智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

一种深度自监督聚类集成算法

杜航原，张晶，王文剑

引用本文:
杜航原, 张晶, 王文剑. 一种深度自监督聚类集成算法[J]. 智能系统学报, 2020, 15(6): 1113–1120.
DU Hangyuan, ZHANG Jing, WANG Wenjian. A deep self-supervised clustering ensemble algorithm[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(6): 1113–1120.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202006050

DOI: 10.11992/tis.202006050

一种深度自监督聚类集成算法

杜航原[1], 张晶[2], 王文剑[1,2]

（1. 山西大学 计算机与信息技术学院, 山西 太原 030006; 2. 山西大学 计算智能与中文信息处理教育部重点实验室, 山西 太原 030006）

摘　要: 针对聚类集成中一致性函数设计问题, 本文提出一种深度自监督聚类集成算法。该算法首先根据基聚类划分结果采用加权连通三元组算法计算样本之间的相似度矩阵, 基于相似度矩阵表达邻接关系, 将基聚类由特征空间中的数据表示变换至图数据表示; 在此基础上, 基聚类的一致性集成问题被转化为对基聚类图数据表示的图聚类问题。为此, 本文利用图神经网络构造自监督聚类集成模型, 一方面采用图自动编码器学习图的低维嵌入, 依据低维嵌入似然分布估计聚类集成的目标分布; 另一方面利用聚类集成目标对低维嵌入过程进行指导, 确保模型获得的图低维嵌入与聚类集成结果是一致最优的。在大量数据集上进行了仿真实验, 结果表明本文算法相比 HGPA、CSPA 和 MCLA 等算法可以进一步提高聚类集成结果的准确性。

关键词: 特征空间; 聚类算法; 一致性函数; 图表示; 相似性度量; 自监督学习; 图数据; 神经网络模型

中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2020)06−1113−08

中文引用格式: 杜航原, 张晶, 王文剑. 一种深度自监督聚类集成算法 [J]. 智能系统学报, 2020, 15(6): 1113–1120.

英文引用格式: DU Hangyuan, ZHANG Jing, WANG Wenjian. A deep self-supervised clustering ensemble algorithm[J]. CAAI transactions on intelligent systems, 2020, 15(6): 1113–1120.

A deep self-supervised clustering ensemble algorithm

DU Hangyuan[1], ZHANG Jing[2], WANG Wenjian[1,2]

(1. College of Computer and Information Technology, Shanxi University, Taiyuan 030006, China; 2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China)

Abstract: In this study, we propose a deep self-supervised clustering ensemble algorithm to obtain the design of a consensus function in a clustering ensemble. In this algorithm, a weighted connected-triple algorithm is applied to the cluster components for estimating the similarity matrix of the samples, based on which the adjacency relation can be determined. Thus, the cluster components can be transformed from data representation in the feature space to graph data representation. On this basis, the consistency integration problem of cluster components is transformed into a graph clustering problem for the graph data representation of cluster components. Further, a graph neural network is used to construct the self-supervised clustering ensemble model. This model uses a graph autoencoder to obtain the low-dimensional embedding of the graph, and the target distribution of the cluster ensemble can be estimated based on the likelihood distribution generated via low-dimensional embedding. The clustering ensemble guides the learning of low-dimensional embedding. The above methods ensure that the low-dimensional embedding and clustering ensemble results obtained by the model are consistent and optimal. Simulation experiments were conducted on a large number of data sets. Results show that the proposed algorithm improves the accuracy of the clustering ensemble result compared with the accuracies obtained using algorithms such as HGPA, CSPA, and MCLA.

Keywords: feature space; clustering algorithm; consistency function; graph representation; similarity measure; self-supervised learning; graphical data; neural network model

收稿日期: 2020−06−29.

基金项目: 国家自然科学基金项目 (61902227, 61673249, 61773247, U1805263); 山西省国际合作重点研发计划项目 (201903D421050); 山西省基础研究计划项目 (201901D211192); 山西省应用基础研究计划项目 (201701D121053); 山西省 1331 工程项目.

通信作者: 王文剑. E-mail: wjwang@sxu.edu.cn.

