ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 2,Mar. 2024,pp. 385－394
DOI:10. 16337/j. 1004-9037. 2024. 02. 011

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

一种半监督金融事件多标签分类方法

杨卓峰[1]，李　昉[2]，李德玉[1,3]

(1. 山西大学计算机与信息技术学院，太原 030006；2. 山西财经大学金融学院，太原 030006；3. 计算智能与中文信息处理教育部重点实验室(山西大学)，太原 030006)

摘　要：随着数字金融服务业的不断发展，互联网和金融服务系统积累了海量文本数据，对金融文本中描述的金融事件自动分类是金融科技的现实需求，也是自然语言处理和机器学习领域广泛关注的方向。目前，深度学习方法已在文本分类中广泛应用，针对文本数据中的金融事件多标签分类中存在的已标注数据缺少、已有深度学习方法消耗资源大以及现有方法未利用金融事件文本的具体特点等问题，通过采用ALBERT和TextCNN等表示工具，引入主体词注意力机制，提出了一种半监督金融事件多标签分类方法。首先，通过无监督数据增强(Unsupervised data augmentation，UDA)方法缓解标注数据量不足的问题；其次，引入了主体词注意力机制，使用ALBERT动态词向量表征方法对文本中的词进行表示；然后，利用TextCNN对文本进行综合语义表示；最后，分别采用交叉熵和KL散度度量标记数据和无标记数据的损失来训练模型。在金融文本数据集上验证了本文所提方法的有效性。

关键词：金融文本；金融事件；多标签分类；半监督方法；注意力机制

中图分类号：TP391　　**文献标志码**：A

Semi-supervised Multi-label Classification Method for Financial Events

YANG Zhuofeng[1], LI Yang[2], LI Deyu[1,3]

(1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China; 2. School of Finance, Shanxi University of Finance and Economics, Taiyuan 030006, China; 3. Key Laboratory of Computational Intelligence and Chinese Information Processing(Shanxi University), Ministry of Education, Taiyuan 030006, China)

Abstract: With the continuous development of the digital financial service industry, the Internet and financial service systems have accumulated a large amount of text data. The automatic classification of financial events described in the financial text is a realistic demand of financial technology, and also a widespread concern in the field of natural language processing and machine learning. At present, the deep learning method has been widely used in text classification. Addressing the issues of lack of labeled data in multi label classification of financial events in text data, frequent resource consumption of existing deep learning methods, and failure to explore the specific characteristics of financial event texts, a semi-supervised multi-label classification method of financial events is proposed by using ALBERT, TextCNN and other presentation tools, introducing the subject word attention mechanism. Firstly, the problem of insufficient labeled data is alleviated through unsupervised data augmentation (UDA) methods;

基金项目：国家自然科学基金(62106130, 62072294)；山西省青年科学研究项目(20210302124084)；山西省高等学校科技创新项目(2021L284)。
收稿日期：2023-06-28；**修订日期**：2023-09-30

Secondly, the subject word attention mechanism is introduced, and the ALBERT dynamic word vector representation method is used to represent the words in the text; Then, TextCNN is used to represent the text comprehensively; Finally, cross entropy and KL divergence are used to measure the loss of labeled data and unlabeled data to train the model. The effectiveness of the proposed method is verified on the financial text dataset.

Key words: financial text; financial event; multi-label classification; semi-supervised method; attention mechanism

引　言

随着数字技术与金融服务在产业上的深度融合,数字金融服务正在快速进入实体经济和每个人的日常生活。瞬息万变的金融市场产生了海量数据,这些数据蕴含大量有价值的信息[1]。金融文本分析中,对文本中描述的事件进行分类是一类典型的任务。过去,金融交易事件的类型通常由金融专家依据事件中发生的行为来确定。这种由专家标注事件类别的方式显然不能适应大数据背景,研究金融事件的自动方法具有现实迫切性。在同一个金融交易事件中,通常关联于多个事件类型标签,如在一些大型并购案中,同时具有“质押”“投资”和“收购”等事件类型标签[2]。因此,金融事件分类是一个多标签分类任务。传统的单标记分类任务局限于处理样本关联的单个标签的问题,因此对于金融事件类型的识别需要通过构建多标签分类方法来解决。准确识别金融事件相关的事件类型对于相关金融数据组织、管理与应用具有重要意义。

随着深度学习的快速发展,许多神经网络模型被用于文本分类任务[3]。Nam等[4]使用以循环神经网络(Recurrent neural network,RNN)模型为基础的序列到序列(Sequence-to-sequence,Seq2Seq)模型去建模,利用RNN的特性来产生标签序列进而学习标签之间的相关性。Yang等[5]提出了引入注意力机制的SGM模型,该模型以序列生成的思想建立标签之间的相关性进而解决多标签分类问题。Lin等[6]通过提出一种具有多级扩展卷积的高级语义单元表示,以及相应的混合注意力机制,用于处理不同层次的文本信息。Yang等[7]构建出一种分层注意力机制的网络(Hierarchical attention network,HAN),在单词和句子层面分别对文档进行建模。Yang等[8]提出一种序列到集合(Sequence-to-set,Seq2set)模型,利用深度强化学习设置框架的新序列,不仅捕获了标签之间的相关性,还减少了对标签顺序的依赖。BERT是自然语言处理领域的里程碑,Yarullin等[9]率先将BERT应用在多标签文本分类中,提出一种基于BERT的多标签文本分类模型。

