基于相关熵的快速聚类算法

李中衡，杨奔，张劲节，刘银川，张雪涛，王飞
（西安交通大学视觉信息处理与应用国家工程实验室，710049，西安）

摘要：针对目前大规模真实数据聚类中存在的效率低和鲁棒性差的问题，提出了一种基于相关熵的快速聚类算法(FCC)。该算法主要分为以下两步：首先对原始数据进行 k 均值操作，得到粗略的样本类别，作为第二步的标签矩阵；其次利用原始数据与其锚点构建的锚点图对应的拉普拉斯矩阵作为图约束来探寻数据间的内在结构，从而得到样本的最终类别。整个聚类过程在相关熵准则而不是传统的欧氏距离框架下进行，可有效抑制真实数据中大量存在的非线性和非高斯分布的噪声对聚类鲁棒性的影响。为了验证提出算法的性能，使用 5 种典型的算法作为对比算法与提出的算法一起在 4 个大规模真实数据集上运行，结果表明，提出的算法可在大部分情况下提高聚类精度，在 WebKB、TDT2 和 Cora 数据集上分别提高 8.58%，6.86%和 1.86%，同时提高聚类效率几倍甚至几十倍；为了验证本算法的鲁棒性，分别加入不同程度的随机噪声和泊松噪声到 WebKB 和 Cora 上，得到 8 个含噪数据集，所有算法均在相同条件下运行于这些噪声数据集上，结果表明，相对于其他对比算法，提出的算法能够保持最优的聚类鲁棒性。

关键词：快速聚类；相关熵；锚点图
中图分类号：TP181 **文献标志码**：A
DOI：10.7652/xjtuxb202106015 **文章编号**：0253-987X(2021)06-0121-10

OSID 码

Fast Correntropy-Based Clustering Algorithm

LI Zhongheng, YANG Ben, ZHANG Jinjie, LIU Yinchuan, ZHANG Xuetao, WANG Fei
(National Engineering Laboratory for Visual Information Processing and Applications,
Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Aiming at the issue of low efficiency and poor robustness in large-scale real-world data clustering, a fast correntropy-based clustering algorithm (FCC) is proposed. FCC is mainly divided into the following two steps: 1) Performing k-means on the original data to obtain rough labels to serve as the label matrix of the second step; 2) Adopting the original data and those anchors to construct the anchor graph and taking Laplacian matrix of the anchor graph as a graph constraint to explore the internal structure of original data so as to obtain the final categories of these samples. Meanwhile, the whole clustering process is carried out under the correntropy instead of the traditional Euclidean distance framework, which can effectively suppress the influence of the large amount of non-linear and non-Gaussian noise in the real-world data on the clustering robustness. To verify the performance of FCC, five state-of-the-art algorithms are developed as baselines to run with FCC on four large-scale real-world data sets. The results show that FCC can improve the clustering accuracy in most cases (by 8.58%, 6.86% and 1.86%

收稿日期：2021-01-15。 作者简介：李中衡(1991—)，男，博士生；王飞(通信作者)，男，教授。 基金项目：国家重大科技攻关资助项目(2018ZX01008103)；中央大学基础研究基金资助项目。
网络出版时间：2021-02-08 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20210208.1132.002.html

respectively on WebKB, TDT2 and Cora) while greatly improving the clustering efficiency (by several or even dozens of times). Furthermore, to verify the robustness of FCC, varying degrees of random noise and Poisson noise are added to WebKB and Cora to obtain 8 noisy data sets, and all algorithms are run on these noisy data sets under the same conditions. Compared with the other baseline algorithms, FCC can maintain the optimal clustering robustness.

Keywords: fast clustering; correntropy; anchor graph

聚类[1-13]也称为无监督学习,可以将无标签数据自动划分为确定的类别,广泛应用于数据挖掘[14]、计算机视觉[15]和模式识别[16]等领域。早期聚类算法研究的热点是如何提高聚类的性能,针对这一问题,研究人员提出了许多聚类算法,如基于矩阵分解的算法[11,12,17-19]、基于 k 均值的算法[1-3,13,20]以及基于谱聚类(SC)的算法[4-6,15,21],这些算法均有效地推动了早期聚类研究的发展。其中,由于基于谱聚类的聚类方法具有不受样本空间形状限制的优点,使得越来越多的研究者开始聚焦于这一领域。

随着近年来数据的爆炸式增长,传统的聚类算法在处理这些大规模数据时已经很难满足人们的需求。例如,谱聚类的计算复杂度为 $O(n^2d+n^3)$[15](n 为样本数,d 为特征维数),其中 $O(n^2d)$是构造 $n\times n$ 的相似度矩阵所需的计算复杂度,$O(n^3)$是对相似度矩阵进行特征分解所需的计算复杂度。当样本数或者样本特征维度急剧增加时,谱聚类的效率会显著降低。特别是当样本数增加到某一临界值时,由于计算机内存的限制,谱聚类将崩溃。因此,如何有效地对大规模数据聚类已逐渐进入大众的视野,并逐渐占据了重要的地位。

对于大规模的无标签数据,特别是多样本点数据或高特征维度数据,研究人员提出了许多快速的聚类算法。例如,基于矩阵分解的方法中,非负矩阵互分解算法[22](FNMTF)和局部保留非负矩阵互分解算法[22](LPFNMTF)直接将因子矩阵约束为聚类指示矩阵,模型的优化部分为只需少量矩阵乘法的快速更新规则,从而可有效降低处理大规模数据时的计算复杂度。类似地,双边 k 均值算法[23](BKM)作为一种基于 k 均值的加速方法,受到 FNMTF 的启发,并在其基础上将吸收因子约束为对角矩阵,进一步提高了对大规模数据的聚类效率。同时,在过去的一段时间里,各种加速谱聚类算法也被陆续提出,如基于二分图的大规模多视图谱聚类算法[9](LMVSCB)、大规模谱聚类算法[24](LSC)、针对大规模数据的谱聚类算法[25](SCLD)、分布式系统中的并行谱聚类算法[26](PSCD)等,这类算法常常通过优化建图模块(采用不同的映射或采样方法)或特征分解模块来降低计算复杂度。尽管这些尝试都能在一定程度上提高大规模数据的聚类效率,但它们仍存在自身的缺陷。例如,当数据维数变大时,基于矩阵分解方法的效率显著降低,而基于加速谱聚类的方法在处理大规模数据时,由于对特征分解的需求,仍然非常耗时。因此,如何进一步提高大规模数据的聚类效率仍然是一个亟待解决的问题。

