基于约束优化传播的改进大规模数据半监督式谱聚类算法*

徐达宇[1,2], 郁莹珺[1], 冯海林[1†], 张旭尧[1]
(1. 浙江农林大学 信息工程学院, 杭州 311300; 2. 信雅达系统工程股份有限公司, 杭州 310053)

摘 要: 针对传统谱聚类算法在聚类过程中所出现的高计算复杂度、噪声敏感,以及聚类簇形态偏斜等问题,结合当前大规模数据聚类的特点与需求,建立基于约束优化传播的改进大规模数据半监督式谱聚类模型。该模型利用先验成对点约束信息构建微型相似性矩阵,在此基础上采用 Gabow 算法提取该微型相似性矩阵所对应连通图的各强连通分支,继而提出面向各强连通分支的新型约束优化传播算法以获取整个数据集的点对相似度,最后通过奇异值分解并运用加速 K-means 算法获得大规模数据的聚类结果。在多个标准测试数据集上的实验表明,相比于该领域其他前期研究成果,该聚类模型具有更高的聚类准确率和更低的计算复杂度,更适合大规模数据的聚类应用。

关键词: 谱聚类; 大规模数据; 点对约束; 相似性传播; 奇异值分解
中图分类号: TP391;TP301.6 **文献标志码:** A **文章编号:** 1001-3695(2018)05-1325-06
doi:10.3969/j.issn.1001-3695.2018.05.010

Constrain optimal propagation-based improved semi-supervised spectral clustering algorithm for large-scale data

Xu Dayu[1,2], Yu Yingjun[1], Feng Hailin[1†], Zhang Xuyao[1]
(1. *School of Information Engineering, Zhejiang A&F University, Hangzhou* 311300, *China*; 2. *Sunyard System Engineering Co., Ltd., Hangzhou* 310053, *China*)

Abstract: Focusing on the problem of high computational complexity, noise sensitivity and the shape deviation of cluster in the clustering process of traditional spectral clustering, and combining the characteristics with the need of current large-scale data clustering, this article established the semi-supervised of large-scale data model based on constrained optimal propagation. First, it constructed the micro similarity matrix by using prior dotted pair constraint information. On this basis, it used the Gabow algorithm to extract the micro similarity matrix corresponding connected graph of each strongly connected component. Then, it proposed a new constrained optimization propagation algorithm for each strongly connected component to obtained the similarity of the point of the whole data set. Finally, it could obtain the clustering results of large scale data by using the singular value decomposition and the accelerated K-means algorithm. Experiments on multiple standard testing data sets show that compared with other previous research results in this field, the proposed clustering model has higher clustering accuracy and lower computation complexity and is more suitable for large-scale data clustering applications.

Key words: spectral clustering; large-scale data; pairwise constraint; affinity propagation; singular value decomposition

0 引言

谱聚类算法在处理具有非欧氏空间性、线性不可分性,以及非凸球性数据的聚类问题时展现出了明显的优势[1],因而被广泛地应用于图像分割[2]、面部识别[3]、特征融合[4]、三维形状检测[5]和蛋白质序列分析[6]等领域。原始谱聚类方法以整个数据集各点对之间的相似性度量矩阵为基础来构建 Laplacian 矩阵,从而计算出矩阵的特征值和特征向量,并根据某种规则选取若干特征向量进行聚类。该算法执行过程简述如下:

a)对于给定的一组包含 n 个 d 维数据样本的数据集 X: $\{x_1^d, x_2^d, \cdots, x_n^d\}$,谱聚类算法将每个数据样本看做无向图 $G=(V,E)$中的顶点 V,并采用特定的数据相似性度量方法获得点对相似度信息,根据样本间的相似度将顶点间的边 E 赋权重值,以此构建亲和矩阵(又称做相似矩阵)W;b)将亲和矩阵的每行元素相加,可得到每个顶点的度值,以所有度值为对角元素构成的对角矩阵称为度矩阵,用 D 表示,再利用 W、D 矩阵计算得出 Laplacian 矩阵(依据不同的计算方法,有规范 Laplacian 矩阵和非规范 Laplacian 矩阵两类形式);c)依托所得 Laplacian 矩阵,采用如最小割集准则[7]、规范割集准则[8]、比例割集准则[9]、多路规范割集准则[10]等不同的划分准则对无向图 G 进行划分并求得 Laplacian 矩阵的特征值和特征向量,

收稿日期: 2016-12-29; **修回日期:** 2017-03-06 **基金项目:** 国家自然科学基金资助项目(61272313);浙江省自然科学基金项目(LQ17G010003);浙江省重大科技专项项目(2015C03008)

作者简介: 徐达宇(1985-),男,浙江杭州人,讲师,博士,主要研究方向为机器学习与人工智能、云计算及大数据挖掘;郁莹珺(1992-),女,浙江临安人,硕士研究生,主要研究方向为农林业物联网;冯海林(1980-),男(通信作者),安徽安庆人,教授,博士,主要研究方向为智能检测技术、大数据、物联网等;张旭尧(1988-),女,安徽马鞍山人,讲师,硕士,主要研究方向为林业信息化(zxyzafu@163.com).