现有的深度学习方法中,模型的训练往往需要大量的标注数据[10],而金融数据的标注往往需要依赖资深金融分析师的参与,增加了标注的人工成本和时间代价。由于数据量增长过快,导致仅有少量数据带有标签,这催生了许多半监督学习方法[11]。半监督学习方法使用无标记数据对仅从标记数据中获得的假设进行修改,以获得更精确的数据分布,构建更好的文本分类器。Jabreel等[12]提出了一种端到端神经网络模型,对输入Twitter文本自动进行多标签情绪分类,该模型由数据驱动,不依赖于外部资源(如,词性标注器和情绪或情感词典)。Zhou等[13]提出了一种半监督文本分类框架,该框架将多头注意力机制与操作风险分类的半监督变分推理(Semi-supervised object recognition and classification,Semi-ORC)相结合。Xie等[14]提出了一种一致性无监督数据增强(Unsupervised data augmentation,UDA)方法,仅利用少量的监督数据便可以达到很好的分类性能。已有的多标签文本分类方法往往专注于研究标签和文本间的关系。而金融文本有其自身的一些特点,大量的金融交易事件往往伴随着不同的交易事件主体,例如“控股股东南通综艺投资有限公司于2019年10月30日将5 000万股进行质押”。在该事件中,事件的主体词“南通综艺投资有限公司”会引出事件的动作“质押”,而事件类型又往往由事件主

体产生的动作决定。因此,针对金融领域中带标签文本数据缺乏问题,对于金融事件识别与分类任务,需要结合金融文本的特点解决以下3个问题:(1)如何从金融文本中充分捕获细粒度的文本语义信息;(2)如何结合金融文本中的特定主体词特点对文本进行更好的语义表示;(3)如何高效利用无标签数据来增强金融事件的识别。为此,本文提出了一种基于半监督学习的金融事件多标签分类方法(Multi-label classification financial text semi-supervised learning,MLC-FTSSL)。

1 金融文本多标签分类方法

给定金融文本数据集$D=\{x^{(i)},y^{(i)}\}_{i=1}^{N}$和标签集$L_s=\{l_1,l_2,\cdots,l_q\}$,其中,$x^{(i)}$表示第$i$个文档,$y^{(i)}$表示该文档对应的标签集。本文提出基于半监督学习的金融文本多标签分类方法,主要包含基于ALBERT的词向量表示、主体词注意力、文本的TextCNN表示、一致性无监督数据增强、模型参数更新以及分类器构建等6个部分,整体框架如图1所示。图中:E、T、U、B、M、C等都在后文进行说明,其中Trm表示transformer内部多层双向transformer机制;Logit表示标签概率;KL表示KL离散度;sigmoid_cross_ entropy表示交叉熵损失;Loss_sup/Loss_un sup以及Loss_total都表示损失函数。