聚类分析在图像处理、机器学习、Web 搜索等众多领域得到了广泛应用, 是机器学习领域一个比较活跃且极具挑战的研究方向。其主要思想是通过计算样本间的相似度把数据集划分成若干

个簇，使得“同一个簇内的样本相似度较高，不同簇间的样本相似度较低[1]”。在实际的聚类分析任务中，由于数据类型和结构分布的复杂性与多样性，单一聚类方法的适用范围、可靠性和稳定性都受到了制约，为此，研究人员结合集成学习技术提出了聚类集成的方法[2-4]。聚类集成的目的是将数据集的多重基聚类集成为统一的、综合的最终聚类结果[5]，其过程主要包括两个阶段，首先使用不同的聚类算法或重复同一种聚类算法生成多个基聚类；其次设计有效的一致性函数对基聚类进行集成，获得最终聚类集成结果[6]。

一致性函数通常用来将基聚类的划分结果进行集成，得到一个统一的聚类结果。现有的一致性函数大体包括以下几种。1) 投票法：其基本思想是首先根据得到的基聚类对样本划分结果进行投票，然后计算每个样本被划分到各个簇中的投票比例，若样本归属于某个簇的投票比例超过一定阈值 (一般大于等于 0.5)，则将样本划分到该簇中[7-8]。2) 证据积累：其基本思想是同一个自然簇中的“样本对”在不同的基聚类中可能属于同一个簇。具体划分过程为：把每个基聚类看作是一个独立的证据，计算“样本对”被分到同一个簇中的次数，从而得到共协矩阵，然后使用基于最小生成树 (mininum spaning tree, MST) 的层次聚类算法得到最终的聚类结果[9-10]。3) 概率积累[11]：其基本思想是使用簇密度来计算基聚类中所有“样本对”间的距离，生成 p-association 矩阵，依据最高寿命标准采用 MST 合并 p-association 矩阵得到最终的聚类结果[12]。4) 超图划分：其基本思想是先将聚类集成问题转化为超图的最小切割问题，在此基础上使用基于图论的聚类算法得到最终的聚类结果[9, 13]。其中用超图表示基聚类，超边表示簇，超边的顶点表示归属于该簇的样本，Strehl 等[9] 提出 HGPA、CSPA 和 MCLA 三种基于超图的划分方法。

1　相关工作

1.1　图神经网络

Scarselli 等[14] 于 2009 年将神经网络和图嵌入模型结合提出了图神经网络 (graph neural network, GNN)，该模型对图的信息结构进行编码，将每个节点用一个低维状态向量表示，从而学习图的表征，能以半监督的方式处理各种类型的图，比如无环图、有向图、无向图等。近年来图神经网络已成为一种应用广泛的图分析方法，主要包括[15]：图递归神经网络 (graph recurrent neural networks, Graph RNNs)、图卷积网络 (graph confolutional networks, GCN)、图自动编码器 (graph autoencoders, GAE) 和图强化学习 (graph reinforcement learning, Graph RL)。其中图卷积网络为半监督方法，利用节点属性和节点标签训练模型参数，图自动编码器为无监督方法，利用自动编码器的降维技术学习图的低维表征[16]。

1.2　图自动编码器

图自动编码器 (graph autoencoders, GAE) 在稀疏自编码器 (sparse autoencoder, SAE)[17] 的启发下被提出，该模型将邻接矩阵或其变体作为节点的原始特征，将自动编码器 (autoencoder, AE) 作为降维方法来学习低维节点表征。简单来说，图自动编码器的基本思想是：首先输入图的邻接矩阵和节点的特征矩阵；然后通过编码器学习节点低维向量表示的均值和方差；最后利用解码器重构图。因此图自动编码器能很好地处理没有监督信息的图，同时学习图的低维表征。

2　图神经网络自监督聚类集成算法

本文提出了一种深度自监督聚类集成算法，该算法首先根据基聚类划分计算样本之间的相似度矩阵以表达样本之间的邻接关系，将基聚类在特征空间的表示转化为图数据表示；然后利用图自动编码器学习图的低维嵌入，由低维嵌入的似然分布监督聚类集成过程，同时通过聚类集成目标指导低维嵌入的学习过程，形成一个聚类集成的自监督优化模型，从而优化聚类集成结果，该算法如图 1 所示。

2.1　基聚类的图数据表示

本文基于已有的基聚类划分计算样本之间的相似性，将相似度矩阵作为邻接矩阵把样本数据的拓扑空间表示转化为图数据表示。以下为利用基聚类计算样本相似度矩阵的过程示例，图 2 给出了 5 个样本 $(x_1, x_2, \cdots, x_5)$ 构成的数据集以及在其上生成的基聚类集合 $\Pi=\{\pi_1, \pi_2\}$，其中 $\pi_1=\{C_1^1, C_2^1, C_3^1\}$，$\pi_2=\{C_1^2, C_2^2\}$。

首先计算相交簇之间的相似度，称为样本之间的一阶全局相似度：

$$w_{ij}=\frac{\left|x_{C_i}\cap x_{C_j}\right|}{\left|x_{C_i}\cup x_{C_j}\right|} \tag{1}$$

式中：C_i、C_j 表示不同基聚类中的任意两个簇；w_{ij} 表示簇 C_i 和簇 C_j 之间的相似度；x_{C_i} 表示簇 C_i 中的任意样本；x_{C_j} 表示簇 C_j 中的任意样本。

