计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900053

融合文本序列和图信息的海关商品 HS 编码分类

杜少华[1]　万怀宇[1]　武志昊[1,2]　林友芳[1,2]
1 北京交通大学计算机与信息技术学院　北京 100044
2 综合交通运输大数据应用技术交通运输行业重点实验室　北京 100044
(18120357@bjtu.edu.cn)

摘　要　海关商品 HS 编码分类是企业和个人进出口贸易的重要国际程序。HS 编码分类可以看作是一个文本分类问题,即给定一段商品的描述,确定商品由 HS 编码表示的所属类别。然而,该任务比一般的文本分类任务更具挑战性,原因是商品描述文本具有特定的层次结构,同时商品描述文本展现出了两个层次上的序列特征,并且商品描述文本还存在关键信息分散且描述形式多样的特点。现有的文本分类方法无法综合考虑以上因素来捕获商品描述文本中的关键信息。对此,文中提出了一种融合文本序列和图信息的神经网络(Text Sequence and Graph Information combination Neural Network,TSGINN)模型,用于解决海关商品 HS 编码分类问题。TSGINN 将 HS 编码分类问题定义为基于词共现网络的子图分类问题,通过图注意力网络建模非连续词之间的关联关系,同时利用分层的长短期记忆网络结合商品文本层次结构捕获多层次的序列信息。在真实海关商品数据集上进行了实验,结果表明 TSGINN 模型的 HS 编码分类效果优于其他分类方法。

关键词:海关商品;HS 编码;文本分类;多层次序列信息;图注意力网络
中图法分类号　TP391

Customs Commodity HS Code Classification Integrating Text Sequence and Graph Information

DU Shao-hua[1],WAN Huai-yu[1],WU Zhi-hao[1,2] and LIN You-fang[1,2]
1 School of Computer and Information Technology,Beijing Jiaotong University,Beijing 100044,China
2 Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Beijing 100044,China

Abstract　Customs commodity HS code classification is an important international procedure for cross-border trade of enterprises and individuals. HS code classification can be regarded as a text classification problem,that is,given a paragraph of description for a commodity,to determine the category of the commodity represented by HS code. However,this task is more challenging than general text classification task. First,commodity description texts are organized with special hierarchical structures. Then commodity description texts present sequential features at two levels. In addition,the key information in the commodity description text is scattered and the description forms are diverse. Most of the existing classification methods cannot comprehensively consider the above factors to capture key information in the commodity description text. In this paper,we proposes a Text Sequence and Graph Information combination Neural Network(TSGINN) to solve the problem of customs commodity HS code classification. The TSGINN defines the HS code classification problem as a subgraph classification problem based on word co-occurrence network,models association between non-contiguous words through graph attention network,and captures multi-level sequential information through hierarchical long short-term memory network. Experiments on the real-world customs datasets show that the classification effect of TSGINN model is better than that of other methods.

Keywords　Customs commodity,HS code,Text classification,Multi-level sequential information,Graph attention network

1　引言

随着跨境贸易的快速增长,对通关商品快速、准确地分类越来越受到相关部门的重视。世界海关组织制定了协调制度(Harmonized System,HS)编码,以数字编码的形式代表和识别跨境贸易的货物。海关商品 HS 编码分类指根据商品描述信息为待分类商品找到最准确的 10 位 HS 类别代码的过程。如果能够高效、准确地实现 HS 编码分类,将有助于海关部门计算关税、贸易统计等工作的顺利进行,也可以帮助企业避免因不规范通关而导致的装运延误、行政处罚等问题,从而加速清关。

商品描述信息是由一系列能够反映商品客观情况的要素组成的文本,这些要素被统称为申报要素。如图 1 所示,某商品描述文本是由商品名称、用途、固定安装配置等一系列的申报要素组成的,各要素之间用"|"分隔,图 1 中商品对应的 10 位 HS 编码是 8903990090,表明该商品属于娱乐用的船舶。

到稿日期:2020-09-07　返修日期:2020-10-04　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:万怀宇(hywan@bjtu.edu.cn)

