基于启发式确定类数的 NJW 谱聚类算法

陈俊芬[1] 张 明[1] 何 强[2]

(河北大学数学与信息科学学院河北省机器学习与计算智能重点实验室 保定 071002)[1]

(北京建筑大学理学院 北京 100044)[2]

摘 要 基于图论理论的 NJW 谱聚类算法的核心思想是将数据点映射到特征空间后再利用 K-means 算法进行聚类,从而得到原始数据的聚类结果。NJW 算法是 K-means 算法的推广,并且在任意形状的数据上都具有较好的聚类效果,从而有着广泛的应用。但是,类数 C 和高斯核函数中的尺度参数 σ 较大程度地影响着 NJW 的聚类性能;另外,K-means 对随机初始值的敏感性也影响着 NJW 的聚类结果。为此,一种基于启发式确定类数的谱聚类算法(记为 DP-NJW)被提出。该算法先根据数据的密度分布确定类中心点和类数,这些类中心点作为特征空间中 K-means 聚类的初始类中心,然后用 NJW 进行聚类。文中通过实验将 DP-NJW 算法和经典聚类算法在 7 个公共数据集上进行测试和对比,其中 DP-NJW 算法在 5 个数据集上的聚类精度高于 NJW 的平均聚类精度,在另 2 个数据集上二者持平。对比 DPC 算法,所提算法在 5 个数据集上也有不俗的聚类精度,而且 DP-NJW 的计算消耗较小,在较大的数据集 aggregation 上表现更为突出。实验结果表明,文中所提的 DP-NJW 算法更具优势。

关键词 谱聚类,NJW 聚类,高斯核函数,尺度参数,聚类精度

中图法分类号 TP181 **文献标识码** A

Heuristically Determining Cluster Numbers Based NJW Spectral Clustering Algorithm

CHEN Jun-fen[1] ZHANG Ming[1] HE Qiang[2]

(Hebei Key Laboratory of Machine Learning and Computational Intelligence, College of Mathematics and Information Sciences, Hebei University, Baoding 071002, China)[1]

(School of Science, Beijing University of Civil Engineering and Architecture, Beijing, 100044, China)[2]

Abstract The main idea of NJW is to project data points into feature space and then to cluster using K-means algorithm, while the clustering results is considered as the clustering of the original data points. However, the number of clusters C and scaling parameter σ of Gaussian kernel function greatly influence the clustering performance of NJW algorithm, whilst the fact that K-means algorithm is sensitive to initial cluster centers also influences the clustering result of NJW algorithm. To this end, an improved NJW algorithmis presented referring to DP-NJW algorithm which heuristically determines cluster-center points and the number of clusters according to density distribution of data points, and then implements NJW algorithm to cluster. Note that, the obtained cluster-center points and the number of clusters in the first stage are the initialization of K-means algorithm in the second stage of the proposed DP-NJW algorithm. In the next section, DP-NJW is compared with state-of-the-art clustering algorithms on the given seven public datasets, where DP-NJW provided higher clustering accuracy than NJW on five datasets and even on the other two datasets. DP-NJW algorithm provided better clustering performance than DPC on the given five datasets. In addition, DP-NJW consumed less computing time than the other two algorithms, and this is especially obvious on the larger aggregation dataset. Overall, DP-NJW algorithm is superior to the state-of-the-art clustering algorithms.

Keywords Spectral clustering, NJW clustering, Gaussian kernel function, Scaling parameter, Clustering accuracy

1 引言

聚类是根据预先设定的相似性度量将数据集分成若干簇或类的过程,其目的是使同一类别的数据相似性最大而不同类别间的数据相似性最小。聚类算法是机器学习领域的非监督学习策略,并且已被广泛应用于模式识别、图像分析、视频分割、文本分类和文本搜索等领域。常见的聚类算法有 K-means 算法、CURE 算法、DBSCAN 算法和 NJW 算法等。

MacQueen 于 1967 年提出了经典的 K-means 聚类算法[1],他总结了 Cox, Fisher 和 Sebestyen 等的研究成果[2-5],给出了 K-means 算法的详细步骤,并用数学方法进行了证明。该算法简单、易实现,且聚类效果很好,因此被广泛应用。但是,其聚类结果依赖于预设的聚类数目 K 和随机初始化的聚类中心,需要通过多次运行算法来找到最优解(通常是局部

本文受河北省自然科学基金(F2016201161),国家自然科学基金(61473111)资助。

陈俊芬(1976—),女,博士,副教授,CCF 会员,主要研究方向为数据分析与机器学习,E-mail:chenjunfen2010@126.com(通信作者);张 明(1993—),男,硕士生,主要研究方向为数据分析和深度学习;何 强(1977—),男,博士,副教授,主要研究方向为机器学习。

