中图法分类号:TP183　　文献标识码:A　　文章编号:1006-8961(2023)04-1069-10

论文引用格式:Geng W F, Wang X, Jing L P and Yu J. 2023. Consensus graph learning-based self-supervised ensemble clustering. Journal of Image and Graphics, 28(04):1069-1078(耿伟峰, 王翔, 景丽萍, 于剑. 2023. 共识图学习驱动的自监督集成聚类. 中国图象图形学报, 28(04):1069-1078)[DOI:10.11834/jig.210947]

共识图学习驱动的自监督集成聚类

耿伟峰,王翔,景丽萍*,于剑

1. 北京交通大学交通数据分析与挖掘北京市重点实验室,北京　100044;
2. 北京交通大学计算机与信息技术学院,北京　100044

摘　要:目的　随着实际应用场景中海量数据采集技术的发展和数据标注成本的不断增加,自监督学习成为海量数据分析的一个重要策略。然而,如何从海量数据中抽取有用的监督信息,并该监督信息下开展有效的学习仍然是制约该方向发展的研究难点。为此,提出了一个基于共识图学习的自监督集成聚类框架。**方法**　框架主要包括3个功能模块。首先,利用集成学习中多个基学习器构建共识图;其次,利用图神经网络分析共识图,捕获节点优化表示和节点的聚类结构,并从聚类中挑选高置信度的节点子集及对应的类标签生成监督信息;再次,在此标签监督下,联合其他无标注样本更新集成成员基学习器。交替迭代上述功能块,最终提高无监督聚类的性能。**结果**　为验证该框架的有效性,在标准数据集(包括图像和文本数据)上设计了一系列实验。实验结果表明,所提方法在性能上一致优于现有聚类方法。尤其是在MNIST-Test(modified national institute of standards and technology database)上,本文方法实现了97.78%的准确率,比已有最佳方法高出3.85%。**结论**　该方法旨在利用图表示学习提升自监督学习中监督信息捕获的能力,监督信息的有效获取进一步强化了集成学习中成员构建的能力,最终提升了无监督海量数据本质结构的挖掘性能。

关键词: 集成聚类;自监督聚类;图表示学习;共识图;伪标签置信度

Consensus graph learning-based self-supervised ensemble clustering

Geng Weifeng, Wang Xiang, Jing Liping*, Yu Jian

1. Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China;
2. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

Abstract: Objective Clustering is focused on machine learning-related data segmentation for multiple datasets. Its applications are in relevant to such domains like image segmentation and anomaly detection. In addition, to simplify complex tasks optimize its performance, clustering is used in data preprocessing tasks of those are data sub-blocks segmentation, pseudo-labels generation, and abnormal points-removal. Self-supervised learning has become an essential technique for massive data analysis. However, it is challenged to extract effective supervision information and analyze the input data. **Method** A consensus graph learning based self-supervised ensemble clustering (CGL-SEC) framework is developed. It consists of three main modules: 1) to construct the consensus graph based on several ensemble components (i. e., the basic clustering methods). 2) to extract the supervision information by learning the consensus graph representation, and 3) its node clustering results, where the subset of nodes with the high-confidence are selected as labeled samples. To optimize the ensemble components and the corresponding consensus graph, t basic clustering methods are re-trained in related the option of samples-labeled and other samples-unlabeled. The final clustering results can be optimized iteratively until

收稿日期:2021-10-13;修回日期:2022-04-19;预印本日期:2022-04-26

*通信作者:景丽萍　lpjing@bjtu.edu.cn

the learning process converges. **Result** A series of experiments are carried out on benchmarks, including both image and textual datasets. Especially, CGL-SEC is 3.85% over baseline in terms of clustering evaluation metric on themodified national institute of standards and technology database (MNIST-Test). First, to optimize data representation and cluster assignment at the same time, deep embedding clustering can be focused on data itself as the supervision information and auto-encoder with the reconstruction loss is pre-trained. The soft cluster assignment of features-embedded is then calculated, and the KL (Kullback-Leibler) divergence is minimized between the soft cluster assignment and the auxiliary target distribution. To improve the performance of the model further, following deep clustering network (DCN) can use hard clustering instead of soft allocation, and local constraints are applied by improved deep embedding clustering (IDEC). The pseudo-label strategy is implemented as a self-supervised learning method that uses the prediction results of the neural network as the label to simulate the supervision information compared to using data itself as the supervision information. Deep-cluster-based K-means clustering is used to generate pseudo-labels to guide the training of convolutional networks. However, the generated pseudo-labels have lower confidence and are prone to trivial solutions in the initial stage of network training. Deep embedding clustering with data augmentation (DEC-DA) and MixMatch-based prediction of data-enhanced samples are used as the supervision information of the original data, which improves the accuracy of the supervision information to a certain extent, but this method is difficult to extend to text and other fields. Deep adaptive clustering-based high-confidence pseudo-label subsets-selected are iteratively trained the network in the prediction results, but low-confidence samples-involved data distribution information is ignored. Pseudo-semi-supervised clustering votes are used to select a subset of high-confidence pseudo-labels, and all samples are used to train semi-supervised neural network. Although the ensemble strategy can improve the confidence of the pseudo-label, the voting strategy is concerned of category representation only without the feature representation of the sample itself, which can reduce the clustering performance in some cases. The ensemble learning is regarded as a representative machine learning method that reflects the ability of "group intelligence", whereas a learning method can improve the overall prediction performance via multiple base learners training and their coordinated prediction results. In pseudo-label-based clustering tasks, it can coordinate multiple base learners to obtain high-confidence pseudo-labels. However, the effectiveness of the supervision information acquisition is still to be resolved. The category information of the sample is considered for current pseudo-label-based ensemble clustering method only when the label is captured and some effective information are ignored like the feature representation of the sample itself and the clustering structure between samples. **Conclusion** Graph neural network is composed of content information of nodes and the structural information between nodes at the same time. To design a self-supervised ensemble clustering method based on consensus graph representation learning, it is required to make full use of sample features and relationships between samples in ensemble learning. To obtain higher confidence pseudo-labels as supervised information and improve the performance of self-supervised clustering, it is necessary to mine global and local information at the same time. We illustrate a learnable data ensemble representation through graph neural network. The confidence of pseudo-labels is improved, and the entire model is trained in self-supervision iteratively. To be summarized: 1) Commonly-used consensus graph learning-integrated clustering framework is developed, which can use multi-level information like clustering-integrated sample characteristics and category structure. 2) Self-supervision method is proposed, which uses graph neural network to mine the global and local information of the consensus graph, and high-confidence pseudo-labels are obtained as supervised information. 3) Experiments are demonstrated that the consensus graph learning ensemble clustering method has its potentials on image and text datasets.