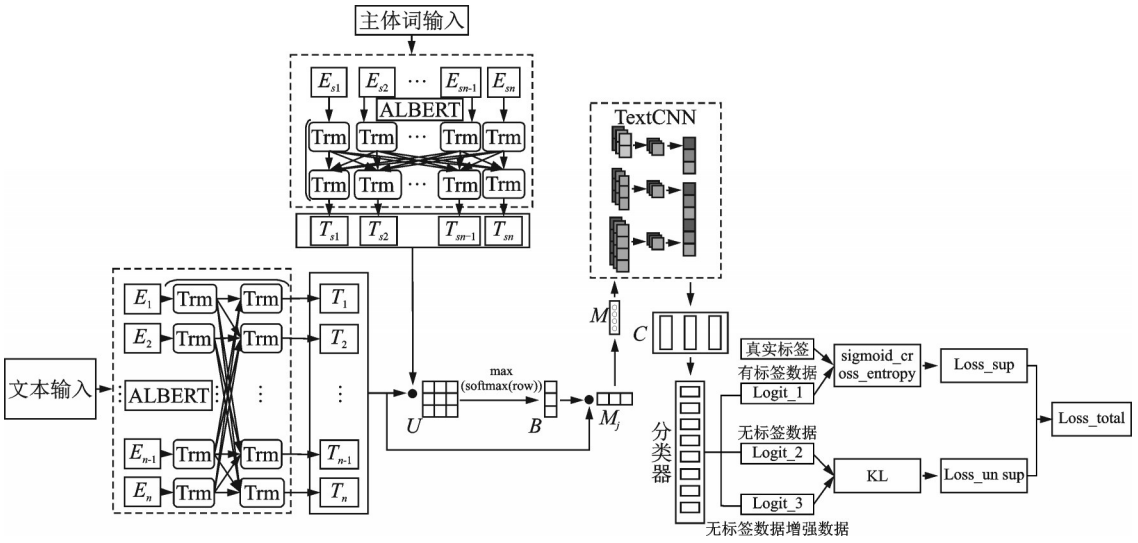

图1 MLC-FTSSL框架示意图

Fig.1 Architecture of MLC-FTSSL

1.1 基于ALBERT的词向量表示

ALBERT是在原始BERT模型[15]的基础上,利用一些新的策略,如:跨层参数共享、嵌入层参数因式分解等[16],减少了模型训练所需的参数总量从而使模型训练速度大大提升。进一步,ALBERT改用了更加有效的句子顺序预测任务,并通过实验测试删除Dropout层,使ALBERT相对BERT参数量更少而达到的效果相当。多头自注意力机制[17]是ALBERT最重要的组成模块之一。通过注意力机制计算句子中每个字词与该句子中所有字词的相互关系,并由此来调整每个字词在句中的权重,从而获得新的文本序列向量表示。由于金融文本具有较强的领域特性,依赖于上下文中的词间信息,因此本文的词向量的表示主要采取ALBERT预训练语言模型。给定一个金融文本$\boldsymbol{X}=[x_1,x_2,\cdots,x_n]$,其中$x_i$表示金融文本中第$i$个词。通过ALBERT输入层对金融文本进行处理后,得到金融文本的序列化嵌入表示为$\boldsymbol{E}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n]$,其中$\boldsymbol{e}_i$表示文本中第$i$个字的序列化字符的嵌入表示。之后序列化嵌入表示文本向量经过ALBERT内多层双向Transformer编码器的训练,得到文本结合上下文语义等的向量表

示 $T=[t_1, t_2, \cdots, t_n] \in \mathbf{R}^{n\times d}$，其中 t_i 表示第 i 个字符经 Transformer 提取后的特征向量表示，d 表示金融文本中每个词对应的词向量维度。

通过对每个金融文本中不同事件的主体词进行汇总，可以得到该文本所对应的一个或多个事件主体词。正是由于不同的事件主体词引出不同的事件动作最终决定了不同的事件类型，所以本文将金融文本中的多个主体词同样通过 ALBERT 输入层对其进行序列化处理，得到主体词文本的序列化嵌入表示 $E_s=\left[e_{s_1}, e_{s_2}, \cdots, e_{s_m}\right]$，其中 e_{s_i} 表示主体词文本中第 i 个字的序列化字符的嵌入表示。之后主体词序列化文本向量同样经过 ALBERT 内多头自注意力机制得到结合前后文语义关系的向量表示 $T_s=[t_{s_1}, t_{s_2}, \cdots, t_{s_m}] \in \mathbf{R}^{m\times d}$，其中 t_{s_i} 表示主体词中第 i 个字符的特征向量表示。

1.2 主体词注意力

不同的金融事件类型往往由事件中不同的动作决定，而不同的事件动作又由事件主体引出。因此，为了更好地从金融文本中提取事件类型区别性信息，需要构建一种特殊的注意力机制，用来学习文本中其他词与主体词间的关联程度，以提升文本综合语义表示效果。

将金融文本的向量表示 $T\in\mathbf{R}^{n\times d}$ 和主体词的向量表示 $T_s\in\mathbf{R}^{m\times d}$ 进行点积，得到匹配矩阵 $U\in\mathbf{R}^{n\times m}$，即

$$U=TT_s^{\mathrm{T}} \tag{1}$$

将 U 按行进行 softmax 运算，得到每个主体词关于金融文本中所有词的相关分布 $Q\in\mathbf{R}^{\mathrm{n}\times\mathrm{m}}$，其第 i 行表示为

$$Q_i=\mathrm{softmax}(U_{i1}, U_{i2}, \cdots, U_{in}) \tag{2}$$

将 Q 按行进行 max 运算，将所有的行最大值对应的词嵌入进行拼接，得到最大相关矩阵 $B\in\mathbf{R}^{n\times 1}$ 为

$$B=\max(Q_1)\oplus\max(Q_2)\oplus\cdots\oplus\max(Q_n) \tag{3}$$

将 B 与金融文本的向量表示 T 进行点积，得到与主体词 S 相关的一个金融文本向量 $M\in\mathbf{R}^{1\times d}$ 为

$$M=B^{\mathrm{T}}T \tag{4}$$

同样地，对所有金融文本执行上述过程，得到 m 个融合主体词信息的金融文本表示，将其拼接可得到结合主体词信息的金融文本集综合语义表示

$$M=[M_1\oplus M_2\oplus\cdots\oplus M_m] \tag{5}$$

1.3 文本的 TextCNN 表示

TextCNN 卷积网络模型[18]内部有输入层、卷积层、池化层和输出层 4 部分，将长度为 n 的文本输入到卷积层中，卷积层通过 k 个不同大小的滑动窗口对文本向量进行卷积操作，在位置 i 处通过卷积核得到卷积特征值来学习文本特征为