最优解）。更多关于 K-means 聚类的相关知识可参见综述文献[6-8]。

谱聚类算法建立在图论中的谱图理论基础上，其本质是将聚类问题转化为图的最优化分问题。由 Person 和 Freeman 于 1998 年提出的 PF 算法[9]是最简单也是应用最广泛的迭代谱聚类算法。PF 算法主要使用数据集的相似度矩阵 W 的最大特征值对应的特征向量来完成聚类，特征向量中非零元素对应的点属于同一类，零元素对应的点属于另外一类。研究表明，使用更多的特征向量将会得到更好的聚类效果。由 Ng，Jordan 和 Weiss 于 2002 年提出的 NJW 算法[10]是多路谱聚类的典型代表，其选取拉普拉斯矩阵的前 h 个最大特征值对应的特征向量，并在这个特征向量空间中利用经典的 K-means 算法进行聚类。谱聚类可被视为一种改进的 K-means 聚类算法，在任意形状的数据分布空间中都具有较好的聚类效果，并且收敛于全局最优解。近些年，为了处理图像、视频、文本等高维数据的聚类问题，学者们提出了基于谱聚类的稀疏子空间分割算法及各种改进算法[11-14]。更多谱聚类算法的优缺点可参见综述文献[15-16]。

相似度矩阵 W 对谱聚类有很大的影响，常用高斯核函数来表示两数据点间的相似度进而得到相似度矩阵 W，但尺度参数 σ 的选取使该函数具有局限性。NJW 算法中预先指定几个尺度参数值，分别执行谱聚类，最后选取使得聚类效果最好的一个 σ 值，这种做法消除了尺度参数选取的人为因素，但增加了运行时间[10]。为了改进单个尺度参数 σ 对所有样本数据的不适合性，Zelnik-Manor 首次将局部尺度引入到高斯核函数中，提出了自修正的谱聚类算法[17]，该算法利用每个样本点到它的第 t 个近邻点的距离作为 σ 值（通常取 $t=7$）。后来的一些研究者将这种局部尺度的自修正思想用于（半监督）谱聚类来自动确定数据点对的相似性[18-20]。Yang 等将某些数据点必须关联和必须不关联两约束引入到局部尺度的自修正谱聚类中，从而提出了 STS^{3}C 谱聚类算法[15]。基于局部尺度的自修正聚类方法降低了对参数和异常点的敏感性，增加了算法的鲁棒性。但是这类方法采用第 t 近邻的距离，其中 t 值仍由人工确定；另外，对每一个数据点都要计算一个 σ 值，这大大增加了计算开支。

如何自动确定聚类数目一直是各种聚类算法需要解决的关键问题之一，Ng 等通过分析特征值的特性指出特征值是 1 的个数等于聚类数目，但是这种方法对噪声很敏感，对各类有重叠的数据不适用[10]。后来，Polito 等根据相邻特征值差的变化来确定聚类数目，该方法虽然有效，但是缺乏理论分析[21]。Zelnik-Manor 和 Perona 从研究特征向量的特性出发，通过梯度下降方法求解一个目标优化来确定聚类数目[17]。显然，该方法的计算复杂度被大大提高。Zeng 等通过一种用于模糊聚类的有效性测量来确定基于粒子群优化的谱聚类数目[22]。自动确定聚类数目的方法还有基于 Rcut 的谱聚类算法[23]和非线性降维算法[24]等。上述方法是基于特征值或者特征向量确定聚类数目，在一些公共数据集上得不到令人满意的聚类数目和聚类结果。

为了改进 NJW 算法中的这几个问题，以进一步提高 NJW 的聚类性能，本文提出了一种启发式确定聚类数目的 NJW 算法，记为 DP-NJW 算法。该算法综合了自动确定聚类中心的 DPC 算法和经典 NJW 算法的优势，改善了 NJW 算法对随机初始值的敏感性。DPC 算法是 2014 年由 Rodriguez 和 ALaio 基于数据点的局部密度提出的一种新颖的聚类算法[25]。DPC 算法中每个数据点用局部密度值和高斯截断值来描述，根据对聚类中心的特征描述找出聚类中心点，进而得到了聚类个数。文中还将所提算法在聚类算法常见且具有挑战性的 7 个数据集上进行测试。本文的主要贡献包括：

（1）NJW 和 DPC 算法都用到了高斯核函数，为了一致性，本文采用 $\exp\{-dis^2(x_i,x_j)/2\sigma^2\}$，其中 $dis(x_i,x_j)$ 是数据点 x_i 和 x_j 的欧氏距离。

（2）使用试错法测试多个预先给定的尺度参数 σ 值，根据 NJW 的聚类精度选择一个最优 σ 值。

（3）对于特征空间中的 h 个特征向量，使用试错法测试多个预给定的 h 值，并根据 NJW 的聚类精度选择最优的 h 值。

（4）结合 DPC 确定聚类中心点的思想与 NJW 聚类思想，提出了启发式的 DP-NJW 算法。

本文第 2 节简单回顾谱聚类 NJW 算法的核心步骤；第 3 节详细描述了本文所提 DP-NJW 算法；第 4 节将 DP-NJW 算法和经典聚类算法在 7 个数据集上进行对比测试，并给出了详细的实验分析；最后给出本文的结论与展望。