Key words: ensemble clustering; self-supervised clustering; graph representation learning; consensus graph; pseudo-label confidence

0 引 言

聚类是根据数据相似性将其划分到若干集合中的一个经典机器学习问题,在兴趣推荐、图像分割和异常检测等领域均有广泛的应用前景(Jain,2010)。除此之外,聚类还常用于划分数据子块、生成伪标签和剔除异常点等数据预处理任务,以简化后续复杂任务或提升后续任务的性能。在大数据时代,随着数据采集技术的发展、数据标注成本的不断增加,自

监督学习成为海量数据分析的一个重要策略。然而，如何从海量数据中抽取有用的监督信息，以及如何在该监督信息下开展有效的学习仍然是制约该方向发展的研究难点。

由于深度神经网络具有强大表示能力，可以将数据转换为更有利于聚类的表示，大量基于深度学习的聚类方法相继提出。例如，深度嵌入聚类（Xie等，2016）方法通过优化样本特征表示和聚类分配隶属度，大幅提升了在复杂数据上的聚类性能。在此基础上，Yang等人（2017）、Guo等人（2017，2018）和Tao等人（2019）等方法的提出使深度聚类成为一个热门的研究领域。现有的大多深度神经网络模型和方法主要依赖监督信息或先验假设来挖掘数据的复杂潜在结构。受该思想的启发，Deep-cluster（Caron等，2018）提出自监督聚类的思想，对学习到的特征进行聚类，将聚类生成的伪标签作为监督信息指导网络更新。伪标签策略在许多无监督、半监督任务中都表现出不错的性能（Caron等，2018；Berthelot等，2019；Chang等，2017）。但是，由于初始阶段的伪标签置信度较低，模型在训练过程中会逐渐产生错误积累，影响最终聚类结果。

集成学习作为一种通过训练多个基学习器并组合它们的预测结果来提高整体预测性能的学习方法，被认为是体现群体智慧能力的代表性机器学习方法（Sagi和Rokach，2018）。在基于伪标签的聚类任务中，集成学习能够整合不同的基学习器，从而获得高置信度的伪标签（Gupta等，2020）。但目前基于伪标签的集成聚类方法在捕获标签时仅考虑了样本的类别信息，忽略了样本本身特征表示和样本间聚类结构等重要信息，难以确保监督信息获取的有效性。

图神经网络能够同时兼顾节点内容信息和节点间的结构信息，图表示学习已成为数据挖掘领域的一个重要分支（Kipf和Welling，2016，2017；Pan等，2018）。受该方法启发，本文设计了基于共识图表示学习的自监督集成聚类方法。核心思路是在集成学习中充分利用样本特征和样本间关系，同时挖掘全局和局部信息，以获得更高置信度的伪标签作为监督信息，提升自监督聚类的性能。本文的主要贡献为3个方面：1）提出一个通用的共识图学习集成聚类框架，充分利用样本特征和类别结构等多层次信息进行集成聚类；2）提出一种自监督训练方法，使用图神经网络挖掘共识图的全局与局部信息，获得高置信度伪标签作为监督信息；3）在图像和文本数据集上进行实验，证明了共识图学习集成聚类方法与最先进的集成聚类方法相比具有显著优势。

1 相关工作

近年提出了大量关于自监督聚类方法。本文聚焦于使用集成学习提升自监督聚类中监督信息的准确性与稳定性。

1.1 自监督聚类

深度嵌入聚类（Xie等，2016）以数据本身作为监督信息，使用重建损失预训练自动编码器，计算嵌入特征的软聚类分配，最小化软聚类分配与辅助目标分布之间的KL（Kullback-Leibler）散度，实现了同时优化数据表示和聚类分配。在此基础上，DCN（deep clustering network）（Yang等，2017）使用硬聚类替代软分配、IDEC（improved deep embedding clustering）（Guo等，2017）应用了局部约束，进一步提升了模型性能。

与以数据本身作为监督信息不同，伪标签策略是用神经网络的预测结果作为标签来模拟监督信息的一种自监督学习方式，避免了深度嵌入聚类等方法收敛时无法达到最佳性能的问题。Deep-cluster（Caron等，2018）使用K均值聚类生成伪标签以指导卷积网络训练，但网络训练初期，生成的伪标签置信度较低，而且容易出现平凡解，这都会影响最终的聚类性能。DEC-DA（deep embedding clustering with data augmentation）（Guo等，2018）和MixMatch（Berthelot等，2019）使用数据增强样本的预测结果作为原始数据的监督信息，在一定程度上提升了监督信息的准确性，但是这种手段难以推广到文本等领域。深度自适应聚类（deep adaptive clutering，DAC）（Chang等，2017）在预测结果中选出了高置信度伪标签子集迭代训练网络，但忽略了低置信度样本中蕴含的数据分布信息。伪半监督聚类（Gupta等，2020）投票选取高置信度伪标签子集，再使用全部样本半监督地训练神经网络。虽然集成策略能够提升伪标签的置信度，但仅使用类别表示的投票策略忽略了样本本身的特征表示，这在某些情况下会严重降低聚类性能（Tao等，2017）。