$$S_i=f(W\cdot M_{i:i+k-1}+b) \tag{6}$$

式中：$W\in\mathbf{R}^{kd}$ 表示卷积核；$M_{i:i+k-1}$ 代表滑动窗口的大小是由输入矩阵第 i 行到第 $i+k-1$ 行；$f(\cdot)$ 表示非线性映射函数；b 为偏置参数。在池化层当中应用 1-MaxPooling 最大池化策略，即从经卷积操作后得到的卷积向量中筛选出一个最大特征值。

$$c_i=\max\{S_i\} \tag{7}$$

最终，通过连接层将所有池化之后的特征值拼接得到文本高层特征向量 $C\in\mathbf{R}^{n-k+1}$，表示为

$$C=[c_1, c_2, \cdots, c_{n-k+1}] \tag{8}$$

式中 n 表示文本序列中字的数量。

对金融文本的观察发现，其中的词汇长度在 1～4 字之间，所以设置 3 种大小的卷积核（token 个数分别为 2、3、4）。对 TextCNN 的训练采用双通道训练方式，即有两个输入矩阵，其中一个用已训练好的词

嵌入表示,且在训练过程中不再发生变化,另外一个也由同样的方式初始化,但会作为参数随着网络的训练过程发生改变。通过双通道训练方式,预先训练的词嵌入利用其他语料库得到了更多的先验知识,而由当前网络训练的词向量则能更好地抓住与当前任务相关的特征。

1.4 一致性无监督数据增强

深度学习方法通常需要大量标记数据才能奏效。半监督学习是利用无标签数据来缓解标记数据缺乏的有效途径之一。一致性UDA[14]是目前深度学习常用的一种半监督方法。该方法基于平滑假设和聚类假设,即具有不同标签的数据点在低密度区域分离,且相似的数据点具有相似的输出(一致性)。这个假设隐含了一种提高模型泛化性能的方法,即通过在无标记数据上构造添加扰动后的预测结果$\tilde{y}$与正常预测结果y之间的无监督正则化损失项来提高模型的泛化能力。对文本分类问题,UDA则采用监督学习中比较优秀的数据增强方法"回译"来代替传统的噪声扰动。回译即将A语言中已有示例x翻译成另一种语言B,然后再翻译回A语言,以获得增强的示例$\hat{x}$。反向翻译可以在保留原句语义的同时生成不同的意译,从而显著提高回答问题的能力。

用x来代表输入文本,用$y(x)$来代表其真实预测标签。本文任务是学习一个模型$p_\theta(y|x)$,这里θ代表模型的参数。用L表示有标签数据,U表示无标签数据。设$\hat{x}$是通过x回译法得到的数据,由一致性假设,$\hat{x}$和x享有共同的真实标签,即$y(x)=y(\hat{x})$。

为了增加语言表达的多样性,本文采用对原始文本翻译为法语,之后翻译为英语,最后翻译回中文来进行数据的"回译"增强。极小化有标签数据和对应的无标签增强数据模型输出间的KL散度。目标函数为

$$\min_\theta J_{\mathrm{UDA}}(\theta)=\mathop{E}_{x\in U}\mathop{E}_{\hat{x}\sim q(\hat{x}|x)}\left[D_{\mathrm{KL}}(p_\theta(y|x)\|p_\theta(y|\hat{x}))\right] \tag{9}$$

式中:$q(\hat{x}|x)$表示无标签数据$\boldsymbol{U}$对应的增强数据;D_{KL}为KL散度。

训练过程中,用交叉熵来度量模型在标签数据上的损失,用KL离度来度量模型在无标签数据上的损失。整体的损失函数为

$$\min_\theta J=\mathop{E}_{x,y(x)\in L}\left[p_\theta(y(x)|x)\right]+\lambda J_{\mathrm{UDA}}(\theta) \tag{10}$$

式中参数λ用于平衡两种损失间的权重,通常设置$\lambda=1$。

1.5 模型参数更新

模型参数更新主要指对ALBERT和TextCNN的参数进行更新。前向传播采用交叉熵损失[19],即

$$H(p,q)=-\sum_{i=1}^{n}p(x_i)\ln q(x_i) \tag{11}$$

式中:$p(x_i)$表示文本标签的真实概率分布;$q(x_i)$表示文本标签的预测概率分布。反向传播时,将交叉熵向下反向传播到TextCNN,从而更新TextCNN模型的超参数$\boldsymbol{W}$、$\boldsymbol{b}$,有

$$\boldsymbol{W}\rightarrow\boldsymbol{W}+\Delta\boldsymbol{W},\boldsymbol{b}\rightarrow\boldsymbol{b}+\Delta\boldsymbol{b} \tag{12}$$

$$\Delta\boldsymbol{W}=\frac{\partial G(\boldsymbol{W},\boldsymbol{b})}{\boldsymbol{W}},\Delta\boldsymbol{b}=\frac{\partial G(\boldsymbol{W},\boldsymbol{b})}{\boldsymbol{b}} \tag{13}$$