并以此为基础,使用类 K-means 算法获得最终聚类结果。

对于聚类算法来说,聚类结果的精确性是评价其优劣的核心概念,原始谱聚类算法是一类非监督式聚类算法,易受样本的分布和数量、划分准则、相似性度量算法选择,以及预分类数等因素的影响,具有较强的不稳定性,算法鲁棒性较差,所以在实际应用过程中往往受到一定的限制。为了解决这一问题,国内外学者进行了大量的谱聚类算法改进研究,提出了一系列半监督式谱聚类算法[3,11~13]。半监督谱聚类算法通过两种先验信息来引导聚类过程,即标号点信息和成对约束信息,由于标号点信息可以转换为成对约束信息,所以通常在半监督聚类中以成对约束信息(Must-Link 与 Cannot-Link)作为先验信息来监督聚类[14]。Must-Link 设定:如果两个样本属于 Must-Link 约束,那么这两个样本在聚类时必须被分配到同一个聚类中。相应地,Cannot-Link 则规定:如果两个样本属于 Cannot-Link 约束,那么这两个样本在聚类时必须被分配到不同聚类之中。即 Must-Link 约束要求两个数据点必须在同一个聚类中,而 Cannot-Link 约束要求两个点不能在同一聚类中。基于半监督方式的谱聚类算法正是利用样本的先验信息和背景知识,达到充分提高无监督聚类性能的目的。

近几年来,随着大规模数据在金融、医疗、生命科学、交通,以及其他社会经济与科学研究领域的不断涌现,使得传统的数据聚类方法在对此类数据进行聚类时遇到了极大的挑战。对于原始谱聚类算法而言,在执行过程中不仅需要 $O(n^2)$ 的时间复杂度来计算出亲和矩阵 W、构建无向图 G 和得出 Laplacian 矩阵,还需要 $O(n^3)$ 的时间复杂度对 Laplacian 矩阵进行特征分解和计算最终结果,如此高的算法复杂度严重妨碍了原始谱聚类模型直接应用于大规模数据集的聚类分析(考虑到一般数据处理设备的 CPU 计算能力与内存容量有限性)。针对这一问题,相关学者也提出了相应的算法加速策略[15~17],此类策略的主要思想是在整个大规模数据集中抽取少量样本点 $\{P\}$,构建样本点之间,以及样本点 $\{P\}$ 与整个数据集 $\{X\}$ 之间的微型点对相似度矩阵,再运用一定的约束传播机制形成整个数据集 $\{X\}$ 的亲和矩阵,在此基础上再进行特征值提取并获得谱聚类结果,但该策略易受样本点选取和约束传播机制设定的影响,聚类精确度难以保证。

针对谱聚类算法在聚类精确度和面向大规模数据应用时模型扩展能力上所出现的缺陷,本文在上述研究成果的基础上,建立基于约束优化传播的改进大规模数据半监督式谱聚类模型。该模型首先利用先验成对约束信息构建微型相似性矩阵,在此基础上采用 Gabow 算法提取该微型相似性矩阵所对应连通图的各强连通分支,继而提出面向各强连通分支的新型约束优化传播算法以获取整个数据集的亲和矩阵和 Laplacian 矩阵,最后结合矩阵奇异值分解(singular value decomposition, SVD)及 K-means 算法获得大规模数据的聚类结果。在多个标准测试数据集上的实验表明,相比于该领域其他前期研究成果,本文所提聚类模型具有更高的聚类准确率和更低的计算复杂度,更适合大规模数据的聚类应用。

1 微型亲和矩阵稀疏表示

建立基于约束优化传播的改进大规模数据半监督式谱聚类模型,总的目标是将给定的由 n 个数据样本构成的大规模数据集 $\boldsymbol{X}\in \mathrm{R}^{d\times n}$ 放入 K 个聚类簇中,在此之前计算出亲和矩阵 $\boldsymbol{W}$ 是关键,而直接对 $\boldsymbol{X}$ 中的所有数据进行点对相似度计算对于大规模数据集来讲并不现实,因此本文将采用算法加速策略来设计 $\boldsymbol{W}$ 以避免高计算耗费。

a)需要完成微型亲和矩阵的稀疏表示。根据先验信息将 $\boldsymbol{X}$ 中部分数据纳入 Must-Link 与 Cannot-Link 两类数据池中,并在这两类约束信息数据池中随机抽取 p 个数据样本构成微型数据集 $\boldsymbol{P}\in \mathrm{R}^{d\times p}$,使其数量上满足 $p<<n$。

b)构造 p 个数据样本与 n 个全体数据之间的相似度矩阵 $\boldsymbol{Z}\in \mathrm{R}^{p\times n}$。由稀疏编码策略[18]与 Nadaraya-Watson 核回归算法[19]可知,对于 $\boldsymbol{X}$ 中的任意一点 $\boldsymbol{x}_i$,其近似值 $\hat{\boldsymbol{x}}_i$ 可按如下公式计算得到

$$\hat{\boldsymbol{x}}_i=\sum_{j=1}^{p} z_{ji}\boldsymbol{P}_j \tag{1}$$

其中:z_{ji} 为矩阵 $\boldsymbol{Z}$ 中第 j 行、第 i 列元素;$\boldsymbol{P}_j$ 为矩阵 $\boldsymbol{P}$ 中第 j 个列向量(即 $\boldsymbol{P}$ 中第 j 个样本点),从而可知 $\boldsymbol{X}\approx \boldsymbol{PZ}$。接下来需要对矩阵 $\boldsymbol{Z}$ 进行稀疏表示。若 $\boldsymbol{P}_j$ 不属于离 $\boldsymbol{x}_i$ 最相近的 r 个数据点,则令 z_{ji} 为 0;若 $\boldsymbol{P}_j$ 属于离 $\boldsymbol{x}_i$ 最相近的 r 个数据点集 $\{I\}$ 中,则 z_{ji} 可由式(2)计算得出:

$$z_{ji}=\frac{\varphi(\boldsymbol{x}_i,\boldsymbol{P}_j)}{\sum_{j'\in\{I\}}\varphi(\boldsymbol{x}_i,\boldsymbol{P}_{j'})} \tag{2}$$