1.2 集成聚类

集成聚类通过组合基学习器的预测来提高整体聚类性能，从本质上看，预测无标签样本的伪标签也是对数据进行划分的一种聚类任务。因此，集成学习也能够用来提升伪标签置信度。集成聚类方法可以分为3类，即基于效用函数的方法、基于共识矩阵的方法和基于图的方法。

基于效用函数的方法将集成聚类转化成最大化效用函数（utility function）问题。效用函数度量了基本划分与共识划分间的相似度，如基于二次互信息的效用函数（Topchy等，2003）、基于K均值的效用函数（Wu等，2015）。随后又提出了多种共识函数（Li等，2007；Lu等，2008）。

基于共识矩阵的方法通过样本对在同一个聚类中出现的次数来构造一个共识矩阵（Fred和Jain，2005），再利用传统聚类生成最终结果。谱集成聚类（Liu等，2015）对共识矩阵使用谱聚类，转化为一个加权K均值聚类问题。伪半监督聚类（Gupta等，2020）在共识矩阵中寻找高置信度的子集来迭代训练基学习器。

基于图的方法以图或超图的形式集成基聚类器信息，再用图分割获得最终的聚类结果。传统的图集成采用启发式算法（Strehl和Ghosh，2003）和随机游走算法（Abdala等，2010）。但由于传统方法仅使用类别特征构建共识图，忽略了对样本原始特征的重用。最近，对抗图自编码器聚类（Tao等，2019）使用图自编码器学习低维的节点嵌入表示，重用了样本原始特征，在低维空间上获得了不错的聚类性能，但其共识图是预先定义的、不可学习的，共识图上的噪声可能会使最终的聚类性能较差。

本文融合图集成方法与自监督方法的优势，通过图神经网络获得可学习的数据集成表示，提高了伪标签的置信度，并自监督迭代训练整个模型。

2 基于共识图学习的自监督集成聚类

在集成学习中，共识图包含多层次的复杂信息。为了充分利用这些信息，本文提出了一个通用的集成聚类框架，如图1所示。

基于共识图学习的自监督集成聚类框架可以分为3个部分：第1部分利用基学习器结果构建共识图；第2部分利用图神经网络分析共识图，捕获节点优化表示和节点的聚类结构；第3部分从聚类中挑选高置信度的节点子集及对应的类标签生成监督信息；在此标签监督下联合其他无标注样本更新集成成员基学习器。模型反复迭代使聚类性能得到了持续的提高。在本框架中，基学习器可以替换成任意无监督聚类模型或半监督模型。

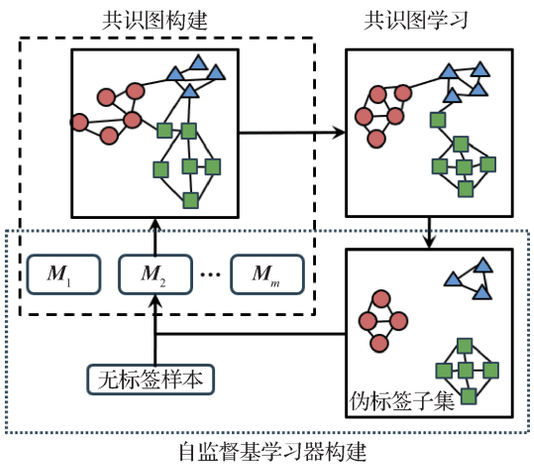

图1 共识图学习自监督集成聚类

Fig. 1 Consensus graph learning-based self-supervised ensemble clustering

2.1 共识图构建

对于给定的无标注数据集$\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$，本文以阶梯网络（Rasmus等，2015）作为基学习器模型$\boldsymbol{M}=\{\boldsymbol{M}_1,\cdots,\boldsymbol{M}_m\}$进行训练，阶梯网络采用多通道的自动编码器抽取样本特征，并在编码器顶层后添加了分类头。然而，阶梯网络是一种半监督模型，同时使用监督损失和无监督损失，但在没有监督损失的情况下会仅输出常数，需要在损失中增加信息最大化损失（Gomes等，2010），具体为

$$\boldsymbol{L}_{\mathrm{IM}}=\boldsymbol{I}(\boldsymbol{X};\boldsymbol{Y})=H(\boldsymbol{Y})-H(\boldsymbol{Y}|\boldsymbol{X}) \tag{1}$$

式中，$\boldsymbol{X}$和$\boldsymbol{Y}$为模型输入和输出，$H(\cdot)$和$H(\cdot|\cdot)$分别是熵和条件熵。最大化$H(\boldsymbol{Y})$鼓励神经网络在输出类别上均匀分布；最小化$H(\boldsymbol{Y}|\boldsymbol{X})$鼓励神经网络对给定的输入进行明确的类分配。

