Computer Science and Application 计算机科学与应用, 2016, 6(5), 303-316
Published Online May 2016 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2016.65037

Hans汉斯

Clustering Algorithm and Its Application in the Classification of Aurora Based on the Manifold Distance

Yangzi Sun, Xuan Wang

School of Physics and Information Technology, Shaanxi Normal University, Xi'an Shaanxi
Email: 18700879455@163.com

Received: May 10th, 2016; accepted: May 28th, 2016; published: May 31st, 2016

Abstract

This paper presents a new Spectral clustering analysis algorithm based on the unsupervised learning. Spectral clustering algorithm has its own unique advantage. For example, it can be clustered in any irregular shape of the sample space, but also be obtained the optimal solution in the global. The article prefers to use the clustering algorithm of the similarity measure as the breakthrough point to improve the traditional similarity measure. I use the manifold distance as the similarity measure instead of the Euclidean distance on the basis of the traditional spectral clustering algorithm (NJW-SC). On the basis the object set and the sample clustering can be clustered. After I set the experimental comparison with the new algorithm and K-means algorithm, traditional spectral clustering algorithm (NJW-SC), the fuzzy clustering algorithm (FCM) on artificial data set, it can be concluded that the new algorithm has been achieved good results in the convex shape of the data sets and on the global consistency. On UCI data sets, I tried to use the artificial labeling evaluation index F-measure numerical calculation to carry out on the clustering quality. At last, I chose the aurora images and tried to use them to verify that spectral clustering algorithm also had very good application in the aurora classification.

Keywords

Spectral Clustering (SC), K-Means Algorithm, Manifold, Laplace Matrix

文章引用: 孙羊子, 王晅. 基于流行距离的聚类算法及其在极光分类中的应用[J]. 计算机科学与应用, 2016, 6(5): 303-316. http://dx.doi.org/10.12677/csa.2016.65037

基于流行距离的聚类算法及其在极光分类中的应用

孙羊子，王　晅

陕西师范大学物理学与信息技术学院，陕西　西安
Email: 18700879455@163.com

收稿日期：2016年5月10日；录用日期：2016年5月28日；发布日期：2016年5月31日

摘　要

本文提出了一种新的基于流行距离的谱聚类算法，这是一种新型的聚类分析算法。不仅能够对任意的非规则形状的样本空间进行聚类，而且能获得全局最优解。文章以聚类算法的相似性度量作为切入点，对传统的相似性测度方法进行改进，将传统谱聚类算法(NJW-SC)中的基于欧氏距离的相似性测度换为基于流行距离的相似性测度，在此基础上对样本对象集进行聚类。之后将新提出来的算法同K-Means算法、传统谱聚类算法、模糊C均值聚类算法在人工数据集上进行实验对比，得出新的算法在非凸形状的数据集和在全局一致性上取得了较好的效果。在UCI数据集上用人工评价指标F-measure对聚类质量进行评价，发现其也优于其他方法。在通过实验数据验证后，我将谱聚类算法应用在实际的数据中，看其是否能取得良好的效果。查阅资料，最终选取了极光图像，通过对极光图像的分类验证了谱聚类算法在极光分类中也有很好的应用。

关键词

谱聚类，K-Means算法，流行距离，拉普拉斯矩阵

1. 引言

近年来人们逐步陷入数据丰富而信息匮乏的尴尬境地，有很多数据分类或聚类问题的困扰。我们国家有一句谚语，说“物以类聚，人以群分”，其实也就是说同类的东西聚在一起，相似度会高一些，而不同的则分开。聚类思想也就一直存在，只是随着科学和人类社会的发展，人们逐渐将它概念化、理论化。从古至今，人类处理大型问题的重要手段之一就是分门别类，治而理之。因此，如何将具有性质相同的对象有效划分到同一个子集中变成了我们要研究的问题。聚类分析是常见的一种信息处理的方法，而聚类就是在聚类分析中常用的数据分析工具[1]。不同于分类的需要在已知类别的训练集基础上构建，聚类是一种探索型的分析方法，不必事先给出分类标准，在聚类之前也不知道要将数据划分成几个什么样的组，依靠的仅仅是数据间的相似性。而对于使用不同的聚类方法或者对于不同的研究学者，聚类结果也不尽相同。

相比于基于监督学习的分类方法，基于非监督的聚类方法有它自己独特的优点[2]：首先，收集并标记大型样本集本身就是个费时费力并且低效的工作，有无数的工作量并且在我们并不一定能得到数据的类别属性；其次，待分类样本的性质会缓慢地随着时间的变化而变化，这种随时间变化的性质在无监督学习的情况下更容易得到，同时会提高机器学习的性能；再次，可以在聚类运行过程中提取出数据的一

些基本特征，在后续分类中有可能会用到，可以为后续步骤提供预处理和特征等有效的前期处理；最后，无监督的学习方法是一种观察式的学习，聚类算法能够展示出数据之间隐藏的人类事先未知却有潜在价值的内部结构和规律，更容易找到一些体现数据间结构的有用信息，就可以根据要求更有针对性的设计性能优良的后续分类器。

本文针对聚类算法作以研究，在其中找到合适的谱聚类算法，通过其与 K-means 算法、传统谱聚类算法、模糊 C 均值聚类算法在人工数据集和 UCI 数据集上进行比较。之后将本文所提出的谱聚类算法应用在极光图像分类数据集上，提出一种新的图像分类思路。

2. 谱聚类算法

2.1. 谱聚类算法基本思想

近些年来基于图论方法的聚类算法取得了明显进展。用图论中的图划分准则改进聚类算法性能的问题，将图论和图形学结合在一起的方法用作聚类过程。基于图论方法的主要思想是：将数据集中的数据点当作是图的顶点，数据对象间的相似度用顶点间的连线来体现。当数据点处于同一个连通分支内且数据点之间的相似度比较高时，我们就把这些数据点看作同一个类，其他情况下都属于不同的类。通过图的连通性来体现基于图论方法的这种聚类方式，基于图论的方式我们应用最多的是谱聚类方法，例如 SM 算法、SLH 算法、NJW 算法等。