式中$G(\boldsymbol{W},\boldsymbol{b})$表示所采用的损失函数。

1.6 分类器

金融文本中的事件往往关联多个类别,为了提高模型分类的准确率,本文设计具有两层全连接层的多层感知器构建分类器。它将金融文本的特征向量进行线性变换,使用sigmoid激活函数将模型输出值转换为输出概率,用于表示某个金融文本具有每个事件类型标签的预测概率,有

$$\hat{\boldsymbol{y}}=\mathrm{sigmoid}\left(\boldsymbol{W}_2\cdot f(\boldsymbol{W}_1\cdot\boldsymbol{V}^{\mathrm{T}})\right) \tag{14}$$

式中：$\boldsymbol{W}_1 \in \mathbf{R}^{b\times 2k}$、$\boldsymbol{W}_2 \in \mathbf{R}^{b}$分别表示全连接层和输出层的可训练参数；$f$表示ReLU线性激活函数；$\boldsymbol{V}$为权重矩阵。

在多标签分类任务中，交叉熵损失变形为

$$\text{Loss} = -\sum_{i=1}^{N}\sum_{j=1}^{L}\left(y_{ij}\ln\hat{y}_{ij} + (1-y_{ij})\ln(1-\hat{y}_{ij})\right) \tag{15}$$

式中：N表示训练文本数；L表示文本对应的标签数；$\hat{y}_{ij}\in[0,1]$表示第i个文本的第j个标签的预测概率；$y_{ij}\in\{0,1\}$表示第i个文本是否具有第j个标签的真实情况，“1”表示“是”，“0”表示“否”。

2 实验设置与分析

2.1 实验参数

实验在CPUInter Core i5-10750H和GPUNVIDIA 2080Ti的基础上，基于Python3.6.9和tensorflow1.15.0搭建的深度学习框架下完成。在所有的数据集上，批次大小设置为128，单词嵌入大小设置为512，标签向量大小设置为32，学习率为0.000 05，epoch次数为50。在ALBERT中编码器和解码器中设置隐藏层神经单元个数为4 096，模型的激活函数选择Gelu。在TextCNN模型中，网络层数为1，隐藏层大小为128，激活函数选择Relu，Dropout率为0.1。

2.2 实验数据及评价指标

本文的数据来源于中国新闻网、人民网和新华网等各大新闻网站近年来的热点金融新闻，共收集金融新闻3万条，其中1.5万条由专家人工标注金融事件类别，剩余1.5万条为无标签数据。将有标签的数据划分为训练集1万条、验证集0.25万条和测试集0.25万条。数据的具体形式如表1所示。

表1 数据具体形式

Table1 Concrete form of data

文本内容	事件类型
兴发集团发布公告，控股股东宜昌兴发集团有限责任公司于2019年11月20日将2 000万股进行质押，质押方为上海浦东发展银行股份有限公司宜昌分行，质押股数占其所持股份比例的8.50%，占公司总股本的2.15%。	质押
104 950 006股(占本公司总股本比例11.53%)的股东烟台恒邦集团有限公司(以下简称“恒邦集团”)计划自本公告披露之日起15个交易日后的6个月内以集中竞价交易或大宗交易的方式减持公司股份不超过27 312 000股，即不超过公司总股本3%。	减持、股份股权转让
没想到2016年10月阿里突然宣布收购好莱坞导演史蒂文·斯皮尔伯格的公司Amblin Partners部分股权，并以联合投资形式参与到Amblin以后的电影项目中。	收购、股份股权转让、投资

每个金融新闻文本具有一个或者多个标签，数据标签个数分布和标签类别数量分布情况统计分别如表2和表3所示。文本采用Precision、Recall和Micro-F_1三项评价指标来衡量模型性能，表达式分别为

表2 样本标签个数分布

Table 2 Sample label number distribution

标签个数	$L=1$	$L=2$	$L=3$
样本数量	4 369	3 190	2 441

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{16}$$

表3 样本类别分布

Table 3 Sample class distribution

标签	质押	股份股权转让	投资	减持	起诉	收购	判决	签署合同	担保	中标
样本数量	3 314	2 778	2 496	3 445	972	1 078	541	697	364	178

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{17}$$

$$\text{Micro-}F_1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{18}$$

式中:TP(True positive)表示被正确分类的正例的数量;FP(False positive)表示负例被错误预测为正例的数量;FN(False negative)表示正例被错误预测为负例的数量。Precision为精确率,表示预测正确的正例样本数量占所有预测结果为正例的样本的比例。Recall为召回率,表示预测正确的正例样本数量占所有真实结果为正例的样本的比例。Micro-F_1先计算所有类别总的精确率和召回率之后再计算F_1值,是一种用于不平衡测试集的重要指标,考虑了所有标签的整体精确率和召回率,是两者的调和平均值。