对于基学习器输出的m个聚类结果，采用类似于证据积累法（Fred和Jain，2005）的方式初始化共识图$\boldsymbol{A}=\{\boldsymbol{X},\boldsymbol{E}\}$，图中每个节点代表一个样本，如果$\boldsymbol{M}$中的超过阈值$t(0<t\leqslant m)$个数的基学习器将样本对划分为同一簇（可以有较高的置信度认为样本对为同一类），则在样本对之间添加边$\boldsymbol{E}$；即

$$(\boldsymbol{x},\boldsymbol{x}')\in\boldsymbol{E}\Leftrightarrow n_{\mathrm{agree}(\boldsymbol{x},\boldsymbol{x}')}\geqslant t$$

$$\boldsymbol{n}_{\text{agree}(\boldsymbol{x},\boldsymbol{x}')}=\left|\left\{\boldsymbol{m}:\boldsymbol{m}\in\boldsymbol{M},\boldsymbol{m}(\boldsymbol{x})=\boldsymbol{m}(\boldsymbol{x}')\right\}\right| \tag{2}$$

2.2 共识图学习

为了挖掘图结构信息、提高伪标签的置信度，本文使用图神经网络学习从嵌入空间到聚类分配空间的映射，充分利用了节点特征作为聚类依据。具体来说，共识图学习先由图神经网络消息传递层计算新的集成表示 $\bar{Z}$，再通过多层感知器计算节点的聚类分配 S（Bianchi 等，2020），即

$$\bar{\boldsymbol{Z}}=MP\left(\boldsymbol{Z},\tilde{\boldsymbol{A}}\right)=ReLU\left(\tilde{\boldsymbol{A}}\boldsymbol{Z}W_m+\boldsymbol{Z}W_s+b\right) \tag{3}$$

$$\boldsymbol{S}=\boldsymbol{MLP}_{\varphi}\left(\bar{\boldsymbol{Z}}\right) \tag{4}$$

式中，W_m, W_s, b, φ 是图神经网络的参数，节点表示 $\boldsymbol{Z}$ 是在基学习器阶梯网络的无噪声通道顶层获得的数据的低维表示，邻接矩阵 $\tilde{\boldsymbol{A}}$ 是共识图 $\boldsymbol{A}$ 经过对称归一化后得到的。具体为

$$\tilde{\boldsymbol{A}}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\in\mathbf{R}^{N\times N} \tag{5}$$

$$\boldsymbol{D}=\operatorname{diag}\left(\boldsymbol{A}1_N\right) \tag{6}$$

网络的损失为

$$L=\underbrace{-\frac{\operatorname{tr}\left(\boldsymbol{S}^{\mathrm{T}}\tilde{\boldsymbol{A}}\boldsymbol{S}\right)}{\operatorname{tr}\left(\boldsymbol{S}^{\mathrm{T}}\tilde{\boldsymbol{D}}\boldsymbol{S}\right)}}_{L_c}+\underbrace{\left\|\frac{\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}}{\left\|\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}\right\|_{\mathrm{F}}}-\frac{\boldsymbol{I}_K}{\sqrt{K}}\right\|_{\mathrm{F}}}_{L_o} \tag{7}$$

式中，$\boldsymbol{A}1_N$ 为全 1 矩阵，图割损失 L_c 计算由聚类分配矩阵 $\boldsymbol{S}$ 给出的图割损失，最小化 L_c 鼓励强连通节点被划分为一簇。正交损失 L_o 鼓励聚类分配是正交的，并且鼓励模型将样本平均分配给每个簇，避免了平凡解的产生，K 为聚类簇数。$\boldsymbol{I}_K$ 可以理解为一个重构的聚类矩阵 $\boldsymbol{I}_K=\hat{\boldsymbol{S}}^{\mathrm{T}}\hat{\boldsymbol{S}}$，其中 $\hat{\boldsymbol{S}}$ 为每个簇分配 n/K 个点。

值得注意的是，图神经网络直接学习从节点特征空间到聚类分配空间的映射，不需要计算图的谱，所以可以应用于流式数据，直接计算新样本的聚类分配，而不需要计算谱分解。

相比于伪半监督聚类（Gupta 等，2020）仅集成基学习器的聚类结果，本文使用共识图学习更有利于聚类的节点集成表示。使用共识图学习有以下优势：1）使用共识图学习优化节点表示不仅考虑了数据的类别表示，还重新引入了数据的嵌入表示。而伪半监督聚类仅考虑了基学习器的结果，忽略了样本本身特征表示，存在严重的信息损失；2）使用共识图学习可以基于全局和局部信息挖掘节点聚类结构，寻找密集强连通子图，进一步提高监督信息（伪标签）的置信度。而以前的方法采用贪心的策略选取伪标签，仅利用了图的局部的信息。

2.3 自监督基学习器构建

尽管基学习器已经提供了较好的基础准确率，但仍然存在部分样本被错误划分的情况。为了尽可能避免使用错误的伪标签训练基学习器，在新的共识图的每一簇中，仅选取高置信度的样本子集标注伪标签。为了最大化每个样本子集的规模，获得更多数量的伪标签，使模型更快地收敛，选取度最大的节点及与其连通的节点标注对齐的伪标签。即

$$\boldsymbol{X}_k=\left\{\boldsymbol{x}\,|\,\boldsymbol{x}\rightarrow\boldsymbol{x}_{\max}^k,\boldsymbol{x}_{\max}^k\in S^k,0\leqslant k<K\right\} \tag{8}$$

式中，K 是聚类簇数，S^k 表示第 k 个簇，$\boldsymbol{x}_{\max}^k$ 是第 k 簇内度最大的节点。

与仅使用高置信度伪标签样本训练学习器不同，本方法使用伪标签子集与无标签样本一起训练半监督基学习器集合 $\boldsymbol{M}=\{\boldsymbol{M}_1,\cdots,\boldsymbol{M}_m\}$，半监督基学习器采用与初始化步骤中相同的阶梯网络。训练后的基学习器将再产生 m 个聚类结果用于构建共识图、优化共识图表示、选取高置信度的伪标签子集并反复迭代更新，即自监督地训练基学习器。