图论是谱聚类算法的理论来源，其聚类想法来自于谱图的划分，算法的主旨目标：把聚类问题转化为图论中的图分割问题来得到合理的聚类结果。将待聚类数据集中的每个数据点视为无向加权图 $G(V,E)$ 中相对应的顶点 V，将无向图顶点与顶点之间的加权边组成集合，为 $E=\{W_{ij}\}$。$E=\{W_{ij}\}$ 表示规定的某一相似性的度量根据不同的公式通过计算最终得到的数据点和数据点间的相似程度，W 则表示由相似程度所组成的相似性矩阵。由于谱聚类划分准则一般分为 2-way 和 k-way，根据使用的划分准则不同，将算法分为迭代谱和多路谱。算法的思路流程如下[3]-[5]：

(1) 通过计算数据点和数据点之间权值的相似性度量，构建表示数据点相似性的相似性矩阵；

(2) 通过计算相似性矩阵或拉普拉斯矩阵的前 k 个最小特征值所对应的特征向量，并且用这些特征向量构建新的数据特征空间，然后按照一定的划分准则对新的数据空间进行划分，具体划分准则如下：

其一，对于 2-way 的划分，将原始样本数据映射到一维空间($k=1$)中；

其二，对于 k-way 的划分，将原始样本数据映射至到 k 维空间。

(3) 然后对特征向量空间中的特征向量进行聚类，聚类过程相应地也分为以下两种情况：

其一，对于 2-way 的划分，在一维空间中依据目标函数的最优化原则进行划分，然后在两个划分好的子图上进行迭代划分；

其二，对于 k-way 的划分，即利用 K-means、FCM、C-means 等经典算法对新的数据点集进行聚类。

2.2. 基于流行距离的谱聚类算法

采用不同的相似性度量方法是区分谱聚类算法的一个重要步骤，在研究过程中学者们使用最多也最简单快捷的相似性度量是基于欧氏距离(Euclidean Distance)的。使用欧氏距离作为相似性测度在球形分布或规律分布的数据集上有良好的效果，但在非凸形状，或者分布未知的、更复杂的数据集上并没有发挥很好的作用。学者们想去寻求其他的相似性测度来应用于更复杂的数据集，所以想方设法去设计一个更具有弹性的相似性测度方法，而不是单纯基于传统欧式距离。

经过一系列理论研究和实验测试，本文将流行距离运用到一种传统谱聚类算法(NJW-SC)中，提出了一种基于流行距离的谱聚类算法。之后在人工数据集和真实数据集(UCI)上分别进行实验，并且将该算法同 K-means 聚类算法、传统谱聚类算法(NJW-SC)和模糊聚类算法(FCM)进行比较，观察实验结果。

2.2.1. 流行距离

实际生活中遇到的聚类问题，数据集分布是复杂甚至无序的。若只用欧氏距离计算，会使得这种相似性度量不能完全反映聚类全局一致性的特点。也就是说通过计算两点之间距离的欧氏距离只是在空间距离上看似最短，但未必是两点之间的最短最优距离，在进行聚类划分时会产生偏差。

由图 1 我们可以看出，我们期望下面的似“U”状的数据可以和中间似“O”形状的数据集可以完全分为两类。即仅仅通过肉眼观察的话，我们很容易会发现数据点 a 与 e 之间的相似度远大于数据点 a 与 f 之间的相似度。但在机器学习中，机器并没有大脑的多重分析思考能力，我们若只利用欧氏距离作为相似性测度的距离度量办法。根据欧氏距离公式计算的话，由于数据点 a 与 e 的欧氏距离大于数据点 a 与 f 的欧氏距离，这会使得数据点 a 与 f 划分为同一类别的概率远比数据点 a 与 e 划分为同一类别的概率要大得多，所以会增大错误划分的概率。在实际应用中采用欧氏距离虽然操作简单，但降低了划分的正确率，严重影响了聚类算法的性能。

我想着去尝试使用一种相似性度量，它能够反映聚类的全局一致性，期望新的相似性度量可以不仅仅依赖于直线距离，在不同的区域加上合适的权重，最终得到我们想要的结果。如图 1，为了得到更完整的聚类效果，我们要尽可能使得位于同一流形上用更多较短边相连接得到的路径长度短于不同流行间直接相连的两点间距离的长度，如果还是不可以的话，在穿过不同流行距离的数据点间加上大一点的权重。反映在图 1 中也就是要使得多个短距离之和小于不同流行两点间的距离，即 $\overline{ab}+\overline{bc}+\overline{cd}+\overline{de}<\overline{ae}$。通过翻阅资料，我们引入流行距离。下面定义流行距离的一些概念：

定义 1：流形上的线段长度：

$$L\left(x_i,x_j\right)=\mathrm{e}^{\rho dist\left(x_i,x_j\right)}-1 \tag{1}$$

其中，$dist\left(x_i,x_j\right)$ 为 x_i 与 x_j 之间的欧氏距离，ρ 为伸缩因子且有 $\rho>1$。用伸缩因子 ρ 来调节两点间线段的长度，以期达到我们对长度的要求。

定义 2：流行距离测度：

基于流形上的线段长度的公式，我们进一步定义了一个新的距离度量，因为它是位于同一流行和不同流行间的长度，所以我们将其称为流形距离。将数据点看作是一个加权无向图 $G=\left(V,E\right)$ 的顶点 V，用 $E=\left\{W_{ij}\right\}$ 来表示边的集合，即 $E=\left\{W_{ij}\right\}$ 就是反映在每一对数据点间流形上的线段长度。令 $p=\left\{p_1,p_2,\cdots,p_l\right\}\in V^l$ 表示在图上一条连接点 p_1 与 p_l 的路径，其中边 $\left(p_k,p_{k+1}\right)\in E,1\le k\le l-1$。令 $P_{i,j}$ 表示连接数据 x_i 与 x_j 所有路径的集合。则 x_i 与 x_j 之间的流形距离度量定义为：

$$S\left(x_i,x_j\right)=\min_{p\in P_{i,j}}\sum_{k=1}^{l-1}L\left(p_k,p_{k+1}\right) \tag{2}$$