另一方面,真实世界数据中存在大量不同分布的噪声和离群点,特别是在高维数据[27-28]中,它们往往对聚类算法的性能有很大影响。如何抑制这些噪声和离群点对聚类性能的影响也是当前的一个热点问题。目前,主流的方法是用 L_1 范数[10]或 $L_{2\text{-}1}$ 范数[29]代替传统的均方误差(MSE),提高聚类的鲁棒性,这在许多应用中已经被证明是有效的。但是,它们仅能有效地抑制线性或高斯分布的噪声,对大量非线性和非高斯分布的噪声及离群点的抑制效果是非常有限的。针对这个问题,GCCF[12](基于相关熵的图正则概念分解算法)将相关熵[30-33]引入到聚类模型中,利用相关熵对非线性和非高斯分布的噪声以及离群点良好的抑制功能,有效地提高了聚类算法的性能,但由于其具有 $O((n^2c+ndc+n^2d)t)$(c 是聚类数,t 是迭代次数)的计算复杂度,很难被应用于大规模数据的聚类中。

针对上述问题,本文提出了一种基于相关熵的快速聚类方法(FCC)。在图正则约束的条件下,最小化 k 均值得到的类别标签与图聚类得到的类别标签之间的误差。由于本文方法兼顾了 k 均值与图聚类的聚类结果,故此也继承了二者各自的优点,可进一步提高聚类的性能。另一方面,由于本文方法最小化两种聚类算法的聚类指示阵($n\times c$),与数据矩阵($n\times d$)或相似度矩阵($n\times n$)相比,更适合于处理大规模数据,特别是大规模高维度的数据。此外,受 GCCF 的启发,FCC 采用相关熵来度量误差,提高了算法对非线性、非高斯噪声和离群点的抑制能力,从而有效地提高了聚类性能。

本文的主要贡献如下:

(1)提出了一种基于相关熵的快速聚类算法，该算法能有效地对真实世界数据(包含大量非线性和非高斯分布的噪声以及离群点)进行聚类。特别是在面对高维数据时，FCC 的效率远远高于现有的快速性方法。

(2)由于引入了相关熵，FCC 的目标函数变为难以直接求解的非凸函数。本文提出了一种基于半二次(HQ)极小化技术的优化策略，可以通过较少的迭代次数提高求解效率。此外，本文还分析了 FCC 的计算复杂度和参数敏感性。

(3)本文在 4 个不同的真实大规模数据集(包含两个多样本点数据集和两个高特征维度数据集)上进行了大量的实验，结果表明，与目前主流的快速聚类算法相比，FCC 能够更快地获得比这些算法更好的聚类性能。

1　相关工作

1.1　相关熵

在信息论学习(ITL)中，为了抑制信号中非线性和非高斯分布的噪声以及离群点引起的不可预测的误差，研究人员提出了相关熵[30]的概念，它可以根据两个随机变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 定义为

$$V(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{E}[\kappa(\boldsymbol{X},\boldsymbol{Y})] = \sum \kappa(\boldsymbol{x},\boldsymbol{y})\mathrm{d}F_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y}) \tag{1}$$

式中：$\boldsymbol{E}[\cdot]$ 表示期望算子；$\kappa(\boldsymbol{x},\boldsymbol{y})$ 是移位不变 Mercer 核；$F_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y})$ 表示 $(\boldsymbol{X},\boldsymbol{Y})$ 的联合分布函数。本文采用的核函数为

$$\kappa(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{G}(\boldsymbol{x}-\boldsymbol{y})=\exp(-(\boldsymbol{x}-\boldsymbol{y})^2/2\delta^2) \tag{2}$$

式中：$\delta>0$ 表示高斯核的核带宽参数。在实际中，由于能获得的样本数量有限，联合分布函数往往是未知的。因此，相关熵通常表示为以下形式

$$\hat{\boldsymbol{V}}(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{G}(x_i-y_i) = \frac{1}{N}\sum_{i=1}^{N}\exp((x_i-y_i)^2/2\delta^2) \tag{3}$$

式中：$\hat{\boldsymbol{V}}(\boldsymbol{X},\boldsymbol{Y})$ 是估计量。一般来说，式(3)最大化时的值称为最大相关熵准则(MCC)，表示两个随机变量之间的误差最小。

1.2　半二次技术

半二次(HQ)技术[27,32-33]由于能够求解 ITL 中的非线性和非凸损失函数，在图像处理、模式识别和鲁棒估计中得到了广泛的应用。对于随机向量 $\boldsymbol{x}\in R^n$，$\boldsymbol{x}$ 的潜在损失函数 $f(\boldsymbol{x})$ 可定义如下

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} f(x_i) \tag{4}$$

式中：x_i 表示 $\boldsymbol{x}$ 的第 i 个元素。引入 x_i 的辅助向量 $\boldsymbol{z}$，可以将式(4)的增强函数定义为

$$f(x_i) = \min_{z_i}\mathscr{Q}(x_i,z_i)+g(z_i) \tag{5}$$

式中：z_i 仅依赖于 $f(x_i)$；$g(z_i)$ 是 $f(x_i)$ 的双势函数；$\mathscr{Q}(x_i,z_i)$ 表示 HQ 函数。本文利用了广泛使用的 HQ 函数的乘法形式，表示如下

$$\mathscr{Q}(x,z) = \sum_{i=1}^{n}\mathscr{Q}(x_i,z_i) = \frac{1}{2}\sum_{i=1}^{n}z_i x_i^2 \tag{6}$$

式中：$\boldsymbol{z}=[z_1,\cdots,z_n]$ 是辅助量。结合式(5)和式(6)，势函数的向量形式如下

$$f(\boldsymbol{x}) = \min_{\boldsymbol{z}}\mathscr{Q}(\boldsymbol{x},\boldsymbol{z}) + \sum_{i=1}^{n}g(z_i) \tag{7}$$

2　FCC 模型

本文提出的算法能够有效地处理包含噪声(特别是非线性和非高斯分布的噪声)以及离群点的大规模数据。FCC 主要由两步组成：利用 k 均值生成粗聚类标签，并将其作为真值输入到第二步；利用第一步输入的标签矩阵，在图正则约束的条件下完成基于锚点图的聚类。因此，FCC 不仅继承了 k 均值和图聚类的优点，而且大大降低了计算复杂度。此外，引入相关熵可以有效地抑制真实数据中的噪声和离群点对性能的影响，从而有效提高聚类算法的性能。