其中:$i=1,2,\cdots,n$ 且 $j\in\{I\}$,$\varphi(\cdot)$ 是带宽为 σ 的高斯核函数,即

$$\varphi(\boldsymbol{x}_i,\boldsymbol{P}_j)=\exp(-\|\boldsymbol{x}_i-\boldsymbol{P}_j\|/2\sigma^2) \tag{3}$$

根据式(2)和(3)可计算得到 $\boldsymbol{Z}$ 的稀疏表示。

2 约束优化传播机制构建

2.1 基于成对约束的强连通分支提取

在完成了微型亲和矩阵 $\boldsymbol{Z}$ 稀疏表示这一初始步骤后,接下来仍需要对矩阵 $\boldsymbol{Z}$ 中的元素进行不断更新,使其能够真实反映数据样本间的亲和关系,并为约束传播建立基础。

a)对于服从 Must-Link 与 Cannot-Link 两类约束的 p 个数据样本,构建其样本点之间的微型相似性矩阵(矩阵规模为 $p\times p$)。构建规则设定为:若两数据点属于 Must-Link 约束,则此两点间的相似度值为 1;若两数据点属于 Cannot-Link 约束,则此两点间的相似度值为 0。

b)采用 Gabow 算法[20]来提取该 $p\times p$ 矩阵所对应无向图的各强连通分支(strongly connected components, SCC),从而将 p 个含有约束信息的数据样本进行分割,产生 M 个相互无交集的岛(islands),每个 island 内的数据属于同一个强连通分支,即满足

$$P\equiv\{\mathrm{island}_1\cup\mathrm{island}_2\cup\cdots\cup\mathrm{island}_M\} \tag{4}$$

若 P 中某一点 P_a 属于某个 island,即 $P_a\in\mathrm{island}_m$,则令 $\mathrm{island}_m^{\sim}$ 为 island_m 的相对补集,并满足条件式(5)。

$$\mathrm{island}_m^{\sim}\equiv P\backslash\mathrm{island}_m\equiv\{P_a^{\sim}\in P\mid P_a^{\sim}\notin\mathrm{island}_m\}$$
$$m=1,2,\cdots,M \tag{5}$$

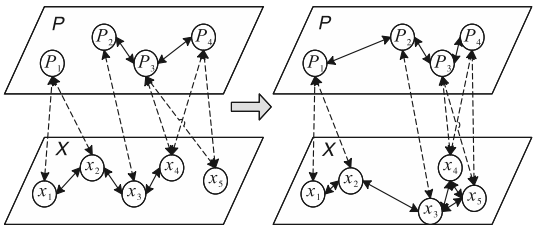
依据式(5)所形成的小样本数据点关系可对微型亲和矩阵 $\boldsymbol{Z}$ 的元素进行更新为

$$\hat{Z}(P_i,P_j)=\begin{cases}1 & \{P_i\in\mathrm{island}_m\mid P_j\in\mathrm{island}_m\}\\ 0 & \{P_i\in\mathrm{island}_m\mid P_j\in\mathrm{island}_m^{\sim}\}\end{cases} \tag{6}$$

进一步地,由 $D_1 = \sum_j \hat{Z}_{ji}$ 可得 $\hat{Z}$ 的度矩阵。依照式(6)所示的亲和矩阵更新过程,可将先验约束信息注入亲和矩阵中,从而能够更好地指导聚类过程,提升聚类精确度。

2.2　约束优化传播算法

在获得了微型亲和矩阵 $\hat{Z}$ 与各强连通分支后,抽取的约束样本间以及约束样本与整个数据集各点间的前期相似关系就已基本建立,但以上信息只囊括了部分数据之间的相似关系,并没有将全体数据之间的相似程度都进行完整表达,亲和矩阵依然稀疏,有许多数据点间的相似信息需要建立或更新,因此需要通过约束传播这一过程实现上述目标。为了在约束传播过程中更为有效与真实地反映数据间的相似关系,本文提出一种新型约束优化传播(constraints optimized propagation,COP)算法。该算法的基本思想是异构空间对象之间的相似性可以相互影响,并且这种影响可以提高或降低两个对象之间的相似程度。具体来说,以本文所需解决的约束传播问题为例,全体数据集 X 与抽取的约束数据集 P 处于两个异构空间,并且整体数据集 X 中的数据 x_i 之间、约束数据集 P 中的数据 P_j 之间,以及 x_i 与 P_j 之间已经建立初始的但不完整的相似关系。下面用如图 1 所示的简化约束优化传播过程来说明这一思想。



(a) 约束传播前　　(b) 约束传播后

图1　约束优化传播过程示意图

在图 1 中,两个异构空间分别为 P 与 X,实连接线表示同一空间内数据间的相似关系,虚连接线表示不同空间数据间的相似关系。在使用所提的优化约束传播算法前(图 1(a)),可以看出在 X 中,x_1、x_2、x_3 和 x_4 相似,x_5 与其他点都不相似,而在 P 中,x_4 和 x_5 都与 P_3 和 P_4 相似。且 P_3 和 P_4 之间也相似;对于 x_2 和 x_3,两点在 X 中相似,而在 P 中,x_2 与 P_1 相似,x_3 与 P_2 相似,但 P_1 与 P_2 不相似。这一初始状态在使用优化约束传播算法后会使数据点间的情况发生较大改变,如图 1(b)所示,x_4 和 x_5 之间的相似关系会因为 P_3 和 P_4 而建立并且不断提升和巩固;x_2 和 x_3 之间原本具有的相似关系会因为 P_1 与 P_2 而减弱,这类关系会随着算法不断循环执行而变得清晰,直到算法收敛,所有数据点间的关系才得以确定。不失一般性,本文所提的 COP 算法描述如下:

已知构造的微型亲和矩阵 $\hat{Z} \in \mathrm{R}^{p\times n}$ 和抽取的约束样本间微型相似矩阵 $S \in \mathrm{R}^{p\times p}$,则全体数据点的亲和矩阵 $W \in \mathrm{R}^{n\times n}$ 可通过如下迭代公式计算得到

$$\begin{cases} \hat{W} = \alpha W + (1-\alpha)\lambda\, \hat{Z}^{\mathrm{T}}\, \hat{S}\hat{Z} \\ \hat{S} = \beta S + (1-\beta)\lambda\, \hat{Z}\hat{W}\hat{Z}^{\mathrm{T}} \end{cases} \tag{7}$$

其中:$\hat{Z}^{\mathrm{T}}$ 为 $\hat{Z}$ 的转置矩阵;α 与 β 为权重系数,用于调整新旧权重之间的变化程度;λ 为衰退系数;α、β 和 λ 的取值都为[0,1],三者的取值将影响 W 与 S 矩阵在每次迭代中的更新幅度;N 为算法迭代数。从式(7)可以看出,两类异构空间数据间的相似度彼此相互影响与作用,并且以一种非线性方式进行作用,真实反映约束传播过程中的非线性映射关系。下面将对式(7)的收敛性进行验证,证明该算法的有效性。

令 W 与 S 矩阵的第 N 次迭代结果为 $\hat{W}^{(N)}$ 与 $\hat{S}^{(N)}$,用归纳法证明式(7)收敛性:

$$\hat{W}^{(N)} - \hat{W}^{(N-1)} = [\alpha W + (1-\alpha)\lambda\, \hat{Z}\hat{S}^{(N)}\hat{Z}^{\mathrm{T}}] - [\alpha W + (1-\alpha)\lambda\, \hat{Z}\hat{S}^{(N-1)}\hat{Z}^{\mathrm{T}}] = (1-\alpha)\lambda\, \hat{Z}(\hat{S}^{(N)} - \hat{S}^{(N-1)})\hat{Z}^{\mathrm{T}}$$

同样地,

$$\hat{S}^{(N)} - \hat{S}^{(N-1)} = [\beta S + (1-\beta)\lambda\, \hat{Z}^{\mathrm{T}}\hat{W}^{(N-1)}\hat{Z}] - [\beta S + (1-\beta)\lambda\, \hat{Z}^{\mathrm{T}}\hat{W}^{(N-2)}\hat{Z}] = (1-\beta)\lambda\, \hat{Z}^{\mathrm{T}}(\hat{W}^{(N-1)} - \hat{W}^{(N-2)})\hat{Z}$$

将 $\hat{S}^{(N)} - \hat{S}^{(N-1)}$ 代入 $\hat{W}^{(N)} - \hat{W}^{(N-1)}$ 中,可得

$$\hat{W}^{(N)} - \hat{W}^{(N-1)} = (1-\alpha)(1-\beta)\lambda^2\, \hat{Z}\hat{Z}^{\mathrm{T}}(\hat{W}^{(N-1)} - \hat{W}^{(N-2)})\hat{Z}\hat{Z}^{\mathrm{T}}$$

令 $\omega = (1-\alpha)(1-\beta)\lambda^2$,$V = \hat{Z}\hat{Z}^{\mathrm{T}}$,则

$$\hat{W}^{(N)} - \hat{W}^{(N-1)} = \omega V^{\mathrm{T}}(\hat{W}^{(N-1)} - \hat{W}^{(N-2)})V = \cdots = \omega^{N-1}V^{N-1}(\hat{W}^{(1)} - \hat{W}^{(0)})V^{N-1}$$

令 $V = [v_{ij}]_{n\times n}$,且 $V = \hat{Z}\hat{Z}^{\mathrm{T}} = \hat{W}^{(0)}$,由式(2)可知,当 $0 < \varphi(x_i, P_j) \leqslant 1$ 时,$0 < v_{ij} \leqslant 1$;当 $\varphi(x_i, P_j) = 0$,$v_{ij} = 0$。因此,$\lim_{N\to\infty} V^{N-1} = 0$,考虑到 $\hat{W}^{(1)} - W$ 是常数矩阵,$\omega < 1$,所以,$\lim_{N\to\infty}(\hat{W}^{(N)} - \hat{W}^{(N-1)}) = 0$。从而证明了式(7)的收敛性和有效性,且亲和矩阵 W 得以最终获得并确定。

