西安交通大学
XI'AN JIAOTONG UNIVERSITY

ICDM 2022：大规模电商图上的风险商品检测

队伍名：GCNGAT
队长：王臻卓
队员：王方、于志洋

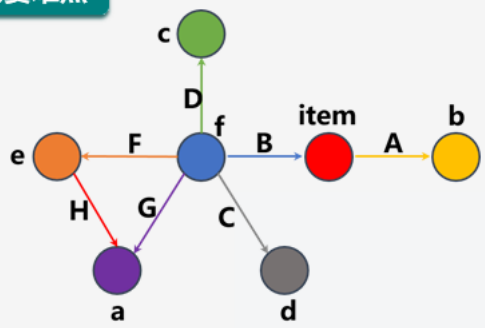
Contents Title

西安交通大学 XI'AN JIAOTONG UNIVERSITY

问题分析

ICDM 2022：大规模电商图上的风险商品检测

比赛难点



电商图元结构

比赛难点：**数据规模巨大**（节点数量达到**1380万**，连边数量达到**1.578亿**）、**数据异质性高**（图中一共包含7种类型的节点，14种类型的连边关系，并且item节点之间并无直接关联）、**训练样本不平衡**（正负样本比例为1比9，标签数据仅有8.5万，存在大量未标注样本）等。

方案简介

1. 针对数据规模巨大的问题，引入**GraphSage**的归纳式学习方式，通过邻居采样的方式，对图神经网络进行训练，避免了传统图神经网络需要利用全图信息进行训练的缺点。

2. 针对图数据的连边异质性问题，利用**R-GCN**作为主要模型，R-GCN对不同类型的连边关系使用不同的聚合函数，并且通过**预处理映射**，将节点特征分别进行嵌入，解决了节点异质性的问题。

3. 由于原图中存在大量的未标注样本，是一个典型的半监督学习任务，通过引入半监督学习中的**一致性正则**机制（SCR），充分利用了未标注样本的价值。

西安交通大学 XI'AN JIAOTONG UNIVERSITY

技术思路

ICDM 2022：大规模电商图上的风险商品检测

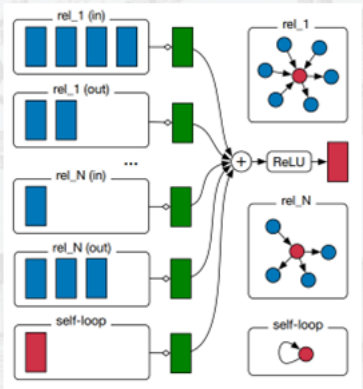
超参数设置

参数名称	设定值
Epoch	10
Batch_size	128
Hidden_dim	64
Learning_rate	0.01

在比赛过程中，我们对epoch，学习率，隐含层节点数等参数进行了多次尝试，发现最佳参数如左图所示。对于学习率，我们采用了StepLR策略进行调整，经过多次调试发现，每4个epoch调整学习率为之前的十分之一最佳。

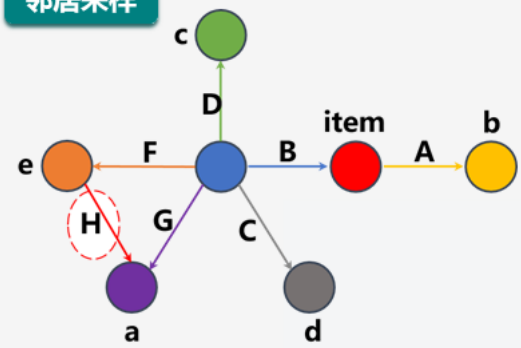
模型选取

在Baseline的基础上，我们分别尝试了RGCN、RGAT、 HGT、HAN等网络，发现RGCN在此问题上具有明显的优势，其参数量更低，单次训练速度快，且相同参数下，RGCN在验证集上的平均精度也更高，所以，我们并未对基础模型做过多的调整。