假设 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}\in R^{n\times d}$ 是一组包含 c 类的无标签大规模数据。在 FCC 的第一步中，需要使用 k 均值为每个样本生成粗类别。更准确地说，这一步需要生成一个指示矩阵 $\boldsymbol{C}\in R^{n\times c}$，它的每一行只有一个元素 1，其余元素都是 0。因此，在后续的图聚类过程中，FCC 还可以继承 k 均值得到的原始数据的内部结构。此外，得到的指示矩阵 $\boldsymbol{C}$ 代替原始数据矩阵 $\boldsymbol{X}$ 作为第二步聚类的输入，可以大大降低需要处理的数据规模，从而提高对大规模数据聚类的效率。

FCC 的第二步需要在原始数据和它们的锚点之间构造一个锚点图[34]，然后利用锚点图的拉普拉斯矩阵作为正则约束来完成图聚类过程。该步骤继承了图聚类不受样本空间形状的限制以及具有良好的数学框架的优点。同时，锚点图的构造大大降低了数据的维度，保证了处理大规模数据时的高效性。

在构造锚点图之前，首先需要生成锚点。目前，主流的锚点生成方法主要有随机采样法和 k 均值两

种。实践证明,使用随机采样和 k 均值生成相同数量的锚点,由 k 均值生成的锚点往往能在后续聚类过程中获得更高的聚类性能,因此本文采用 k 均值生成锚点。

设 $\boldsymbol{U}=[u_1,u_2,\cdots,u_m]^{\mathrm{T}}\in R^{n\times d}$ 为用 k 均值生成的 $\boldsymbol{X}$ 的锚点,锚点图矩阵可用下式表示

$$H_{ij}=\frac{\psi(x_i,u_j)}{\sum_{k\in\Delta_i}\psi(x_i,u_j)} \tag{8}$$

式中:$\boldsymbol{H}$ 为原始数据与其锚点之间的锚点图矩阵;x_i 和 u_j 分别为 $\boldsymbol{X}$ 和 $\boldsymbol{U}$ 中的第 i 行和第 j 行;Δ_i 表示 $\{1,2,\cdots,m\}$ 的一个子集,$\{1,2,\cdots,m\}$ 为 $\boldsymbol{U}$ 中 x_i 的 k 近邻的个数;$\psi(a,b)$ 可以表示为

$$\psi(a,b)=\exp\frac{-(a-b)^2}{2\sigma^2} \tag{9}$$

其中 σ 是一个名为高斯核带宽(GKB)的自由参数,它决定着式(9)的鲁棒性。

此时,数据 $\boldsymbol{X}$ 与其锚点之间的锚点图矩阵对应的相似度矩阵 $\boldsymbol{W}$ 可表示为

$$\boldsymbol{W}=\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}} \tag{10}$$

设 $\boldsymbol{D}$ 是 $\boldsymbol{W}$ 的度矩阵,则 $\boldsymbol{D}$ 是满足 $d_{ij}=\sum_{j=1}^{n}w_{ij}$ 的对角矩阵。因此,拉普拉斯矩阵可表示为 $\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$。显然,找到 $\boldsymbol{L}$ 最小的 p 个特征向量意味着找到 $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ 最大的 p 个特征向量。

设 $\hat{\boldsymbol{H}}=\boldsymbol{D}^{-1/2}\boldsymbol{H}$,对 $\hat{\boldsymbol{H}}$ 进行奇异值分解(SVD),如下所示

$$[\boldsymbol{\Sigma},\boldsymbol{\Pi},\boldsymbol{\Lambda}^{\mathrm{T}}]=\mathrm{svd}(\hat{\boldsymbol{H}}) \tag{11}$$

式中:$\boldsymbol{\Sigma}$ 表示 $\hat{\boldsymbol{H}}$ 的左奇异向量;$\boldsymbol{\Lambda}$ 表示 $\hat{\boldsymbol{H}}$ 的右奇异向量;$\boldsymbol{\Pi}=\mathrm{diag}(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_k)(\varepsilon_1\geqslant\varepsilon_2\geqslant\cdots\geqslant\varepsilon_k\geqslant0)$ 是 $\hat{\boldsymbol{H}}$ 的奇异值集。很明显,$\boldsymbol{\Sigma}$ 是 $\boldsymbol{W}$ 的奇异向量集,$\boldsymbol{\Sigma}$ 可以表示为

$$\boldsymbol{\Sigma}=\hat{\boldsymbol{H}}\boldsymbol{\Lambda}\boldsymbol{\Pi}^{-1} \tag{12}$$

因此,与 $\boldsymbol{L}$ 的 p 个最小特征向量相对应的 $\boldsymbol{W}$ 的 p 个最大特征向量 $\boldsymbol{\Sigma}_p$ 可以按如下公式计算

$$\boldsymbol{\Sigma}_p=\hat{\boldsymbol{H}}\boldsymbol{\Lambda}_p\boldsymbol{\Pi}_p^{-1} \tag{13}$$

式中:$\boldsymbol{\Lambda}_p$ 是 $\boldsymbol{\Lambda}$ 的前 p 列;$\boldsymbol{\Pi}_p$ 是集合 $(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p)$ 对应的对角矩阵。

结合第一步得到的粗标签 $\boldsymbol{C}$ 和本部分基于图约束的聚类方法,可以定义如下聚类公式

$$\min_{\boldsymbol{Z}_p}\|\boldsymbol{\Sigma}_p\boldsymbol{Z}_p-\boldsymbol{C}\|_F^2+\lambda\mathrm{tr}(\boldsymbol{Z}_p^{\mathrm{T}}\boldsymbol{\Sigma}_p^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\Sigma}_p\boldsymbol{Z}_p) \tag{14}$$

式中:$\boldsymbol{Z}_p=\boldsymbol{R}^{p\times c}$ 是辅助变量矩阵。

此外,为了处理大量的各种分布的噪声,特别是各种非线性和非高斯的噪声以及实际数据中的离群点,在式(14)中引入相关熵,即可得到 FCC 的目标函数如下

$$\max_{\boldsymbol{Z}_p}\sum_{i=1}^{n}\boldsymbol{G}\left(\sqrt{\sum_{j=1}^{c}((\boldsymbol{\Sigma}_p\boldsymbol{Z}_p)_{ij}-\boldsymbol{C}_{ij})^2}\right)-\lambda\mathrm{tr}(\boldsymbol{Z}_p^{\mathrm{T}}\boldsymbol{\Sigma}_p^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\Sigma}_p\boldsymbol{Z}_p) \tag{15}$$

当式(15)达到最大值时,$\boldsymbol{\Sigma}_p\boldsymbol{Z}_p$ 和 $\boldsymbol{C}$ 之间的误差最小。此时,可以直接从 $\boldsymbol{\Sigma}_p\boldsymbol{Z}_p$ 获得最终的数据类别。