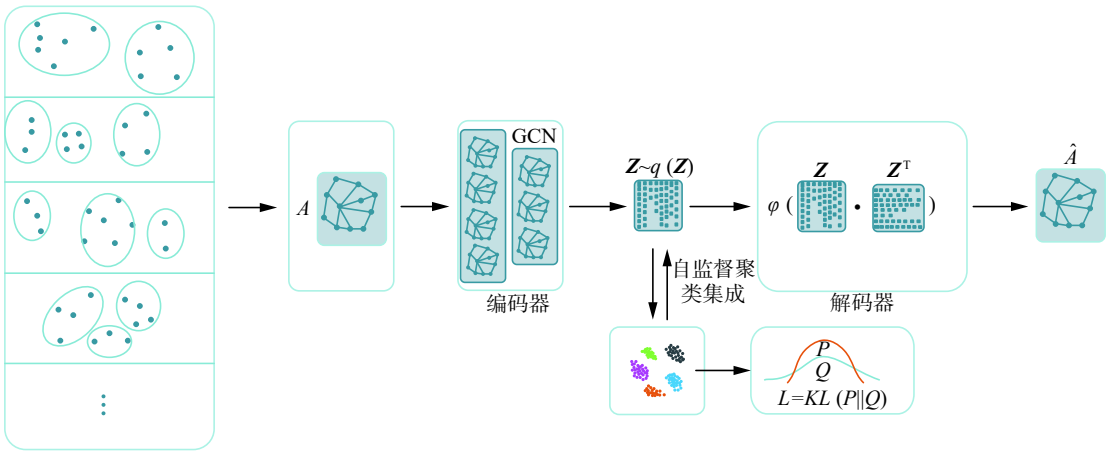


图 1　自监督聚类集成模型

Fig. 1　Self-supervised clustering ensemble model

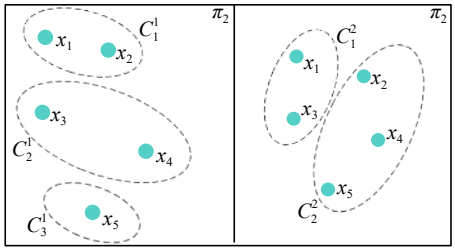


图 2　样本的基聚类结果

Fig. 2　Clustering results

图 3 为利用基聚类集合 Π 生成的样本一阶全局相似性关系图, 该图中存在两个不相交的簇能同时与第 3 个簇相连, 从而构成三元组。图 4 表示簇 C_1^1、C_3^1 和 C_2^2 之间的三元组。

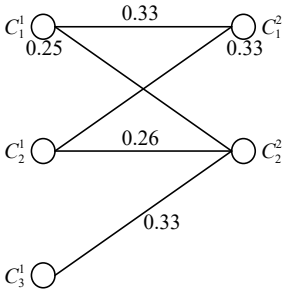


图 3　样本的一阶全局相似性关系

Fig. 3　First-order global similarity relation of samples

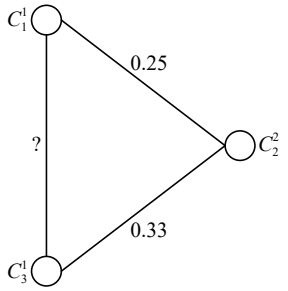


图 4　三元组关系

Fig. 4　Connected triple

式 (1) 只能计算存在交集的簇间的相似度, 忽略了不相交的簇间的相似性关系。为此, 加权连通三元组 (weighted connected-triple, WCT) 算法[18] 通过利用三元连通关系来计算不相交簇之间的相似度, 即样本间的二阶全局相似度。WCT 对簇 C_i、C_j、C_k 组成的三元组中 C_i、C_j 的之间的相似度估算值为

$$\mathrm{WCT}_{ij}^{k}=\min(w_{ik},w_{jk}) \tag{2}$$

式中: w_{ik} 为簇 C_i 和簇 C_k 之间的相似性; w_{jk} 为簇 C_j 和簇 C_k 之间的相似性。WCT 算法根据簇 C_i、C_j 之间存在的所有三元组 $(1,2,\cdots,q)$ 对 C_i、C_j 之间相似度的估算值为

$$\mathrm{WCT}_{ij}=\sum_{k=1}^{q}\mathrm{WCT}_{ij}^{k} \tag{3}$$

簇 C_i、C_j 的相似度, 即样本的二阶全局相似度计算为

$$\mathrm{sim}^{\mathrm{WCT}}(C_i,C_j)=\frac{\mathrm{WCT}_{ij}}{\mathrm{WCT}_{\max}}\times\mathrm{DC} \tag{4}$$