3　聚类形成与算法复杂度分析

3.1　聚类形成

在获得全体数据集的亲和矩阵 W 后,计算其度矩阵 $D_2 \in \mathrm{R}^{n\times n}$,结合度矩阵 D_1,依据式(8)获得最终的微型亲和矩阵 Z^* 为

$$Z^* = D_1 W D_2 \tag{8}$$

对矩阵 Z^* 进行奇异值分解,令

$$Z^* = A\Sigma B^{\mathrm{T}} \tag{9}$$

其中:$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_p)$;$\sigma_i$ 为 Z^* 的奇异值,且满足 $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_p \geqslant 0$;矩阵 A 为 Z^* 的左奇异向量,即 $A = [a_1, a_2, \cdots, a_p] \in \mathrm{R}^{p\times p}$,同理矩阵 $B = [b_1, b_2, \cdots, b_p] \in \mathrm{R}^{n\times p}$ 为 Z^* 的右奇异向量。根据奇异值分解理论可知,B 是矩阵 $(Z^*)^{\mathrm{T}}Z^*$ 的特征向量,而 A 是 $Z^*(Z^*)^{\mathrm{T}}$ 的特征向量,从而可以用 $O(p^3)$ 的时间复杂度计算得到 A,并通过公式 $B = \Sigma^{-1}A^{\mathrm{T}}Z^*$ 计算出 B。在此基础上,本文再对 B 中的行向量使用加速 K-means 算法[21]获得最终的数据聚类结果。本文将所提的模型命名为 COP-SC,其算法伪代码如下所示。

COP-SC 聚类模型

输入:大规模数据集 X,约束集 P,参数 α、β 和 λ,迭代数 N。

输出:K 个聚类簇。

1 for $i = 1:n$; $j = 1:p$

　　由式(2)(3)计算矩阵 Z

end

2 采用 Gabow 算法和式(6)更新 Z 得到 $\hat{Z}$,计算 $\hat{Z}$ 的度矩阵 D_1

3 for $i=1:N$

由式(7)计算得到亲和矩阵 W

end

4 计算 W 的度矩阵 D_2,由式(8)得到 Z^*

5 由式(9)计算 A 和 B

6 利用 B 和加速 K-means 算法得到 K 个聚类簇

3.2 算法复杂度分析

基于以上分析与论述可知,使用本文所提的基于约束优化传播的改进大规模数据半监督式谱聚类模型,可以完全避免直接计算整个数据集的亲和矩阵以及 Laplacian 矩阵的特征值与特征向量。因此,可将原始谱聚类算法的时间复杂度由原先的 $O(n^3+n^2)$ 降低到现有的 $O[p^3+(p^2+Np+1)n]$。当 $p<<n$ 时,本文所提模型 COP-SC 相比于原始谱聚类算法可以有效降低计算复杂度,大幅压缩了算法执行所需的计算空间与时间。

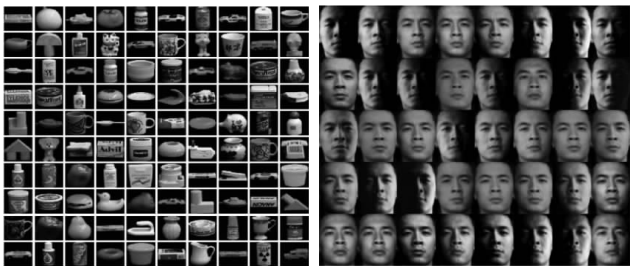
4 实验与分析

4.1 实验数据

在实验数据来源方面,本文采用四组高维大规模基准测试数据集[22,23]对所提模型进行有效性论证,具体包括一组形状图像数据集 COIL100、一组人脸识别数据集 CMU PIE,以及两组手写数字数据集 MNIST 和 USPS。四组数据的基本信息如表 1 所示,图 2 给出了 COIL100 和 CMU PIE 的部分数据实例。

表 1 实验数据描述

数据集	数据量(n)	维数	类别数
COIL100	7 200	1 024	100
CMU PIE	11 554	1 024	68
MNIST	70 000	784	10
USPS	9 298	256	10



(a) COIL100数据实例 (b) CMU PIE的部分数据实例

图2 部分数据实例

4.2 实验设置

在对比模型选取方面,为了全面、客观地评价本文所提模型 COP-SC 的有效性和可靠性,将采用该领域研究成果中两类共七个模型进行对比实验。第一类中包含两个基准谱聚类算法,即 NCut[24] 和 NSDR[25],两个算法均未采用加速策略,从而可以验证使用加速策略与约束传播机制的谱聚类算法在聚类精度和计算耗费上的优劣性;第二类比较算法包含五个改进谱聚类算法,分别是 LI-ASP 谱聚类算法[15]、LSC-WPR 谱聚类算法[17]、LSC-K 谱聚类算法[18]、Nyström 谱聚类算法[26] 及 Graclus[27],这五个算法都采用了相应的算法加速策略与约束传播机制,从而可以更直接反映本文所提 COP-SC 模型的优劣性。在聚类精度评估标准方面,本文采用在聚类领域常用的两个评价标准:

聚类精度 ACC(accuracy of clustering)与归一化互信息 NMI(normalized mutual information)。

a)ACC 准则计算公式为

$$\mathrm{ACC}=\frac{\sum_{i=1}^{n}\delta(c_i^h,\mathrm{map}(c_i^l))}{n} \tag{10}$$