2.3 对比实验设置

本文选择如下8种方法进行对比实验,所有实验方法均采用原文代码。

(1)Seq2seq-RNN[4]。该方法利用递归神经网络最大化多标签分类中的子集精度,通过降低标签频率将每个子集映射到一个序列,并用用于序列预测的RNN解决多标签任务。

(2)SGM+GE[5]。该方法训练类似于seq2seq-RNN的RNN模型,但使用一种新的全局解码器结构(Global embedding,GE),该结构基于所有标签计算加权全局嵌入,而不是在每个时间步中只计算顶部标签。

(3)SU4MLC[6]。该方法称为基于语义单元的多标签文本分类(Semantic-unit-based dilated convolution for multi-label text classification),提出了一种基于序列到序列学习的多标签文本分类模型。该模型生成具有多级扩展卷积的高级语义单元表示,以及相应的混合注意力机制,并利用该机制提取单词级别和语义单元级别的信息。

(4)Hierarchical seq2seq[20]。该方法提出了一种新的用于多标签文本分类的序列到序列模型,基于“并行编码,串行解码”策略。该模型结合了卷积神经网络和并行自关注作为编码器提取细粒度局部邻域信息和全局来自源文本的交互信息。

(5)Seq2set[8]。该方法提出了一种利用深度强化学习设置框架的新序列,利用集合的无序性不仅捕获了标签之间的相关性,还减少了对标签顺序的依赖。

(6)LSAN[21]。该方法称为特定标签的注意力网络(Label-specific attention network),提出了一个标签特定注意网络来学习新的文档表示。利用标签语义信息构建标签特定文档表示的标签和文档。利用自注意机制识别标签特定的文档表示。设计了自适应融合策略可以有效地集成上述两者,并输出压缩的文档表示来构建多标签文本分类器。

(7)MLC-LWL[22]。该方法提出了一种多标签文本分类模型,结合了潜在的逐词标签信息。该模型应用标注主题模型构建有效的逐词标注信息,并通过门控网络将词所携带的标注信息与标注上下文信息相结合。

(8)LELC[23]。该方法提出了一种基于多层注意机制和标签相关性的多标签文本分类模型,共同学习标签嵌入和标签之间的关联,并利用注意机制选择标签相关特征。

2.4 对比方法实验结果

为保证实验结果的稳定性,采用10次实验的均值,结果如表4所示,其中最优的结果以黑体字体高

亮标出，结果柱状图如图2所示。由实验结果可知：

（1）MLC-FTSSL的综合指标值优于其他所有方法，说明了MLC-FTSSL对金融文本多标签分类的有效性。AlBERT作为整体模型的基线模型，融合了多模块的MLC-FTSSL模型与其对比发现，最终显著提升了分类性能，Micro-F_1值提升了0.055。Seq2seq-RNN、SGM＋GE、SU4MLC和Hierarchical seq2seq等方法大多采用序列生成的思想来解决多标签分类问题，往往会由于后一个标签依赖于前一个标签导致标签错误传播，从而造成模型分类性能下降。

（2）本文模型MLC-FTSSL相比Seq2seq-RNN、SGM＋GE、SU4MLC和Hierarchical seq2seq方法，Micro-F_1最高提升了0.068，最低提升了0.02，这是由于本文方法用ALBERT和TextCNN对文本进行动态语义表示，在很大程度上缓解了序列依赖这一问题，使模型性能得到提升。

（3）在LSAN、MCL-LWL、LELC中学习标签和文本的联系会使模型整体的Micro-F_1值有所提升，说明通过学习标签的语义表示并有效地集成文档和标签的语义可以提高模型的分类性能。本文模型MLC-FTSSL相比以上3种方法均有所提升。通过本文对金融文本的具体特性进行分析发现，不同的金融事件类型倾向于由事件主题词直接决定，所以本文构造了金融特定领域的金融文本主体词注意力机制，实验结果也展示了该模块可以帮助模型更好地学习金融文本语义表示，从而提升模型的分类效果。

表4 对比方法实验结果

Table 4 Comparison of experimental results

Method	Precision	Recall	Micro-F_1
Albert	0.837 3	0.745 8	0.788 9
Seq2seq-RNN	0.823 9	0.734 1	0.776 4
SGM＋GE	0.844 6	0.756 8	0.791 3
SU4MLC	0.839 2	0.744 9	0.789 2
Hierarchical seq2seq	0.853 8	0.764 3	0.806 5
Seq2set	0.864 6	0.787 1	0.824 1
LSAN	0.860 1	0.769 5	0.812 3
MLC-LWL	0.867 3	0.787 9	0.826 4
LELC	0.871 1	0.788 7	0.827 9
MLC-FTSSL	**0.886 7**	**0.805 1**	**0.843 9**