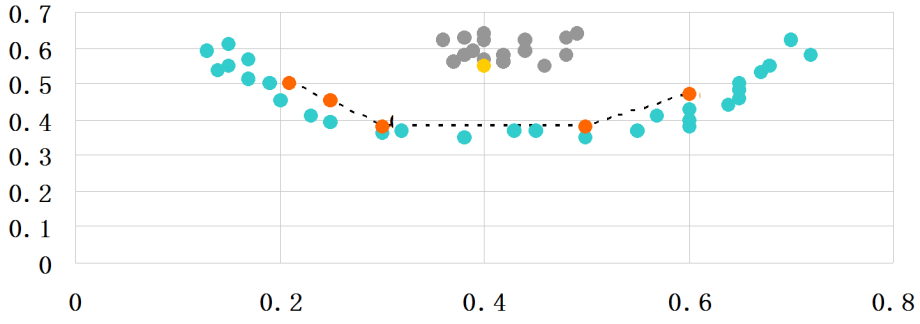


Figure 1. Euclidean distance defects on reflecting global consistency
图 1. 欧氏距离在体现全局一致性上的缺陷

其中，$L\left(p_k, p_{k+1}\right)$ 表示的是两点间流形上的线段长度，即定义 1 中的公式(1)。将(1)带入到(2)中，可以得到：

$$S\left(x_i, x_j\right)=\min_{p\in P_{i,j}}\sum_{k=1}^{l-1}\left[\rho^{dist\left(p_k, p_{k+1}\right)}-1\right] \tag{3}$$

2.2.2. 基于流行距离的谱聚类算法流程

为了在复杂的数据集和非凸形状的数据集中获得更好的聚类结果，我在原有谱聚类算法(NJW-SC)的基础上提出了一种基于流行距离的谱聚类算法，将测度距离由欧式距离替换为流行距离。具体的实现流程如下：

Input：n 个数据点 $\{x_l\}_{l=1}^{n}$，聚类类别数 k，伸缩因子 ρ

Output：数据点的划分 $c_1, c_2, \cdots, c_k$。

Step 1：通过用流行距离的相似性度量来构造相似度矩阵 $S\in R^{n\times n}$，其中流行距离的计算方法如公式(3)所示。

Step 2：构造拉普拉斯矩阵

$$L=D^{-1/2}SD^{-1/2} \tag{4}$$

其中 D 为对角度矩阵

$$D_{ii}=\sum_{j-1}^{n}S_{ij} \tag{5}$$

Step 3：求拉普拉斯矩阵 L 的前 k 个最大特征值所对应的特征向量 $v_1, v_2, \cdots, v_k$，并且构造矩阵 $V=\left[v_1, v_2, \cdots, v_k\right]\in R^{n\times k}$，其中 v_l 为列向量。

Step 4：单位化 V 的行向量，得到矩阵 Y，其中

$$Y_{ij}=\frac{V_{ij}}{\left(\sum_j V_{ij}^2\right)^{1/2}} \tag{6}$$

Step 5：将上式得到矩阵中的每一行都可以看作是 R^k 空间中的任意一点，之后再利用 K-means 这样的经典聚类算法或者其它一些算法将矩阵的所有行聚成 k 类，也就是将初始数据集中所有数据点聚为 k 类。

Step 6：如果矩阵 Y 中的第 i 行在聚类过程中是属于第 j 类，那么将初始数据点中的点 x_j 就划分到第 j 类，最终完成聚类的过程。

3. 实验及其分析

为了验证本文所提出的基于流行距离的聚类算法的聚类性能，我在此将该算法应用在 4 个人工数据集和 7 个真实 UCI 数据集的聚类问题上，并分析实验结果。同时加入对比试验，将我自己提出的算法(MDNJW)与 K-means 聚类算法、传统谱聚类算法(NJW-SC)和模糊聚类算法(FCM)进行比较，验证提出算法的可行性。

3.1. 对人工数据集聚类

我选取 two-spiral2、smile、blobls and circle 和 two_moons 和 4 个人工数据样本来进行实验，并加入对比实验。如图 2~图 5 是在 4 种聚类算法应用在 4 个不同人工数据集上得到的聚类结果。其中(a)是 K-means 聚类算法，(b)是传统谱聚类算法(NJW-SC)，(c)是模糊聚类算法(FCM)，(d)是我自己提出的算法(MDNJW)。