2　NJW 谱聚类算法

谱聚类的核心思想是构建样本点的相似度矩阵（图），并将图切割成 K 个子图，使得各个子图内的相似度最大而子图间的相似度最弱。根据不同的准则函数及谱映射方法，谱聚类算法有着不同的具体实现方法，这些谱聚类方法的总体框架为定义一个描述成对数据点相似度的亲和矩阵，构造拉普拉斯矩阵并计算特征值和特征向量，选择合适数量的特征向量进行聚类，其结果就是原始数据点的聚类。谱聚类具有坚实的理论基础，相对于其他聚类方法具有许多优势和更广泛的应用空间。

经典的谱聚类 NJW 算法由 Ng 等[10]于 2002 年提出，该算法的主要步骤如下：

Step1　计算相似度矩阵 W，进而获得拉普拉斯矩阵 L。

Step2　计算矩阵 L 的特征值和特征向量，并选取前 h 个最大特征值对应的特征向量来构造矩阵 $X=[x_1,x_2,\cdots,x_h]_{n\times h}$。

Step3　将矩阵 X 的行向量标准化为单位向量，得到矩阵 Y。

Step4　采用 K-means 或者其他聚类方法对 n 个特征点（Y 的每一行是一个特征点）进行聚类，得到 K 个类簇。

谱聚类 NJW 是将原始数据映射到特征空间后再进行 K-means 聚类，从而被视为一种改进的 K-means 聚类算法。因此，NJW 比 K-means 更适合于高维非凸分布的数据和数据分布不均匀的情况。

3　改进的 DP-NJW 谱聚类算法

为了减弱 NJW 算法对 K-means 随机初始值的敏感性，本文提出了新的改进 NJW 算法，记为 DP-NJW 算法。该算法先在数据空间中根据每个点的局部密度和对某个点的欧氏距离来确定最优中心点和类别数，这些聚类中心点和类别数是在特征空间中采用 K-means 对特征数据点聚类时的聚类初始值。由于聚类初始值已经接近最优中心点位置，因此接

下来 K-means 的聚类过程只需要迭代一次或者几次即可。所提 DP-NJW 算法中聚类中心的初始值是固定的，也不需要多次重复实验来寻找(局部)最优解。其主要步骤如下：

(1)已知数据集 $X=\{x_1,x_2,\cdots,x_N\}$，其中 $x_i\in R^n$，用 z-score 对数据进行标准化后仍记为 $X=\{x_1,x_2,\cdots,x_N\}$。先根据式(1)计算数据点对 (x_i,x_j) 的欧氏距离 d_{ij}，再根据式(2)计算两点间的高斯距离 w_{ij}，进而构成相似矩阵 W。

$$d_{ij}=\|x_i-x_j\| \tag{1}$$

$$w_{ij}=\exp\{-dis^2(x_i,x_j)/2\sigma^2\} \tag{2}$$

(2)对 W 的每一行求和，生成一个对角矩阵 D。

(3)将 W 主对角线上的元素 1 换成 0，并对每一行求和，生成一个行向量 $\rho=[\rho_1,\rho_2,\cdots,\rho_N]$ 来表示每个数据点的局部密度值。最大的局部密度值 ρ_u 对应的数据点为 x_u，其距离 δ_u 定义为该点到数据集合中数据点的最远距离；第二大局部密度值 ρ_l 对应的数据点为 x_l，其距离 $\delta_l=d_{lu}$；对于第三大局部密度值 ρ_v 对应的数据点 x_v，定义其距离 $\delta_v=\min(d_{vu},d_{vl})$；其余点依此类推。

(4)根据 ρ 和 δ 值(一共是 N 对)画出平面决策图，选出 ρ 和 δ 值较大的点作为聚类中心点，同时也得到了聚类数目 K。

(5)构造规范的拉普拉斯矩阵 $L=D^{-1/2}WD^{-1/2}$，并求其特征值与特征向量。选取前 h 个最大的特征值对应的特征向量形成矩阵 $B_{N\times h}$。

(6)对矩阵 B 的行进行归一化，即 $\tilde{b}_{ij}=b_{ij}/\sqrt{\sum_{j=1}^{h}b_{ij}^2}$，得到矩阵 $\tilde{B}$。

(7)将 $\tilde{B}$ 的每一行看成特征空间中的一个数据点 $\tilde{x}_i$。根据步骤(4)确定的聚类个数 K 和聚类中心位置，利用 K-means 算法对这 N 个特征数据点进行聚类。