3 优化与分析

为了求解 FCC 的目标函数,本节提出了一种新的交互式迭代求解策略。在优化之前,首先需要利用半二次最小化技术将式(15)转化为凸函数如下

$$\max_{\boldsymbol{Z}_p}\sum_{i=1}^{n}\left(\frac{e_i}{\sigma^2}\left(\sum_{j=1}^{c}((\boldsymbol{\Sigma}_p\boldsymbol{Z}_p)_{ij}-C_{ij})^2\right)-\phi(e_i)\right)-\lambda\mathrm{tr}(\boldsymbol{Z}_p^{\mathrm{T}}\boldsymbol{\Sigma}_p^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\Sigma}_p\boldsymbol{Z}_p) \tag{16}$$

式中:e_i 是 HQ 技术的辅助量。由于式(16)是一个凸函数,故可直接优化求解。具体来说,式(16)需要在固定 $\boldsymbol{Z}_p$ 的条件下最大。此时,e_i 可通过以下公式获得

$$e_i=-\boldsymbol{G}\left(\sqrt{\sum_{j=1}^{c}((\boldsymbol{\Sigma}_p\boldsymbol{Z}_p)_{ij}-C_{ij})^2}\right)=-\exp\frac{\sum_{j=1}^{m}((\boldsymbol{\Sigma}_p\boldsymbol{Z}_p)_{ij}-C_{ij})^2}{2\sigma^2} \tag{17}$$

式中:σ 是影响相关熵鲁棒性的自由参数,表示为

$$\sigma=\sqrt{\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{m}((\boldsymbol{\Sigma}_p\boldsymbol{Z}_p)_{ij}-C_{ij})^2} \tag{18}$$

此时,式(16)可以表示为如下形式

$$\max_{\boldsymbol{Z}_p}\sum_{i=1}^{n}\left(\frac{e_i}{\sigma^2}\sum_{j=1}^{c}((\boldsymbol{\Sigma}_p\boldsymbol{Z}_p)_{ij}-C_{ij})^2\right)-\lambda\mathrm{tr}(\boldsymbol{Z}_p^{\mathrm{T}}\boldsymbol{\Sigma}_p^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\Sigma}_p\boldsymbol{Z}_p) \tag{19}$$

等价于以下优化问题

$$\min_{\boldsymbol{Z}_p}\sum_{i=1}^{n}\left(r_i\sum_{j=1}^{c}((\boldsymbol{\Sigma}_p\boldsymbol{Z}_p)_{ij}-C_{ij})^2\right)-\lambda\mathrm{tr}(\boldsymbol{Z}_p^{\mathrm{T}}\boldsymbol{\Sigma}_p^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\Sigma}_p\boldsymbol{Z}_p) \tag{20}$$

式中:$r_i=-e_i/\sigma^2$。在此基础上,式(15)的优化问题转化为只需交互迭代求解 $\boldsymbol{R}=\mathrm{diag}(r_1,r_2,\cdots,r_n)^{\mathrm{T}}$ 和 $\boldsymbol{Z}_p$,使式(20)最小化。

更新 $\boldsymbol{R}$:$\boldsymbol{R}$ 是对角矩阵,对角线元素为 $R_{ii}=r_i$,其中 r_i 可表示为

$$r_i=-\frac{e_i}{\sigma^2}=\exp\left(\sum_{j=1}^{c}((\boldsymbol{\Sigma}_p\boldsymbol{Z}_p)_{ij}-C_{ij})^2/2\sigma^2\right)/\sigma^2 \tag{21}$$

更新 $\boldsymbol{Z}_p$:为了求解 $\boldsymbol{Z}_p$,式(20)可以重新写成如下矩阵形式,其中 $\boldsymbol{Y}=\boldsymbol{\Sigma}_p\boldsymbol{Z}_p$

$$\min_{\boldsymbol{Y}} \operatorname{tr}((\boldsymbol{Y}-\boldsymbol{C})^{\mathrm{T}}\boldsymbol{R}(\boldsymbol{Y}-\boldsymbol{C}))+\lambda\operatorname{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{Y}) \tag{22}$$

根据矩阵迹的性质，如 $\operatorname{tr}(\boldsymbol{AB})=\operatorname{tr}(\boldsymbol{BA})$ 和 $\operatorname{tr}(\boldsymbol{A}^{\mathrm{T}})=\operatorname{tr}(\boldsymbol{A})$，式(22)可变换为以下形式

$$\min_{\boldsymbol{Y}} \operatorname{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Y})-2\operatorname{tr}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Y})+\operatorname{tr}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{C})+\lambda\operatorname{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{Y}) \tag{23}$$

对应的拉格朗日函数可定义为

$$\mathscr{L}=\operatorname{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Y})-2\operatorname{tr})\boldsymbol{C}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Y})+\operatorname{tr}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{C})+\lambda\operatorname{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{Y})+\operatorname{tr}(\boldsymbol{\varphi}\boldsymbol{Y}^{\mathrm{T}}) \tag{24}$$

式中：$\Phi=[\Phi]_{ij}\in R^{n\times c}$ 是拉格朗日乘数。然后，可以导出拉格朗日函数的偏导数

$$\frac{\partial\mathscr{L}}{\partial Y}=2\boldsymbol{R}\boldsymbol{Y}-2\boldsymbol{R}^{\mathrm{T}}\boldsymbol{C}+2\lambda\boldsymbol{L}\boldsymbol{Y}+\Phi \tag{25}$$

转换如下

$$2(RY)_{ij}Y_{ij}-2(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{C})_{ij}Y_{ij}+2\lambda(\boldsymbol{L}\boldsymbol{Y})_{ij}Y_{ij}+\Phi_{ij}Y_{ij}=0 \tag{26}$$

利用 KKT 条件(如 $\Phi_{ij}Y_{ij}=0$)，式(26)可转化为

$$2(\boldsymbol{R}\boldsymbol{Y})_{ij}Y_{ij}-2(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{C})_{ij}Y_{ij}+2\lambda(\boldsymbol{L}\boldsymbol{Y})_{ij}y_{ij}=0 \tag{27}$$

因此，可得到如下更新规则

$$y_{ij}\leftarrow Y_{ij}\frac{(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{C}+\lambda\boldsymbol{W}\boldsymbol{Y})_{ij}}{(\boldsymbol{R}\boldsymbol{Y}+\lambda\boldsymbol{D}\boldsymbol{Y})_{ij}} \tag{28}$$

由于 $\boldsymbol{Y}=\boldsymbol{\Sigma}_p\boldsymbol{Z}_p$，$\boldsymbol{Z}_p$ 可通过下式计算

$$\boldsymbol{Z}_p=\boldsymbol{\Sigma}_p^{-1}\boldsymbol{Y} \tag{29}$$

此外，$\boldsymbol{Z}_p$ 可按照以下规则进行归一化

$$\boldsymbol{Z}_p\leftarrow\boldsymbol{Z}_p[\operatorname{diag}(\sum_{j=1}^{c}(\boldsymbol{Z}_p)_{ij}^2)]^{-1} \tag{30}$$