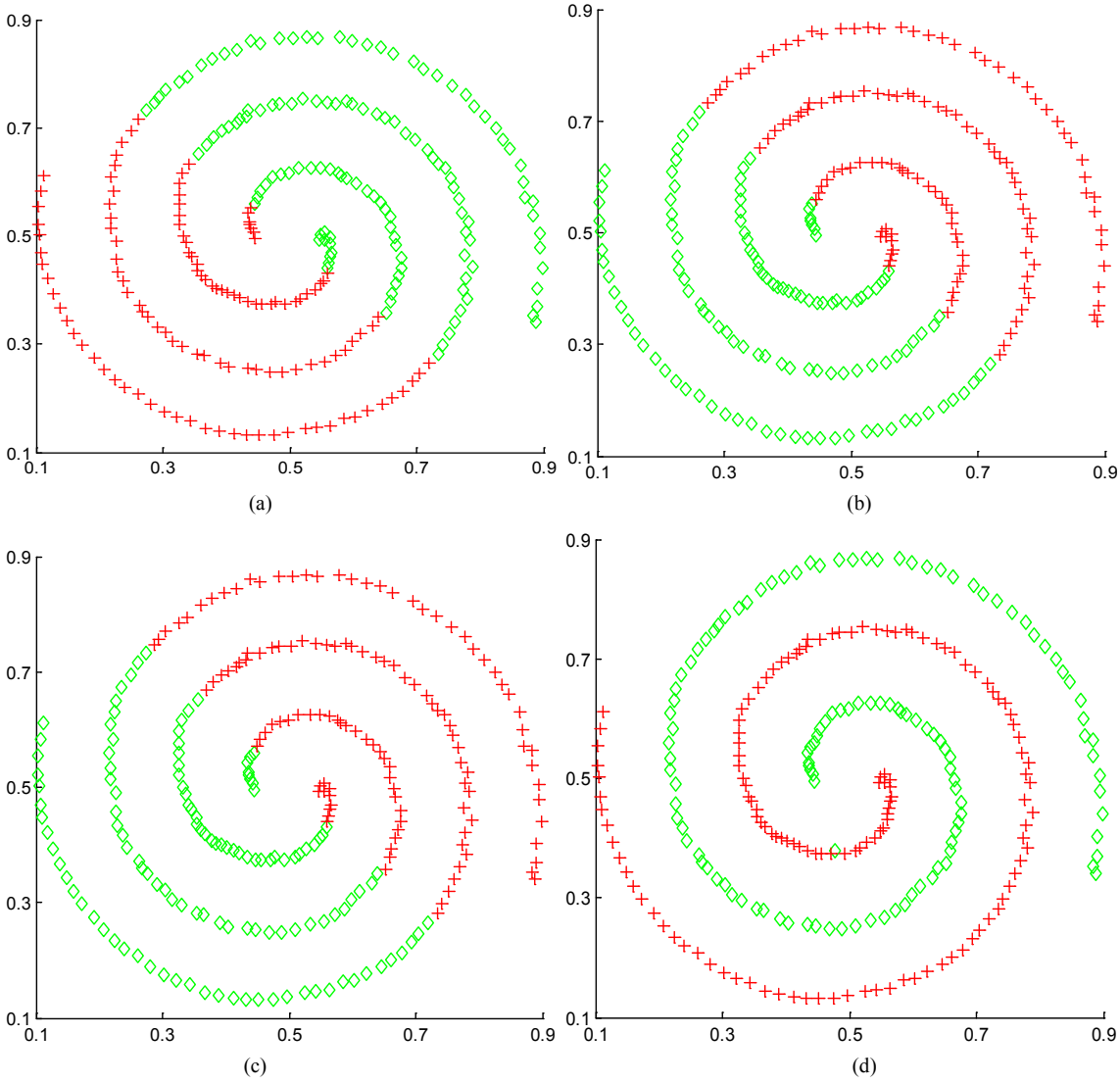
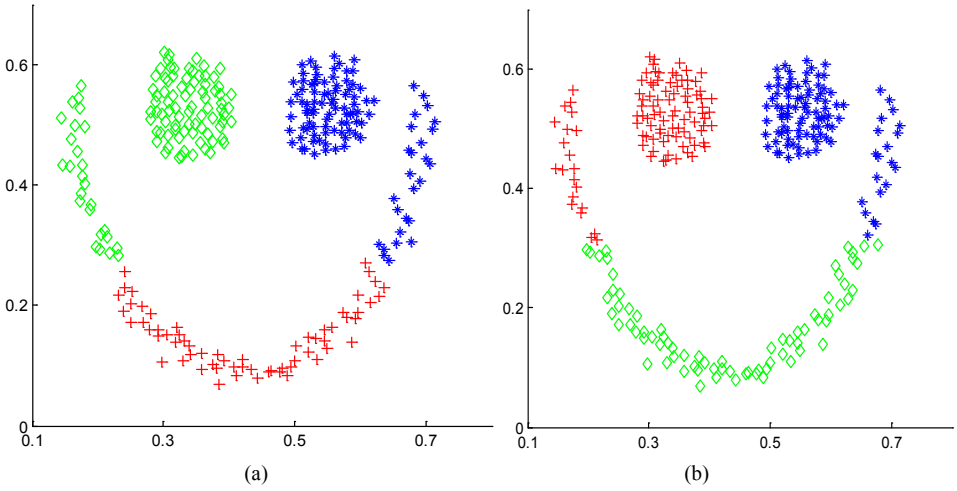


Figure 2. Contrast of four algorithms on artificial data sets two-spiral2
图 2. 人工数据集 two-spiral2 的 4 种算法对比



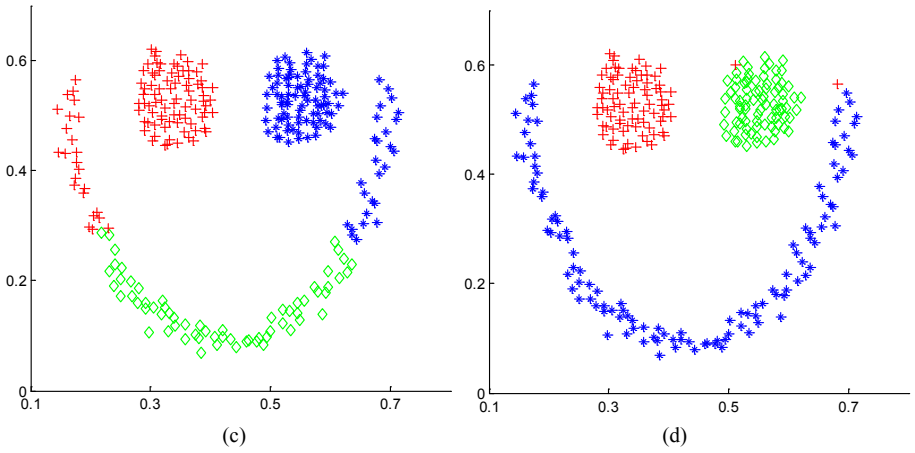


Figure 3. Contrast of four algorithms on artificial data sets smile
图 3. 人工数据集 smile 的 4 种算法对比