式 (4) 中 $\mathrm{WCT}_{\max}$ 是任意两个簇 C_p、C_q 之间 WCT_{pq} 的最大值, $\mathrm{DC}\in(0,1]$ 表示置信度。图 5 是由图 3 生成的样本间的二阶全局相似性关系图。

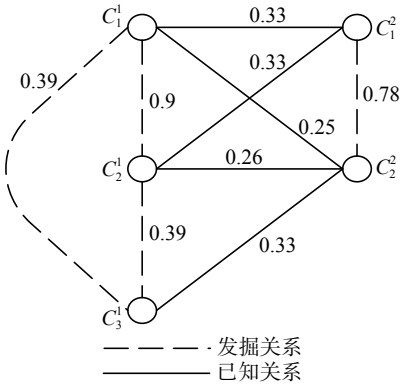


图 5　样本的二阶全局相似性关系

Fig. 5　Second-order global similarity relation of samples

计算得到样本之间的二阶全局相似度之后，依据基聚类划分结果获得样本与簇之间的二分图关系，图 6 为样本 $(x_1,x_2,\cdots,x_5)$ 与簇 C_1^1、C_2^1、C_3^1、C_1^2、C_2^2 之间的二分图关系，图中数据点与簇相连表示该数据点在基聚类中被划分到该簇中。

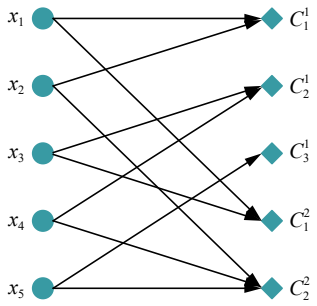


图 6　样本与簇二分关系

Fig. 6　Bipartite graph of the samples and dustering

利用式 (5) 计算样本之间的相似度：

$$\mathrm{sim}(x_i,x_j)=\begin{cases}1\ ,\quad x_i=x_j\\ \dfrac{\mathrm{DC}}{|\mathfrak{R}_{x_i}||\mathfrak{R}_{x_j}|}\displaystyle\sum_{R_{x_i}\in\mathfrak{R}_{x_i}}\sum_{R_{x_j}\in\mathfrak{R}_{x_j}}\mathrm{sim}\left(R_{x_i},R_{x_j}\right),\ \text{其他}\end{cases}\tag{5}$$

式中：$\mathrm{DC}\in(0,1]$ 表示置信度；R_{x_i}、R_{x_j} 分别表示包含样本 x_i、x_j 的簇；$\mathfrak{R}_{x_i}$、$\mathfrak{R}_{x_j}$ 分别表示包含样本 x_i、x_j 的簇的集合。

综上，得到样本之间的相似度矩阵表示样本邻接关系，从而将基聚类特征空间表示转化至图数据表示为 $G=(V,E)$，其中 $V=\{v_i\}$ 是图的节点的集合，$E=\{e_{ij}\}$ 表示两个节点之间的边。

2.2　自监督聚类集成算法

利用图自动编码器与自监督聚类集成模型[19]对基聚类的图数据表示 G 进行图聚类，即利用图自动编码器学习 G 的低维嵌入并由似然分布监督聚类集成，同时利用聚类集成目标指导低维嵌入学习过程，得到最终聚类集成结果。

本文采用的自监督聚类集成算法同时优化图自动编码器和聚类集成过程，得到总体目标函数：

$$L=L_r+\gamma L_c\tag{6}$$

式中：L_r、L_c 分别为图自编码器的重构损失和聚类过程中的聚类损失，超参数 $\gamma>0$。

在自监督聚类集成过程中，首先训练图自动编码器得到低维嵌入 $Z=(z_1,z_2,\cdots,z_m)$，图自动编码器由图卷积网络 (graph convolutional network, GCN) 编码器和简单的内积解码器组成[20]，编码器定义为

$$Z=\boldsymbol{A}'\mathbf{ReLU}(\boldsymbol{A}'\boldsymbol{X}W_0)W_1\tag{7}$$

式中：$\mathbf{ReLU}(\cdot)=\max(0,\cdot)$ 和 $\boldsymbol{A}'=\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$ 是对称归一化的邻接矩阵，$\boldsymbol{D}$ 为度矩阵，$\boldsymbol{X}$ 输入为单位矩阵。

解码器定义为

$$\hat{\boldsymbol{A}}=\mathrm{sigmoid}\left(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right)\tag{8}$$

由此得到以下重构损失：

$$L_r=\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\boldsymbol{A}}_{ij}-\boldsymbol{A}_{ij}\right\|_2^2\tag{9}$$

