ASD 0402 文献汇报

1. Block Modeling-Guided Graph Convolutional Neural Networks

1.1 背景

- GCN对于 图异质性 问题的解决方法通常为：
  - 聚合高阶邻域信息

    direct neighborhoods may be heterophily-dominant, but the higher-order neighborhoods are homophily-dominant and thereby provide more valuable information *改变原始结构，造成信息损失*
  - 在异质邻居中传递“标记”信息

    As the nodes aggregate information from neighbors, they get positive messages from neighbors with the same class while negative messages from neighbors with different classes. *计算代价高*
- 文章考虑了一种更为直观的思想：**不同类的邻居以不同方式聚合**
  块建模是用来描述网络的结构规律（包括同质性、异质性或其混合物），有很大的潜力来解决这个问题。然而，虽然块建模通过所谓的块矩阵（它描述了由一条边连接的两个块中的节点的可能性）来描述了类之间的规则关系，但它仍然不能用于指导图卷积框架中的分类聚合。

1.2 提出 BM-GCN 框架

- 软标签得到节点**块标签**（假设部分已知）
  引入 MLP 学习节点的块标签，用学习的 软标签 导出 块矩阵
  We introduce a multilayer perception (MLP) into the whole learning framework and use it to learn soft labels for all nodes using attribute information. And then we use the soft labels to derive the block matrix.
- 基于得到的块矩阵推导**分类聚合机制**
  对导出的块矩阵计算 相似度 ，表征连接模式中不同块的相似度，作为聚合指标构建图卷积操作，实现同质数据与异质数据的 分类聚合
  We propose to create a new block similarity matrix based on the derived block matrix, which can characterize the similarity degree between different blocks in the connecting pattern of blocks. By doing so, we can use this new matrix as an aggregation indicator to construct the graph convolutional operation and finally realize the classified aggregation for both homophilic and heterophilic graphs.
- 符号说明

Notations	Explanations
$\mathcal{G}$	A graph
$\mathcal{V}$	The set of nodes in graph $\mathcal{G}$
$\mathcal{E}$	The set of edges in graph $\mathcal{G}$
$\mathcal{T}_{\mathcal{V}}$	Training node set
$\mathcal{N}_i$	The set of neighbor of node v_i
A	Adjacency matrix
X, X_i	Attribute matrix, attribute vector of node v_i
Y, Y_i	Label matrix, one-hot label vector of node v_i
B, B_i	Soft label matrix, soft label vector of node v_i
H, $h_{i,j}$	Block matrix, an element in H
Q, $q_{i,j}$	Block similarity matrix, an element in Q
Z	Node representations

Table 1: Notations and Explanations.

- 关键计算
  - 同质率 h

Homophily Ratio. The homophily ratio (Pei et al. 2020) can measure the overall homophily level in a graph. It counts the ratio of same-class neighbor nodes to the total neighbor nodes in a graph, defined as

$$h = \frac{1}{|\mathcal{V}|} \sum\nolimits_{v_i \in \mathcal{V}} \frac{|\{v_j \| v_j \in \mathcal{N}_i, Y_j = Y_i\}|}{|\mathcal{N}_i|} \tag{1}$$

where $\mathcal{N}_i$ is the neighbor set of node v_i. In this work, we use homophily ratio h to determine whether a graph is homophilic or heterophilic.

  - 块矩阵（描述块间连接概率）H

Block Matrix. Given the labels $Y \in \mathbb{R}^{n \times c}$ for all nodes and the adjacency matrix $A \in \{0,1\}^{n \times n}$, the block matrix is defined as

$$H = \left(Y^T AY\right) \oslash \left(Y^T AE\right) \tag{2}$$

where E an all-ones matrix with the same size as Y, and $\oslash$ the Hadamard (element-wise) division operation. Block matrix models the linked possibility of nodes in any two blocks. In this work, blocks represent the classes of labels in a graph. From the node-wise level, $H_{i,j}$ is the probability that a node in the i-th class connects with a node in the j-th class.

1.3 方法

1.3.1 学习块矩阵

- 用已有的节点属性学习全部的块标签

  BM-GCN adopts the way of learning the unknown labels from the given data (the known labels and the attribute network) to compute the block matrix.
- 为什么?