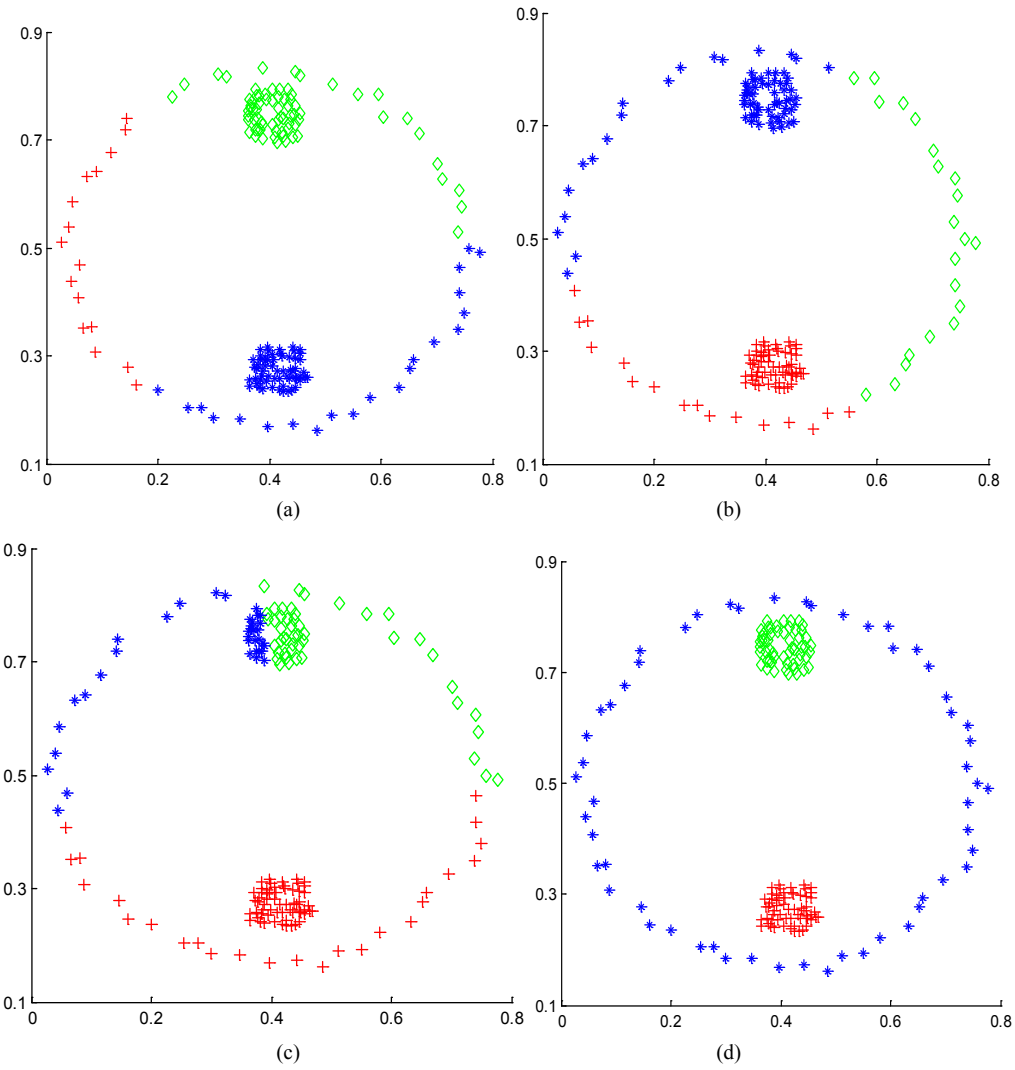


Figure 4. Contrast of four algorithms on artificial data sets blobls and circle
图 4. 人工数据集 blobls and circle 的 4 种算法对比

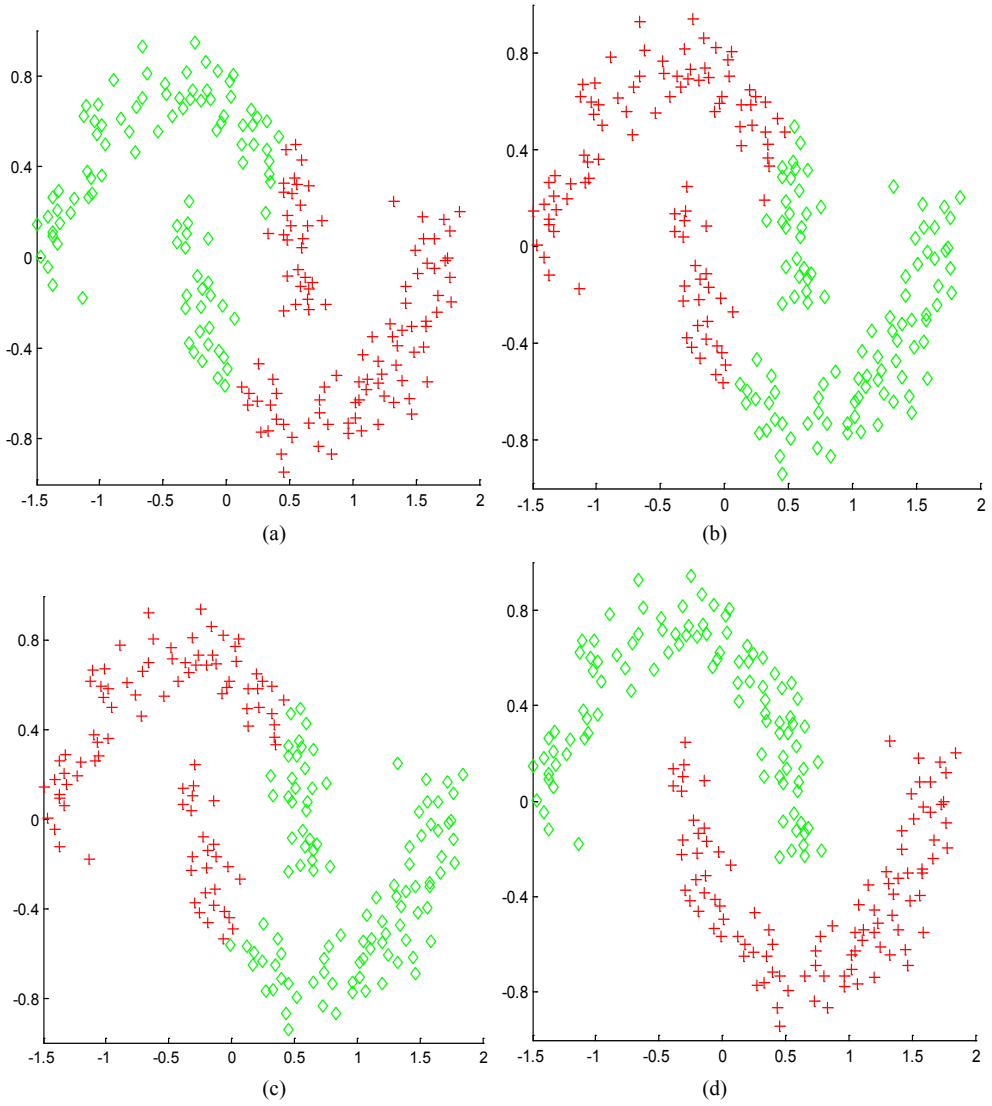


Figure 5. Contrast of four algorithms on artificial data set two_moons
图 5. 人工数据集的 two_moons 4 种算法对比

从图 2~图 5 中，我们明显可以看出对于各种不同数据结构的人工数据集，基于流行距离的谱聚类算法(即 MDNJW)都能获得较好的聚类效果。

3.2. 对 UCI 数据集聚类

UCI 数据集是一个用于机器学习的常用标准测试数据集，是 University of California Irvine 提出的真实数据集。在我自己的论文中，为了进一步考察我所提出来的算法是否优于之前的其他算法，所以再次在 UCI 数据集上进行聚类分析。表 1 列出我所选用的数据集以及数据特征。