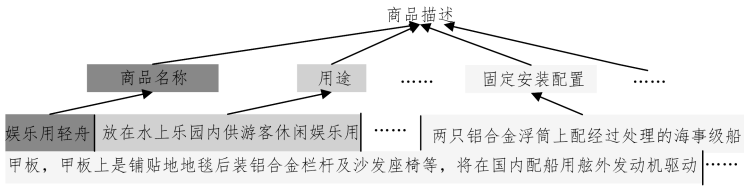

图 1 商品描述信息示例

Fig. 1 Example of commodity description

HS 编码分类本质上是一个文本形式的分类问题，与一般的文本分类问题相比，其难点在于：

(1)商品描述文本具有特定的层次结构，即由词组成申报要素，再由申报要素组成商品描述。因此，需要模型能够结合文本的层次结构进行分类。

(2)商品描述文本在两个层次上呈现出序列性。首先申报要素内部词语按照上下文语义顺序排列，其次多个申报要素之间按照领域规则顺序排列。这就需要模型能够结合商品描述文本的结构特点捕获文本中两个层次的序列信息。

(3)描述形式多样。如图 2 所示，两个描述信息差异很大的商品却拥有相同的 HS 编码。此外，商品描述中细微的差别也可能导致分类结果不同。即某些商品可能仅因某个关键指标不同而对应于完全不同的 HS 编码。因此，模型需要捕捉关键词并深入理解其含义才能正确分类。

(4)关键信息分散。例如，图 2 中的商品 1 中“摩擦片”“变速箱”“制动器”等关键词在描述信息中的位置并不紧邻甚至相隔较远，需要将这些分散的关键词联系起来，捕捉它们之间非连续的关联关系才能正确分类。

商品描述信息	HS编码
商品1：摩擦片|可用于多品牌多型号的小轿车 排量:1.4L至2.0L|4档以下车子的自动变速箱中是通过离合器和制动器来固定三元件中的任意一或两元件来实现不同的传动比的|成套散件|TY/209708-180,M-4S等|无成套散件装配后的编号|LINTEX|50-200NM|209708-180,M-4S等	8708409199 变速箱零件
商品2：变速箱油冷器|品牌轿车专用|通过水箱里的循环水,将变速箱中的油冷却,非4档及4档以下|不是成套散件或毛坯|不是通用零部件|无编号|无品牌|250NM|216WC50X0AS 216WC50X0A|未构成整机特征	8708409199 变速箱零件

图 2 商品描述信息差异大但属于同一类别

Fig. 2 Different description belong to the same class

可见，如何结合商品描述文本的层次结构，挖掘商品描述文本中的关键信息，以捕获文本语义特征，是 HS 编码分类任务的一项极大的挑战。

由于 HS 编码自动分类问题面临诸多挑战，目前国内外尚无对此类任务的相关高水平研究。但是，HS 编码分类问题本质上是一种文本形式的数据分类问题，目前学者们在分类以及文本分类问题上已经做了大量的研究。常见的深度学习方法如卷积神经网络(Convolutional Neural Network，CNN)、循环神经网络(Recurrent Neural Network，RNN)等可以有效地捕获连续的上下文依赖关系，但是它们丢失了语料库中全局的词共现信息和句子中词之间的非连续关联关系。目前图神经网络在知识图谱、信息抽取等领域都取得了很好的效果，并且可以有效地处理具有丰富关系结构的任务。它可以结合语料库中的全局词共现信息，捕获词之间的非连续关联关系。然而，文本在图结构中丢失了原有的顺序信息，同时由于文本本身并不存在图结构，因此目前将文本形式数据建模为图结构进而捕获文本中关键信息的方法仍在探索之中。

为了应对上述挑战，本文提出了一种融合文本序列和图信息的神经网络模型。将基于语料库中词共现信息构建的全局图结构作为全语料文本图，然后将商品所在子图从全语料文本图中抽取出来，从而将 HS 编码分类问题转换为子图分类问题，再利用捕获图信息的子图学习模块和捕获序列信息的序列学习模块分别建模非连续词关联关系和分层的上下文依赖，以应对 HS 编码分类任务在两个层次上呈现序列性且关键信息分散的特点。最终获得文本不同层次、不同结构的特征并将其融合用于分类。本文的主要贡献如下：

