智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

加权PageRank改进地标表示的自编码谱聚类算法

储德润， 周治平

引用本文:
储德润, 周治平. 加权PageRank改进地标表示的自编码谱聚类算法[J]. 智能系统学报, 2020, 15(2): 302–309.
CHU Derun, ZHOU Zhiping. An autoencoder spectral clustering algorithm for improving landmark representation by weighted PageRank[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(2): 302–309.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201904021

您可能感兴趣的其他文章

公理化模糊共享近邻自适应谱聚类算法
Shared nearest neighbor adaptive spectral clustering algorithm based on axiomatic fuzzy set theory
智能系统学报. 2019, 14(5): 897–904 https://dx.doi.org/10.11992/tis.201810002

结合谱聚类的标记分布学习
Label distribution learning based on spectral clustering
智能系统学报. 2019, 14(5): 966–973 https://dx.doi.org/10.11992/tis.201809019

基于PageRank的主动学习算法
Active learning through PageRank
智能系统学报. 2019, 14(3): 551–559 https://dx.doi.org/10.11992/tis.201804052

基于加权聚类集成的标签传播算法
Label propagation algorithm based on weighted clustering ensemble
智能系统学报. 2018, 13(6): 994–998 https://dx.doi.org/10.11992/tis.201806011

结合稀疏表示与约束传递的半监督谱聚类算法
A semi-supervised spectral clustering algorithm combined with sparse representation and constraint propagation
智能系统学报. 2018, 13(5): 855–862 https://dx.doi.org/10.11992/tis.201703013

稀疏样本自表达子空间聚类算法
Sparse sample self-representation for subspace clustering
智能系统学报. 2016, 11(5): 696–702 https://dx.doi.org/10.11992/tis.201601005

微信公众平台

关注微信公众号，获取更多资讯信息

DOI: 10.11992/tis.201904021
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190828.1756.008.html

加权 PageRank 改进地标表示的自编码谱聚类算法

储德润, 周治平
(江南大学 物联网技术应用教育部工程研究中心, 江苏 无锡 214122)

摘 要: 针对传统谱聚类算法在处理大规模数据集时, 聚类精度低并且存在相似度矩阵存储开销大和拉普拉斯矩阵特征分解计算复杂度高的问题。提出了一种加权 PageRank 改进地标表示的自编码谱聚类算法, 首先选取数据亲和图中权重最高的节点作为地标点, 以选定的地标点与其他数据点之间的相似关系来逼近相似度矩阵作为叠加自动编码器的输入。然后利用聚类损失同时更新自动编码器和聚类中心的参数, 从而实现可扩展和精确的聚类。实验表明, 在几种典型的数据集上, 所提算法与地标点谱聚类算法和深度谱聚类算法相比具有更好的聚类性能。

关键词: 机器学习; 数据挖掘; 聚类分析; 地标点聚类; 谱聚类; 加权 PageRank; 自动编码器; 聚类损失
中图分类号: TP18 **文献标志码**: A **文章编号**: 1673−4785(2020)02−0302−08

中文引用格式: 储德润, 周治平. 加权 PageRank 改进地标表示的自编码谱聚类算法 [J]. 智能系统学报, 2020, 15(2): 302–309.
英文引用格式: CHU Derun, ZHOU Zhiping. An autoencoder spectral clustering algorithm for improving landmark representation by weighted PageRank[J]. CAAI transactions on intelligent systems, 2020, 15(2): 302–309.

An autoencoder spectral clustering algorithm for improving landmark representation by weighted PageRank

CHU Derun, ZHOU Zhiping
(Engineering Research Center of Internet of Things Technology Applications Ministry of Education, Jiangnan University, Wuxi 214122, China)

Abstract: Several problems, such as low clustering precision, large memory overhead of the similarity matrix, and high computational complexity of the Laplace matrix eigenvalue decomposition, are encountered when using the traditional spectral clustering algorithm to deal with large-scale datasets. To solve these problems, an autoencoder spectral clustering algorithm for improving landmark representation by weighted PageRank is proposed in this study. First, the nodes with the highest weight in the data affinity graph were selected as the landmark points. The similarity matrix was approximated by the similarity relation between the selected ground punctuation points and other data points. The result was further used as the input of the superimposed automatic encoder. At the same time, the parameters of the automatic encoder and cluster center were updated simultaneously using clustering loss. Thus, extensible and accurate clustering can be achieved. The experimental results show that the proposed autoencoder spectral clustering algorithm has better clustering performance than the landmark and depth spectral clustering algorithms on several typical datasets.
Keywords: machine learning; data mining; cluster analysis; landmark spectral clustering; spectral clustering; weighted pagerank; autoencoder; clustering loss

聚类是数据挖掘、模式识别等许多研究领域中的基本问题之一, 聚类分析的目的是将给定的数据集划分为紧凑的聚类, 使聚类内的数据对象比不同的聚类中的数据对象更加相似。其中谱聚类可以适应更广泛的几何形状, 并检测非凸模式和线性不可分离的簇, 而不存在局部最优问题,

收稿日期: 2019−04−09. **网络出版日期**: 2019−08−29.
通信作者: 储德润. E-mail: CDR0727@163.com.