之后对聚类的正确率进行评价，验证提出算法的可行性，我使用的是基于聚类簇的评价指标 F-measure [6]-[8]。对于聚类结果簇 i 和簇 j 的计算准确率 P (Precision Rate)、召回率 R (Recall Rate)和 F-measure 公式如下：

$$P(i,j)=\frac{n_{ij}}{n_j} \tag{7}$$

$$R(i,j)=\frac{n_{ij}}{n_i} \tag{8}$$

$$F(i,j)=\frac{2*P(i,j)*R(i,j)}{P(i,j)+R(i,j)} \tag{9}$$

上面三个式子中：n_i 为簇 i 中所包含的数据样本数目；而 n_j 为簇 j 中所包含的数据样本数目；n_{ij} 表示一个错误的划分，本该属于簇 j 却被错划为簇 i 的数据样本数目。

公式(4)~(10)是对数据集进行聚类后最终所得到的结果，通过计算它最大 F-measure 的加权平均来评价该聚类算法的整个聚类结果。我们将该检测指标称为聚类结果的总体正确率 F-measure，记为 F：

$$F=\sum_i \frac{n_i}{n}\max_j\left\{F(i,j)\right\} \tag{10}$$

由于 F 是正确率的评价指标，故其取值范围是[0,1]，F 值越大表明簇内越紧密，簇间分离度增大，正确率增高，聚类结果越完善。

图 6 是 4 种聚类算法在所选定的真实 UCI 数据集上的聚类效果 F-measure 的值(伸缩性参数 ρ 分别采用 0.1；0.5；1；0.1；500；10；100)。

表 2 是 4 种聚类算法在所选定的 7 个 UCI 数据集上的聚类结果 F-measure 的比较。其中每一个算法在同一个数据集上的顺序依次是 K-means 聚类算法，传统谱聚类算法(NJW-SC)，模糊聚类算法(FCM)，我自己提出的算法(MDNJW)。

Table 1. The experiment selected data characteristics of the data set
表 1. 实验选取的数据集的数据特征

数据集	样本个数	维数	类别数
Soybean	47	35	4
waveform	5000	21	3
Sonar_all_data	208	60	2
iris	150	4	3
glass	214	9	6
Libras_movement	360	89	15
zoo	101	16	7

Table 2. The value of F-measure on 7 data sets using 4 algorithms
表 2. 4 种算法在 7 个数据集上的 F-measure 值

数据集	K-means	NJW-SC	FCM	MDNJW
Soybean	0.67	0.79	0.78	0.82
waveform	0.53	0.51	0.54	0.56
Sonar_all_data	0.57	0.55	0.56	0.67
iris	0.88	0.87	0.89	0.93
glass	0.53	0.54	0.55	0.57
Libras_movement	0.44	0.47	0.45	0.53
zoo	0.88	0.86	0.81	0.87

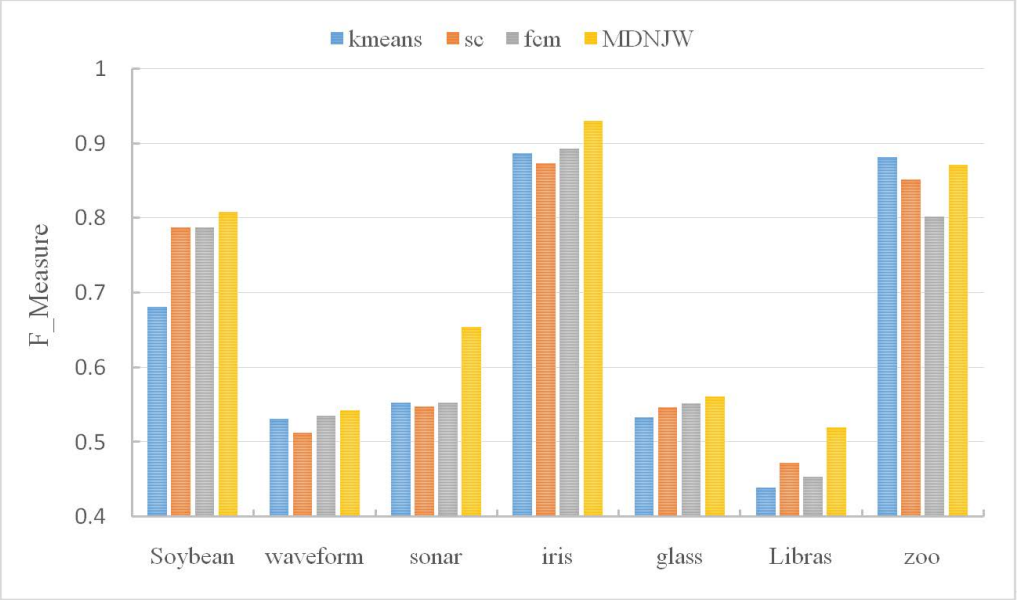


Figure 6. The comparison of the value of F-measure on 7 data sets using 4 algorithms
图 6. 4 种算法在 7 个数据集上的 F-measure 的比较

3.3. 鲁棒性分析

为了进一步分析 4 种算法的优劣，我尝试对四种算法的鲁棒性进行考察分析，根据参考文献[9]中提出的方法对上述四种算法在 UCI 数据集上的鲁棒性进行比较。具体如下公式：

$$b_m = \frac{P_m}{\max_k P_k} \tag{11}$$

其中分子 P_m 是我们用某一算法获得的聚类正确率，分母是 P_k 是解决这个问题所有聚类算法得到的最大聚类正确率，我们用两数相除得到某一算法在某一数据集上的相对性能值。

用 m^* 来表示在某个数据集上能实现的最好聚类结果的算法，相对性能我们记作 $b_{m^*}=1$，那么其他算法在该数据集上的相对性能取值范围就为 $0 \le b_m \le 1$。b_m 的值越大，表示算法 m 在该数据集上的所有算法中的相对性能越好，聚类效果越好。我所进行的这种鲁棒性分析方法，不单纯是分析某一算法在某一数据集上的单独作用，而是将所有数据集在该算法上的相对性能之和。因此，用相对性能之和 b_m 来对这个算法的鲁棒性进行客观和量化评价，b_m 总和(sum)值越大，这个算法就拥有更好的鲁棒性。4 种聚类算法在 UCI 数据集上的鲁棒性比较和鲁棒性之和如表 3。