其中:c_i^h 与 c_i^l 分别表示真实类标签与模型计算后获得的类标签;map(·)为置换函数,用于匹配真实类标签与获得的类标签;$\delta(\cdot)$为二元布尔函数,当且仅当 $x=y$ 时,$\delta(x,y)=1$,其他情况 $\delta(x,y)=0$。ACC 准则反映出模型计算所得的分类情况与真实分类之间的匹配程度,值越大表示越精确。

b)NMI 准则计算公式为

$$\mathrm{NMI}(X,Y)=\frac{MI(X,Y)}{\sqrt{H(X)H(Y)}} \tag{11}$$

其中:X 与 Y 为两个随机向量;$MI(X,Y)$ 表示两者的互信息;$H(X)$ 与 $H(Y)$ 分别表示 X 与 Y 的信息熵。

对于聚类算法而言,X 与 Y 分别为真实类标签向量与计算所得类标签向量,即 $C_i^l=\{c_1^l,c_2^l,\cdots,c_k^l\}$ 与 $C_j^h=\{c_1^h,c_2^h,\cdots,c_{k'}^h\}$,$k$ 与 k' 分别为真实类别数与计算所得类别数。针对聚类评价问题,NMI 准则可表示为

$$\mathrm{NMI}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{k'}n_{ij}\log\left(\frac{nn_{ij}}{n_i^l n_j^h}\right)}{\sqrt{\sum_{i=1}^{k}n_i^l\log\left(\frac{n_i^l}{n}\right)\sum_{j=1}^{k'}n_j^h\log\left(\frac{n_j^h}{n}\right)}} \tag{12}$$

其中:n_i^l、n_j^h 分别为 C_i^l 与 C_j^h 中的数据点数量;n_{ij} 表示 C_i^l 与 C_j^h 中同时具有的数据点数量。NMI 取值为[0,1],与 ACC 准则类似,NMI 值越大表示聚类越精确,效果越好。

在模型参数设置方面,针对上述四个数据集,经过多次实验反复调整与对比,将本文所提模型 COP-SC 的最优参数分别设定为 $\alpha=0.3$,$\beta=0.8$,$\lambda=0.8$,以及确保收敛所需的迭代数 $N=30$,并且为了满足 $p<<n$ 这一条件,p 的取值分别为整个数据集数据量 n 的 5%、10%、15% 和 20%,其他对比模型的参数设置都以其文献中最优参数设置为准。考虑到实验数据的大规模性,所有实验均在具备 4 核 3.4 GHz CPU 与 16 GB 内存的工作站上完成,所用软件为 MATLAB 2016a。

4.3 实验结果与分析

为了排除约束样本选择不同带来的算法执行的不确定性,本文中每个实验都运行 10 次,取平均值作为最后输出。表 2、3 给出了八个聚类模型在四个测试数据集上运行所得结果的 ACC 值与 NMI 值。表 4 给出了所有模型的执行时间,黑色加粗部分为每组最优值。

表 2 各模型聚类结果的 ACC 值 /%

	模型	$p=5\%$	$p=10\%$	$p=15\%$	$p=20\%$
COIL100 加速模型	COP-SC	**54.73**	**56.32**	**60.12**	**65.78**
	Nyström	45.53	46.18	46.53	46.92
	LI-ASP	44.58	49.88	51.14	52.80
	LSC-WPR	48.05	53.83	55.64	59.36
	LSC-K	47.87	52.20	53.92	55.59
	Graclus	45.33	48.62	52.70	56.28
	基准模型				
	NCut	41.59	51.41	51.46	51.09
	NSDR	42.31	53.22	55.98	57.45

续表 2

模型		$p=5\%$	$p=10\%$	$p=15\%$	$p=20\%$
CMU-PIE 加速模型	COP-SC	26.87	**31.60**	**33.00**	**36.99**
	Nyström	18.69	20.56	23.54	25.51
	LI-ASP	16.78	17.61	18.88	20.07
	LSC-WPR	**27.58**	28.01	29.94	30.34
	LSC-K	25.87	26.01	27.87	28.51
	Graclus	22.84	25.86	28.35	31.44
	基准模型				
	NCut	21.90	22.09	21.98	22.07
	NSDR	6.82	6.91	6.22	4.99
MNIST 加速模型	COP-SC	**75.21**	**82.54**	**83.85**	**84.11**
	Nyström	47.94	54.35	54.72	60.36
	LI-ASP	60.63	64.78	65.02	66.06
	LSC-WPR	71.20	73.46	76.24	77.81
	LSC-K	70.54	72.98	74.54	75.37
	Graclus	69.19	71.54	75.30	76.46
	基准模型				
	NCut	N/A	N/A	N/A	N/A
	NSDR	N/A	N/A	N/A	N/A
USPS 加速模型	COP-SC	**70.60**	72.65	**77.55**	**84.86**
	Nyström	58.73	59.57	60.02	60.53
	LI-ASP	57.51	**74.35**	65.97	68.11
	LSC-WPR	61.98	69.44	71.78	76.57
	LSC-K	61.75	67.34	69.56	71.71
	Graclus	63.56	70.69	73.24	83.46
	基准模型				
	NCut	53.65	66.32	66.38	65.91
	NSDR	54.58	68.65	72.21	74.11

表 3　各模型聚类结果的 NMI 值　　%

	模型	$p=5\%$	$p=10\%$	$p=15\%$	$p=20\%$
COIL100 加速模型	COP-SC	73.58	**80.71**	**83.85**	**86.76**
	Nyström	71.60	72.14	73.44	73.59
	LI-ASP	70.56	72.66	73.54	74.61
	LSC-WPR	**74.79**	78.21	78.94	80.31
	LSC-K	72.60	76.18	76.32	76.45
	Graclus	72.66	77.48	78.50	79.29
	基准模型				
	NCut	72.34	72.37	72.54	72.23
	NSDR	73.40	74.78	75.74	76.66
CMU-PIE 加速模型	COP-SC	39.26	**45.35**	**54.11**	**59.83**
	Nyström	36.49	37.15	40.80	42.80
	LI-ASP	25.51	27.49	28.02	31.09
	LSC-WPR	**40.72**	42.81	44.89	46.72
	LSC-K	32.63	34.26	35.27	38.61
	Graclus	38.61	42.45	43.03	48.63
	基准模型				
	NCut	39.50	39.71	39.66	39.93
	NSDR	14.91	14.93	12.54	8.79
MNIST 加速模型	COP-SC	70.58	**75.39**	76.50	**77.07**
	Nyström	45.55	48.30	48.40	50.76
	LI-ASP	56.16	61.92	64.21	65.26