(8)输出最终聚类中心点和每个数据点的聚类标号。

根据上述描述过程，本文所提 DP-NJW 算法如算法 1 所示。

算法 1 启发式确定聚类个数的 DP-NJW 聚类算法

输入：数据集 $X=\{x_1,x_2,\cdots,x_N\}$

输出：聚类后的类标向量 gnd，聚类中心点 $c_1,c_2,\cdots,c_K$

初始化：聚类中心点 $c_1,c_2,\cdots,c_K$ 和类别数 K

```
for 对每一个数据点
    计算 ρ,δ 值
end
    画出决策平面图；
    确定中心点和类别数；
    计算拉普拉斯矩阵 L 并对其进行特征分解；
    减少特征向量的维度；
repeat
    对降维后的特征向量进行 K-means 聚类；
until 聚类中心点不再变动或达到了最大循环次数
```

DP-NJW 算法和 NJW 算法的对比分析：本文所提的 DP-NJW 算法相对 NJW 算法来说，增加的计算量包括向量 ρ 排序和每一个数据点到比它密度大的数据点的最小欧氏距离的计算。但是由于初始中心点很接近最优聚类中心点，因此在聚类阶段 K-means 迭代较少次就能收敛到最优聚类中心；另外，由于初始聚类中心点固定，因此 DP-NJW 算法更稳定，不需要重复多次聚类来寻找一个最优的聚类结果。第 4 节的实验对比也会证实此处的分析。

4 实验与分析

本文实验均在聚类算法常见且具有挑战性的 7 个数据集上进行，详细信息如表 1 所列。其中，Iris 和 Tae 是文献[26]中测试 NJW 算法的小型公共数据集，其余均是文献[25]中测试 DPC 算法的公共数据集，这些数据集虽然数据量不大，但数据分布形状和疏密程度对聚类算法非常具有挑战性。本次实验均在计算机内存 4.0GB，GPU2.3GHz 的硬件条件以及 Matlab2014a 环境下运行。

表 1 7 个数据集的相关信息

数据集	样本数	属性个数	类别数
Aggregation	788	2	7
Compound	399	2	6
Flame	240	2	2
Iris	150	4	3
Jain	373	2	2
Spiral	312	2	3
Tae	151	5	3

聚类效果的评价指标包括聚类结果可视化、聚类精度和互信息[27]，其中后两种客观指标的取值范围为[0,1]，数值越接近 1，说明聚类准确度越高，反之则聚类准确度越低。

已知数据集 X 包含 n 个数据点，label 是真实类标，gnd 是聚类输出的类标，K 是类别数，则聚类精度定义如下：

$$ca=\frac{1}{n}\sum_{i=1}^{K}\max_j|X(label=i)\cap X(gnd=j)| \tag{3}$$

其中，$X(label=i)$ 是真实类标为 i 的类簇，$X(gnd=j)$ 是聚类类标为 j 的类簇，$\max_j|X(label=i)\cap X(gnd=j)|$ 表示第 j 类别中属于第 i 真实类的最大样本个数，j 的取值范围为[1,K]。文献[28]将此度量定义为纯度(Purity)。

聚类中互信息用来衡量同一个数据集的不同划分之间的相似程度。算法 A,B 的聚类结果相当于数据集 X 上两个不同的划分，记为 P^a,P^b。$P^a=\{C_1^a,C_2^a,\cdots,C_{k_a}^a\}$，$P^b=\{C_1^b,C_2^b,\cdots,C_{k_b}^b\}$。其中，$n_i^a$ 表示 C_i^a 中数据的个数，n_j^b 表示 C_j^b 中数据的个数，n_{ij}^{ab} 表示 C_i^a 和 C_j^b 中相同数据的个数，则 P^a,P^b 的互信息为：

$$I(P^a,P^b)=\sum_{i=1}^{k_a}\sum_{j=1}^{k_b}(n_{ij}^{ab}/n)\log(nn_{ij}^{ab}/n_i^an_j^b) \tag{4}$$

标准化的互信息为：

$$MI(P^a,P^b)=I(P^a,P^b)/\sqrt{H(P^a)*H(P^b)} \tag{5}$$

$$H(P^a)=-\sum_{i=1}^{k_a}(n_i^a/n)\log(n_i^a/n) \tag{6}$$

$$H(P^b)=-\sum_{j=1}^{k_b}(n_j^b/n)\log(\frac{n_j^b}{n}) \tag{7}$$

当划分 P^a,P^b 差别很小时，$MI(P^a,P^b)$ 接近于 1，反之接近于 0。

4.1 尺度参数 σ 对 NJW 聚类效果的影响

通过第 2 节的文献综述可知，合适的尺度参数 σ 使高斯核函数对原始数据更能发挥非线性映射的作用，使得映射点更紧凑地簇合在一起，而不相似的数据点更远离对方。本文采用文献[10]中预先指定几个尺度参数值的方法来选择最优 σ 值。