从上表可以看出，除过在数据集 zoo 之外，基于流行距离的谱聚类算法(即 MDNJW)都拥有最大的相对性能，理所应当它也获得了 b_m 总和最大的值，b_m 第二大的值是传统的谱聚类算法(NJW-SC)。所以我提出的算法由于改进了传统的谱聚类算法容易局部最优的缺点，获得了好于传统谱聚类算法的鲁棒性。

图 7 展示了在 7 个 UCI 数据集上每一种算法相对性能的分布情况。在每一种算法中，相对性能的 7 个值堆叠起来，最终堆叠成的结果就是该算法的鲁棒性之和，可以一目了然的看到。可以看出我提出算法有更好的鲁棒性。

4. 基于流行距离的谱聚类算法在极光分类中的应用

4.1. 极光分类发展现状

极光图像的分类和特征提取研究是个交叉学科应用课题，起初极光图像的分类和特征提取研究，都

Table 3. The robust of the 4 algorithms on UCI data set
表 3. 是 4 种算法在 UCI 数据集上的鲁棒性

数据集	K-means	NJW-SC	FCM	MDNJW
Soybean	0.842	0.974	0.974	1.000
waveform	0.978	0.943	0.986	1.000
Sonar_all_data	0.846	0.838	0.846	1.000
iris	0.953	0.939	0.961	1.000
glass	0.950	0.975	0.983	1.000
Libras_movement	0.845	0.909	0.872	1.000
zoo	1.000	0.966	0.910	0.989
sum	6.414	6.545	6.531	6.989

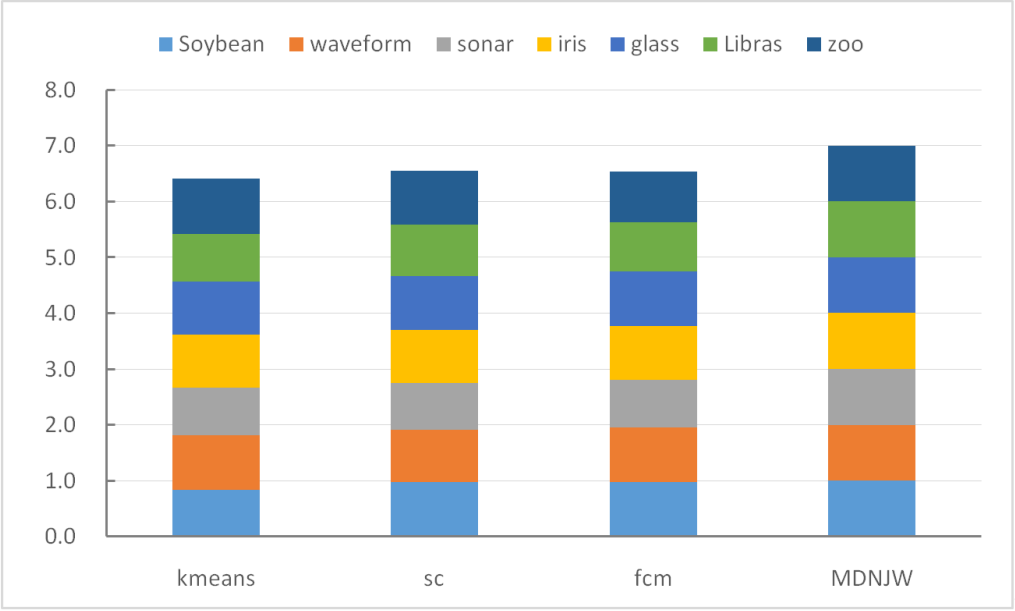


Figure 7. The comparison of the robust of the 4 algorithms on UCI data set
图 7. 4 种算法在 UCI 数据集上的鲁棒性比较

是用眼睛先观察，再用手工完成标记和分类。之后，Carl Stormer 通过对大量的极光数据分析在 1955 年提出了对极光图像分类的一种方法，即将所有的极光图像分为三大类，依次是无放射状的结构、有放射状的结构和火焰形状结构。之后所有学者的分类方法都是基于此。2004 年学者 Syrjäsuo 和他的研究团队首次的将数字图像处理和机器视觉技术引入到了极光分类的研究领域，从此，自动的极光分类方法产生。自动的分类技术主要的目的就是利用极光图像的纹理特征对弧形极光，斑块型极光和欧米伽型极光进行自动分类。2009 年，学者胡泽俊引入了形态学的概念，将极光图像分为了冕状形态极光和弧光状的极光两大类，然后对冕状的再进行细分，即热点冕状极光、辐射冕状极光和帷幔型的冕状三种极光。在我之后的论文中，就利用了学者们之前研究的方法对极光图像进行大致的分类。

4.2. 极光图片处理

在我所选用的图片集中，将极光大致划分为以下几类：分别是弧状极光(arc)、帷幔状极光(drapery)、

热斑点状极光(hotspot)和射线状极光(radial)。

每一类的特点：

弧状(arc)，这种极光包含一个或多个极光弧，如图 8(a)。

帷幔状(drapery)，射线结构明显，并且多层重叠排列，看起来平稳，分布广，如图 8(b)。

热斑点状(hotspot)，极光结构较为复杂，既包含光纤结构也包含光斑，如图 8(c)。

射线状(radial)，光线由中心向四周呈辐射发散状，射线结构由中心指向四周，如图 8(d)。

由于极光图片较为多，所以我在四类中各选取了 50 张差异比较小的图片，一共 200 张组成一个全新的极光图片集，标号后待用。

4.3. 实验过程

4.3.1. 图片的特征提取

对得到的 200 幅图片进行特征提取，首先经过 Radon 变换，固定 θ 的值求其在 r 方向上的均值方差。将 θ 值在[0,179]变换得到 180 个均值方差值，通过最大值循环移位的方法将方差最大的值放在第一位。每幅图生成 180 个数，为 1×180 的矩阵，200 幅图生成一个 200×180 的矩阵，生成我之后所要用的数据集。并在矩阵第一行上标注每张图片所属的类，待之后使用。