利用以上更新规则迭代更新 $\boldsymbol{R}$ 和 $\boldsymbol{Z}_p$，直到目标函数收敛，完成聚类的全过程。此时，样本的类别可以直接从 $\boldsymbol{Y}$ 中得到。详细的 FCC 算法如下。

算法：基于相关熵的快速聚类算法(FCC)

输入：原始数据 $\boldsymbol{X}$、锚点数量 m、前 p 个引导特征向量、类的数量 c 和参数 λ。

输出：所有样本的聚类类别 $\boldsymbol{Y}$。

1：使用 k 均值生成 m 个锚点。

2：利用式(8)构造锚点图。

3：计算度矩阵 $\boldsymbol{D}$ 和拉普拉斯矩阵 $\boldsymbol{L}$。

4：设 $\boldsymbol{Z}_p\in R^{p\times c}$ 为随机正矩阵，σ 由式(18)得到。

5：while 不收敛 do

6：通过解式(21)更新 $\boldsymbol{R}$。

7：通过解式(28)、式(29)和式(30)更新 $\boldsymbol{Z}_p$。

8：end while

9：由 Y 直接获取样本类别。

FCC 的计算复杂度主要由生成锚点、构造锚点图和迭代优化几个部分组成。这些部分的复杂性分别如下。

(1)利用 k 均值从 n 个样本中生成 m 个锚点，需要 $O(nmdt_1)$ 的复杂度，其中 d 是样本的特征维数，t_1 是 k 均值的迭代次数。

(2)利用式(8)在 n 个样本和 m 个锚点之间构建锚点图时，需要 $O(nmd)$ 的复杂度。

(3)优化时需要的复杂度为 $O((nc^2+npc)t_2)$。准确地说，求解 $\boldsymbol{R}$ 需要 $O(nc^2)$ 的复杂度，更新 $\boldsymbol{Z}_p$ 需要 $O(npc)$ 的复杂度，t_2 是目标函数收敛时的迭代次数。

因此，FCC 的总体计算复杂度为 $O(nmd(t_1+1)+(nc^2+npc)t_2)$。由于在处理大规模数据时 m、d、p、c、t_1 和 t_2 比 n 小得多，因此 FCC 的复杂度可以近似为 $O(n)$。特别是当数据的维数较高时，FCC 在求解目标函数时与数据维数无关，因而仍能保持较低的计算复杂度。

4　实　验

4.1　数据集

本文使用 WebKB[35]、TDT2[12]、Mnist[36] 和 Cora[37] 4 个常用的真实数据集测试了 FCC 和其他典型算法的聚类性能。其中，WebKB 是一种广泛应用于聚类算法的真实文本数据集，本文采用随机采样得到 827 个样本点共 7 类数据，组成新的 WebKB 数据集来测试所有算法的性能；TDT2 包含了11 201个主题文档，本文删去了少于 30 个文档的那些类别，组成了新的包含9 394个文档共 30 类的新的 TDT2 数据集；Mnist 是一个包含60 000个训练样本和10 000个测试样本的手写体数字数据集，本文在训练集中针对每一类随机采样，组成了具有6 996个样本共 10 类的新的 Mnist 数据集；Cora 数据集为计算机科学领域的研究论文集，类似地，我们利用随机采样得到了包含7 511个样本共 10 类的新的 Cora 数据集作为本文中所使用的算法的测试数据。各数据集具体情况见表 1。

表 1　数据集描述

Table 1　Description of all datasets

数据集	类型	样本数	特征维度	类别数
WebKB	Text	827	4 134	7
TDT2	Text	9 394	36 771	30
Mnist	Image	6 996	784	10
Cora	Text	751	6 234	10

4.2 对比方法

为了验证 FCC 的性能，本文选取了几种典型的快速以及鲁棒的聚类算法 FNMTF、BKM、LPNMTF、LSSC 和 KMM 作为对比，对比算法的详细情况总结如下。

FNMTF[22]：一种基于矩阵分解的快速聚类算法，它将因子矩阵直接约束为聚类指示矩阵，并提出了一种只包含少量矩阵乘法的优化算法。

BKM[23]：一种快速的双边 k 均值算法，在约束因子矩阵为聚类指示矩阵的基础上，进一步将吸收因子矩阵限制为对角矩阵。

LPFNMTF[22]：一种新的局部保留正则化的 FNMTF 方法，该方法通过增加流形正则化来实现对两个分解因子矩阵的几何约束。

LSSC[10]：一种稀疏的鲁棒大规模聚类算法，由于 L_1 范数的稀疏性和鲁棒性，该模型利用了 L_1 范数约束带有图正则的聚类模型。

KMM[13]：一种最新的高精度扩展 k 均值方法，它将问题形式化为一个优化问题，这些子问题采用交替优化策略分别求解。

4.3 评估指标

(1)聚类精度[38]：衡量簇与类之间的一一对应关系，可以表示为

$$C=\sum_{i=1}^{n} f(x_i,\mathrm{map}(y_i))/n \tag{31}$$

式中：x_i 为第 i 个样本的真实标签；y_i 为相应的习得标签；$\mathrm{map}(y_i)$ 为使用 Kuhn-Munkres 算法[39]从学习标签到真值标签的转换，$f(x,y)$ 取值如下

$$f(x,y)=\begin{cases}1, & x=y\\ 0, & \text{其他}\end{cases} \tag{32}$$

(2)归一化互信息[40]：通常用于度量聚类结果的相似度，其定义如下

$$S=\frac{I(x_i,y_i)}{\sqrt{E(x_i)E(y_i)}} \tag{33}$$

式中：$I(x_i,y_i)$ 表示互信息；$E(\cdot)$ 表示信息熵。归一化互信息也可以表示为

$$S=\frac{\sum_{i=1}^{c}\sum_{j=1}^{c} h_{ij}\log\left(\frac{h_{ij}}{h_i h_j}\right)}{\sqrt{\left(\sum_{i=1}^{c} h_i\log\frac{h_i}{h}\right)\left(\sum_{j=1}^{c} h'_j\log\frac{h_j}{h}\right)}} \tag{34}$$

式中：c 为类别个数；h_i 为通过聚类方法学习到的第 i 个簇的样本数；h'_j 为第 j 个真实类别含有的样本数；h_{ij} 为 x_i 与 y_i 相交的数量。