首先用 z-score 对数据进行标准化，然后运行 NJW 发现多数情况下尺度参数 σ 取值较小时聚类效果很好。图 1 是设置 σ 在[0.05,2]范围内取值时 NJW 的聚类精度变化图。注

意，此处的实验结果是对于每一个 σ 值，NJW 算法在一个数据集上运行 50 次的最高聚类精度。

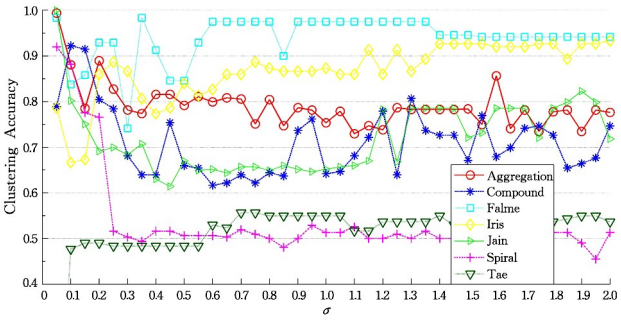


图 1　算法 NJW 在不同 σ 值下的聚类精度变化图

总之，NJW 算法在 Aggregation，Jain 和 Spiral 上，$\sigma=0.05$ 时能得到最高的聚类精度；而在 Compound 上，$\sigma=0.1$ 时能得到最高的聚类精度；在 Flame 和 Iris 上，多个 σ 处都能得到最好的聚类精度；但在 Tae 上，无论 σ 如何变化，NJW 的聚类精度都在 0.55 以下。

4.2　特征向量个数对 NJW 聚类效果的影响

实验中发现，在其他参数不变的情况下，所选取的特征向量的个数也会影响聚类的结果。为了减少参数，文献[10]中的 NJW 算法用同一个参数来表示特征向量的个数和特征空间中 K-means 聚类的个数。本节设计了两组实验进行聚类对比，聚类精度和互信息如表 2 所列，其中一组实验是选取前 10 个最大特征值对应的特征向量，而另一组实验设置特征向量的个数等于聚类个数。

实验结果显示，第一组实验的聚类结果在 3 个数据集上高于第二组，在 2 个数据集上持平，而只在 Spiral 上的聚类效果低于第二组。另外，李金泽等所提的自适应 NJW 算法[26]中所取特征向量的个数也等于聚类数目，在 Iris 和 Tae 数据集上的聚类精度分别是 0.920 和 0.424。通过对比可以发现，NJW 算法基于 10 个特征向量时的聚类精度是最高的。

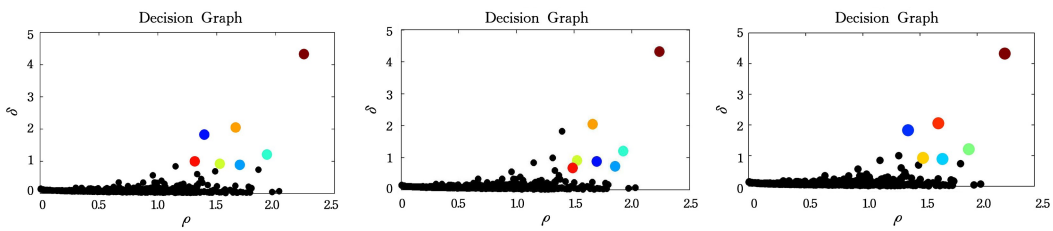
表 2　特征向量个数对 NJW 聚类性能的影响

数据集	特征向量的个数	聚类精度	互信息
Aggregation	10	0.994	0.98
	7	0.994	0.98
Compound	10	0.917	0.848
	6	0.865	0.797
Flame	10	0.9875	0.907
	2	0.858	0.445
Iris	10	0.920	0.771
	3	0.833	0.636
Jain	10	1	1
	2	1	1
Spiral	10	0.920	0.773
	3	1	1
Tae	10	0.552	0.115
	3	0.417	0.072

因此，本文实验在特征空间中选择特征向量时直接选择前 10 个最大特征值对应的特征向量，即在特征空间中 K-means 对 10 维的特征数据进行聚类。此处的数字 10 是根据实验观察随机选择的，并没有理论支持，以后将对此开展进一步的工作。