  Considering that the soft labels should come from the original data while the topology may be not trustworthy in heterophilic graphs, here BM-GCN uses node attributes alone to generate soft labels.
- MLP接受节点属性信息，输出软标签

$$\bar{B} = \sigma(MLP(X))$$
$$B = softmax(\bar{B})$$

- 损失函数

  In order to ensure the reliability of the soft label B,BM-GCN first pre-trains the MLP layer with the training ground-truth labels for several iterations. Specifically, the pre-training process aims to minimize the loss function

$$L_{MLP} = \sum_{v_i \in T_v} f(B_i, Y_i)$$

f为交叉熵损失, Y为真值标签

- 组合标签

  BM-GCN MLP maximizes the use of available ground-truth labels by assembling the two kinds of labels

  H描述任意两个块（类别）中两个节点通过边连接的可能性分布。

$$Y_s = \{Y_i, B_j | \forall v_i \in T_v, \forall v_j \notin T_v\}$$
$$H = (Y_s^T A Y_s) \oslash (Y_s A E)$$

1.3.2 块相似度矩阵

> 在异质图中，边倾向于连接不同类别的节点，不同类别之间的可能性值可能大于同一类别内的可能性值。因此，为了实现块引导的分类聚合，需要修改块矩阵H中的元素值，使这些元素能够反映在图卷积操作中不同类别节点之间传播规则的潜在有价值信息。

- 提出了基于块矩阵的新的相似度矩阵 Q，表明**具有相似结构连接模式的块（类）彼此之间会有更多的信息传播**

$$Q = HH^T$$

- 考虑**同一类内部**的节点信息交换会更加频繁，BM-GCN提高了**同一类内的信息传播率**

$$Diag(Q) \leftarrow \alpha Diag(Q)$$
$$\alpha为增强因子$$

1.3.3 块引导图卷积

- 基于新创建的块相似度矩阵Q，BM-GCN可以为不同的类组合分配不同的信息传播规则。此外，软标签可以指示这两个节点属于哪个类组合。这样，信息传播过程可以共同由软标签B和块相似矩阵Q决定，具体为：

> 考虑c个类别，节点v_i和v_j的软标签:

$$B_i = \{b_i^1, b_i^2, ..., b_i^c\}$$
$$B_j = \{b_j^1, b_j^2, ..., b_j^c\}$$

> 节点对i j的类别组合概率化为：

$$p(\varphi(v_i) = Y_r, \varphi(v_j) = Y_t) = b_i^r b_j^t, \quad r, t \in \{1, 2, .., c\}$$

- 块相似矩阵Q表示任意两类之间的信息传播概率，即**两类越相似，信息传播就越多**。因此，节点vi和节点vj之间的**传播概率**可以看作是Q中元素的期望:

> 节点vi和vj之间的传播概率同时由节点的软标签和块相似矩阵Q引导。对于图中的所有节点对，这些对的传播概率可以表示为权重矩阵

$$\omega_{ij} = \sum_{r=1}^{c} \sum_{t=1}^{c} q_{r,t} b_i^r b_j^t$$
$$\Omega = BQB^T$$

- 用Ω细化拓扑结构：

$$A' = \Omega \odot (A + \beta I)$$

- 归一化A'：

$$\tilde{a}_{i,j} = \frac{exp(a'_{i,j})}{\sum_{v_s \in N} exp(a'_{i,s})}$$

- 用新的归一化的$\tilde{A}$替换GCN中使用的归一化图拉普拉斯算子，这样BM-GCN中的图卷积操作是在软标签和块相似矩阵Q的引导下，最终可以实现分类聚合机制。属于不同软标签组合的节点对将有不同的信息交换，信息传播率由Q确定，新的图卷积层可以写为：

$$Z^{(k)} = Z^{(k-1)} W_1^{(k)} + \tilde{A} Z^{(k-1)} W_2^{(k)}$$

1.4 模型优化

- BM-GCN采用半监督损失函数：

$$L_{GCN} = \sum_{v_i \in T_v} f(Z_i, Y_i)$$

- **BM-GCN**将块相似度学习过程和块引导图卷积操作过程集成到一个统一的框架中。在MLP层和图卷积操作中加入损失函数，最终的损失函数可以写成

$$L_{final} = \lambda L_{GCN} + (1 - \lambda) L_{MLP}$$