被广泛用于数据挖掘和图像分割。

谱聚类算法虽然性能良好, 但由于其计算复杂度高, 很难适用于大规模的数据集。近些年来, 研究人员为了加快谱聚类算法的计算速度和降低谱聚类算法对于大规模数据集的计算复杂度问题, 探索研究了一系列的方法来提高谱聚类算法对于大规模数据集的可拓展性。Li 等[1] 提出了一种精确的、可扩展的 Nystrom 算法, 使用最近的随机低秩矩阵逼近算法对内部子矩阵进行近似的 SVD, 降低随机采样引起的不稳定性和采样误差。赵晓晓等[2] 结合稀疏表示和约束传递, 提出一种结合稀疏表示和约束传递的半监督谱聚类算法, 进一步提高了聚类准确率。Ding 等[3] 提出一种基于 hmrf 模型的半监督近似谱聚类算法, 利用 hmrf 半监督聚类与近似加权核 k 均值之间的数学联系, 利用近似加权核 k-均值计算 hmrf 谱聚类的最优聚类结果。He 等[4] 提出了一种有效的大规模数据谱聚类方法, 谱聚类的复杂度比现有的 Nystrom 近似在大规模数据上的复杂度要低。显著加快谱聚类中的特征向量逼近和效益预测速度。林大华等[5] 针对现有子空间聚类算法没有利用样本自表达和稀疏相似度矩阵, 提出了一种新的稀疏样本自表达子空间聚类方法, 所获得的相似度矩阵具有良好的子空间结构和鲁棒性。Yang 等[6] 提出了一种新的基于层次二部图的谱聚类方法, 采用无参数而有效的邻域分配策略构造相似矩阵, 避免了调整相似性矩阵的需要, 大大降低了算法的计算复杂度。虽然众多改进的谱聚类算法在一定程度上减小了谱聚类的时间复杂度, 但仍然要对拉普拉斯矩阵进行特征分解, 对于大规模的数据集来说内存消耗过大, 具有巨大的空间复杂度。最近, 深度学习[7] 在计算机视觉、语音识别和自然语言处理中得到了广泛的研究。一些深度学习的方法已经被提出来用于数据聚类[8]。Tian 等[9] 利用叠加稀疏自编码器得到原始图的非线性嵌入, 然后对嵌入表示进行 k-均值聚类, 得到聚类结果, 用叠加的自动编码器代替特征分解, 有效地降低了计算复杂度。Shao 等[10] 提出了一种新的快速图谱聚类深度线性编码方案, 该方法同时学习了线性变换和编码, 并利用深层结构进一步细化了识别特征。Song 等[11] 提出了一种新的基于自编码网络的聚类方法, 通过设计数据与聚类中心之间距离的约束, 得到了一种更适合于聚类的稳定而紧凑的表示形式。这种深层次的结构可以学习到强大的非线性映射, 数据可以很好地在变换后的空间中进行分割。上述基于深度自编码的聚类算法可以提高大规模应用的效率, 但它们都需要使用整个数据集的相似度矩阵作为自动编码器的输入, 保存了所有数据点的相似性, 由于内存消耗大, 不适用于大规模数据集。

本文采用加权 PageRank 算法, 选取数据亲和图中最具代表的点作为地标点。以选定的地标点与其他数据点之间的相似关系来逼近相似度矩阵作为叠加自动编码器的输入, 通过采样几个数据点来代替所有数据点。用叠加式自动编码器代替拉普拉斯矩阵的特征分解, 然后进行 k-means 聚类。同时进一步改进聚类和嵌入表示, 通过迭代优化基于 Kullback Leibler(KL) 散度的聚类损失, 将聚类和表示联合更新, 从而能够获得更强大的表示和更精确的聚类结果。

1　相关算法理论

1.1　基于加权 PageRank 算法的地标选择

PageRank 算法[12] 是使用最广泛的页面排序算法之一。但是原始 PageRank 算法也存在一些问题, 假设两个节点 u 和 c 彼此指向但没有指向其他节点, 并且存在第 3 个节点 w 指向其中一个节点。这个循环将累积秩, 但不会将任何秩分配到前两个节点, 因为没有输出链路。为了处理这个问题, 通过一个迭代过程来近似节点 u 的 PageRank 值 PR (u)。则有式(1):

$$\mathrm{PR}(u)=(1-d)+d\sum_{v\in B(u)}\frac{\mathrm{PR}(v)}{N_v} \tag{1}$$

其中, d 是一个阻尼因子, 通常设置为 0.85[12]。

加权 PageRank 算法在文献 [13] 中提出, 加权 PageRank 算法根据节点的重要性为节点分配排序值。这种重要性是根据传入链路和传出链路的权重来分配的, 其中, 链路表示各自的基于内容的关系, 分别以 $w^{\mathrm{in}}_{(a,b)}$ 和 $w^{\mathrm{out}}_{(a,b)}$ 表示。$w^{\mathrm{in}}_{(a,b)}$ 是 (a,b) 的传入链路权重, 它是根据到节点 b 的传入链路数和到节点 a 的所有参考节点的传入链路数来计算的, 其公式为