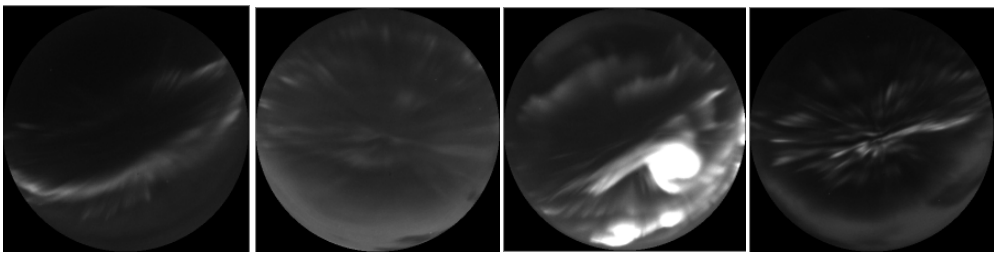
4.3.2. 实验结果

将特征提取后的数据集保存在一个矩阵中，其中矩阵的第一列是已知的该图片的分类类别，用 F-measure 为聚类结果的正确率做评价，并与传统谱聚类算法(NJW-SC)，模糊 C 均值聚类算法(FCM)进行比较，验证实验效果。表 4 是聚类结果的评价指标，其中第一行是伸缩性参数 ρ 的取值。

4.3.3. 加噪处理

在这部分中，为了验证所提出的方案的抗噪性能，在极光图像中加入两种常见的噪声，第一种是常见的高斯噪声，其均值为 0，方差为 0.01；第二种是采用密度为 0.05 的椒盐噪声。

高斯噪声是一种常见噪声，其概率密度函数满足高斯分布，例如随机变量 z 满足高斯分布，则高斯函数



(a) 弧状(arc) (b) 帷幔状(drapery) (c) 热斑点状(hotspot) (d)射线状(radial)

Figure 8. The classification of the aurora images
图 8. 极光图片的分类

Table 4. The comparison of F-measure value of the aurora image classification data sets using 3 algorithms
表 4. 3 种算法在极光图像分类数据集上的 F-measure 值的比较

算法	0.1	1	10	100	1000	10,000
MDNJW	/	0.3150	0.3900	0.3850	0.4050	0.4050
NJW	0.3150	0.4250	0.4600	0.4550	0.4550	/
fcm			0.3650			

Table 5. The value of F-measure after adding noise using 2 algorithms
表 5. 2 种算法在加噪后的 F-measure 值

噪声	高斯噪声	椒盐噪声
MDNJW	0.3850	0.3700
NJW	0.4350	0.4150

$$p(z)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right] \tag{12}$$

在图像处理中，z 表示的是图像的灰度值，μ 表示的是数学期望值，σ 表示的是方差。当 z 服从高斯分布的时候，落在$\left[(\mu-\sigma),(\mu+\sigma)\right]$范围内的灰度值图像大概为 70%，落在$\left[(\mu-2\sigma),(\mu+2\sigma)\right]$范围内的大概有 90%。

椒盐噪声：椒盐噪声其实也就是我们经常说的双脉冲噪声，此噪声在图像上的随机分布特别像胡椒和盐粉微粒，它一般产生或者出现在图像传感器，传输信道，图像的解码处理等这些步骤中。因为图像的切割，往往都会产生出椒盐噪声干扰。所以在下面的实验中加入密度为 0.05 的椒盐噪声来干扰预处理过的极光图像。

根据上面的实验结果发现在 MDNJW 和当 NJW 算法中，当 ρ 分别取 1000 和 10 时获得最好的聚类结果，所以我在 ρ 取 1000 和 10 时加入这两种噪声，得到的实验聚类结果的评价指标的 F-measure 如表 5。

4.3.4. 结果分析

从上面所得出的实验数据看，谱聚类算法的结果好于 K-means 聚类算法，我所提出的算法相比于其他算法，获得了相对较为良好的聚类结果，验证了基于流行距离的谱聚类算法在极光分类中有一定的作用。但是无论是我所提出的算法还是其他经典算法，在极光分类中都没有获得特别好的聚类效果，远不如监督学习的聚类算法那么优异的分类效果，还有很大的改进空间。之后可以在两个方面继续努力，一个是特征提取；另一个是谱聚类算法的改进。ρ 值的调节可以影响到谱聚类算法的聚类效果，在调试过程中对 ρ 值进行不断调整，以期得到更高正确率的聚类结果。

5. 结束语

由于基于欧式距离的谱聚类算法在非凸形状上容易陷入局部最优解，不能达到较好的聚类效果。为了提高聚类质量，将基于流行距离的相似性测度方法用在谱聚类算法中，在人工数据集和 UCI 真实数据集上得以应用，并将提出的方法与传统的三种算法进行比较。之后将所提出的方法应用在极光图像的分类中，虽然没有取得非常好的聚类效果，但提供了一种新思路和方法。在我提出的算法中也有待提高的方面，能够降低计算量的方法，如果能够在不伤害全局一致的情况下降低运算复杂度，找到其他的可以满足性能优越、代价又小的相似性测度。

参考文献 (References)

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

[2] Duda, R.O., Hart, P.E. and Stork, D.G. (2001) Pattern Classification. 2nd Edition, John Wiley & Sons, New York.

[3] Che, W.F. and Feng, G.C. (2012) Spectral Clustering: A Semi-Supervised Approach. *Neuro Computing*, **77**, 119-228.

[4] Zhao, F., Liu, H. and Jiao, L. (2011) Spectral Clustering with Fuzzy Similarity Measure. *Digital Signal Processing*, **21**, 56-63. http://dx.doi.org/10.1016/j.dsp.2011.07.002

[5] Alzate, C., Johan, A. and Suykens, K. (2012) Hierarchical Kernel Spectral Clustering. *Pattern Recognition*, **35**, 24-35.

[6] 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.

[7] Fiedler, M. (1975) A Property of Eigenvectors of Non-Negative Symmetric Matrices and Its Application to Graph Theory. *Czechoslovak Mathematical Journal*, **25**, 619-633.

[8] 贾建华. 谱聚类集成算法研究[M]. 天津: 天津大学出版社, 2011.

[9] Geng, X., Zhan, D.C. and Zhou, Z.H. (2005) Supervised Nonlinear Dimensionality Reduction for Visualization and Classification. *IEEE Transactions on Systems*, *Man*, *and Cybernetics*, *Part B*: *Cybernetics*, **35**, 1098-1107. http://dx.doi.org/10.1109/TSMCB.2005.850151