需要强调的是，本文模型在迭代过程中，通过不断精炼伪标签，以提高伪标签置信度。前一轮模型迭代产生的高置信度伪标签子集经过对齐后，成为下一轮迭代的基学习器的监督信息。由于剔除了低置信度的伪标签，所以并非所有样本都具有监督信息，因此，基学习器采取半监督分类器充分挖掘未标注样本和伪标签样本的语义信息。

CGL-SEC 算法完整的伪代码如下：

输入：数据 $\boldsymbol{X}$，聚类簇数 K；

输出：聚类分配 $\boldsymbol{S}$。

1）For $j\in\{1,2,\cdots,m\}$ Do；

2）初始化第 j 基学习器阶梯网络 $\boldsymbol{M}_j$ 参数；

3）使用无监督损失训练基学习器 $\boldsymbol{M}_j$；

4）End For；

5）For $it\in\{1,2,\cdots,n\_iter\}$ Do；

6）初始化共识图 $\boldsymbol{A}$；

7）初始化图神经网络；

8）使用式（7）中损失训练图神经网络；

9）图神经网络前向传播，输出聚类分配矩阵 $\boldsymbol{S}$；

10） For $k\in\{1,2,\ldots,K\}$ Do；

11） 筛选伪标签；

12） End For；

13） For $j \in \{1, 2, \cdots, m\}$ Do；

14） 对齐$\boldsymbol{M}_j$的输出伪标签；

15） 最小化监督损失和无监督损失之和训练基学习器$\boldsymbol{M}_j$；

16） End For；

17) End For；

18) 图神经网络前向传播，输出聚类分配矩阵$\boldsymbol{S}$；

19) Return $\boldsymbol{S}$。

3 实 验

为验证本文方法的性能，在5个标准数据集上进行实验。本文的代码开源发布在：https://github.com/ggg929627701/GNN-kingdra。

3.1 实验数据集

MNIST-Test（modified national institute of standards and technology database），是MNIST数据集的测试集，共10类10 000个样本。STL10（self-taught learning 10）包含10类13 000幅图像。每个样本为96×96像素的彩色RGB图像。USPS（United States Postal Service）包含10类手写体数字，共9 298个样本，每个样本为16×16像素的灰度图像。Reuters包含4类英语新闻文本数据，是一个类别不均衡数据集，最大的类占数据集的43%。20News包含20类英文新闻文档，每个类有902个样本。本文采用与Hu等人（2017）相同的数据预处理方式。

3.2 基准方法

实验将提出的基于共识图学习的自监督集成聚类与多种聚类算法进行比较。包括：1）3类传统的聚类方法。即K均值聚类（K-means）（MacQueen，1967）、谱聚类（normalized-cut spectral clustering，SC）（Shi和Malik，2000）和基于ζ函数的凝聚聚类（zeta function based agglomerative clustering，AC）（Zhao和Tang，2008）。2）3类最先进的深度聚类方法。即使用信息最大化损失的自增强聚类（information maximizing self-augmented training，IMSAT）（Hu等，2017）、深度嵌入聚类（deep embedded clustering，DEC）（Xie等，2016）和深度聚类网络（deep clustering networks，DCN）（Yang等，2017）。3）4类集成聚类方法。即谱集成聚类（spectral ensemble clustering，SEC）（Liu等，2015）、K均值共识聚类（K-means-based consensus clustering，KCC）（Wu等，2015）、对抗图自编码器聚类（adversarial graph auto-encoder，AGAE）（Tao等，2019）和伪半监督聚类（Kingdra）（Gupta等，2020）。由于在Tao等人（2019）的研究中，变分图自编码器（variational graph auto-encoder，VGAE）（Kipf和Welling，2016）和（adversarially regularized variational graph autoencoder，ARVGE）（Pan等，2018）等图嵌入聚类方法已与AGAE对比了性能，本文只对比性能最好的AGAE方法。同样，在Liu等人（2015）的研究中，CSPA、HGPA和MCLA等集成聚类方法已与KCC和SEC对比了性能，本文不再对其进行比较。

3.3 实验结果

实验在3个图像基准数据集和两个文本数据集上进行，实验结果如表1和表2所示，展示了聚类准确率（accuracy，ACC）和归一化互信息（normalized mutual information，NMI）指标。

如表1和表2所示，诸如K均值聚类、谱聚类等传统聚类方法往往比深度聚类方法性能差。与DEC等先进的深度聚类方法相比，本文方法在多个数据集上实现了更高的聚类准确率。实际上，除了Reuters之外，本文算法都优于其他方法。在Reuters数据集上略低于DEC，是因为基学习器信息最大化损失鼓励聚类平均分配，但Reuters是一个类不均衡数据集，最大的一类占总样本数的43%，这使得模型性能受到了影响。然而，可以使用任意聚类算法替换本框架中基学习器来适应不同数据集。

与投票集成方法Kingdra相比，本方法在所有数据集上都获得了更好的聚类性能。这表明本文提出的图集成方法比投票集成方法更具性能优势。在MNIST-Test数据集上，图集成方法的准确率相比于投票集成方法提升了3.85%。

与AGAE预先定义共识图相比，本文方法在迭代训练中持续更新共识图表示，在STL和USPS数据集上的准确率（ACC）与归一化互信息（NMI）高出了3%以上。这体现了可学习的共识图的优越性。