(1)将海关商品 HS 编码分类问题定义为子图分类问题，并提出基于商品描述文本特点的文本图构建方法。

(2)提出了 TSGINN 模型，利用子图学习模块和序列学习模块分别对非连续的词关联关系和多层次序列性进行建模，从而捕获融合了图信息和序列信息的语义特征。

(3)在两个真实的海关通关数据集上进行实验，验证了本文模型的分类效果优于现有的其他分类方法。

2 相关工作

海关商品 HS 编码分类问题本质上是一种文本数据分类问题，因此对于该问题，我们可以参考分类尤其是文本分类领域的相关方法。现有的文本分类方法主要分为两种：基于序列的方法和基于图的方法。

基于序列的方法主要捕获文本序列特征进行分类。文献[1-2]利用单词级或字符级卷积神经网络来捕获局部 n-gram 语义用于分类。文献[3-4]通过增加网络的深度来捕获更长距离的语义。文献[5]提出了一种简单有效的文本分类方法，利用 n-gram 扩充词袋学习局部连续语义，同时利用层次 softmax 进行高效训练。文献[6-9]利用循环神经网络学习文本中的上下文依赖，从而捕获序列信息。为了提高这类模型的表达灵活性，文献[10-12]将注意力机制引入到文本分类来学习文档的特征表示。为了进一步提高模型对于词嵌入的学习，文献[13]基于自注意力机制学习句子内部的词依赖关系，进而捕获句子的内部结构。文献[14-15]分别在 CNN 和 RNN 的基础上，利用胶囊网络并用神经元向量代替单个神经元节点对空间信息建模。基于序列的方法主要捕获句子中上下文的依赖关系，这种依赖关系往往是顺序的、连续的，但其

忽略了语料库中全局的词共现信息及非连续词之间的关联关系。

基于图的方法主要根据语料库的全局共现信息构建文本的图结构，并利于图神经网络捕获图中丰富的结构信息。文献[16]利用点互信息(Point-Wise Mutual Information，PMI)和词频逆文档频率(Term Frequency-Inverse Document Frequency，TF-IDF)计算边权重，从而构建了一个包括单词节点和文档节点的大型异构文本图，并利用图卷积来学习词和句子的表示。文献[17]构造了基于语义、句法和顺序的文本图张量来描述文档，同时将单个图上的图神经网络学习推广到图张量上。然而，这类基于图的方法丢失了文本中原有的序列信息，同时基于 PMI 等统计信息的边权重建模方法限制了神经网络对单词之间关联性的自动学习，使得图神经网络的学习效果受到节点间边权值的影响。

综上，无论是基于序列还是基于图的模型都无法同时捕获文本中非连续词关联性和局部上下文序列性，且商品描述文本具有特定的层次结构，而上述方法无法结合文本结构捕获不同层次的特征，这也使得这些方法不能很好地在 HS 编码分类任务中发挥作用。

海关商品 HS 编码分类任务的核心在于，如何有效地结合商品描述文本固有的层次结构来捕获文本的语义特征，以进行分类。本文利用图注意力网络[18]来建模非连续词之间的关联关系，并通过分层的长短期记忆网络(Long Short-Term Memory，LSTM)来捕获文本中不同层次的序列信息，从而得到融合了图信息和序列信息的语义特征，有效解决了海关商品 HS 编码分类问题。

