文章编号：2095-2163(2023)01-0100-05 中图分类号：TP311 文献标志码：A

一种基于 GAT 的小样本均衡补偿文本主题分类模型

王琦菲，张大为
（辽宁师范大学 计算机与信息技术学院，辽宁 大连 116000）

摘 要：针对小样本不均衡数据采用 GAT 模型主题分类效果不佳的问题，本文提出一种基于 GAT 的样本均衡补偿模型（BC-GAT），优化 GAT 模型的构建方法，对数据集中小比例样本进行均衡补偿。本文通过合理运用 EDA 算法和网络爬虫算法，使数据集中小比例样本的扩充更加合理和高效，使 GAT 模型更加适合小样本不均衡主题分类。实验表明，BC-GAT 模型小比例样本识别准确率在 90%以上，可以有效解决实际应用中存在的极小样本和数据倾斜问题。

关键词：BC-GAT；小样本；数据倾斜；主题分类

A small sample balanced compensation text topic classification model based on GAT

WANG Qifei, ZHANG Dawei
(College of Computer and Information Technology, Liaoning Normal University, Dalian Liaoning 116000, China)

【Abstract】 This paper proposes a GAT-based sample balancing compensation model (BC-GAT) to optimize the construction method of the GAT model for the balancing compensation of small-scale samples in the dataset, in order to address the problem that the GAT model is not effective in classifying topics with small unbalanced samples. By reasonably using EDA algorithm and web crawler algorithm, this paper makes the expansion of small-scale samples in the dataset more reasonable and efficient and makes the GAT model more suitable for small sample unbalanced topic classification. The experiments show that the accuracy of the BC-GAT model for small-scale sample recognition is above 90%, which can effectively solve the problems of very small samples and data skewing in practical applications.

【Key words】 BC-GAT; small sample; data skew; subject classification

0 引 言

随着深度学习和大数据技术的发展，文本分类取得了巨大的成功，但在实际应用中，常常存在小样本和类别不均衡现象。小样本学习（Few-shot Learning）指的是通过有限的数据样本使模型获得较为稳定分类效果的机器学习方法[1]。最早在计算机视觉领域被提出，在图像领域也取得了许多较好的研究成果，受限于文本特征提取比图像更加困难，在自然语言处理领域发展较为缓慢。近年来在文本领域的小样本学习研究工作主要集中在基于微调、基于数据增强和基于迁移学习等领域。迁移学习是目前最为前沿的方法，主要包括基于度量学习、基于元学习和基于图神经网络等方法。

基于图神经网络的方法由于性能较好、易于解释，已经成为一种广泛应用的方法。Yao L[2]认为传统的深度学习文本分类方法，例如 RNN（Recurrent Neural Networks，简称 RNN）、LSTM（Long Short Term Memory，简称 LSTM）等存在忽略词共现和需要大数量训练样本的问题，只优先考虑文本的局部信息和远距离文本信息，而不考虑全局的词共现信息，可能导致词共现中包含的长距离和不连续的语义信息缺失，需要大量的数据样本进行训练，致使小样本文本分类难以达到理想效果。考虑到上述不足，Yao 提出 TextGCN（Text Graph Convolutional Networks，简称 TextGCN）模型，通过构建一个大型的异构文本图，将单词和文档作为图的点，边由词节点和词节点的权重、词节点和文档节点的权重两部分构成，使用两

基金项目：国家自然科学基金（20200037，20200084）；辽宁省科技厅-博士科研启动基金计划项目（20210301）。

作者简介：王琦菲（1996-），女，硕士研究生，主要研究方向：软件工程、NLP；张大为（1971-），男，学士，副教授，主要研究方向：软件工程、数据挖掘。

通讯作者：张大为 Email：daweiz@lnnu.edu.cn

收稿日期：2022-04-02

层图卷积网络实现文本分类，该模型首次将图神经网络从图像领域迁移到文本领域，并取得良好的分类效果；Huang[3] 针对 TextGCN 在整个语料库中建图需要消耗大量存储空间的问题进行改进，对每个输入的文本数据单独构图，并引入滑动窗口记作 p，将文本中的单词只与左右 p 个单词相连，而非全部单词节点相连，该方法减少了单个文本与整个语料库之间的依赖，降低了存储空间的消耗；Zhang[4] 为了引入文本之间非连续和长距离的单词交互，提出 TextING 分类模型，通过图神经网络模型来学习局部结构的单词表示；针对 GCN（Graph Convolutional Networks，简称 GCN）模型采用对称的拉普拉斯矩阵，不能直接用于有向图，无法为每个邻居分配不同的权重问题，Velikovi[5] 提出图注意力网络 GAT，为不同的节点分配不同的权重，同时训练时不依赖具体的网络结构，该结构只依赖成对的相邻节点，可以很好地解决 GCN 的缺陷。