$$w^{\mathrm{in}}_{(a,b)}=\frac{i_b}{\sum_{c\in R_a}i_c} \tag{2}$$

其中, i_b 是节点 b 的传入链路数; i_c 是节点 c 的传入链路数; R_a 是节点 a 的参考节点集 (基于内容的最近邻域)。$w^{\mathrm{out}}_{(a,b)}$ 是 (a,b) 的传出链路权重, 它是根据节点 a 的所有参考节点的传出链路数计算的, 其公式为

$$w^{\mathrm{out}}_{(a,b)}=\frac{o_b}{\sum_{c\in R_a}o_c} \tag{3}$$

其中，o_b 是节点 b 的传出链路数；o_c 是节点 c 的传出链路数。

wpr(b) 的计算需要迭代过程来调整近似到理论真值，wpr 表示节点的加权 PageRank 权重，所有节点的 wpr 初始值初始化为 $1/n$，n 为节点数。在每个迭代中，每个节点 b 的 wpr(b) 计算如下：

$$\mathrm{wpr}(b)=(1-d)\mathrm{wpr}(a)+d\sum_{a\in R(b)}\mathrm{wpr}(a)w^{\mathrm{in}}_{(a,b)}w^{\mathrm{out}}_{(a,b)} \tag{4}$$

在每次迭代中，所有节点的 wpr 值都是基于式 (4) 减小的[13]。对于所有节点当满足以下的停止准则时，迭代过程停止：

$$\frac{\mathrm{wpr}(b)_{\mathrm{iter}-1}-\mathrm{wpr}(b)_{\mathrm{iter}}}{\mathrm{wpr}(b)_{\mathrm{iter}}}\leqslant\beta \tag{5}$$

其中，分数是前一次迭代与当前迭代之间的归一化差值，β 是一个预定义的停止阈值，这里设置 $\beta=10^{-3}$。最后当式 (5) 成立时，选择 wpr 最高的 p 节点作为地标点。

2　改进地标表示的自编码谱聚类算法

2.1　构造新的拉普拉斯矩阵

在文献 [14-15] 中，提出了基于地标点的谱聚类算法来加速谱聚类，在给定 n 个数据点的数据集的情况下，选择 $p\ll n$ 个具有代表性的数据点作为地标的特征点，将原始数据点表示为这些地标的线性稀疏组合，然后利用基于地标的表示来高效地计算数据的谱嵌入，并通过理论分析证明其性能优于 Nyströom 和快速谱聚类方法，有效地利用稀疏表示矩阵逼近了整个图的相似度矩阵。它表示选定的 p 个地标点与 n 个数据点之间的成对相似性。

给定数据矩阵 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2\cdots\boldsymbol{x}_n\}\in\mathbf{R}^{n\times d}$，它可以近似为 $\boldsymbol{X}\approx\boldsymbol{U}\boldsymbol{Z}$。$\boldsymbol{U}$ 每一列都可以看作捕捉数据中较高层次特征的基向量。选定的地标点可以看作是基向量。假设对于任何数据，已经有了地标矩阵 $\boldsymbol{U}$。对于任意点 $\boldsymbol{x}_i$，它的近似 $\hat{\boldsymbol{x}}_i$ 计算为

$$\hat{\boldsymbol{x}}_i=\sum_{j=1}^{p}\boldsymbol{z}_{ji}\boldsymbol{u}_j \tag{6}$$

根据稀疏编码策略，假设当 $\boldsymbol{x}_i$ 更接近 $\boldsymbol{u}_j$ 时，矩阵 $\boldsymbol{Z}$ 的第 j 个元素应该更大。为了强调这一假设，创建 $\boldsymbol{Z}$ 亲和稀疏矩阵的稀疏表示，通过选择 $r<p$ 个最近的地标点，代替 p 个地标点 $(\boldsymbol{U}\in\mathbf{R}^{d\times p})$，例如，如果 $\boldsymbol{u}_j$ 不是 r 的最近的地标之一，则 $\boldsymbol{z}_{ji}$ 的值为 0，生成稀疏表示矩阵 $\boldsymbol{Z}$。设 $\boldsymbol{U}_{\langle i\rangle}\in\mathbf{R}^{d\times r}$ 表示 $\boldsymbol{U}$ 的子矩阵，由 $\boldsymbol{x}_i$ 最近的 r 个地标点组成，然后计算出每个 $\boldsymbol{z}_{ii}$ 元素如下：

$$\boldsymbol{z}_{ji}=\frac{\Phi(\boldsymbol{x}_i,\mathrm{u}_j)}{\sum\limits_{j'\in U_{\langle i\rangle}}\Phi(\boldsymbol{x}_i,\mathrm{u}_j)},\quad i\in 1,2,\cdots,n,\mathrm{and},j\in U_{\langle i\rangle} \tag{7}$$

