Note

abstract

现有文章的缺陷

- 缺少**动态的融合机制**，来进行融合与提纯图结构、节点的信息
- **不能**从鲁棒目标分布中**提取信息**

propose a Deep Fusion Clustering Network（**DFCN**）

- SAIF模块，a interdependency learning-based **S**tructure and Attribute **I**nformation **F**usion

network training

- reliable target distribution generation measure
- triplet self-supervision strategy

  三元组自监督策略

Introduction

深度聚类**deep clustering**

- 目标：

  训练一个可以用来学习特征的表示并将数据划分为几个不相连的组的神经网络，without intense manual guidance
- 应用
  - 异常检测，anomaly detection
  - 社交网络分析，social network analysis
  - 人脸识别，face recognition

决定模型性能的主要因素

1. 优化目标，optimization objective
2. 特征的提取，fashion of feature extraction

现有的5种聚类方法

1. 基于空间聚类的方法，subspace clustering-based methods
2. 基于生成对抗网络的方法，generative adversarial network-based methods
3. 基于谱的聚类方法，spectral clustering-based methods
4. 基于高斯混合的方法，gaussian mixture model-based methods
5. 自优化的方法，self-optimizing-based methods

我们的方法是**自优化**的方法

过去聚焦于探索数据**原始特征空间的属性**

未来聚焦于**几何结构信息与其属性信息的集成**

现有模型的缺点

- 没有信息的融合与连接
- 目标分布仅使用了源信息

solution

- 设计一个动态信息合成模块，对从AE、GAE中提取出来的特征进行表示融合

  SAIF，Structure and Attribute information Fusion

  步骤

  1. 合成了两种sample embedding，包括local和global
  2. 估计两个像本点和聚类中心的相似性
  3. 设计了一个三方自监督机制

  还设计了一个improved graph autoencoder(IGAE)

主要的贡献

1. 提出了一个DFCN

   SAIF模块，structure and attribute information fusion，用于融合AE和GAE

   三方自监督学习策略
2. 提出了一个IGAE

   improve the generalization capability of the proposed method
3. SOTA

Related work

attributed graph clustering

target distribution generation

The Proposed Method

fusion-based autoencoders

- Input of the Decoder

  融合了AE和IGAE两者的输出latent representation

- Improved graph antoencoder

  同时优化特征的权重矩阵以及邻接矩阵

  损失函数如下

  $L_{IGAE} = L_w + \gamma L_\alpha$

  $L_w = \frac{1}{2N}||AX - Z||_F^2$

  $L_\alpha = \frac{1}{2N}||A - A||_F^2$

structure and attribute information fusion

- a cross-modality dynamic fusion mechanism

  4个步骤

  1. 将AE和IGAE的embedding联合在一起

     $Z_I = \alpha Z_{AE} + (1-\alpha) Z_{IGAE}$

  2. 进行一个图卷积的操作，massage passing operation

     $Z_L = AZ_I$

     local structure enhanced Z_I

  3. 自相关的机制，self-correlated learning

     计算归一化后的自相关矩阵

     $S_{ij} = \frac{e^{(Z_L Z_L^T)_{ij}}}{\sum_{k=1}^{N} e^{(Z_L Z_L^T)_{ik}}}$

     全局相关性

     $Z_G = SZ_L$

  4. fusion机制

     $Z = \beta Z_G + Z_L$

     其中β初值为0

  我们的cross-modality dynamic fusion mechanism 考虑了样本相关性，包括局部和全局

- triplet self-supervised strategy

  为了生成更多的**引导**

  需要聚合更多的信息

  两个步骤

  1. 计算了第i个sample和第j个预先计算的聚类中心在fused embedding，使用Student's分布作为核

$$q_{ij} = \frac{(1+||z_i-u_j||^2/v)^{-\frac{v+1}{2}}}{\sum_j(1+||z_i-u_j||^2/v)^{-\frac{v+1}{2}}}$$

2. 为了提高可信度，使得所有的sample都驱动到聚类中心

   $0 =< p_{ij} <= 1$是生成目标分布的元素，表示第i个样本对第j个中心的概率

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

   triplet clustering loss by adapting the KL散度

   $$L_{KL} = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{(q_{ij}+q'_{ij}+q''_{ij})/3}$$