ICDM 2022：大规模电商图上的风险商品检测

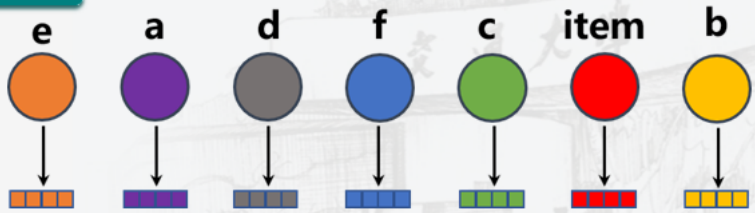
邻居采样



对H连边信息的采样

我们发现，item 的一阶邻居只包括 f 与 b 类型的节点，且 item 到 f 节点的最大出度为 1，到 b 节点的最大出度也为 1，也即一个 item 节点至多与一个 f 节点相邻，也至多与一个 b 节点相邻，所以 item的一阶邻居必须全部采样。item 的二阶邻居包括了 a，c，d，e 四种类型的节点，通过观察比赛所给电商图可以发现，若采样只包括 item 的二阶邻居，则对于 e 节点与 a 节点之间的关联信息并没有充分利用，所以，在邻居采样环节，我们对 item 的 3 阶邻居进行了采样，并且，为了防止邻居爆炸，第 3 阶节点仅包括 e 与 a 之间的 H 连边信息。

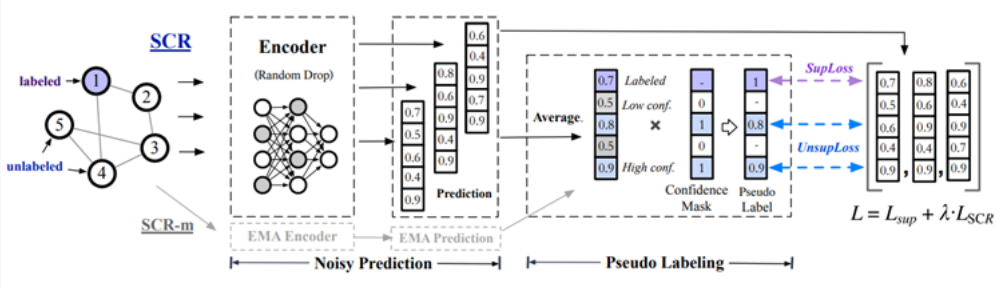
对节点异质性的处理



由于比赛所给电商图包括了 7 种不同类型的节点，所以，对于不同类型的节点特征，在送入 RGCN 之前，应根据其种类使用不同的 Linear 层，将它们分别进行嵌入，然后再进行图上的信息聚合。

ICDM 2022：大规模电商图上的风险商品检测

一致性正则化



由于比赛所给的数据集中，仅有 85562 个标注样本，所以，电商图中存在着大量的未标注样本，于是，我们引入了 SCR 正则化技术，对未标注样本进行了利用。我们发现，在初赛数据集上，SCR 正则化可以带来比较明显的效果提升。

消融实验

Method	AP_score	Accuracy
RGCN(baseline)	0.9563	0.9798
RGCN+projection_layer	0.9602	0.9812
RGCN+ projection_layer +Three-hop neighbor sampling	0.9625	0.9823
RGCN+ projection_layer +Three-hop neighbor sampling+SCR	0.9635	0.9827

我们将训练样本以 8 比 2 的比例划分为训练集（train_set）和验证集(val_set)，并在验证集上测试了各技术点对平均精度和分类准确率的提升。

ICDM 2022：大规模电商图上的风险商品检测

一些其他的尝试

1. 由于比赛数据集异质性较高，item 与 item 节点之间并无直接的连边关系，所以，我们曾尝试过对原始数据集增加 item 到 item 的直接连边关系，实验发现此类做法可以加快模型的收敛，但对最终精度并无明显的提升。

2. 对于训练样本不平衡的问题，我们曾引入 Focal loss 以及对正样本的过采样方法，经验证发现，对精度提升也无明显的效果。

3. 模型方面，我们曾尝试过对不同 hop 的邻居加入注意力机制，但最终精度也没有明显的变化。

训练过程中的注意事项

1. 由于 PyG 框架的特性，无法通过随机种子固定训练过程的随机性，所以，最终模型往往需要尝试多次训练。

2. 使用官方代码（format_pyg.py）构建出 session1 的图后，节点名称的存储顺序 是['b', 'f', 'a', 'item', 'c', 'e', 'd']，转为同质图后，节点特征的存储顺序是按照['b', 'f', 'a', 'item', 'c', 'e', 'd']的先后顺序进行堆叠。 而 session2 的图构建出来后，节点名称的存储顺序是['b', 'f', 'item', 'a', 'c', 'e', 'd']，转为同质图后，节点特征的存储顺序和 session1 不一样。

西安交通大学 XI'AN JIAOTONG UNIVERSITY

参赛总结与收获

0

关联比赛：ICDM 2022 : 大规模电商图上的风险商品检测

全部评论(0)