通过观察 GAT 模型在真实和公开数据集的分类效果，可以看出对数据样本的均衡性有很高的要求。在实际应用中小样本和不均衡的文本数据分类的需求随处可见，导致 GAT 的分类效果难以体现。本文以主观作业的主题分类为例，针对作业规模较小、负主题样本较小的情况，提出采用均衡补偿方法进行少样本补偿的 BC-GAT 主题分类方法。通过公开和真实数据集的实验表明，主题分类效果得到了显著改善。

1　相关技术

本文采用基本 GAT 模型作为小样本文本数据主题分类器，通过前馈神经网络计算注意力系数，e_{ij} 表示 j 节点对 i 节点的注意力系数，计算公式(1)：

$$e_{ij} = \alpha([Wh_i; Wh_j]),\ j \in N_i \tag{1}$$

其中，N_i 表示 i 节点的邻居节点；$\propto$ 表示图注意力计算函数；h_i 表示输入层的节点特征；$\boldsymbol{W}$ 表示矩阵。

为了使系数在不同节点之间的相互比较，使用 softmax 函数将所有邻居节点的注意力系数归一化作为特征权重，通过将邻居节点加权求和的方式得到新的特征。节点 j 到 i 的注意力系数 α_{ij} 计算方式为

$$\alpha_{ij} = \frac{\exp(LeakyReLU(\alpha^{\mathrm{T}}[Wh_i \parallel Wh_j]))}{\sum_{k \in N_i} \exp(LeakyReLU(\alpha^{\mathrm{T}}[Wh_i \parallel Wh_j]))} \tag{2}$$

其中，$LeakyReLU$ 为激活函数，“$\parallel$”为向量拼接操作。

输出层节点 h_i' 计算方式为

$$h_i' = \sigma\left(\sum_{j \in N_i} \alpha_{ij} W h_j\right) \tag{3}$$

本文对数据增强采用 EDA 算法和网络爬虫两种方式，其中 EDA 算法主要采用同义词替换、随机插入、随机交换、随机删除 4 种方式[6]；而网络爬虫方法则利用 Requests 库中的 GET 方式向浏览器发出关键词搜索请求，获取相应的网页信息并对其进行过滤，利用 Lxml 网页解析器对网页信息进行解析从而实现文本扩充。

补偿样本关键词的提取方法采用 TextRank 算法，将文本抽象为词图模型记作 $G=(E,V)$，V 是由候选关键词矩阵组成的节点集，E 是利用共现窗口构建图中两节点之间的边，迭代计算每个顶点的权值，收敛时权值排名在前的点即为文本关键词。每个顶点权值的计算公式为

$$TR(v_i) = (1-d) + d \times \sum_{v_j \in \ln(V_i)} \frac{1}{Out(v_j)} \times TR(v_j) \tag{4}$$

其中，d 为阻尼系数；v_i 和 v_j 均为词语节点；$\ln(v_i)$ 是指向词语节点 v_i 的词语节点集合；$Out(v_j)$ 是词语节点 v_j 指向的词语节点集合。

2　BC-GAT 方法

本文提出一种基于 GAT 的小样本均衡补偿文本主题分类模型 BC-GAT（Balanced compensation-Graph Attention Network，简称 BC-GAT），旨在解决 GAT 模型针对数据倾斜样本分类结果倾向大比例样本集的问题。基本思想是对输入 GAT 模型中具有数据倾斜特征原始数据进行均衡补偿，在不引入干扰数据的前提下，有效提高小样本不均衡样本的分类正确率。

分析小样本数据集分类问题中存在的情况，给出如下解决方案：

(1)针对样本中特定类别极少现象，可以对小比例样本进行同源扩充或非同源扩充；

(2)针对样本中特定类别缺失现象，无法对类别缺失的样本进行同源扩充，只能非同源扩充。

BC 包含两种：同源扩充方法可以采用 EDA 算法直接进行数据增强，非同源扩充方法可以通过 TextRank 算法提取全部样本中靠后的关键词，通过关键词爬取相关文本内容并进行过滤，然后投入到欠均衡样本集，使其达到数据均衡的效果。

均衡补偿(BC)样本的具体方法描述如下：

算法 1 BC 算法

输入 不均衡的训练集和测试集

输出 均衡的训练集和测试集

(1)利用 TextRank 算法提取全部样本中排名靠后的关键词;

(2)针对样本特征选择同源扩充或非同源扩充方式,从而实现小比例样本的数据增强;