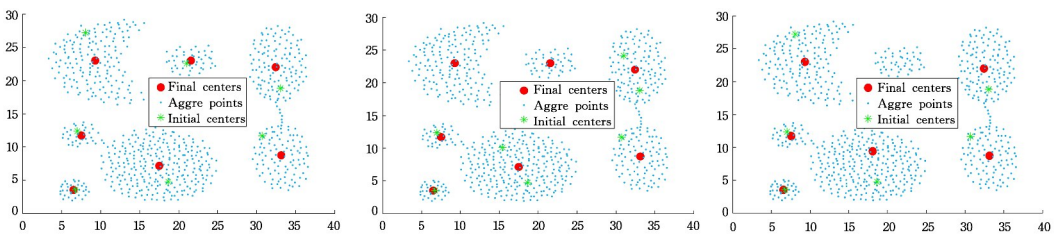
4.3　DP-NJW 算法确定聚类中心点

DP-NJW 算法中参数 σ 和特征向量个数直接采用了 NJW 算法在 4.1 节和 4.2 节中探讨的结果。本节主要探讨如何启发式地确定最优聚类中心点和类别数，这也是 DP-NJW 比 NJW 算法更优的核心。在 Aggregation 和 Iris 上进行两组实验，每组实验都选择 3 种不同的初始聚类中心点，然后执行 DP-NJW 算法，实验的可视化结果如图 2 和图 3 所示。



(a)较大的 δ 确定 7 个中心点　(b)较大的 ρ 确定 7 个中心点　(c)较大的 δ 确定 6 个中心点

图 2　DP-NJW 在 Aggregation 上的 3 个决策图



(a)图 2(a)的聚类　(b)图 2(b)的聚类　(c)图 2(c)的聚类

图 3　DP-NJW 在 Aggregation 上的 3 种初始类中心点的最终聚类中心可视化

平面决策图 2 中，ρ 和 δ 值较大的一些数据点显示在右上角，由此给了我们启示：可以选择这些数据点来作为聚类中心点。图 2(a)和图 2(b)中所选择的 7 个聚类中心点略有不同，而图 2(c)中选择了 6 个聚类中心点，前 4 个中心点的 δ 值较大，远离其他数据点，而另外 3(或者 2)个中心点的 δ 值较小，不太容易被辨识出来。另外，也可以根据 $\gamma=\rho*\delta$ 来选择 7 个较大的 γ 值对应的数据点作为聚类中心点。对应到图 3(a)中 7 个初始中心点分别在 7 个类簇内，而图 3(b)中有 2 个类簇分别包含 2 个初始中心点，另外 2 个类簇中没有初始中心点。虽然初始聚类中心略有不同而最终取得了同样好的聚类精度，但是运行时间的长短不一，这 3 种类中心变化的详细信息如表 3 所列，其中不同的类中心用黑体标识出来。相同

的实验在 Iris 上进行，3 种类中心变化的详细信息如表 4 所列，聚类中心点的变化可视图如图 4 所示。

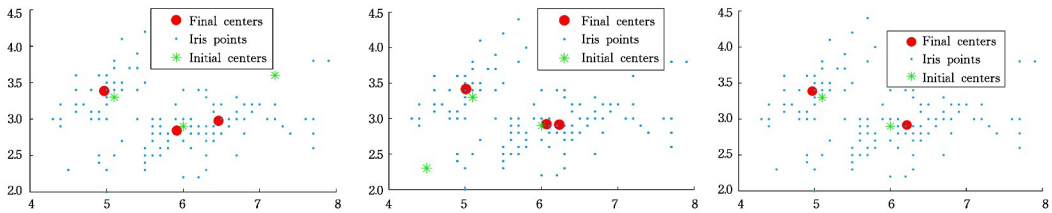
这两组实验表明，DP-NJW 算法的第一阶段提供的初始类中心比较精准，靠近最终的类中心位置，同时也表明 DP-NJW 算法对初始聚类中心位置具有一定的鲁棒性。如果启发式确定的类数目少于真正的类数目，那么一定有聚类错误发生。比如这两组实验中的第三种聚类中心数均小于真实的类数，其聚类精度大大降低(见表 3 和表 4)。另外，聚类精度与数据分布有很大关系，Aggregation 上的聚类困难主要来自各个类数据密度不均匀，有大类和小类之分，对于数据分布密度小的簇类，在决策图中不容易辨识出其类中心。Iris 各类分布虽然均匀，但第二类和第三类的数据重叠度很高，很难被分开。另外，类中心初始值越接近于最优中心点，DP-NJW 聚类结果就越好，也能在短时间内收敛到最优结果。

表 3 在 Aggregation 上 3 种不同初始类中心的 DP-NJW 聚类性能

初始类中心标号	聚类精度	互信息	运行时间/s
40,186,343,550,612,723,769	0.992	0.975	8.402
186,256,343,550,612,633,769	0.992	0.975	10.133
40,186,343,550,612,769	0.935	0.86	16.250

表 4 在 Iris 上 3 种不同初始类中心的 DP-NJW 聚类性能比较

初始类中心标号	聚类精度	互信息	运行时间/s
24,79,110	0.880	0.664	9.033
24,42,79	0.827	0.572	29.836
24,79	0.633	0.464	4.105