其中，$\Phi(\cdot)$ 是一个具有尺度参数 σ 的高斯核函数，$\Phi(\boldsymbol{x}_{\mathrm{i}},\boldsymbol{u}_{\mathrm{j}})=\exp\left(-\left\|\boldsymbol{x}_{\mathrm{i}}-\boldsymbol{u}_{\mathrm{j}}\right\|/2\sigma^2\right)$ 是最常用的高斯核函数之一，其中，σ 控制着每个数据点的局部邻域。根据核尺度参数 σ 的自调整策略，在本文所提算法中，设置 $\sigma^2=\sigma_i\sigma_j,\forall\Phi(\boldsymbol{x}_i,\boldsymbol{u}_j)$，其中 σ_i 是 i 的邻居点的平均距离，也是 $\boldsymbol{Z}\in\mathbf{R}^{p\times n}$ 的稀疏表示中的非零元素。对于 $\boldsymbol{W}$ 亲和矩阵，认为 $\boldsymbol{W}=\hat{\boldsymbol{Z}}^{\mathrm{T}}\hat{\boldsymbol{Z}}$，其中 $\hat{\boldsymbol{Z}}=\boldsymbol{D}^{-1/2}\boldsymbol{Z}$ 是 $\boldsymbol{D}=\sum_j\boldsymbol{Z}_{ji}$ 度矩阵的归一化 $\boldsymbol{Z}$。

如果计算所有数据点的相似性矩阵并将其作为堆叠式自动编码器的输入，对于大型数据集来说空间消耗非常高。代替计算给定数据与所有其他数据点的相似性，本文算法只考虑选择的地标点与其他数据点之间的相似性。在这里重写了如下拉普拉斯矩阵的形式：

$$\boldsymbol{L}_{\mathrm{new}}=\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} \tag{8}$$

在这里使用 $\boldsymbol{S}$ 作为叠加自编码器的输入，从前文可以知道归一化相似矩阵可以用稀疏表示矩阵 $\boldsymbol{Z}$，为了进一步降低计算复杂度，在这里使用一种简单的方法计算度矩阵 $\boldsymbol{D}_2$，矩阵的主要对角线元素一般计算如下：

$$\boldsymbol{d}_{2ii}=\sum\nolimits_i\boldsymbol{w}_{ij}=\sum\nolimits_{\mathrm{i}}\hat{\boldsymbol{z}}_i^{\mathrm{T}}\hat{\boldsymbol{z}}_j \tag{9}$$

对式 (11) 进行运算演化可以得到：

$$\boldsymbol{d}_{2ii}=\sum_{j=1}^{n}\boldsymbol{w}_{ij}=\sum_{j=1}^{n}\hat{\boldsymbol{z}}_i^{\mathrm{T}}\hat{\boldsymbol{z}}_j=\hat{\boldsymbol{z}}_i^{\mathrm{T}}\sum_{j=1}^{n}\hat{\boldsymbol{z}}_j=\hat{\boldsymbol{z}}_i^{\mathrm{T}}\hat{\boldsymbol{Z}}^s \tag{10}$$

其中，$\hat{\boldsymbol{Z}}^s$ 是一个 $p\times1$ 的向量，第 k 个元素是 $\hat{\boldsymbol{Z}}$ 中第 k 行元素的和。

$$\boldsymbol{D}_2=\mathrm{diag}\left(\hat{\boldsymbol{Z}}^{\mathrm{T}}\hat{\boldsymbol{Z}}^s\right) \tag{11}$$

在这里新的拉普拉斯矩阵的构造如下：

$$\boldsymbol{L}_{\mathrm{new}}=\boldsymbol{D}_2^{-1/2}\boldsymbol{W}\boldsymbol{D}_2^{-1/2}=\boldsymbol{D}_2^{-1/2}\hat{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{Z}\hat{\boldsymbol{D}}_2^{-1/2} \tag{12}$$

$$\boldsymbol{S}=\boldsymbol{D}_2^{-1/2}\hat{\boldsymbol{Z}}^{\mathrm{T}} \tag{13}$$

最后，将相似度矩阵 $\boldsymbol{S}$ 输入到自动编码器中，最后运行 k-means 聚类，从而得到聚类结果。

2.2　自编码器与聚类优化

自动编码器是一种用于有效编码和降维的无监督学习的人工神经网络。自动编码器的目的是使原始输入 x_i 和新的嵌入表示 y_i 的重构输出 m_i 之间的重构损失最小化[16-19]。重构损失定义如下：

$$L_r=\sum_{i=1}^{N}l(x_i,g(f(x_i;\theta_1);\theta_2)) \tag{14}$$

其中，$\{\theta_1,\theta_2\}=\{h_1,b_1,h_2,b_2\}$ 是自动编码器的参数。

虽然用自编码器配合得到聚类结果，有效地降低了计算复杂度，减小了空间复杂度。但是，由于需要计算和保存所有数据点的相似关系，因此内存消耗进一步增加。学习表示与聚类分离，

嵌入表示是不可靠的, 对聚类有负面影响。为了将聚类和学习表示联合更新, 并且能够获得更强大的表示和更精确的聚类结果, 将这两个过程集成到一个具有基于 KL 散度的聚类损失函数的单一框架中, 并迭代地优化了自动编码器和聚类中心的参数。聚类损失被定义为分布在 P 和 Q 之间的 KL 散度, 其中 Q 是由 t-SNE 测量的软标签的分布, P 是由 Q 导出的目标分布。聚类损失可以用来同时更新堆叠式自动编码器和聚类中心的参数, 如式(15)所示:

$$L_c = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{15}$$

其中, q_{ij} 是由 t-SNE 测量的嵌入表示 y_i 和聚类中心 c_j 的相似性。

$$q_{ij} = \frac{\left(1+\|y_i - c_j\|^2\right)^{-1}}{\sum_j \left(1+\|y_i - c_j\|^2\right)^{-1}} \tag{16}$$

p_{ij} 是由 q_{ij} 确定的目标分布:

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j \left(q_{ij}^2 / \sum_i q_{ij}\right)} \tag{17}$$

因为, p_{ij} 是由 q_{ij} 确定的目标分布, 所以 L_c 的最小化可以看作是自我训练的一种形式。重构损失保证了嵌入空间能够保持数据的局部结构, 因此将重构损失和聚类损失作为优化参数的目标函数。

$$L = L_c + \lambda L_r \tag{18}$$

其中, $\lambda(\lambda \geqslant 0)$是一个控制嵌入空间失真程度的因子。

利用最小批量随机梯度下降可以得到自动编码器的聚类中心和参数的优化, 计算出 L_c 关于嵌入表示 y_i 和聚类中心 c_j 的梯度。

$$\frac{\partial L_c}{\partial y_i} = 2\sum_{j=1}^{k} \left(1+\|y_i - c_j\|^2\right)^{-1} (p_{ij} - q_{ij})(y_i - c_j) \tag{19}$$

$$\frac{\partial L_c}{\partial c_j} = 2\sum_{i=1}^{k} \left(1+\|y_i - c_j\|^2\right)^{-1} (q_{ij} - p_{ij})(y_i - c_j) \tag{20}$$

$\partial L_c/\partial y_i$ 可以传递到叠加的自动编码器, 并用于反向传播以计算参数 $\partial L_c/\partial M_1$、$\partial L_c/\partial M_2$ 的梯度。编码器权重的更新公式如下:

$$M_1 = M_1 - \frac{\eta}{m} \sum_{i=1}^{m} \left(\frac{\partial L_c}{\partial M_1} + \lambda \frac{\partial L_r}{\partial M_1}\right) \tag{21}$$

译码器的权重公式更新如下:

$$M_2 = M_2 - \frac{\eta}{m} \sum_{i=1}^{m} \left(\frac{\partial L_r}{\partial M_2}\right) \tag{22}$$

聚类中心 c_j 的更新公式如下:

$$c_j = c_j - \frac{\eta}{m} \sum_{i=1}^{m} \frac{\partial L_c}{\partial c_j} \tag{23}$$

式中: m 是小批量的样本数; η 是学习速率。

2.3　所提算法流程

加权 PageRank 改进地标表示的自编码谱聚类算法步骤:

输入　数据集 X, 目标聚类数 k, 地标数目 p, 最近邻地标数目 r, 停止阈值 β, 迭代数目 t, 停止阈值 h。

输出　实际聚类数目 y, 聚类损失 L_c, 重构损失 L_r。

1) 根据式 (2)~(4) 计算 wpr(b) 的值, 迭代计算 wpr 直到满足式 (5), 根据最高的 wpr 值选择 p 个地标点。

2) 通过式 (7) 计算稀疏表示矩阵 $\boldsymbol{Z}$, 所选地标点 p 和具有 r 个最近邻地标点的矩阵 $\boldsymbol{U}$。

3) 通过式 (12) 计算新的拉普拉斯矩阵 $\boldsymbol{L}_{new}$。

4) 使用 $\boldsymbol{D}_2^{-1/2}\hat{\boldsymbol{Z}}^T$ 作为自动编码器的输入。接着对叠加式自动编码器进行训练, 得到嵌入表示 h_i。

5) 在最后一个隐藏层的嵌入表示 h_i 上进行 k-means 聚类, 以初始化聚类中心 c_j。根据式 (22)、(23) 更新自动编码器的参数和聚类中心 c_j。

2.4　算法复杂度分析

假设从 n 个数据点中选择 p 个地标, 则所提算法采用加权 PageRank 选择地标点, 该步骤的时间复杂度为 $O(tpn)$, 其中 t 为迭代次数。为了构造新的相似矩阵, 只需根据方程计算稀疏表示矩阵 $\boldsymbol{Z}$,这一步的时间复杂度为 $O(pn)$。堆叠式自动编码器和优化算法的时间复杂度为 $O(nD^2+ndk)$, k 是簇数, D 是隐藏层的最大单元数, d 是嵌入层的维数。通常 $k \leqslant d \leqslant D$, 所以时间复杂度可写为 $O(tpn+pn+nD^2)$。基于地标点的谱聚类算法[15] 的时间复杂度为 $O(tpn+pn+p^2n+p^3)$, 又 $p \ll n$, 则 $p^3 \ll p^2n$, 故 $p^2n+p^3 \approx p^2n$, 所以基于地标点谱聚类算法的时间复杂度可写为 $O(tpn+pn+p^2n)$, 其中 p 与 D 的取值相近且远小于 n, 因此, 可推断本文方法与基于地标点谱聚类算法时间复杂度相当, 但由于算法改进中通过采样少数几个数据点来推断数据集的全局特征, 空间复杂度有所降低。

3　实验与结果分析

3.1　实验环境及性能指标

在本文算法实验中, 选取了几个数据量较大的数据集进行实验, 它们分别为手写数字数据集 USPS、Pendigits、MNIST、英文字母表数据集 LetterRec 和 UCI 数据库中的数据集 CoverType, 表 1 给出了数据集的详细特征。本文算法实验是在 Pyt-

hon 3.6, 计算机的硬件配置为 Intel Core i7-7700 CPU 3.60 GHz、16 GB 内存的平台下进行。

表 1　实验数据集的特征

Table 1　Characteristics of experimental datasets

数据集	数据个数	类数	维数
USPS	9 280	10	256
Pendigits	10 992	10	16
MNIST	70 000	10	784
LetterRec	20 000	26	16
Covtype	581 012	7	54