在时间性能方面，以MNIST数据集为例，模型在单个TITAN Xp GPU加速的条件下，每次迭代（模型每获得一次伪标签）约花费40 ± 3 min，完整训练过程需要50次迭代。图神经网络训练时间复杂度与节点数和边数成正比，MNIST是10类手写数字分

表 1　在图像标准数据集上的聚类准确率和归一化互信息

Table 1　Clustering ACC and NMI on image datasets

	MNIST-Test		STL10		USPS	
	ACC	NMI	ACC	NMI	ACC	NMI
K-means	0.531 6	0.499 4	0.848 4	0.786 0	0.668 0	0.627 0
SC	0.660 0	0.704 0	0.563 5	0.568 5	0.562 0	0.540 0
AC	0.810 0	0.693 0	0.821 7	0.754 4	0.657 0	0.798 0
IMSAT	N/A	N/A	0.941 0	N/A	N/A	N/A
DEC	0.856 0	0.830 0	0.841 3	0.851 8	0.688 0	0.683 0
DCN	0.802 0	0.786 0	0.893 0*	0.816 3*	0.762 0*	0.767 0*
SEC	0.568 7	0.515 7	0.797 6*	0.783 0*	0.681 5*	0.652 9*
KCC	0.602 6	0.465 1	0.727 8*	0.787 3*	0.560 7*	0.638 5*
AGAE	N/A	N/A	0.932 5*	0.874 1*	0.735 7*	0.741 5*
Kingdra	0.939 3	0.909 6	0.951 0	0.888 0	0.773 2	0.798 0
CGL-SEC	**0.977 8**	**0.942 0**	**0.963 1**	**0.892 7**	**0.797 3**	**0.821 5**

注：加粗字体表示各列最优结果，带*的结果摘自 Tao 等人（2019）的研究，下划线的结果摘自 Gupta 等人（2020）的研究，N/A 表示对应方法不适用于相关数据集。

表 2　在文本标准数据集上的聚类准确率和归一化互信息

Table 2　Clustering ACC and NMI on text datasets

方法	20News		Reuters	
	ACC	NMI	ACC	NMI
K-means	0.153 9	0.185 5	0.540 4	0.412 8
SC	0.248 6	0.214 7	0.663 3	0.339 1
AC	0.239 7	0.202 4	0.436 6	0.011 1
IMSAT	0.311 0	N/A	0.710 0	N/A
DEC	0.308 0	0.299 6	**0.736 8**	**0.497 6**
Kingdra	0.439 0	0.414 7	0.705 0	0.393 0
CGL-SEC	**0.442 7**	**0.461 7**	0.723 2	0.416 4

注：加粗字体表示各列最优结果，下划线的结果摘自 Gupta 等人（2020）的研究，N/A 表示对应方法不适用于相关数据集。

类数据集，在训练过程中 90% 的同类的样本对之间均会产生边，这形成了一个稠密图。常见的图数据集比如 Cora、Citeseer 等的拓扑关系往往是稀疏的，这使得图神经网络在这些稀疏图上的训练时间较少。本文期望通过拓扑关系是否稠密来衡量节点集合是否紧致，进一步从语义角度定义节点伪标签的置信度，这也导致了一定的时间性能代价。

3.4　消融实验

为了说明每个模块对于最终性能的贡献，设计了消融实验，分别对比本文使用的单独的基学习器（base learner，BL）、投票集成多个基学习器的聚类结果（base-learner-vote，BL-Vote）和不进行共识图学习，仅投票选取伪标签的自监督迭代聚类（voted self-supervised clustering，V-SEC）的性能，结果如表 3 所示。可以看出，基学习器提供了一个较好的基础准确率，但投票集成仅提供了 1%～2% 的性能增益，

表 3　在 5 个标准数据集上的消融实验

Table 3　Ablation experiment on 5 standard datasets

模块	MNIST-Test		STL10		USPS		20News		Reuters	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
BL	0.893 7	0.837 8	0.871 8	0.768 4	0.718 9	0.680 8	0.384 0	0.330 5	0.682 0	0.315 0
BL-Vote	0.916 0	0.886 4	0.915 0	0.817 2	0.731 2	0.729 8	0.405 0	0.401 2	0.690 0	0.347 2
V-SEC	0.939 3	0.909 6	0.951 0	0.888 0	0.773 2	0.798 0	0.439 0	0.414 7	0.705 0	0.393 0
CGL-SEC	**0.977 5**	**0.942 0**	**0.963 1**	**0.892 7**	**0.797 3**	**0.821 5**	**0.442 7**	**0.461 7**	**0.723 2**	**0.416 4**

注：加粗字体表示各列最优结果。

最大的性能增益来自自监督训练方式。共识图学习进一步为自监督训练提供了更大的增益，使性能相比于基学习器提高了4%～10%。

与V-SEC相比，CGL-SEC在MNIST-Test和Reuters上的准确率(ACC)与归一化互信息(NMI)都增加了2%以上，表明本文提出的基于图神经网络的集成学习方法足够通用，在图像和文本数据集上都获得了较大的性能提升。

图2描述了在MNIST-Test数据集上本方法和Kingdra方法的伪标签数量和准确率与迭代次数的关系。图2表明，本方法的图神经网络选出了数量更少但准确率更高的伪标签，基于图神经网络的集成策略比投票策略更好地抑制了伪标签置信度的下降。然而，伪标签的减少并没有让模型的收敛速度减慢，本方法只减少了少量的低置信度伪标签，但使得模型性能获得了较大地提升。