(a)第一个初始类中心的聚类 (b)第二个初始类中心的聚类 (c)第三个初始类中心的聚类

图 4 DP-NJW 在 Iris 上的 3 种初始类中心的最终聚类中心可视图

4.4 DP-NJW,NJW,DPC 算法的比较

本节将对 DP-NJW 算法与 NJW 算法和 DPC 算法在 7 个数据集上的聚类性能进行对比。因前面实验结果中互信息的变化与聚类精度的变化是一致的，所以此组实验中的聚类性能主要指聚类精度和运行时间。此组实验中，NJW 算法在每个数据集上的聚类性能包括：1)运行 50 次的平均聚类精度和运行时间；2)最高聚类精度和 50 次累加运行时间。DP-NJW 算法和 DPC 算法均是一次的聚类性能，实验结果如图 5 和图 6 所示。

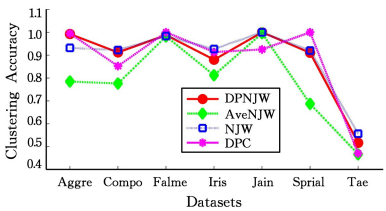


图 5 DP-NJW,NJW 和 DPC 算法的聚类精度对比

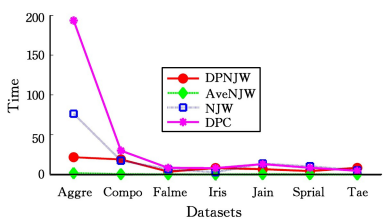


图 6 DP-NJW,NJW 和 DPC 算法的运行时间对比

根据图 5 和图 6 可知，本文的 DP-NJW 算法的聚类精度高于 NJW 的平均聚类精度，只在 Flame 和 Jain 上有一样的精度。观察 NJW 的最高聚类精度曲线发现，DP-NJW 算法也只在 Iris 和 Tae 这两个数据集上的精度不如 NJW 的，这说明 NJW 算法的波动性较大，不如 DP-NJW 算法稳定。在 Spiral 数据集上，DPC 算法的聚类精度最高，但其也是最耗时的，尤其在较大的数据集 Aggregation 上的运行时间大大增加。由此可见，本文所提的 DP-NJW 算法具有聚类精度高、计算简便的优点，降低了 NJW 算法对初始类中心的敏感性，也大大降低了 DPC 算法的计算复杂度。

结束语 NJW 算法对 K-means 随机初始值的敏感性大大影响着聚类结果，即取定一个 σ 值，需要多次运行 NJW 算法才能得到如图 2 所示的聚类精度，这不仅很耗时，还容易陷入局部最优，比如在 Iris 数据集上的聚类精度是 0.833，达不到文献[26]中的 0.903。本文所提的 DP-NJW 算法根据数据分布的特点来确定聚类中心点和类别数，能为后续的 K-means 聚类提供很好的初始类中心，弥补了多次运行寻找最优聚类结果的缺点。实验结果也充分证实了所提 DP-NJW 算法比 NJW 算法更稳健，能在较短时间内找到(局部)最优解的聚类性能。虽然通过 ρ 和 δ 值能找出聚类中心点，但是对分布稀疏的簇类以及分布密度不均匀的簇类仍具有挑战性，需进一步进行研究。另外，特征空间中特征向量的维度也是聚类识别中的一个重要环节，在接下来的工作中将对此进行进一步的研究。

参 考 文 献

[1] MACQUEENJ B. Some Methods for Classification and Analysis of Multivariate Observations[C]// Proceedings of Berkeley Symposium on Mathematical Statistics and Probability. 1967: 281-297.

[2] MACQUEENJ B. The Classification Problem[Z]. Western Management Science Institute Working Paper, 1962.

[3] COXD R. Note on Grouping[J]. Journal of the American Statistical Association, 1957, 52(280): 543-547.

[4] FISHERW D. On Grouping for Maximum Homogeneity[J]. Journal of the American Statistical Association, 1958, 53(284): 789-798.

[5] SEBESTYENG S. Decision Making Process in Pattern Recognition[M]. New York: Macmillan, 1962.

[6] JAINA K. Data Clustering：50 Years Beyond K-means[J]. Pattern Recognition Letters，2010，31(8)：651-666.

[7] BOCK H H. Clustering Methods：A History of K-means Algorithms[M]// Selected Contributions in Data Analysis and Classification. Springer Berlin Heidelberg，2007：161-172.

[8] 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，2012，20(7)：21-24.

[9] PERONAP，FREEMAN W. A Factorization Approach to Grouping [C]// Proceedings of European Conference on Computer Vision (ECCV). 1998：655-670.

[10] NGA Y，JORDANM I，WEISS Y. On Spectral Clustering：Analysis and an Algorithm[C]// Proceedings of the 14th Advances in Neural Information Processing System. 2002：849-856.

[11] LI C G，YOU C，VIDAL R. Structured Sparse Subspace Clustering：A Joint Affinity Learning and Subspace Clustering Framework[J]. IEEE Transactions on Image Processing，2017，26(6)：2988-3001.