总体而言, 提出的方法在 ACC 和 NMI 上均优于其他所列的对比方法。

本文算法与文献 [1] 中基于 nystrom 的方法, 文献 [9] 中基于叠加稀疏自动编码器聚类方法 GraEn, 文献 [10] 中基于深度学习编码的快速图聚类方法 DLC, 文献 [14-15] 基于地标点采样的快速谱聚类方法 LSC-R 和 LSC-K, 文献 [17] 中深层嵌入聚类方法 DEC 和文献 [19] 基于改进可拓展的 nystrom 方法 RSNy 这些算法进行实验对比。

本文实验的具体的实验参数设置如下: 基于地标点采样的快速谱聚类方法 LSC 有两个参数需要设置, 分别为地标点 p 的数目, 最近邻地标点 r 的数目, 在这里, 我们统一设置 $p=1\,000$ 和 $r=5$。在加权 PageRank 算法中, 我们设置收敛阈值 $\beta=10^{-3}$。编码器被设置为维数为 $p-500-500-2\,000-10$ 的多完全连接层, 适用于所有的数据集。解码器是维数为 $10-2\,000-500-500-p$ 的编码器的镜像。并且, 将最小批量设置为 256 个, 初始学习速率 η 设为 0.1。停止阈值设为 $h=10^{-3}$, 同时将控制嵌入表示空间失真程度的 λ 值设为 0.1。

3.2　实验结果分析

为了验证本文算法的性能, 采用文献 [20] 中聚类准确率 ACC 和归一化互信息 NMI 两种聚类度量指标来对聚类结果进行评估和分析比较。ACC 和 NMI 的取值范围都在 0～1 之间, 较高的值表示较好的聚类结果。

根据表 2 可以看出, 基于聚类准确率 (ACC) 的实验结果表明, 提出的加权 PageRank 改进地标表示的自编码谱聚类算法在 USPS、Pendigits、MNIST、LetterRec、Covtype 数据集上相较于基于 nystrom 的方法、基于地标点采样的快速谱聚类方法、基于深度学习自编码的快速图聚类方法都取得了最好的结果, 所提方法在聚类准确率方面都优于以上所提其他方法。特别是在 MNIST 数据集上, GraEn、DLC、DEC 和提出的方法这些基于深度学习自编码的谱聚类方法的聚类结果均明显高于传统 nystrom 的方法、RSNy 方法和基于地标点的方法 (LSC-R, LSC-K), 其中相较于传统 nystrom 的方法, 本文所提的方法在 MNIST 数据集上聚类准确率方面提高了 35.91%。相较于基于地标点的快速谱聚类方法 LSC-R, 提高了 30.71%。可以看出, 所提方法相较于这些方法聚类精度的提高是巨大的。相较于其他几种快速谱聚类方法, 所提方法在 MNIST 数据集上分别提高了 18.73% 和 16.91%, 总体而言也是相当可观。同时, 从表中可以看出基于地标点的方法在 USPS、Pendigits、LetterRec 和 Covtype 数据集上的聚类准确率比基于自编码的方法在一定程度上还要好。但是所提的方法相较于这些方法而言, 实验得到的聚类准确率方面均是最优的。

表 2　不同数据集的聚类准确率 (ACC) 的比较

Table 2　Comparison of clustering accuracy of different data sets

算法	USPS	Pendigits	MNIST	LetterRec	Covtype
Nystrom	0.681 4	0.739 4	0.537 0	0.301 1	0.223 1
RSNy	0.706 1	0.754 7	0.708 8	0.294 9	0.224 2
LSC-R	0.752 4	0.790 4	0.589 0	0.292 2	0.247 5
LSC-K	0.775 3	0.819 9	0.727 0	0.303 3	0.255 0
GraEn	0.693 1	0.736 4	0.818 2	0.287 2	0.221 8
DLC	0.726 3	0.769 2	0.836 7	0.295 7	0.243 1
DEC	0.749 9	0.790 4	0.865 1	0.297 5	0.249 7
本文算法	**0.789 7**	**0.846 1**	**0.896 1**	**0.309 3**	**0.268 2**

同时, 根据表 3 可以看出所提的方法与其他几种方法相比在 NMI 上均得到了提高, 并且相较于其他几种方法是最优的。同样在 MNIST 数据集上, 所提的方法相较于传统 nystrom 方法和基于地标的快速谱聚类方法都取得了大幅的提高。相对于传统 nystrom 方法和基于地标点的谱聚类方法 LSC-R, 所提的方法在 NMI 上分别提高了 40.31% 和 29.22%, 在其他数据集上, 所提的方法相较于其他方法而言也均有不同幅度的提高。从表 2 和表 3 中均可以看出, 对于最原始的基于自编码的方法 GraEn。在 MNIST 数据集上, 它的 ACC 和 NMI 相较于传统 nystrom 方法和基于地标点的谱聚类方法而言均有所提高。与 DEC 方法相比, DEC 方法的 ACC 和 NMI 分别提高了 4.69% 和 8.96%, 结果表明, 联合更新聚类中心和学习嵌入表示可以提高聚类结果, 这也适用于其他 4 个数据集。尽管在 USPS 数据集上, 基于自编码的方法的聚类精度没有基于地标点方法的聚类精度好, 但是在其他几个高维大规模数据集上基于自编码的方法在聚类精度方面要更好, 说明基于自编码的方法和所提的方法更加适用于高维的大规模数据。