3　问题定义

下文针对本文涉及到的概念给出了形式化定义，并对所研究的问题进行了形式化描述。

定义 1(申报要素)　申报要素是一段真实反映商品客观性质的文本。设 DE_n^m 为商品 n 的第 m 个申报要素，$w_{n,m}^t$ 为第 m 个申报要素中的第 t 个单词，则申报要素 DE_n^m 可以表示为：

$$DE_n^m=\{w_{n,m}^1,w_{n,m}^2,\cdots,w_{n,m}^t,\cdots\} \tag{1}$$

定义 2(商品描述)　商品描述指用于商品 HS 编码分类的全部描述信息，由一系列申报要素组成，申报要素之间由“|”分隔。设 CD_n 为商品 n 的描述信息，表示为：

$$CD_n=\{DE_n^1|DE_n^2|\cdots|DE_n^m|\cdots\} \tag{2}$$

本文研究的海关商品 HS 编码分类问题可描述为：针对某商品 n，已知商品描述 CD_n，分类判定该商品所属的 10 位 HS 编码。

4　模型

针对 HS 编码分类问题，本文方法主要分为两步，首先进行文本图构造，其次将图结构送入 TSGINN 网络进行特征捕获并分类。

4.1　文本图构造

文本的图构造可以分为两步：1)基于语料库中词共现信息构建全局的图结构作为全语料文本图；2)从全语料文本图中抽取商品子图。

(1)全语料文本图构造。将语料库中的单词作为节点，使用特定的窗口收集词共现信息，构建全语料文本图 $G=(V,E)$，其中 $|V|=N$ 为整个语料库的词数，$v\in V$ 表示节点集中的一个词节点。本文将每个申报要素作为一个窗口，词节点之间的边由两个词是否在同一个窗口中共现来产生，若词 i,j 在窗口内共现，则认为词节点 v_i 和 v_j 之间有连边，记为 e_{ij}，反之则无连边，由此得到边集 E。此外，我们约定对于任何 v，都有 $(v,v)\in E$。图 3 的左部分为全语料文本图的示意图，其中椭圆表示词节点，连线表示节点间的共现关系。词节点不同的灰度背景代表其属于不同的申报要素。实线表示申报要素内部词节点的连边，虚线表示跨申报要素的词节点间的连边。

(2)商品子图抽取。如图 3 所示，在全语料文本图中将某商品描述中包含的词节点及其之间的边抽取出来，得到商品子图。对于商品 n，抽取其商品子图 $SG_n=(V_n,E_n)$ 的过程可以分为以下两步：