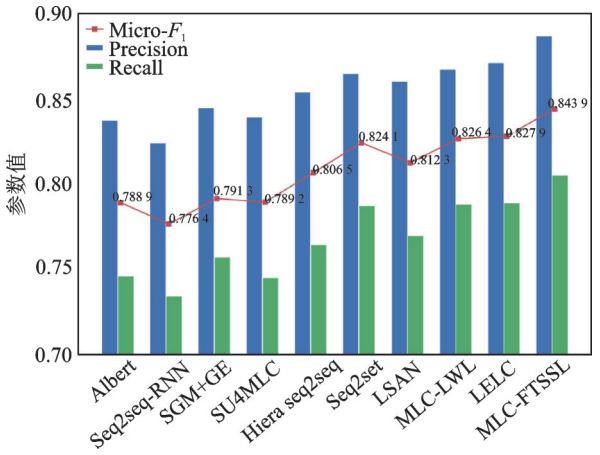

图2 对比方法实验结果柱状图

Fig.2 Bar chart of comparison of experimental results

2.5 消融实验

本文模型MLC-FTSSL主要是由4个相互协同工作的关键模块构成，包括ALBERT预训练模型（ALBERT）、基于金融文本主体词的注意力模块（SubAtt）、TextCNN预训练模型（TextCNN）以及一致性无监督数据增强方法（UDA），本节通过消融实验来验证每个模块的效果。

本文分别在数据集上测试UDA-主体词注意力机制-Word2vec-TextCNN（MLC-FTSSL-ALBERT）、UDA-主体词注意力机制-ALBERT（MLC-FTSSL-TextCNN）、UDA-ALBERT-TextCNN（MLC-FTSSL-SubAtt）、主体词注意力机制-ALBERT-TextCNN（MLC-FTSSL-UDA）、UDA-主体词注意力机制-ALBERT-TextCNN（MLC-FTSSL）等方案的分类性能，实验结果如表5所示，其柱状图如图3所示。

由实验结果可以看出，MLC-FTSSL模型相较于MLC-FTSSL-ALBERT模型，Micro-F_1值提升了0.034 7，可见使用ALBERT动态词向量表征方法优于word2vec静态词向量表征方法，显著提升了分类器的性能。与MLC-FTSSL-TextCNN模型对比，发现在引入TextCNN模块对最终文本进行高维语义表示后对模型的分类性能也具有一定的提升。MLC-FTSSL-SubAtt模型与MLC-FTSSL模型对比，

Micro-F_1值降低了0.013 2,去掉主体词注意力机制后模型整体分类性能有所下降,说明金融领域主体词对金融文本的深层表示起着重要的作用。通过学习特定领域的潜在特征表示,可以提升特定领域的下游任务性能。对比MLC-FTSSL模型与MLC-FTSSL-UDA模型,Micro-F_1值提升了0.026 7,可见引入无监督数据增强方法UDA,通过对无监督数据以及无监督数据的增强数据进行一致性训练,可以学习到无监督数据中的信息,也正是由于大量无监督数据信息的学习从而使得模型的效果有明显的提升。

消融实验表明,AlBERT、Text-CNN、SubAtt和UDA四个模块对于分类性能的提升起到了不同的作用,本文模型MLC-FTSSL结合了4个模块的优势,构建了一个多模块层次递进、相互配合的特定领域的多标签分类模型,最终在已有的金融文本数据集上提升了多标签分类的性能,验证了模型的有效性。

表5 消融实验结果

Table 5 Comparison of ablation results

Method	Precision	Recall	Micro-F_1
MLC-FTSSL-ALBERT	0.854 6	0.768 5	0.809 2
MLC-FTSSL-TextCNN	0.875 5	0.798 3	0.835 1
MLC-FTSSL-SubAtt	0.873 9	0.791 7	0.830 7
MLC-FTSSL-UDA	0.858 1	0.780 1	0.817 2
MLC-FTSSL	0.886 7	0.805 1	0.843 9

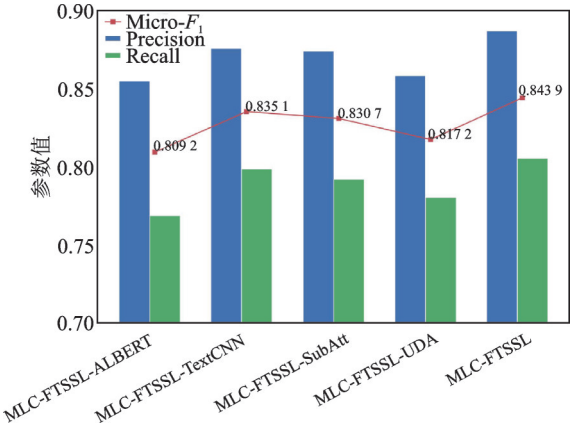

图3 消融实验结果柱状图

Fig.3 Bar chart of comparison of ablation results

3 结束语

本文提出了一种针对金融文本的事件多标签分类方法。为了缓解有标签的训练数据量少的问题,引入了UDA方法;为缓解BERT资源消耗过大的问题,引入了ALBERT动态词向量表示;为在文本表示中强化事件主体词信息,设计了主体词注意力机制,并与ALBERT结合进行文本综合语义表征;通过采用轻量化模型TextCNN,最大程度地保留不同抽象层次的语义信息,最终学习到文本的高层特征语义表示,从而更有效地对金融事件进行多标签分类。在已有的金融文本数据集上的实验结果验证了本文方法的有效性。对金融文本数据的观察发现,在金融文本事件存在多个标签时,观察不同标签在文本中的位置会发现语义关联较强的标签往往在文本中距离较近,同时多个共同出现的不同事件标签之间往往具有层次关系或隶属关系。因此下一步工作中将会探索不同标签之间的关联关系,进一步改善模型的性能。