算法流程

algorithm 1 Deep Fusion Clustering Network

输入：

- 特征矩阵**X**
- 邻接矩阵**A**
- 目标分布更新区间**T**
- 迭代次数**I**
- 聚类个数**K**
- 超参数γ和λ

输出：聚类的结果**O**

步骤：

1. 初始化AE、IGAE的参数，并进行融合，获得Z_{AE}、Z_{IGAE}以及$\widetilde{Z}$
2. 初始化聚类中心u，使用K-means算法 在$\widetilde{Z}$的数据上
3. for i=1 to **I** do
   - 更新Z_I和Z_L，$Z_I = \alpha Z_{AE} + (1-\alpha)Z_{IGAE}$，$Z_L = \widetilde{A}Z_I$
   - 更新 normalized self-correlation matrix **S**，计算deep clustering embedding $\widetilde{Z}$ $S_{ij} = softmax(Z_L Z_L^T)$, $Z_G = SZ_L$,$\widetilde{Z} = \beta Z_G + Z_L$
   - 计算soft assignment distributions Q，Q', Q'' 在数据$\widetilde{Z}$、Z_{IGAE}和Z_{AE}
   - if i%T==0
     - 计算目标分布P，通过Q

       $$p_{ij} = \frac{q_{ij}^2/\sum_i q_{ij}}{\sum_{j'}(q_{ij'}^2/\sum_i q_{ij'})}$$

   - 使用P来优化Q,Q',Q''

     $$L_{KL} = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{(q_{ij}+q'_{ij}+q''_{ij})/3}$$

   - 计算L_{AE} L_{IGAE} L_{KL}
   - 更新整个网络通过最小化loss

     $$L = \underbrace{L_{AE} + L_{IGAE}}_{Reconstruction} + \underbrace{\lambda L_{KL}}_{Clustering}$$

     - L_{AE}是AE的重建误差，MSE
4. 获得聚类结果**O**，通过最终的$\widetilde{Z}$，利用K-means算法

   return **O**

joint loss and optimization

$$L = \underbrace{L_{AE} + L_{IGAE}}_{Reconstruction} + \underbrace{\lambda L_{KL}}_{Clustering}$$

Experiments

- Benchmark Datasets

  总共6个著名的数据集

  三个图数据集

  - ACM
  - DBLP
  - CITE

  三个非图数据集

  - USPS
  - HHAR
  - REUT

  非图数据集没有邻接矩阵A，我们使用 heat kernel method

- Experiment Setup
  - 训练过程

    pytorch 平台

    一块NVIDIA 2080TI GPU

    1. 单独预训练AE和IGAE，30个iterations，通过最小化重建误差
    2. 在united framework中训练100个iterations
    3. 在triplet self-supervised strategy下，训练最少200iterations，直到收敛

    做10次重复性实验

  - 参数设置

    Adam优化器

    learning rate 1e-3

    - USPS
    - HHAR

    learning rate 1e-4

    - REUT
    - DBLP
    - CITE

    learning rate 5e-5

    - ACM

    batch size 256， early stop

    超参数

    - $\gamma = 0.1$
    - $\lambda = 10$

非图数据集，nearest neighbors nums = 5

- evaluation metric

  四种指标

  - Accuracy(ACC)
  - Normalized Mutual Information(NMI)
  - Average Rand Index(ARI)
  - macro F1-score(F1)

- 和SOTA模型进行对比

  一共为10种模型，在6个数据集上进行对比，用了4种指标

  - K-means，经典的浅层聚类方法
  - AE，DEC，IDEC，代表了autoencoder-based 聚类方法，通过训练一个autoencoder
  - GAE/VGAE，ARGA，DAEGC，是经典的图卷积神经网络方法，这些方法的embedding是通过GCN来获取的
  - $SDCN_Q$和SDCN是将AE和GCN的方法优点进行结合的聚类

对比实验

- IGAE的效率
  - 只用特征信息GAE_{L_w}
  - 只用邻接矩阵GAE_{L_α}

- 分析SAIF
  - 证明cross-modality dynamic fusion机制能提升
  - triplet self-supervised strategy 能提升

- 利用双源信息的影响
  - 证明加入AE和IGAE一起会更好

- 分析超参数λ
  - 超参数对性能有提升
  - 该方法的性能比较稳定，即使是一个大范围的波动
  - 当λ为10的时候性能较好
- 可视化

  将输出的$\tilde{Z}$映射到2d空间，通过t-SNE算法

Conclution

- 提出了DFCN
- 核心SAIF模块利用graph和node特征通过dynamic cross-modality fusion和triplet self-supervised strategy
- IGAE