续表 3

模型		$p=5\%$	$p=10\%$	$p=15\%$	$p=20\%$
MNIST 加速模型	LSC-WPR	**71.23**	73.45	**76.69**	76.99
	LSC-K	69.11	72.37	74.38	75.50
	Graclus	68.20	70.11	68.36	70.75
	基准模型				
	NCut	N/A	N/A	N/A	N/A
	NSDR	N/A	N/A	N/A	N/A
USPS 加速模型	COP-SC	**76.32**	**78.29**	**81.69**	82.45
	Nyström	72.32	69.98	74.17	74.18
	LI-ASP	71.27	70.48	74.28	75.21
	LSC-WPR	75.54	75.86	79.73	**83.95**
	LSC-K	73.37	73.89	77.08	77.06
	Graclus	74.03	75.94	78.24	79.98
	基准模型				
	NCut	73.06	70.20	73.27	72.81
	NSDR	74.13	72.54	76.50	77.27

表 4　各模型聚类时间　　/s

	加速模型						基准模型	
	COP-PC	Nyström	LI-ASP	LSC-WPR	LSC-K	Graclus	NCut	NSDR
COIL100								
$p=5\%$	**5.16**	6.46	5.33	6.19	5.6	7.79	429.41	456.33
$p=10\%$	**5.88**	10.5	5.93	6.76	6.61	8.52	429.41	460.08
$p=15\%$	**6.51**	28.4	6.84	7.67	8.16	9.66	429.41	463.39
$p=20\%$	**7.84**	56.92	8.21	9.03	10.43	11.38	429.41	466.84
CMU-PIE								
$p=5\%$	**4.57**	8.16	5.28	6.84	5.77	6.18	2 023.93	2 125.66
$p=10\%$	**5.97**	33.88	6.45	8.01	7.69	7.55	2 023.93	2 139.66
$p=15\%$	**6.76**	101.93	7.48	9.04	11.33	8.75	2 023.93	2 151.87
$p=20\%$	**8.24**	235.81	8.56	10.12	14.04	10.02	2 023.93	2 165.21
MNIST								
$p=5\%$	**49.88**	62.79	51.81	60.17	54.43	55.59	N/A	N/A
$p=10\%$	57.64	102.06	**55.53**	65.71	64.25	61.62	N/A	N/A
$p=15\%$	66.48	276.05	**62.44**	74.55	79.32	71.26	N/A	N/A
$p=20\%$	**75.08**	553.26	79.80	87.77	101.38	85.67	N/A	N/A
USPS								
$p=5\%$	**7.88**	8.33	8.35	7.99	7.22	7.58	556.95	558.67
$p=10\%$	**8.65**	13.55	9.75	8.72	8.83	12.33	556.95	593.50
$p=15\%$	**9.82**	36.64	10.27	9.89	10.53	18.34	556.95	597.77
$p=20\%$	**10.59**	53.43	11.40	11.65	14.45	26.82	556.95	602.22

从表 2 和 3 可以看出,除了 NCut 和 NSDR 两个模型在 MNIST 数据集上无法获得计算结果外(由于 MNIST 数据集规模过于庞大,直接计算其亲和矩阵已超出设备处理能力),其他所有模型在四组测试数据集上都获得了相应的 ACC 值与 NMI 值。从各个模型聚类结果的总体情况来看:a)随着 p 取值的增大,所有加速模型运行结果的 ACC 值和 NMI 值都得到了相应的提升,说明利用更多的约束信息有利于构建更符合整个数据集特征的亲和矩阵,这也导致在很多情况下,使用了加速策略和约束传播机制的模型,其聚类性能并不一定优于未使用的基准模型;b)从使用加速策略和约束传播机制的五个模型聚类结果对比来看,绝大部分情况下,本文所提出的 COP-SC 模型在四个测试数据集上的聚类精确度都优于其他模型,尤其是 ACC 值;LSC-WPR 模型在个别数据集的 NMI 值上表现优良,甚至在有些情况下优于 COP-SC 模型,这主要得益于其引入了约束样本的预选择机制,提升了其亲和矩阵构建的精确性。

从表 4 可以看出,除了 LI-ASP 模型在 MNIST 部分测试集上的算法执行时间短于其他所有模型,本文所提出的 COP-SC 模型在剩余的三个数据集上所用时间都短于 LI-ASP、LSC-WPR、LSC-K 以及 NCWE 模型,从而验证了本文在模型构建过程中所使用的一系列算法加速机制(构建微型亲和矩阵、实施约束优化传播策略,及引入奇异值分解和加速 K-means 算法)能够切实提升算法谱聚类的执行速度,降低运行时间,并且保证结果的精确性。

综上所述,综合衡量聚类精度与计算时间复杂度两大指标,本文所提出的模型 COP-SC 在聚类性能上具有更为优越的表现,证明了该模型的有效性和可靠性。

5 结束语

本文针对原始谱聚类算法在面向大规模数据聚类过程中出现的计算复杂度过高、鲁棒性差、可扩展性弱等问题,提出了基于约束优化传播的改进大规模数据半监督式谱聚类模型 COP-SC。该模型利用加速策略与约束优化传播机制,能在降低算法复杂度的同时最大限度地利用约束信息来保障聚类结果的精确度。通过在四组大规模测试数据集上采用 ACC、NMI 以及算法执行时间三个聚类有效性评价指标与该领域其他典型聚类算法进行对比实验,证明了本文所提模型在聚类精确度与算法复杂度上具有更为优良的综合性能。

参考文献:

[1] 管涛,杨婷. 谱聚类广义模型与典型算法分析[J]. 模式识别与人工智能,2014,27(11):1015-1025.