参考文献:

[1] OUSSOUS A, BENJELLOUN F Z, LAHCEN A A, et al. Big data technologies: A survey[J]. Journal of King Saud University-Computer and Information Sciences, 2018, 30(4): 431-448.

[2] SHENG J, GUO S, YU B, et al. CasEE: A joint learning framework with cascade decoding for overlapping event extraction [EB/OL]. (2021-07-04) [2023-05-10]. https://doi.org/10.48550/arXiv.2107.01583.

[3] LIU W, SHEN X, WANG H, et al. The emerging trends of multi-label learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(11): 7955-7974.

[4] NAM J, MENCIA E L, KIM H J, et al. Maximizing subset accuracy with recurrent neural networks in multi-label classification [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2017: 5419-5429.

[5] YANG P, XU S, WEI L, et al. SGM: Sequence generation model for multi-label classification[C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: [s.n.], 2018: 3915-3926.
[6] LIN J Y, SU Q, YANG P C, et al. Semantic-unit-based dilated convolution for multi-label text classification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2018: 4554-4564.
[7] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, Callifornia, USA: [s.n.], 2016: 1480-1489.
[8] YANG P, MA S, ZHANG Y, et al. A deep reinforced sequence-to-set model for multi-label text classification[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: [s.n.], 2018: 5252-5258.
[9] YARULLIN R, SERDYUKOV P. BERT for Sequence-to-sequence multi-label text classification[M]. Berlin, Germany: Springer, 2021: 187-198.
[10] CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for chinese natural language processing [EB/OL]. (2020-04-29) [2023-05-10]. https://doi.org/10.48550/arXiv.2004.13922.
[11] 武红鑫, 韩萌, 陈志强, 等. 监督和半监督学习下的多标签分类综述[J]. 计算机科学, 2022, 49(8): 12-25.
WU Hongxin, HAN Meng, CHEN Zhiqiang, et al. Overview of supervised and semi supervised learning based multi label classification[J]. Computer Science, 2022, 49(8): 12-25.
[12] JABREEL M, MORENO A. A deep learning-based approach for multi-label emotion classification in tweets[J]. Applied Sciences, 2019, 9(6): 1123.
[13] ZHOU F, ZHANG S, YANG Y. Interpretable operational risk classification with semi-supervised variational autoencoder[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.]: ACL, 2020: 846-852.
[14] XIE Q, DAI Z, HOVY E, et al. Unsupervised data augmentation for consistency training[J]. Advances in Neural Information Processing Systems, 2020, 33: 6256-6268.
[15] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2018-10-11)[2023-05-10]. https://doi.org/10.48550/arXiv.1810.04805.
[16] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: A lite BERT for self-supervised learning of language representations [EB/OL]. (2019-9-26)[2023-05-10]. https://doi.org/10.48550/arXiv.1909.11942.
[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2017: 6000-6010.
[18] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. (2014-08-25) [2023-05-11]. https://doi. org/10.48550/arXiv.1408.5882.
[19] NAM J, KIM J, EL MENCÍA, et al. Large-scale multi-label text classification-revisiting neural networks[C]//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Germany: Springer, 2014: 437-452.
[20] YANG Z, LIU G. Hierarchical sequence-to-sequence model for multi-label text classification[J]. IEEE Access, 2019, 7: 153012-153020.
[21] XIAO L, HUANG X, CHEN B L, et al. Label-specific document representation for multi-label text classification[C]//Proceedings of the Conference on Empirical Methods in Nature Language Processing and the 9th International Joint Conference on Natural Language Processing. USA: ACL, 2019: 466-475.
[22] CHEN Z, REN J. Multi-label text classification with latent word-wise label information[J]. Appl Intell, 2021, 51(2): 966-979.
[23] LIU H T, CHEN G, LI P P, et al, Multi-label text classification via joint learning from label embedding and label correlation [J]. Neurocomputing, 2021, 460: 385-398.

作者简介:

杨卓峰(1999-),男,硕士研究生,研究方向:文本挖掘和自然语言处理,E-mail: 17303431234@163.com。

李旸(1988-),女,博士,副教授,研究方向:自然语言处理、机器学习等,E-mail: liyangprimrose@163.com。

李德玉(1965-),**通信作者**,男,博士,教授,CCF高级会员,研究方向:数据挖掘,E-mail: lidy@sxu. edu. cn。

(编辑:刘彦东)