(3)聚类纯度[41]：正确聚类的样本百分比通常用聚类纯度表示，定义为

$$P=\sum_{i=1}^{c}\frac{p_i}{pq_i}\max_j q_i^j \tag{35}$$

式中：P 为聚类纯度；p 为样本总数；p_i 为第 i 个簇的样本数；q_i^j 为分配给第 j 个簇的第 i 个类别的数目。

4.4 结果

所有算法均在 256 GB 的 3.50 GHz Intel(R) Xeon(R)CPU E5-1620 v3 和 Matlab 2019a(64 位)平台上运行，聚类结果和计算时间分别如表 2 和表 3 所示。其中“—”表示计算时间超过 5 h，由于其他算法均能在几千秒甚至几十秒内完成聚类全过程，故此计算时间超过 5 h 后我们不再关注该算法的性能。结合表 2 和表 3 可以看到，与这些算法相比，FCC 不仅可以获得更好的聚类性能，而且可以保持最短的计算时间。特别地，由于这些数据集都是真实世界的数据，因此包含了大量不同分布的噪声和离群点。FCC 通过引入相关熵可以有效地抑制非线性和非高斯分布的噪声以及离群点对聚类性能的影响，从而可以更快地处理包含大量噪声和离群点的大规模真实世界数据。

4.5 参数分析

两个参数主要影响 FCC 的性能和效率，分别是锚点数 m 和正则项系数 λ。本文选取锚点集合为 $[c+1,c+3,c+5,c+7,c+9,50]$，图 1 给出了 FCC 在

表 2 所有方法在不同数据集上的聚类结果

Table 2 Clustering results comparison of all methods on different datasets

方法	聚类精度				归一化互信息				聚类纯度			
	WebKB	TDT2	Mnist	Cora	WebKB	TDT2	Mnist	Cora	WebKB	TDT2	Mnist	Cora
FNMTF	0.318 0	0.173 7	0.303 6	0.183 8	0.046 7	0.235 1	0.192 4	0.042 0	0.702 5	0.329 6	0.319 3	0.273 0
BKM	0.351 9	0.375 6	0.134 4	0.239 7	0.033 3	0.302 5	0.030 2	0.025 0	0.702 5	0.430 1	0.138 1	0.258 3
LPFNMTF	0.180 2	0.051 6	0.120 4	0.170 4	0.011 0	0.014 9	0.002 6	0.025 9	0.702 5	0.210 3	0.123 1	0.255 7
LSSC	0.679 6	0.453 4	0.327 8	0.241 0	0.007 3	0.279 3	0.172 3	0.004 9	0.702 5	0.504 7	0.330 5	0.241 0
KMM	0.695 3	—	0.533 2	0.243 7	0.002 6	—	0.524 4	0.012 6	0.702 5	—	0.542 8	0.247 7
本文	0.781 1	0.522 0	0.362 1	0.262 3	0.174 2	0.309 9	0.256 5	0.056 0	0.781 1	0.529 0	0.377 1	0.283 6

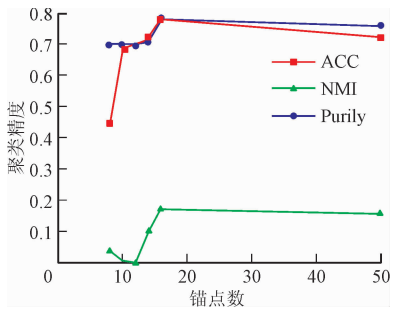
表 3　所有方法在不同数据集上的聚类时间

Table 3　Computational time comparison of all methods on different datasets

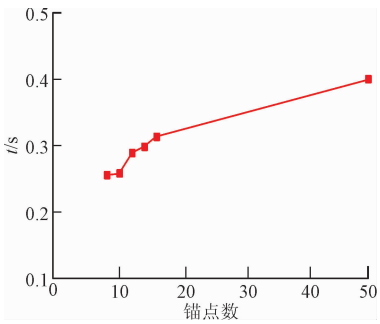
数据集	聚类时间/s					
	FNMTF	BKM	LPFNMTF	LSSC	KMM	FCC
WebKB	3.461 1	8.211 2	4.608 3	0.317 8	12.530 7	0.315 0
TDT2	1 629.486 3	2 656.953 1	1 814.701 7	76.487 7	—	62.035 7
Mnist	12.432 1	7.956 1	25.179 1	6.514 0	41.044 5	6.117 6
Cora	5.360 0	10.076 0	7.134 2	0.302 0	14.036 8	0.211 9

不同锚点数的 WebKB 和 Cora 上运行的实验结果。可以看出，当锚点数较少时，随着锚点数的增加，聚类性能逐渐提高。当锚点数增加到一定程度时，聚类性能相对稳定，不再大幅度增加。同时，随着锚点数量的不断增加，计算时间也在不断增加。因此，选择合适数量的锚对聚类性能和效率有很大的影响。

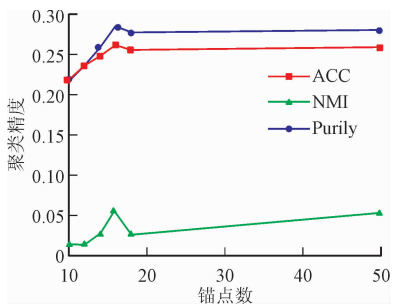
图 2 展示了具有不同 λ 的 FCC 的聚类结果和计算时间，其中 λ 属于 $[10^0, 10^1, 10^2, 10^3, 10^4, 10^5]$ 集合。可以看出，总体来说，当 λ 变化时，聚类精度和聚类纯度相对稳定，最大互信息波动较大。另外，计算时间随着 λ 的增加总体保持稳定波动。因此，合适的 λ 往往会带来较好的聚类性能和短的计算时间。