(3)将扩充后的样本并写入文本数据库。

对均衡补偿后的样本进行分词、去停用词等一系列预处理。把整个语料库转换为一个有向图 $G=(V,E)$。V 表示点的集合,即由单词与文档构成;E 为边,由词节点和词节点的权重,词节和文档节点的权重两部分构成。通过图注意力网络赋予节点相应的权重来获取节点之间的依赖信息,利用公式(1)计算节点的注意力系数;使用 softmax 函数将所有邻居节点的注意力系数归一化,利用公式(2)邻居节点的注意力系数,将所有邻居节点的注意力系数加权求和,得到新的特征;通过 Softmax 层获得文本类别的概率分布,并输出文档节点的类别标签。

3 实验与结果分析

3.1 数据集

本文实验使用以下两个数据集。

(1)MR 数据集。用于二元情感分类的电影评论数据集,每个评论只包含一句话,且每条评论都有情感正负标记,且正负样本数量绝对均衡;

(2)文本主观类型作业。选取辽宁师范大学 2015 级至 2019 级软件工程和软件工程设计两门课中的文本类型作业作为实验数据,包括随笔写作、需求描述、实验报告 1、实验报告 2 和综合型实验报告,其中每组作业在 40-60 份之间,合计 1 240 份作业。为了方便主题分类结果的实验对比,由 5 位评阅人对每份作业的主题贴合度进行评价,取平均值作为作业的实际主题标注结果,作业详情见表 1。

表 1 文本主观作业

Tab. 1 Text subjective assignment

名称	作业要求
随笔写作	对大学中学习和生活的反思感悟。
需求描述	软件产品的需求描述。
实验报告 1	需求分析的实验报告。
实验报告 2	软件过程实验报告。
综合型实验报告	针对坦克大战游戏的实验报告。

3.2 结果分析

本文实验采用 Pytorch 深度学习框架,Python 编程语言实现,分词工具采用 jieba,并使用 GPU 环境加速模型训练。其中 GAT 模型的学习率 0.02,dropout 率为 0.5,迭代次数为 100 次。评价指标采用准确率(*Percision*)、召回率(*Recall*)和 F 值。

3.2.1 MR 数据集实验结果与分析

为了验证 GAT 模型对小样本不均衡数据集的分类效果,选取 MR 数据集中 400 个正样本和 400 个负样本作为训练集,100 个正样本和 100 个负样本作为测试集;取 400 个正样本、40 个负样本作为训练集,100 个正样本、10 个负样本为测试集,并将上述两个样本集投入 GAT 模型进行分类,结果见表 2。

表 2 应用于 MR 数据集的 GAT 模型实验结果

Tab. 2 Experimental results of GAT model applied to MR data set

实验	样本类型(数量)	精确度	召回率	F 值
实验一	正样本(100)	0.666 7	0.640 0	0.653 1
	负样本(100)	0.653 8	0.680 0	0.666 7
	总样本(200)	0.660 3	0.660 0	0.659 9
实验二	正样本(100)	0.916 7	0.990 0	0.951 9
	负样本(10)	0.500 0	0.100 0	0.166 7
	总样本(110)	0.878 8	0.909 1	0.880 5

由表 2 可知,当正负样本分布比例为 1∶1 时,正样本、负样本和总样本的准确率、召回率和 F 值均在 60%以上。但当正负样本分布比例为 10∶1 时,负样本的召回率和精准率过低,即没有被正确识别。通过 GAT 在公开数据集的实验效果可以证明,GAT 模型虽然可以适用于小样本数据集的分类,但是无法解决由于样本数据不均衡导致的过拟合问题。这种不均衡现象主要体现在不同类别的样本数量上存在极大的差距,不均衡的数据集使模型难以达到对数据的最佳拟合,即少数量类别样本被误分到多数量类别样本中,即少数量类别样本没有办法被正确识别。

采用 EDA 算法对表 2 中实验二的小比例样本进行扩充,即将原始正负样本比例为 10∶1 的数据均衡成 1∶1,实验结果见表 3。

表 3 应用于 MR 数据集的 BC-GAT 方法实验结果

Tab. 3 Experimental results of BC-GAT model applied to MR data set

样本类型	精确度	召回率	F 值
正样本	0.892 5	0.830 0	0.860 1
增强后的负样本	0.824 7	0.888 9	0.855 6
总样本	0.860 4	0.857 9	0.858 0