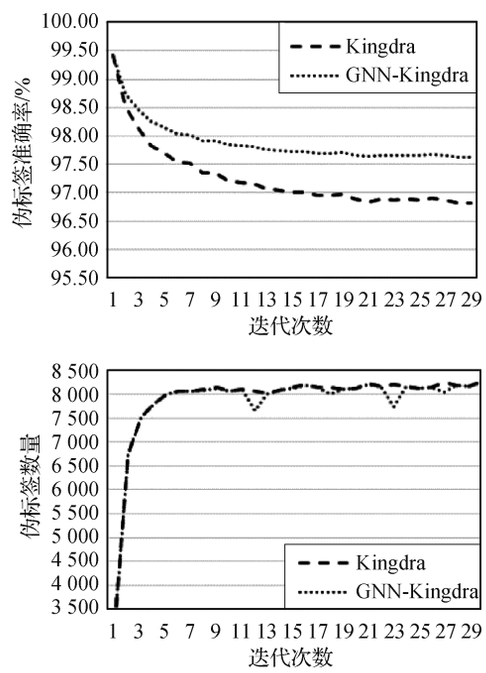

图2　伪标签数量、准确率与迭代次数的关系

Fig. 2　The number and accuracy of pseudo-labels vary with the number of iterations

3.5　收敛性和泛化性

在许多聚类算法中，分阶段特征学习和聚类、深度嵌入聚类以及数据增强的深度嵌入聚类，会出现模型性能在训练阶段早期达到最大值，但随后降低并收敛到稳定值的现象。这是由于网络在微调阶段发生了过拟合，随着训练迭代轮次的增加，重建损失使网络逐渐关注数据中与类别无关的信息。这使得基于自动编码器的聚类模型不得不增加停止迭代的条件(Guo等，2018)。

基于伪标签的聚类方法可以有效避免收敛时达不到最佳聚类性能的问题。图3描述了聚类性能指标与迭代次数的关系，可以观察到，随着迭代次数的增加，准确率与归一化互信息逐步收敛到最大值。因为随着训练的进行，高置信度伪标签的数量会逐渐增加，更多的样本参与计算伪监督损失，指导网络学习数据中与类别相关的信息，有效避免了过拟合。

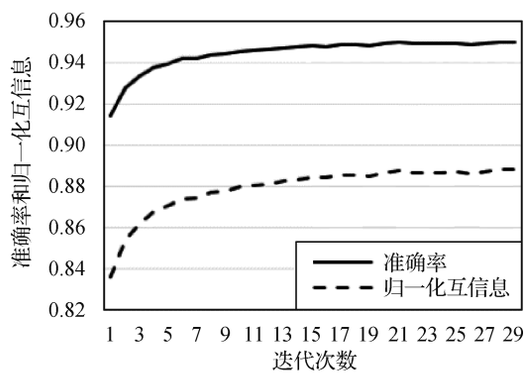

图3　STL数据集在训练过程中准确率和归一化互信息

Fig. 3　Accuracy and normalized mutual information during training on the STL dataset

为证明模型的泛化性能，在MNIST-Test上训练本文模型，并在MNIST和MNIST-Test上分别测试，结果如表4所示。其中，在新样本测试集(MNIST)上的准确率仅比训练集上的测试结果低0.007，尤其是模型的训练集仅有10 000个样本，而测试集含有60 000个新样本，这表明模型具有很好的泛化能力。

表4　训练集为MNIST-Test，测试集分别为MNIST和MNIST-Test的模型性能

Table 4　The performance of the model when training set is MNIST-Test and test set is MNIST or MNIST-Test respectively

测试集	ACC	NMI
MNIST-Test	0.977 8	0.942 3
MNIST	0.970 8	0.924 8

4　结　论

本文希望解决自监督聚类中如何从海量数据中

抽取有用的监督信息，以及如何在该监督信息下开展有效的学习这两方面问题，提出了一种基于共识图学习的自监督集成聚类框架。总结如下：1）在集成聚类任务中，以往的共识图的构建方法往往忽略了样本特征信息和共识图自身的可学习性，本文使用图神经网络融合了样本特征与类别表示等多层次信息，更好地从数据中抽取了高质量的自监督信息。2）提出了基于共识图学习的自监督集成聚类框架，利用从共识图中提取的伪标签监督模型学习知识，并在多个图像、文本标准数据集上取得了令人满意的结果。3）实验与结果表明，共识图表示学习可以提升伪标签的置信度，进而提升深度集成聚类的性能。

通过实验探究，发现了所提方法存在的问题。在大规模数据集上，本方法在图神经网络训练阶段耗时较长。时间性能较低几乎是所有基于图的聚类方法的限制。

因此，如何提高图集成聚类的效率将是本领域未来的研究方向。此外，对于集成聚类领域，如何构建多样性的基学习器也是一个具有挑战性的问题。

参考文献（References）

Abdala D D, Wattuya P and Jiang X Y. 2010. Ensemble clustering via random walker consensus strategy//Proceedings of the 20th International Conference on Pattern Recognition. Istanbul, Turkey: IEEE: 1433-1436 [DOI: 10.1109/icpr.2010.354]

Berthelot D, Carlini N, Goodfellow I, Oliver A, Papernot N and Raffel C. 2019. MixMatch: a holistic approach to semi-supervised learning//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: [s.n.]: #454

Bianchi F M, Grattarola D and Alippi C. 2020. Spectral clustering with graph neural networks for graph pooling//Proceedings of the 37th International Conference on Machine Learning. [s. l.]: PMLR: 874-883

Caron M, Bojanowski P, Joulin A and Douze M. 2018. Deep clustering for unsupervised learning of visual features//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 139-156 [DOI: 10.1007/978-3-030-01264-9_9]