[12] LI C G，YOU C，VIDAL R. On Geometric Analysis of Affine Sparse Subspace Clustering[J]. IEEE Journal on Selected Topics in Signal Processing，2018，12(6).

[13] LI C G，ZHANG J J，GUO J. Constrained Sparse Subspace Clustering with Side Information[C]// Proceedings of the 24th International Conference on Pattern Recognition (ICPR). 2018：2093-2099.

[14] LIANG J Q，HAN Y H，HU Q H. Semi-Supervised Image Clustering with Multimodal Information[J]. Multimedia Systems，2016，22：149-160.

[15] LUXBURGU V. A Tutorial on Spectral Clustering[J]. Statistics & Computing，2007，17(4)：395-416.

[16] 金建国. 聚类方法综述[J]. 计算机科学，2014，41(S2)：288-293.

[17] ZELNIK-MANORL，PERONA P. Self-Tuning Spectral Clustering[C]// Proceedings of the 16th Advances in Neural Information Processing System. 2004：1601-1608.

[18] WANG F，ZHANG C S. Robust Self-Tuning Semi-Supervised Learning[J]. Neurocomputing，2007，70(16)：2931-2939.

[19] YANG C，ZHANG X，JIAO L，et al. Self-Tuning Semi-Supervised Spectral Clustering [C] // International Conference on Computational Intelligence & Security. 2008：1-5.

[20] KUMARV，HAHNJ，ZOUBIR A M. Band Selection for Hyperspectral Images Based on Self-Tuning Spectral Clustering[C]// Proceedings of the 21st EuropeanSignal Processing Conference (EUSIPCO). 2013.

[21] POLITOM，PERONAP. Grouping and Dimensionality Reduction by Locally Linear Embedding[C]// Proceedings of International Conference on Neural Information Processing Systems：Natural & Synthetic. 2001：1255-1262.

[22] ZENGC P，ZHOUA M，ZHANGG X. Self-adaptive Spectral Cluster Number Detecting with Particle Swarm Optimization Algorithm[C]// Evolutionary Computation. IEEE，2016：4607-4611.

[23] CHAN P，SCHLAG M，ZIEN S. Spectral K-Way Ratio-Cut Partitioning and Clustering[C]// Proceedings of Conference on Design Automation. 1993：749-754.

[24] TENENBAUM J B，SILVAV D，LANGFORD J C. A Global Geometric Framework for Nonlinear Dimensionality Reduction [J]. Science，2000，290(5500)：2319-2323.

[25] RODRIGUEZ A，LAIO A. Clustering by Fast Search and Find of Density Peaks[J]. Science，2014，344(6191)：1492-1496.

[26] 李金泽，徐喜荣，潘子琦，等. 改进的自适应谱聚类 NJW 算法[J]. 计算机科学，2017，44(S1)：424-427.

[27] 周海松，黄德才. 密度自适应的半监督谱聚类算法[J]. 计算机科学，2016，43(12)：209-212.

[28] 杨虎，付宇，范丹. 噪音特征对聚类内部有效性的影响[J]. 计算机科学，2018，45(7)：22-30.

(上接第 473 页)

[6] KASHIF J，SAMMEN M，BABRI HAROON A. A two-stage Markov blanket based feature selection algorithm for text classification [J]. Neurocomputing，2015，157：91-104.

[7] 李军怀，付静飞，蒋文杰，等. 基于 MRMR 的文本分类特征选择方法[J]. 计算机科学，2016，43(10)：225-228.

[8] 黄源，李茂，吕建成，等. 一种基于开方检验的特征选择方法[J]. 计算机科学，2015，42(5)：54-56.

[9] ZHENG Z，LEI W，HUAN L，et al. On Similarity Preserving Feature Selection[J]. IEEE Transactions on Knowledge & Data Engineering，2013，25(99)：1.

[10] DEAN J，GHEMAWAT S. MapReduce：simplified data processing on large clusters[J]. Communications of the ACM，2008，51(1)：107-113.

[11] MASHAYEKHY L，NEJAD M，GROSU D，et al. EnergyAware-Scheduling of MapReduce Jobs[J]. IEEE International Congress on Big Data，2014，26(10)：32-39.

[12] HAN L，SUN X Z，WU Z C，et al. Optimization Study on Sample Based Partition on MapReduce[J]. Journal of Computer Research and Development，2013，50(Suppl.)：77-84.

[13] GUNTHER N，PUGLIA P，TOMASETTE K. Hadoop Superlinear Scalability[J]. Communications of the ACM，2015，58(4)：1542-7730.

[14] FEI X，LI X F，SHEN C. Parallelized text classification algorithm for processing large scale TCM clinical data with MapReduce[C]// IEEE International Conference on Information and Automation. IEEE，2015：1983-1986.

[15] LIU J，ZHU A，QIN C. Estimation of theoretical maximum speedup ratio for parallel computing of grid-based distributed hydrological models [J]. Computers & Geosciences，2013，60(10)：58-62.