式中：n 为样本的数量；$\boldsymbol{A}$ 为图 G 的邻接矩阵；$\hat{\boldsymbol{A}}$ 为图 G 的重构邻接矩阵。

然后，在 Z 上执行 k-means 聚类获得初始类别中心 $\mu=(\mu_1,\mu_2,\cdots,\mu_n)$，使用学生 t 分布测量第 i 个样本的嵌入点 z_i 与第 u 个簇的类别中心 μ_u 之间相似性[21]：

$$q_{iu}=\frac{(1+\|z_i-\mu_u\|^2)^{-1}}{\sum_i\left(1+\|z_i-\mu_k\|^2\right)^{-1}}\tag{10}$$

式中：q_{iu} 表示样本 i 被分配到簇 u 的概率，$Q=[q_{iu}]$ 表示似然分布。样本 i 与簇 u 之间的相似性定义为

$$p_{iu}=\frac{q_{iu}^2/\sum_i q_{iu}}{\sum_k\left(q_{ik}^2/\sum_i q_{ik}\right)}\tag{11}$$

式中 $P=[p_{iu}]$ 表示目标分布。聚类集成的目标函数定义为

$$L_c=KL(P\|Q)=\sum_i\sum_u p_{iu}\log\frac{p_{iu}}{q_{iu}}\tag{12}$$

通过最小化分布 Q 和分布 P 之间的 KL 散度使嵌入点更接近类别中心。在这一过程中，由分布 Q 计算得到分布 P，而分布 P 监督分布 Q 的更新。

2.3　模型优化

在训练过程中，基于 L_c 关于 u 和 z 的梯度，使用随机梯度下降 (stochastic gradient descent, SGD) 对簇中心 μ 和低维嵌入 Z 进行同步更新[22]，如式 (13) 所示。目标分布 P 在训练过程中监督分布 Q 的更新，同时依赖于每次迭代时更新的分布 Q。由于目标不断变化会阻碍学习和收敛，在每次迭代中用 Q 更新 P 会导致自训练过程的不稳定性，因此本文设置了一个迭代间隔 T，每 T 次迭代更新一次 P，以避免上述可能出现的不稳定性。

$$\begin{aligned}\frac{\partial L}{\partial z_i}&=\frac{\alpha+1}{\alpha}\sum_u\left(1+\frac{\|z_i+\mu_u\|^2}{\alpha}\right)^{-1}\times(p_{iu}-q_{iu})(z_i-\mu_u)\\ \frac{\partial L}{\partial \mu_u}&=-\frac{\alpha+1}{\alpha}\sum_i\left(1+\frac{\|z_i-\mu_u\|^2}{\alpha}\right)^{-1}\times(p_{iu}-q_{iu})(z_i-\mu_u)\end{aligned}\tag{13}$$

综上所述，本文算法流程如算法 1 所示。

算法 1　自监督聚类算法

输入　基聚类 $\Pi=\{\pi_1\pi_2\cdots\pi_n\}$, 类别数 k, 迭代次数 Iter, 目标分布 P 更新间隔 T, 超参数 γ;

输出　最后聚类集成结果。

1) 使用式 (1)~(5) 计算相似度矩阵 $\boldsymbol{A}$;

2) 最小化式 (9) 得到自动编码器的低维嵌入 Z, 基于 Z 计算初始类别中心 μ;

3) for l=0 to Iter – 1 do

使用式 (10) 基于 Z 和 μ 计算分布 Q;

if l% T==0 then

使用式 (11) 基于分布 Q 计算分布 P;

end if

使用式 (12) 计算聚类集成损失函数 L_c;

最小化式 (6) 更新整个算法

end if

4) 迭代完成后得到聚类集成结果。

3　实验结果与分析

3.1　数据集

为验证算法的有效性, 本文选用 UCI 数据库中的 5 个真实数据集进行测试, 表 1 给出了实验中选用的 UCI 数据集描述。

表 1　UCI 数据集描述

Table 1　Description of the UCI data sets

数据集	样本量	特征维度	目标簇数
Wine	178	13	3
Glass	214	9	6
Wdbc	569	30	2
Yeast	1484	8	10
Segment	2310	19	7

在仿真实验中选取 HGPA[9]、CSPA[9]、MCLA[9]、基于投票的聚类集成算法[13] 和谱聚类集成算法[23] 与本文方法进行对比, HGPA、CSPA、MCLA 算法属于基于超图的聚类集成算法, 其原理是根据基聚类构造超图, 对超图进行分割; 基于投票的聚类集成算法根据基聚类的划分进行投票, 得到样本的集成结果; 谱聚类集成算法利用基聚类得到样本的图数据表示, 对图进行分割。