表 3　不同数据集的归一化互信息 (NMI) 的比较

Table 3　Comparison of normalized mutual information (NMI) for different data sets

算法	USPS	Pendigits	MNIST	LetterRec	Covertype
Nystrom	0.634 0	0.668 1	0.480 2	0.391 2	0.074 2
RSNy	0.654 9	0.681 3	0.625 5	0.371 8	0.074 4
LSC-R	0.754 1	0.776 7	0.591 1	0.373 4	0.083 1
LSC-K	0.791 5	0.780 8	0.722 2	0.396 3	0.090 2
GraEn	0.662 0	0.655 5	0.747 3	0.371 2	0.071 9
DLC	0.735 1	0.703 5	0.780 4	0.383 5	0.078 6
DEC	0.741 3	0.735 1	0.836 9	0.385 1	0.085 6
本文算法	**0.799 3**	**0.798 3**	**0.883 3**	**0.401 6**	**0.093 4**

总体而言, 提出的方法在 ACC 和 NMI 上均优于其他所列的对比方法。

图 1 和图 2 显示了选取的 3 个不同的数据集中, 随着地标点数量 p 从 100~1 000 变化, 基于地标点的谱聚类算法 LSC-R 和 LSC-K 与所提方法的聚类指标 ACC 和 NMI 的对比实验。从图 1 中可以看出, 随着地标点个数的增加, 3 种方法的 ACC 也随之增加, 总体均呈现增幅形式, 最后均趋于稳定, 并且所提算法在 3 个数据集上的 ACC 均高于其他两种对比算法的 ACC。

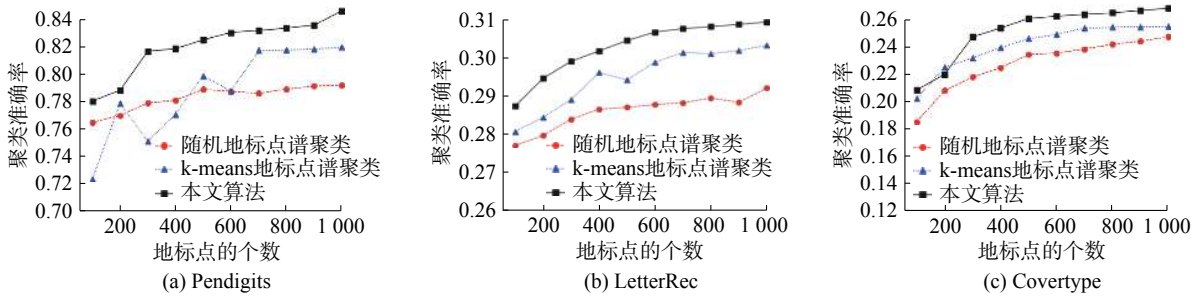

图 1　不同数据集上不同地标点的聚类准确率的比较

Fig. 1　Comparison of clustering accuracy of different punctuation points on different datasets

由图 2 可知, 所提的方法在 LetterRec 和 Covtype 数据集上在初始阶段, 随着地标点的增加 NMI 存在低于其他方法的情况, 这是因为所选 LetterRec 数据集的字符图像基于 26 种不同的字体, 26 种字体中的每一个字母都被随机扭曲。Covertype 数据集是一个从地图变量预测森林覆盖类型的数据集, 它们都主要是在荒野地区发现的, 所以覆盖类型在实际地理上是非常接近的, 相对于其他手写数字数据集而言, 这两个数据集数据特性更加复杂。并且在选取较少地标点时, 采用加权 PageRank 算法选择地标点与随机选择地标点和基于 k-menas 算法选择地标点相比并不能展现出较大的优势, 所以在选取较少地标点时聚类性能存在低于其他方法的情况, 但在总体情

况下，随着选取地标点的增多，所提算法展现出较快的增长优势和较好的聚类性能，在地标点达到 1 000 左右时趋于稳定，展现出均优于对比算法的聚类性能。

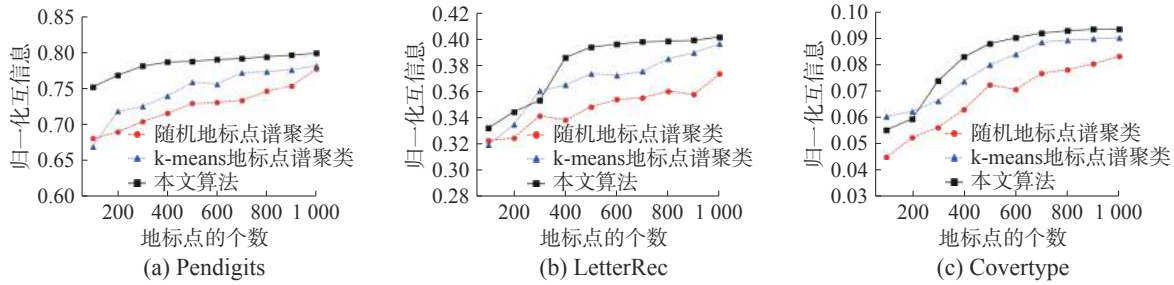

图 2　不同数据集上不同地标点的归一化互信息的比较

Fig. 2　Comparison of normalized mutual information of different punctuation points on different dataset