[2] Du Hui, Wang Yuping, Dong Xiaopan, *et al*. Texture image segmentation using spectral clustering[C]//Proc of HCI International Posters' Extended Abstracts. [S. l.]: Springer International Publishing, 2015:671-676.

[3] Ahn I, Kim C. Face and hair region labeling using semi-supervised spectral clustering-based multiple segmentations[J]. IEEE Trans on Multimedia,2016,18(7):1414-1421.

[4] Xiao Xiang, Liu Le, Hu Haifeng. Discriminative feature fusion with spectral method for human action recognition[C]//Proc of Chinese Conference on Biometric Recognition. [S. l.]: Springer International Publishing,2015:641-648.

[5] Bergamasco L C C, Oliveira R A P, Wechsler H, *et al*. Content-based image retrieval of 3D cardiac models to aid the diagnosis of congestive heart failure by using spectral clustering[C]//Proc of the 28th International Symposium on Computer-Based Medical Systems. Piscataway, NJ: IEEE Press,2015:183-186.

[6] Paccanaro A, Casbon J A, Saqi M A S. Spectral clustering of protein sequences[J]. Nucleic Acids Research,2006,34(5):1571-1580.

[7] Flake G W, Tarjan R E, Tsioutsiouliklis K. Graph clustering and minimum cut trees[J]. Internet Mathematics, 2003, 1(4): 385-408.

[8] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(8):888-905.

[9] Chan P K, Schlag M D F, Zien J Y. Spectral K-way ratio-cut partitioning and clustering[J]. IEEE Trans on Computer Aided Design of Integrated Circuits and Systems,1994,13(9):1088-1096.

[10] Bolla M Z. Multiway cuts and spectra[M]//Spectral Clustering and Biclustering: Learning Large Graphs and Contingency Tables. [S. l.]: Wiley,2013:44-95.

[11] Ding Shifei, Jia Hongjie, Zhang Liwei, *et al*. Research of semi-supervised spectral clustering algorithm based on pairwise constraints [J]. Neural Computing and Applications,2014,24(1):211-219.

[12] Lu Canyi, Yan Shuicheng, Lin Zhouchen. Convex sparse spectral clustering: single-view to multi-view[J]. IEEE Trans on Image Processing,2016,25(6):2833-2843.

[13] 赵凤焦,李成,刘汉强,等. 半监督谱聚类特征向量选择算法[J]. 模式识别与人工智能,2011,24(1):48-56.

[14] Sourati J, Erdogmus D, Dy J G, *et al*. Accelerated learning-based interactive image segmentation using pairwise constraints[J]. IEEE Trans on Image Processing,2014,23(7):3057-3070.

[15] Cao Jiangzhong, Chen Pei, Dai Qingyun, *et al*. Local information-based fast approximate spectral clustering[J]. Pattern Recognition Letters,2014,38(1):63-69.

[16] Kang U, Meeder B, Papalexakis E E, *et al*. HEigen: spectral analysis for billion-scale graphs[J]. IEEE Trans on Knowledge and Data Engineering,2014,26(2):350-362.

[17] Rafailidis D, Constantinou E, Manolopou-los Y. Scalable spectral clustering with weighted PageRank[C]//Proc of International Conference on Model and Data Engineering. [S. l.]: Springer International Publishing,2014:289-300.

[18] Chen Xinlei, Cai Deng. Large scale spectral clustering with landmark-based representation[C]//Proc of the 25th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press,2011:313-318.

[19] Eubank R L. Applied nonparametric regression[J]. Technometrics, 1991,35(2):225-226.

[20] Gabow H N. Path-based depth-first search for strong and biconnected components[J]. Information Processing Letters, 2000, 74(3): 107-114.

[21] Drake J, Hamerly G. Accelerated K-means with adaptive distance bounds[C]//Proc of the 5th NIPS Workshop on Optimization for Machine Learning. 2012.

[22] MATLAB codes and datasets for feature learning[EB/OL]. http://www.cad.zju.edu.cn/home/dengcai/Data/data.html.

[23] Datasets for "the elements of statistical learning" [EB/OL]. http://www.stat.stanford.edu/~tibs/ElemStatLearn/data.html.

[24] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: analysis and an algorithm[C]//Proc of the 14th International Conference on Neural Information Processing Systems Natural Synthetic. Cambridge, MA: MIT Press,2002:849-856.

[25] Chen Weifu, Feng Guocan. Spectral clustering: a semi-supervised approach[J]. Neurocomputing,2012,77(1):229-242.

[26] Fowlkes C, Belongie S, Fan C, *et al*. Spectral grouping using the Nyström method[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2004,26(2):214-225.

[27] Dhillon I S, Guan Yuqiang, Kulis B. Weighted graph cuts without eigenvectors a multilevel approach[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2007,29(11):1944-1957.