3.2　评价指标

本文使用准确率 (accuracy, ACC)、调整兰德系数 (adjusted rand index, ARI) 以及标准互信息 (normalized mutual information, NMI)3 个指标对各算法的聚类集成结果进行评价与分析。准确率对样本的真实标签和生成标签的匹配程度进行度量:

$$\mathrm{ACC}=\frac{\sum_{i=1}^{n}\delta(s_i,\mathrm{map}(r_i))}{n} \tag{14}$$

式中: r_i、s_i 分别表示数据 x_i 聚类得到的标签和真实标签; n 表示数据总个数; map 表示最佳类标的重现分配; δ 表示指示函数, 表示为

$$\delta(x,y)=\begin{cases}1, & x=y\\ 0, & \text{其他}\end{cases} \tag{15}$$

调整兰德系数度量真实标签和聚类标签相似性, 定义如下:

$$\mathrm{ARI}=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\binom{n_{ij}}{2}-\eta}{\frac{1}{2}(\rho+\vartheta)-\eta} \tag{16}$$

式中: $\rho=\sum_{i=1}^{I}\binom{n_i}{2}$、$\vartheta=\sum_{j=1}^{J}\binom{n_j}{2}$、$\eta=\frac{2\rho\vartheta}{n(n-1)}$。标准互信息用来度量真实样本簇划分与聚类结果之间的相近程度:

$$\mathrm{NMI}=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}n_{ij}\mathrm{lb}\frac{nn_{ij}}{n_in_j}}{\sqrt{\sum_{i=1}^{I}n_i\mathrm{lb}\frac{n_i}{n}\sum_{j=1}^{J}n_j\mathrm{lb}\frac{n_j}{n}}} \tag{17}$$

式中: n_{ij} 表示聚类结果的第 i 个簇中包含原数据集类标签为 j 的样本总数; n_i 表示聚类结果的第 i 个簇的样本总数; n_j 表示原数据集类标签为 j 的样本总数; n 表示样本总数; I 和 J 分别表示聚类得到的簇个数和原数据集的类个数。

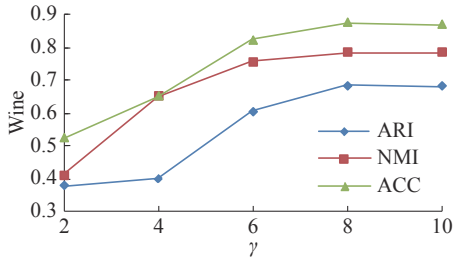
ACC、ARI、NMI 的取值范围为 [0,1], 且取值越接近 1, 表明算法在数据集中获得的聚类效果越好。

3.3　实验结果与分析

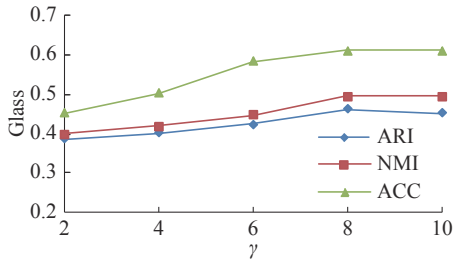
对本文算法进行如下设置: 将置信度 DC 的值设置为 0.9, 收敛阈值 tol 设置为 0.1%, 目标分布 P 的"更新间隔设置 T"设置为 5。为了确定超参数 γ 的取值, 在实验中将 γ 分别设置为 2、4、6、8、10, 在不同 γ 下获得的 ARI、NMI 和 ACC 值如图 7 所示。

由图 7 可以看出, 当超参数 γ<8 时, 本文算法在 5 个数据集上的 ARI、NMI 和 ACC 取值随着 γ 值的增加整体上呈上升趋势。当超参数 γ=8 时, 取值达到峰值, 聚类效果最佳。而当 γ 的取值继续增加到 10 时, 取值没有呈现明显的增长趋势。这是因为在自监督聚类集成过程中, 通过最小化目标损失函数 L 达到优化图自动编码器与聚类集

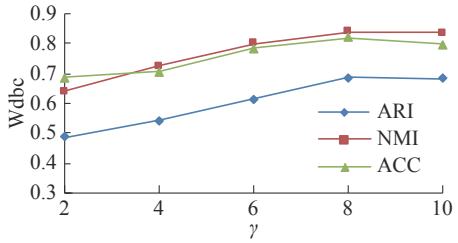
成模块的目的，通过 L 的定义可知，如果 γ 的值过小，会使聚类损失对低维嵌入学习过程影响过大从而导致空间扭曲；而当 γ 取值过大时，会影响低维嵌入空间对聚类损失的优化，从而导致聚类集成结果变差。分别使用本文算法与 5 种对比算法在 5 个数据集上进行聚类划分，各方法聚类集成结果的 ARI、NMI 和 ACC 值如表 2~4所示。