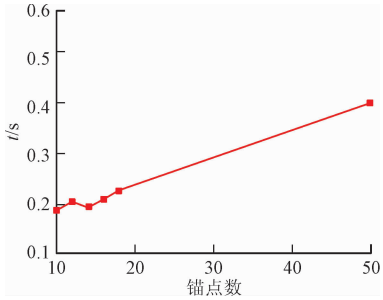
(a)不同锚点数在 WebKB 上对应的聚类结果



(b)不同锚点数在 WebKB 上对应的计算时间



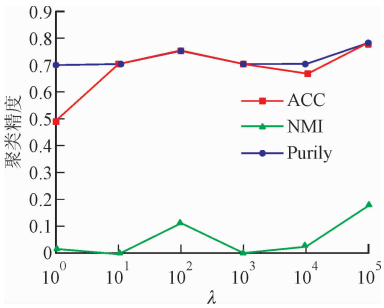
(c)不同锚点数在 Cora 上对应的聚类结果



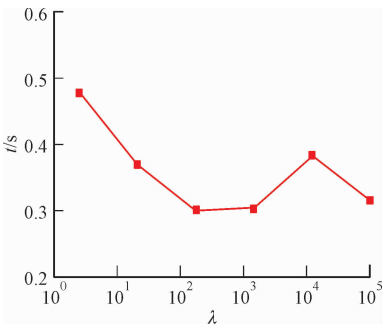
(d)不同锚点数在 Cora 上对应的计算时间

图 1　不同锚点在 WebKB 和 Cora 上聚类性能对比

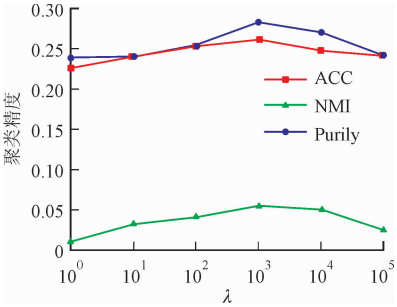
Fig. 1　Clustering performance with different number of anchors on WebKB and Cora



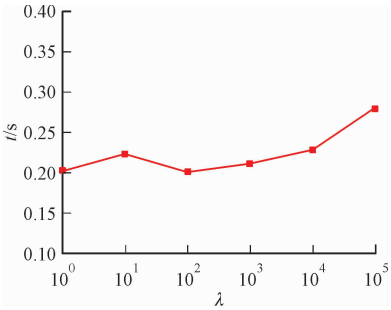
(a)不同 λ 在 WebKB 上对应的聚类结果



(b)不同 λ 在 WebKB 上对应的计算时间



(c)不同 λ 在 Cora 上对应的聚类结果



(d)不同 λ 在 Cora 上对应的计算时间

图 2 不同 λ 在 WebKB 和 Cora 上聚类性能对比

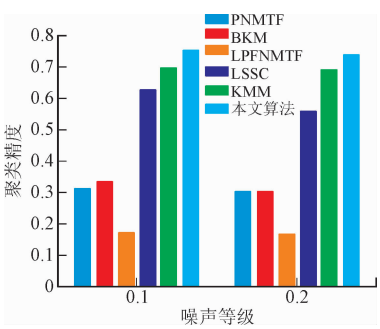
Fig. 2 Clustering performance with different λ on WebKB and Cora

4.6 鲁棒性分析

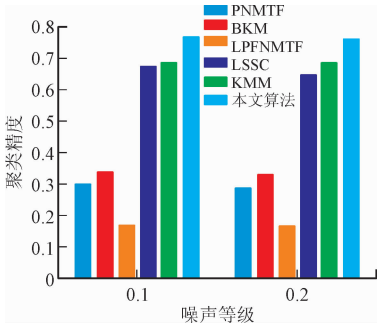
为了进一步验证本算法对噪声(特别是非线性和非高斯噪声)的抑制效果,本文以 WebKB 和 Cora 数据集为例,对其分别加入 10%和 20%的随机噪声和泊松噪声,构成含噪数据集作为鲁棒性测试的数据集。图 3 给出了不同算法在不同噪声条件下的聚类精度对比,从中可以看到,相对于其他对比算法,特别是鲁棒的算法 LSSC 和 KMM,本文方法均可在各种噪声条件下达到优于它们的聚类精度且性能几乎不随噪声程度的增加明显变化,故此本文算法是一种能够高效处理含噪真实数据的鲁棒性算法。

5 结 论

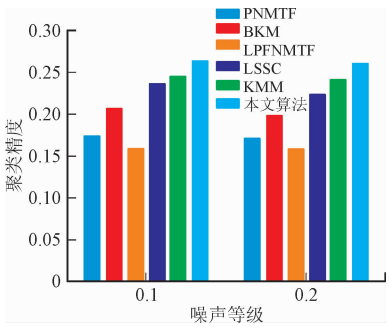
本文提出了一种基于相关熵的快速聚类算法(FCC),以 k 均值生成的标签矩阵代替原始数据作为图聚类的输入,以锚点图代替传统图,降低了数据的维数,可以有效处理大规模的真实数据。同时,通过相关熵有效降低了实际数据中的噪声以及离群点对算法性能的影响,极大地提高了聚类算法的性能。在真实世界大规模数据集上的实验表明,相对于那些典型的聚类算法,FCC 能够在最短时间内达到良好的聚类性能。



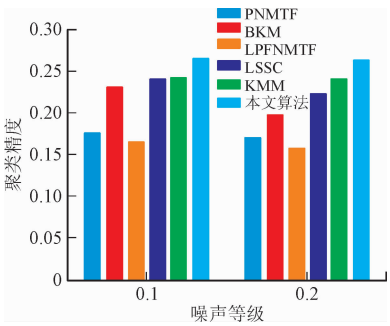
(a)随机噪声腐蚀状态下算法在 WebKB 上的聚类精度



(b)泊松噪声腐蚀状态下算法在 WebKB 上的聚类精度



(c)随机噪声腐蚀状态下算法在 Cora 上的聚类精度



(d)泊松噪声腐蚀状态下算法在 Cora 上的聚类精度

图 3 不同噪声条件在不同数据集上的聚类精度对比

Fig. 3 Clustering accuracy of all methods on different datasets with different noise levels

参考文献:

[1] 刘涛, 尹红健. 基于半监督学习的 K 均值聚类算法研究[J]. 计算机应用研究, 2010, 27(3): 913-916.
LIU Tao, YIN Hongjian. Semi-supervised learning

based on K-means clustering algorithm [J]. Application Research of Computers, 2010, 27(3): 913-916.

[2] 喻金平，郑杰，梅宏标. 基于改进人工蜂群算法的 K 均值聚类算法 [J]. 计算机应用，2014，34(4)：1065-1069，1088.
YU Jinping, ZHENG Jie, MEI Hongbiao. K-means clustering algorithm based on improved artificial bee colony algorithm [J]. Journal of Computer Applications, 2014, 34(4): 1065-1069, 1088.

[3] 林涛，赵璨. 最近邻优化的 k-means 聚类算法 [J]. 计算机科学，2019，46(S2)：216-219.
LIN Tao, ZHAO Can. Nearest neighbor optimization k-means clustering algorithm [J]. Computer Science, 2019, 46(S2): 216-219.

[4] 胡卓娅，翁健. 基于人工蜂群算法的自适应谱聚类算法 [J]. 重庆理工大学学报(自然科学)，2020，34(3)：137-144.
HU Zhuoya, WENG Jian. Adaptive spectral clustering algorithm based on artificial bee colony algorithm [J]. Journal of Chongqing University of Technology (Natural Science), 2020, 34(3): 137-144.