Chang J L, Wang L F, Meng G F, Xiang S M and Pan C H. 2017. Deep adaptive image clustering//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5880-5888 [DOI: 10.1109/iccv.2017.626]

Fred A L N and Jain A K. 2005. Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (6): 835-850 [DOI: 10.1109/tpami.2005.113]

Gomes R, Krause A and Perona P. 2010. Discriminative clustering by regularized information maximization//Proceedings of the 23rd International Conference on Neural Information Processing Systems. Vancouver British, Canada: Curran Associates Inc.: 775-783

Guo X F, Gao L, Liu X W and Yin J P. 2017. Improved deep embedded clustering with local structure preservation//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: IJCAI.org: 1753-1759 [DOI: 10.24963/ijcai.2017/243]

Guo X F, Zhu E, Liu X W and Yin J P. 2018. Deep embedded clustering with data augmentation//Proceedings of the 10th Asian Conference on Machine Learning. Beijing, China: PMLR: 550-565

Gupta D, Ramjee R, Kwatra N and Sivathanu M. 2020. Unsupervised clustering using pseudo-semi-supervised learning//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net: #1520

Hu W H, Miyato T, Tokui S, Matsumoto E and Sugiyama M. 2017. Learning discrete representations via information maximizing self-augmented training//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR: 1558-1567 [DOI: 10.48550/arXiv.1702.08720]

Jain A K. 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31 (8): 651-666 [DOI: 10.1016/j.patrec.2009.09.011]

Kipf T N and Welling M. 2016. Variational graph auto-encoders [EB/OL]. [2021-10-13]. https://arxiv.org/pdf/1611.07308.pdf

Kipf T N and Welling M. 2017. Semi-supervised classification with graph convolutional networks//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net:

Li T, Ding C and Jordan M I. 2007. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization//Proceedings of the 7th IEEE International Conference on Data Mining (ICDM). Omaha, USA: IEEE: 577-582 [DOI: 10.1109/icdm.2007.98]

Liu H F, Liu T L, Wu J J, Tao D C and Fu Y. 2015. Spectral ensemble clustering//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia: ACM: 715-724 [DOI: 10.1145/2783258.2783287]

Lu Z W, Peng Y X and Xiao J G. 2008. From comparing clusterings to combining clusterings//Proceedings of the 23rd National Conference on Artificial Intelligence. Chicago, USA: AAAI Press: 665-670 [DOI: 10.5555/1620163.1620175]

MacQueen J B. 1967. Some methods for classification and analysis of multivariate observations//The 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, USA: University of California Press: 281-297 [DOI: 10.1.1.308.8619]

Pan S R, Hu R Q, Long G D, Jiang J, Yao L N and Zhang C Q. 2018. Adversarially regularized graph autoencoder for graph embedding//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: IJCAI. org: 2609-2615 [DOI: 10.24963/ijcai.2018/362]

Rasmus A, Valpola H, Honkala M, Berglund M and Raiko T. 2015. Semi-supervised learning with ladder networks//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal Canada: MIT Press: 3546-3554

Sagi O and Rokach L. 2018. Ensemble learning: a survey. WIREs Data Mining and Knowledge Discovery, 8(4): #1249 [DOI: 10.1002/widm.1249]

Shi J B and Malik J. 2000. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8): 888-905 [DOI: 10.1109/34.868688]

Strehl A and Ghosh J. 2003. Cluster ensembles-a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research, 3: 583-617 [DOI: 10.1162/153244303321897735]

Tao Z Q, Liu H F and Fu Y. 2017. Simultaneous clustering and ensemble//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI Press: 1546-1552 [DOI: 10.1609/aaai.v31i1.10720]

Tao Z Q, Liu H F, Li J, Wang Z W and Fu Y. 2019. Adversarial graph embedding for ensemble clustering//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: IJCAI.org: 3562-3568 [DOI: 10.24963/ijcai.2019/494]

Topchy A, Jain A K and Punch W. 2003. Combining multiple weak clusterings//Proceedings of the 3rd IEEE International Conference on Data Mining. Melbourne, USA: IEEE: 331-338 [DOI: 10.1109/icdm.2003.1250937]

Wu J J, Liu H F, Xiong H, Cao J and Chen J. 2015. K-means-based consensus clustering: a unified view. IEEE Transactions on Knowledge and Data Engineering, 27(1): 155-169 [DOI: 10.1109/tkde.2014.2316512]

Xie J Y, Girshick R B and Farhadi A. 2016. Unsupervised deep embedding for clustering analysis//Proceedings of the 33nd International Conference on Machine Learning. New York, USA: JMLR. org: 478-487

Yang B, Fu X, Sidiropoulos N D and Hong M Y. 2017. Towards K-means-friendly spaces: simultaneous deep learning and clustering//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR. org: 3861-3870 [DOI: 10.1145/3149166.3149171]

Zhao D L and Tang X O. 2008. Cyclizing clusters via Zeta function of a graph//Proceedings of the 21st International Conference on Neural Information Processing Systems. Vancouver British, Canada: Curran Associates Inc.: 1953-1960 [DOI: 10.1.1.144.2539]

作者简介

耿伟峰，男，硕士研究生，主要研究方向为机器学习、聚类和集成学习。E-mail: 21120348@bjtu.edu.cn

景丽萍，通信作者，女，教授，主要研究方向为机器学习理论和方法及在人工智能领域中的应用。E-mail: lpjing@bjtu.edu.cn

王翔，男，博士研究生，主要研究方向为无监督聚类和多源数据学习。E-mail: wangxbjtu@gmail.com

于剑，男，教授，主要研究方向为机器学习理论及相关应用。E-mail: jianyu@bjtu.edu.cn