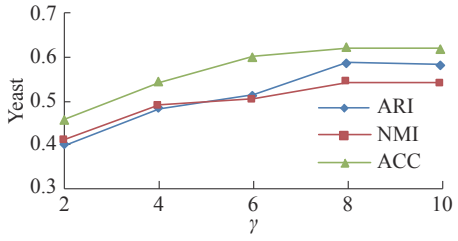
(a) 超参数 γ 的取值对数据集 Wine 的影响



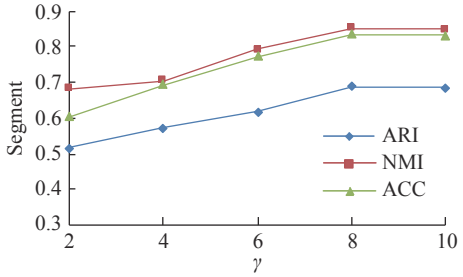
(b) 超参数 γ 的取值对数据集 Glass 的影响



(c) 超参数 γ 的取值对数据集 Wdbc 的影响



(d) 超参数 γ 的取值对数据集 Yeast 的影响



(e) 超参数 γ 的取值对数据集 Segment 的影响

图 7　超参数 γ 的取值对数据集聚类的影响

Fig. 7　Effect of parameter γ on data sets

表 2　不同算法的 ARI 比较

Table 2　Comparison of different algorithms with respect to ARI

数据集	本文算法	HGPA	CSPA	MCLA	Voting	Spectral
Wine	**0.684**	0.301	0.369	0.370	0.370	0.422
Glass	**0.461**	0.103	0.130	0.171	0.110	0.184
Wdbc	**0.687**	0.000	0.135	0.490	0.491	0.501
Yeast	**0.587**	0.132	0.210	0.362	0.210	0.380
Segment	**0.689**	0.351	0.353	0.390	0.363	0.510

表 3　不同算法的 NMI 比较

Table 3　Comparison of different algorithms with respect to NMI

数据集	本文算法	HGPA	CSPA	MCLA	Voting	Spectral
Wine	**0.785**	0.381	0.394	0.426	0.426	0.564
Glass	**0.495**	0.219	0.241	0.321	0.240	0.256
Wdbc	**0.842**	0.001	0.093	0.465	0.465	0.650
Yeast	**0.542**	0.122	0.195	0.291	0.196	0.352
Segment	**0.852**	0.471	0.483	0.558	0.492	0.658

表 4　不同算法的 ACC 比较

Table 4　Comparison of different algorithms with respect to ACC

数据集	本文算法	HGPA	CSPA	MCLA	Voting	Spectral
Wine	**0.874**	0.467	0.485	0.587	0.549	0.682
Glass	**0.612**	0.314	0.342	0.401	0.365	0.385
Wdbc	**0.820**	0.010	0.157	0.694	0.658	0.741
Yeast	**0.621**	0.256	0.351	0.415	0.354	0.582
Segment	0.834	0.620	0.689	0.754	0.710	**0.852**

表 2~4 中加粗的数字分别表示不同算法在每个数据集上取得的最优值。由这些实验结果可以看出，本文算法在各数据集上获得的 ARI、NMI 和 ACC 值整体优于其他 5 种算法。其中，本文算法在 5 个数据集上的 ARI 和 NMI 取值相较于其他算法均有明显提升，仅在 Segment 数据集的 ACC 取值稍逊于 Spectral 算法。这是由于本文算法生成的基聚类的图数据表示完整反映了样本的全局相似关系，使用了处理属性缺失的图数据更具有优势的图神经网络，并且自监督模型使图自

动编码器中的信息传递和数据映射服从最终聚类集成目标, 从而使产生的低维嵌入有利于获得最优的聚类集成结果。而 HGPA、CSPA、MCLA 方法采用的图分割算法趋向于将图形划分为大小相似的簇, 对于真实分布大小不一的簇聚类结果较差; 投票法只依赖基聚类的划分结果, 没有充分挖掘样本之间的相似性关系; 谱聚类集成算法的聚类集成结果很大程度上受到相似度矩阵计算结果的影响。

4 结束语

本文针对聚类集成的一致性函数设计问题, 提出了一种基于图神经网络的聚类集成算法, 主要工作包括:

1) 根据基聚类划分结果计算样本之间的相似性, 完整地反映了样本之间的一阶与二阶相似性, 在此基础上, 用相似度矩阵表示样本之间的邻接关系, 将基聚类在拓扑空间中的表示转化为图数据表示, 通过图聚类获得聚类集成的最优解;

2) 利用图神经网络模型构造自监督聚类集成算法, 图自动编码器在聚类集成结果中很好地保留了样本在空间中的结构, 并且自监督优化模型使图自动编码器中的低维学习服从聚类集成的目标, 从而得到最优的聚类集成结果;

3) 在数据集上进行了仿真实验, 结果表明本文算法提升了聚类集成结果的准确性。

然而, 本文算法的空间复杂度较大, 在未来的工作中, 考虑如何降低算法的空间复杂度并且将算法应用于实际问题。

参考文献:

[1] HAN Jiawei, KAMBER M, PEI Jian. Data mining: concepts and techniques[M]. 3rd ed. Amsterdam: Elsevier, 2012: 223−259.