[5] 原虹. 基于稀疏子空间聚类的文本谱聚类算法研究 [J]. 电子技术与软件工程，2020(13)：156-157.

[6] 葛君伟，杨广欣. 基于共享最近邻的密度自适应邻域谱聚类算法 [J/OL]. 计算机工程. [2020-12-20]. https://doi.org/10.19678/j.issn.1000-3428.0058893.

[7] BANERJEE A, MERUGU S, DHILLON I, et al. Clustering with Bregman divergences [C]// Proceedings of the 2004 SIAM International Conference on Data Mining. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2004: 1705-1749.

[8] CAMASTRA F, VERRI A. A novel kernel method for clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 801-805.

[9] LI Yeqing, NIE Feiping, HUANG Heng, et al. Large-scale multi-view spectral clustering via bipartite graph [C]// Proceedings of the 29th National Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI, 2015: 2750-2756.

[10] ZHANG Ruiqi, LU Zhiwu. Large scale sparse clustering [C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. California, USA: IJCAI, 2016: 2336-2342.

[11] LI Xuelong, CUI Guosheng, DONG Yongsheng. Graph regularized non-negative low-rank matrix factorization for image clustering [J]. IEEE Transactions on Cybernetics, 2017, 47(11): 3840-3853.

[12] PENG Siyuan, SER W, CHEN Badong, et al. Correntropy based graph regularized concept factorization for clustering [J]. Neurocomputing, 2018, 316: 34-48.

[13] NIE Feiping, WANG Chenglong, LI Xuelong. K-multiple-means: a multiple-means clustering method with specified K clusters [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM, 2019: 959-967.

[14] BERKHIN P. A survey of clustering data mining techniques [M]// Grouping Multidimensional Data. Cham, Germany: Springer, 2006: 25-71.

[15] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm [C]// Proceedings of the 14th International Conference on Neural Information Processing Systems. Palo Alto, CA, USA: AAAI, 2001: 849-856.

[16] JAIN A K. Data clustering: 50 years beyond k-means [J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

[17] XU Wei, GONG Yihong. Document clustering by concept factorization [C]// Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2004: 202-209.

[18] LEE D, SEUNG H S. Algorithms for non-negative matrix factorization [C]// Proceedings of the 13th International Conference on Neural Information Processing Systems. Vancouver, Canada: NIPS, 2000: 556-562.

[19] CHEN Mulin, LI Xuelong. Concept factorization with local centroids [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020(10): 1-7.

[20] HUANG Jin, NIE Feiping, HUANG Heng. Spectral rotation versus k-means in spectral clustering [C]// Proceedings of the 27th AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI, 2013: 431-437.

[21] NIE Feiping, WANG Xiaoqian, JORDAN M I, et al. The constrained Laplacian rank algorithm for graph-based clustering [C]// Proceedings of the 30th AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI, 2016: 1969-1976.

[22] WANG Hua, NIE Feiping, HUANG Heng, et al. Fast nonnegative matrix tri-factorization for large-scale data co-clustering [C]// Proceedings of the 22nd International Joint Conference on Artificial Intelligence.

California, USA: IJCAI, 2011: 1553-1558.

[23] HAN Junwei, SONG Kun, NIE Feiping, et al. Bilateral k-means algorithm for fast co-clustering [C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI, 2017: 1969-1975.

[24] CAI Deng, CHEN Xinlei. Large scale spectral clustering via landmark-based sparse representation [J]. IEEE Transactions on Cybernetics, 2015, 45(8): 1669-1680.

[25] SHINNOU H, SASAKI M. Spectral clustering for a large data set by reducing the similarity matrix size [C]// Proceedings of the 6th International Conference on Language Resources and Evaluation. Luxemburg: ELRA, 2008: 201-204.

[26] CHEN W Y, SONG Yangqiu, BAI Hongjie, et al. Parallel spectral clustering in distributed systems [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(3): 568-586.

[27] HE Ran, HU Baogang, ZHENG Weishi, et al. Robust principal component analysis based on maximum correntropy criterion [J]. IEEE Transactions on Image Processing, 2011, 20(6): 1485-1494.

[28] PENG Chong, KANG Zhao, HU Yunhong, et al. Robust graph regularized nonnegative matrix factorization for clustering [J]. ACM Transactions on Knowledge Discovery from Data, 2017, 11(3): 1-30.

[29] YAN Hui, LIU Siyu, YU P S. From joint feature selection and self-representation learning to robust multi-view subspace clustering [C]// Proceedings of the 2019 IEEE International Conference on Data Mining. Piscataway, NJ, USA: IEEE, 2019: 1414-1419.

[30] PRINCIPE J C. Information theoretic learning: Renyi's entropy and kernel perspectives [M]. Cham, Germany: Springer, 2010: 123-130.

[31] HE Ran, ZHENG Weishi, HU Baogang. Maximum correntropy criterion for robust face recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(8): 1561-1576.

[32] LU Canyi, TANG Jinhui, LIN Min, et al. Correntropy induced L2 graph for robust subspace clustering [C]// Proceedings of the 2013 IEEE International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2013: 1801-1808.

[33] ZHOU N, XU Y, CHENG H, et al. Maximum correntropy criterion-based sparse subspace learning for unsupervised feature selection [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 29(2): 404-417.

[34] NIE Feiping, ZHU Wei, LI Xuelong. Unsupervised large graph embedding [C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI, 2017: 2422-2428.

[35] CAI Deng, HE Xiaofei, WU Xiaoyun, et al. Non-negative matrix factorization on manifold [C]// Proceedings of the 2008 8th IEEE International Conference on Data Mining. Piscataway, NJ, USA: IEEE, 2008: 63-72.

[36] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[37] SEN P, NAMATA G, BILGIC M, et al. Collective classification in network data [J]. AI Magazine, 2008, 29(3): 93.

[38] WU Mingrui, SCHÖLKOPF B. A local learning approach for clustering [C]// Proceedings of the 19th International Conference on Neural Information Processing Systems. Vancouver, Canada: NIPS, 2006: 1529-1536.

[39] JOHNSON D S, PAPADIMITRIOU C H, STEIGLITZ K. Combinatorial optimization: algorithms and complexity [J]. The American Mathematical Monthly, 1984, 91(3): 209.

[40] STREHL A, GHOSH J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions [J]. Journal of Machine Learning Research, 2002, 3(12): 583-617.

[41] MANNING C D, RAGHAVAN P, SCHÜTZE H. An introduction to information retrieval [M]. New York, USA: Cambridge University Press, 2008: 139-162.

(编辑 杜秀杰 陶晴)