由表 2 和表 3 可知,使用 GAT 模型训练正负样

本分布比例为 10∶1 的数据集时，尽管正样本和总样本的准确率、召回率和 F 值均在 0.8 以上，但是负样本的准确率、召回率和 F 值为 0.5，0.1 和 0.16，显然此时的实验结果并不可靠。因为数据集中最为关注的负样本并没有被正确识别，总样本识别准确率较高也是由于数据集中正样本被有效识别导致的。而使用 BC-GAT 方法进行训练之后，负样本的准确率、召回率、F 值分别为 0.82、0.88 和 0.85，负样本的准确率提升了 0.3，此时绝大多数的负样本被有效识别，证明 BC-GAT 方法中通过对负样本进行数据增强，使正负样本的数量达到平衡这一思想是有效的，可以解决小样本不均衡文本分类问题。

3.2.2　文本主观类型作业实验结果与分析

在实际应用中经常存在数据集类别分布极度不均衡的情况，以高校课堂中被广泛使用的文本主观作业为例，符合主题的作业一般占据大部分，不符合主题的作业往往占据小部分。本文对文本主观类型作业进行实验，以此证明模型的有效性。分别抽取某一年级的全部作业作为测试集，其他年级的作业作为训练集，并将其分别投入 GAT 模型和 BC-GAT 模型中进行实验，结果见表 4 和表 5。

表 4　随笔、实验报告 2 和综合型实验报告的识别准确率

Tab. 4　The identification accuracy rate of essay, experimental report 2 and comprehensive report

样本类型		GAT 模型			BC-GAT(EDA)			BC-GAT(网络爬虫)		
		精准率	召回率	F 值	精准率	召回率	F 值	精准率	召回率	F 值
随笔描述	正样本	0.969 2	1	0.984 4	0.927	1	0.962	0.954 5	1	0.976 7
	负样本	1	0.333 3	0.5	1	0.849	0.918	1	0.911 8	0.953 8
	总样本	0.970 6	0.969 7	0.962 4	0.952	0.948	0.947	0.970 5	0.969 1	0.968 7
实验报告 2	正样本	0.981 1	0.981 1	0.981 1	0.981 5	1	0.990 7	0.981 1	0.981 1	0.981 1
	负样本	0.5	0.5	0.5	1	0.95	0.974 4	0.972 2	0.972 2	0.972 2
	总样本	0.963 6	0.963 6	0.963 6	0.986 6	0.986 3	0.986 2	0.977 5	0.977 5	0.977 5
综合实验报告	正样本	0.982 1	1	0.991	1	0.981 8	0.990 8	1	0.981 8	0.990 8
	负样本	0	0	0	0.975 6	1	0.987 7	0.964 3	1	0.981 8
	总样本	0.964 6	0.982 1	0.973 3	0.989 7	0.989 5	0.989 5	0.982 2	0.987 8	0.987 9

表 5　需求描述和实验报告 1 的识别准确率

Tab. 5　The recognition accuracy of requirement description and experimental report 1

样本类型	BC-GAT(网络爬虫)		
	精准率	召回率	F 值
需求正样本	1	0.981 8	0.990 8
需求负样本	0.964 3	1	0.981 8
需求总样本	0.982 2	0.987 8	0.987 9
实验报告 1 正样本	0.981 8	1	0.990 8
实验报告 1 负样本	1	0.970 6	0.985 1
实验报告 1 总样本	0.988 8	0.988 6	0.988 6

由表 4 可知，当样本中存在特定类别极少现象时，无论是采用 EDA 算法还是网络爬虫算法对小比例类别样本进行均衡补偿，其负样本识别的准确率、召回率、F 值比采用 GAT 模型有着明显的提升。同时，BC-GAT 方法得出的正样本、负样本和总样本的准确率均在 0.9 以上，证明 BC-GAT 方法可以有效规避 GAT 模型在迭代过程中由于类别不均衡所导致的过拟合问题，可以作为小样本不均衡的文本主题分类方法。

两种均衡补偿的方法均有各自的优势和适用范围，EDA 算法简单有效，但是针对样本中存在特定类别缺失现象时，难以采用 EDA 算法将进行均衡补偿，可以采用网络爬虫算法。由表 5 可知，正样本、负样本和总样本的识别准确率均在 0.9 以上，该设定使模型即使在没有负样本的情况下也可以有效实现主题分类，更具有灵活性。

4　结束语

本文提出一种不均衡样本分类的优化模型 BC-GAT，通过对数据集中小比例类别样本进行均衡补偿的方式对 GAT 模型进行优化。通过公开和真实数据集的实验表明，BC-GAT 模型中正负样本识别准确率均在 0.9 以上，可以有效解决实际应用中存在的小样本且数据倾斜问题。后续的研究工作对 GAT 模型进行改进，从而提高 BC-GAT 模型的准确率。

（下转第 111 页）