4　结束语

本文提出了一种加权 PageRank 改进地标表示的自编码谱聚类算法，该算法利用加权 PageRank 算法选取数据亲和图中最具代表性的点作为地标点，然后以选定的地标点与其他数据点之间的相似度矩阵作为叠加自动编码器的输入，以减少内存消耗，用叠加式自动编码器代替拉普拉斯矩阵的特征分解，降低了时间复杂度。通过迭代优化基于 KL 散度的聚类损失对聚类进行了进一步细化，结合重构损失和聚类损失考虑了数据的局部结构，同时更新了自动编码器和聚类中心的参数。证明了该方法对不同类型数据集的有效性，实验结果表明，与传统谱聚类算法、基于地标点的谱聚类算法和其他基于深度学习自编码的方法相比，在几个大规模的数据集上，提出的方法具有更好的聚类性能。在未来，致力于研究更加有效的地标点采样方法，结合优化的自动编码器从而更好的提高谱聚类的聚类性能。

参考文献：

[1] LI Mu, BI Wei, KWOK J T, et al. Large-scale nyström kernel matrix approximation using randomized SVD[J]. IEEE transactions on neural networks and learning systems, 2015, 26(1): 152–164.

[2] 赵晓晓, 周治平. 结合稀疏表示与约束传递的半监督谱聚类算法 [J]. 智能系统学报, 2018, 13(5): 855–863.
ZHAO Xiaoxiao, ZHOU Zhiping. A semi-supervised spectral clustering algorithm combined with sparse representation and constraint propagation[J]. CAAI transactions on intelligent systems, 2018, 13(5): 855–863.

[3] DING Shifei, JIA Hongjie, DU Mingjing, et al. A semi-supervised approximate spectral clustering algorithm based on HMRF model[J]. Information sciences, 2018, 429: 215–228.

[4] HE Li, RAY N, GUAN Yisheng, et al. Fast large-scale spectral clustering via explicit feature mapping[J]. IEEE transactions on cybernetics, 2019, 49(3): 1058–1071.

[5] 林大华, 杨利锋, 邓振云, 等. 稀疏样本自表达子空间聚类算法 [J]. 智能系统学报, 2016, 11(5): 696–702.
LIN Dahua, YANG Lifeng, DENG Zhenyun, et al. Sparse sample self-representation for subspace clustering[J]. CAAI transactions on intelligent systems, 2016, 11(5): 696–702.

[6] YANG Xiaojun, YU Weizhong, WANG Rong, et al. Fast spectral clustering learning with hierarchical bipartite graph for large-scale data[J/OL]. Pattern recognition letters: (2018-06-22). https://www.sciencedirect.com/science/article/abs/pii/S016786551830271X. DOI: 10.1016/J.PATREC.2018.06.024.

[7] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436–444.

[8] SHAHAM U, STANTON K, LI H, et al. SpectralNet: spectral clustering using deep neural networks[J]. arXiv:1801.01587, 2018.

[9] TIAN Fei, GAO Bin, CUI Qing, et al. Learning deep representations for graph clustering[C]//Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Québec City, Canada, 2014: 1293−1299.

[10] SHAO Ming, LI Sheng, DING Zhengming, et al. Deep linear coding for fast graph clustering[C]//Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 3798−3804.

[11] SONG Chunfeng, LIU Feng, HUANG Yongzhen, et al. Auto-encoder based data clustering[C]//Proceedings of the 18th Iberoamerican Congress on Pattern Recognition. Havana, Cuba, 2013: 117−124.

[12] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the web. SIDL−WP−1999−0120[R]. Technical Report, California: Stanford Digital Libraries, 1999.

[13] XING W, GHORBANI A. Weighted PageRank algorithm[C]//Proceedings of Second Annual Conference on Communication Networks and Services Research. Fredericton, Canada, 2004: 305−314.

[14] CHEN Xinlei, CAI Deng. Large scale spectral clustering with landmark-based representation[C]//Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. San Francisco, USA, 2011: 313−318.

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

[16] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798–1828.

[17] XIE Junyuan, GIRSHICK R B, FARHADI A. Unsupervised deep embedding for clustering analysis[C]//Proceedings of the 33rd International Conference on Machine Learning. New York, USA, 2016: 478−487.

[18] LI Mu, ZHANG Tong, CHEN Yuqiang, et al. Efficient mini-batch training for stochastic optimization[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 661−670.

[19] LI Mu, KWOK J T, LU Baoliang. Making large-scale Nyström approximation possible[C]//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 631−638.

[20] 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.

作者简介:

储德润, 硕士研究生, 主要研究方向为数据挖掘。

周治平, 教授, 博士, 主要研究方向为智能检测、网络安全, 发表学术论文 20 余篇。