based Similarity Matrix Construction for Spectral Clustering. *Neurocomputing*, **151**, 835-844. <https://doi.org/10.1016/j.neucom.2014.10.012>
- [14] Yan, J., Cheng, D.B., Zong, M. and Deng, Z.Y. (2014) Improved Spectral Clustering Algorithm Based on Similarity Measure. In: Luo, X., Yu, J.X. and Li, Z., Eds., *Advanced Data Mining and Applications. ADMA 2014. Lecture Notes in Computer Science*, Vol. 8933, Springer, Cham, 641-654. https://doi.org/10.1007/978-3-319-14717-8_50
- [15] Anil, K.J. (2010) Data Clustering: 50 Years beyond K-Means. *Pattern Recognition Letters*, **31**, 651-666. <https://doi.org/10.1016/j.patrec.2009.09.011>
- [16] Ng, A.Y., Jordan, M.I. and Weiss, Y. (2002) On Spectral Clustering: Analysis and an Algorithm. In: *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, 894-856.
- [17] Yang, P., Zhu, Q.S. and Huang, B. (2011) Spectral Clustering with Density Sensitive Similarity Function. *Knowledge-Based Systems*, **24**, 621-628. <https://doi.org/10.1016/j.knosys.2011.01.009>
- [18] Zhang, X., Jiao, L., Liu, F., et al. (2008) Spectral Clustering Ensemble Applied to SAR Image Segmentation. *IEEE Transactions on Geosciences and Remote Sensing*, **46**, 2126-2136. <https://doi.org/10.1109/TGRS.2008.918647>
- [19] Cai, X.Y., Dai, G.Z. and Yang, L.B. (2009) A New Self-Adaptive Clustering Algorithm. China Science and Technology Press, Beijing, 329-334.
- [20] Fiedler, M. (1975) A Property of Eigenvectors of Non-Negative Symmetric Matrices and Its Application to Graph Theory. *Czechoslovak Mathematical Journal*, **25**, 619-633.

知网检索的两种方式:

1. 打开知网页面 <http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD>
下拉列表框选择: [ISSN], 输入期刊 ISSN: 2161-8801, 即可查询
2. 打开知网首页 <http://cnki.net/>
左侧“国际文献总库”进入, 输入文章标题, 即可查询

投稿请点击: <http://www.hanspub.org/Submission.aspx>

期刊邮箱: csa@hanspub.org