[2] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究 [J]. 软件学报, 2008, 19(1): 48–61.
SUN Jigui, LIU Jie, ZHAO Lianyu. Clustering algorithms research[J]. Journal of software, 2008, 19(1): 48–61.

[3] JUDD D, MCKINLEY P K, JAIN A K. Large-scale parallel data clustering[J]. IEEE transactions on pattern analysis and machine intelligence, 1998, 20(8): 871–876.

[4] BHATIA S K, DEOGUN J S. Conceptual clustering in information retrieval[J]. IEEE transactions on systems, man, and cybernetics, part B (cybernetics), 1998, 28(3): 427–436.

[5] FRIGUI H, KRISHNAPURAM R. A robust competitive clustering algorithm with applications in computer vision[J]. IEEE transactions on pattern analysis and machine intelligence, 1999, 21(5): 450–465.

[6] FERN X Z, LIN Wei. Cluster ensemble selection[J]. Statistical analysis and data mining, 2008, 1(3): 128–141.

[7] 罗会兰. 聚类集成关键技术研究 [D]. 杭州: 浙江大学, 2007.
LUO Huilan. Research on key technologies of clustering ensemble[D]. Hangzhou: Zhejiang University, 2007.

[8] FRED A L N. Finding consistent clusters in data partitions[C]//Proceedings of the 2nd International Workshop on Multiple Classifier Systems. Cambridge, UK, 2001: 309−318.

[9] STREHL A, GHOSH J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions[J]. Journal of machine learning research, 2003, 3: 583–617.

[10] FRED A L N, JAIN A K. Data clustering using evidence Accumulation[C]//Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02). Quebec City, Canada, 2002: 276−280.

[11] WANG Xi, YANG Chunyu, ZHOU Jie. Clustering aggregation by probability accumulation[J]. Pattern recognition, 2009, 42(5): 668–675.

[12] 杨草原, 刘大有, 杨博, 等. 聚类集成方法研究 [J]. 计算机科学, 2011, 38(2): 166–170.
YANG Caoyuan, LIU Dayou, YANG Bo, et al. Research on cluster aggregation approaches[J]. Computer science, 2011, 38(2): 166–170.

[13] ZHOU Zhihua, TANG Wei. Clusterer ensemble[J]. Knowledge-based systems, 2006, 19(1): 77–83.

[14] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on neural networks, 2009, 20(1): 61–80.

[15] WU Z , PAN S , CHEN F. A comprehensive survey on graph neural networks[J]. IEEE transactions on neural networks and learning systems, 2019(02): 4–24.

[16] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th International Con-

ference on Machine Learning. Helsinki, Finland, 2008: 1096−1103.

[17] TIAN Fei, GAO Bin, CUI Qing, et al. Learning deep representations for graph clustering[C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec City, Canada, 2014: 1293−1299.

[18] IAM-ON N, BOONGOEN T, GARRETT S. LCE: a link-based cluster ensemble method for improved gene expression data analysis[J]. Bioinformatics, 2010, 26(12): 1513–1519.

[19] WANG Chun, PAN Shirui, HU Ruiqi, et al. Attributed graph clustering: a deep Attentional embedding approach[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China, 2019: 3670−3676.

[20] KIPF T N, WELLING M. Variational graph auto-encoders[J/OL]. Available: http://axrxiv.org/abs/1611.07308. 2016.

[21] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(86): 2579–2605.

[22] XIE J, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[J]. Computer science, 2015: 478–487.

[23] VON LUXBURG U. A tutorial on spectral clustering[J]. Statistics and computing, 2007, 17(4): 395–416.

作者简介：

杜航原，副教授，博士，主要研究方向为机器学习、社会网络。主持和参与国家级、省部级科研项目 7 项。发表学术论文 10 余篇。

张晶，硕士研究生，主要研究方向为数据挖掘与机器学习。

王文剑，教授，博士生导师，博士，国家自然科学基金委信息学部自动化学科会评专家，中国人工智能学会理事、中国人工智能学会机器学习专委会常务委员、知识工程与分布智能专委会委员、粗糙集与软计算专业委员会委员，中国计算机学会人工智能与模式识别专委会委员，中国计算机学会太原分部监督委员会主席、ACM 太原分部副主席，并担任多个国际国内学术会议的程序委员会主席或委员，主要研究方向为机器学习与数据挖掘。主持国家自然科学基金项目 4 项。发表学术论文 150 余篇。