1)将商品描述信息 CD_n 中的所有词 $w_{n,m}^t$ 作为节点加入到该商品子图的节点集 V_n 中；

2)对于全语料文本图 G 的边集 E 中的边 e_{ij}，若 $v_i,v_j\in V_n$，则将边 e_{ij} 加入商品子图的边集 E_n。

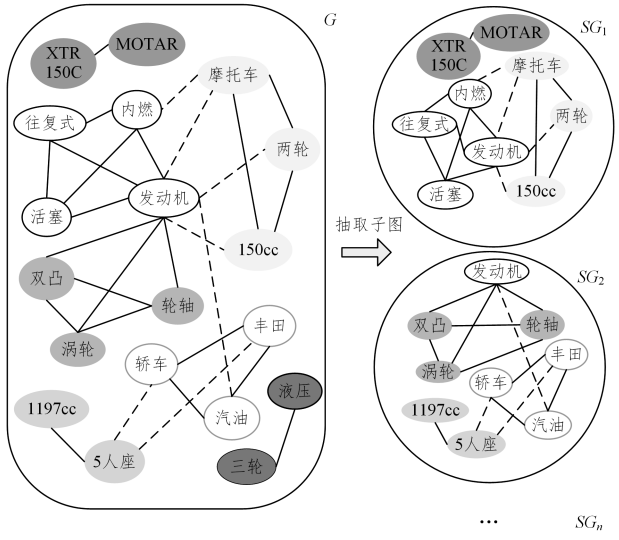

图 3　从全语料文本图中抽取商品子图示例

Fig. 3　Example of subgraph extraction from the corpus text graph

采用上述方法构建文本图，给原来文本中不连续、不相邻的词之间赋予了连边，因此我们可以通过后续的图神经网络来捕获非连续词节点间的关联性。

4.2　TSGINN 模型

图 4 展示了本文提出的融合文本序列信息和图信息的商品 HS 编码分类网络的总体结构。

TSGINN 由 3 个部分组成，分别是建模文本非连续词关联性的子图学习模块、捕获文本局部序列信息的序列学习模块以及用于信息融合并分类的融合分类模块。

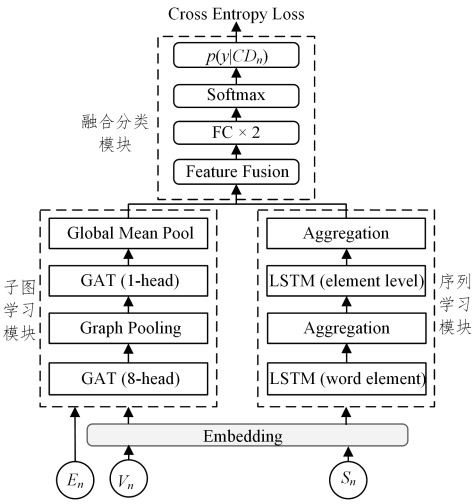

图 4 TSGINN 模型结构图

Fig. 4 Architecture of TSGINN model

4.2.1 子图学习模块

如图 4 所示，本模块中商品子图 SG_n 的输入包括词节点集 V_n 和边集 E_n。我们使用 Embedding 层对预训练得到的词向量进行微调，得到词节点向量集 $\boldsymbol{X}_G^n=\{\boldsymbol{x}_n^i \mid \boldsymbol{x}_n^i \in \mathbb{R}^d, v_i \in V_n\}$，其中 d 是词向量维度。

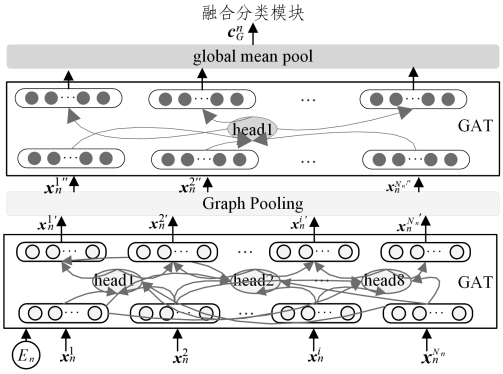

图 5 子图学习模块结构

Fig. 5 Architecture of subgraph learning module

子图学习模块的网络结构如图 5 所示。首先我们利用图注意力层学习词节点间的边权值。图注意力层先对输入特征进行线性变换 $\boldsymbol{z}_n^i=\boldsymbol{\omega}\boldsymbol{x}_n^i$，其中 $\boldsymbol{\omega}\in\mathbb{R}^{d'\times d}$ 是各节点共享的参数矩阵；接着与相邻节点的特征向量 $\boldsymbol{z}_n^j$ 拼接，并与一个可学习的权重向量 $\boldsymbol{a}$ 做点积；最后通过 LeakyReLU 激活函数得到两节点间的注意力分数。设 $v_i, v_j \in V_n$，$v_j \in \Gamma_n^i$，Γ_n^i 是子图 SG_n 中节点 i 的所有邻居节点的集合，则节点 v_i 与节点 v_j 的注意力分数 γ_n^{ij} 为：

$$\gamma_n^{ij}=LeakyReLU(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{z}_n^i \parallel \boldsymbol{z}_n^j]) \tag{3}$$

使用 softmax 对注意力分数进行标准化，得到：

$$\alpha_n^{ij}=\frac{exp(\gamma_n^{ij})}{\sum\limits_{v_k\in\Gamma_n^i} exp(\gamma_n^{ik})} \tag{4}$$

其中，α_n^{ij} 即为学习到的节点 v_i 与节点 v_j 的关联程度。接着加权聚合邻居节点信息，更新词节点的嵌入表示为：

$$\boldsymbol{x}_n^{i\prime}=\sum_{v_j\in\Gamma_n^i}\alpha_n^{ij}\boldsymbol{z}_n^j \tag{5}$$

为了让模型关注不同方面的信息，我们采用多头注意力方式建模节点间的关联关系。使用 K 头注意力后得到的节点表示如式(6)所示：

$$\boldsymbol{x}_n^{i\prime}=ELU(concat_{k\in[1,K]}(\boldsymbol{x}_{n,k}^{i\prime})+\boldsymbol{b}) \tag{6}$$

其中，$\boldsymbol{x}_{n,k}^{i\prime}$ 表示经过第 k 个注意力机制后得到的节点向量，$\boldsymbol{b}\in\mathbb{R}^{(K\times d)}$ 为可学习的偏置向量，$concat$ 表示向量拼接操作。

图 5 中，第一层图注意力网络采用多头图注意力机制，第二层图注意力网络采用单头图注意力直接聚合节点二阶邻居内的信息，从而得到每个词节点的最终嵌入表示。

通过图注意力网络，我们实现了对词节点间关联性的隐式建模，即对于节点间的边权值不再使用统计的方法，而是利用注意力机制的优势自适应地学习节点间的关联程度，并将其作为边权值，进而聚合邻居节点信息。

为了减少商品描述中无用信息的干扰，增强模型泛化能力，如图 5 所示，我们在两个图注意力层之间增加了图池化操作。采用 Gao 等提出的 Graph Pooling(gPool)[19]对文本图进行下采样。gPool 将节点向量 $\boldsymbol{x}_n^{i\prime}$ 投影到一个可学习的投影向量 $\boldsymbol{p}$，并将得到的投影值排序，然后将前 k 个投影值对应的节点和它们之间的连边保留下来，其余的节点和边舍弃，从而实现图池化。计算投影值的公式如下：

$$score=tanh\left(\frac{\boldsymbol{x}_n^{i\prime}\odot\boldsymbol{p}}{\|\boldsymbol{p}\|_2}\right) \tag{7}$$

其中，$\odot$ 表示点积操作。

通过上述步骤我们得到了词节点的特征表示，接着需要将商品子图中所有节点信息汇聚在一起，得到整个子图的嵌入表示来预测与整个子图相关的标签。本文采用全局平均池化的方法将商品子图中所有的节点信息平均聚合成一个向量，从而得到整个商品子图 SG_n 的嵌入表示 $\boldsymbol{c}_G^n$，该嵌入表示可用于后续的融合分类模块。

4.2.2 序列学习模块

商品描述文本数据同时还存在局部序列性，而仅将文本数据构建成图结构则丢失了原有的位置和序列信息。因此，本模块采用分层的长短期记忆网络来捕获商品文本不同层次的序列信息。

序列学习模块中，商品 n 的输入是一个词序列矩阵 $\boldsymbol{S}_n\in\mathbb{R}^{M\times T}$，其中 M 为商品描述中的申报要素个数，T 为申报要素中的单词数。此处我们采用补 0 的方式来填充矩阵中的空白元素。如图 6 所示，$\boldsymbol{S}_n$ 中的每一行都对应着一个申报要素所包括的词序列，例如第 m 个词序列可表示为 $DE_n^m=\{w_{n,m}^1, w_{n,m}^2, \cdots, w_{n,m}^t, \cdots, w_{n,m}^{\mathrm{T}}\}$，每条词序列经过 Embedding 层编码后均得到一个词嵌入矩阵 $\boldsymbol{X}_S^{n,m}=\{\boldsymbol{x}_{n,m}^1, \cdots, \boldsymbol{x}_{n,m}^t, \cdots, \boldsymbol{x}_{n,m}^{\mathrm{T}}\}\in\mathbb{R}^{T\times d}$，其中 d 是词向量维度。将所有申报要素对应的词嵌入矩阵 $\{\boldsymbol{X}_S^{n,m}\}_{m=1}^M$ 按顺序拼接起来，即得到一个三维的词嵌入矩阵 $\boldsymbol{X}_S^n\in\mathbb{R}^{M\times T\times d}$。

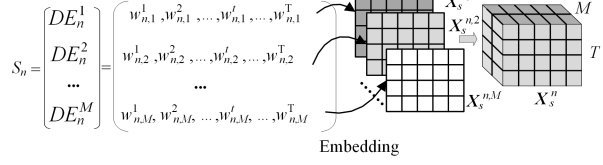

图 6 词序列矩阵 $\boldsymbol{S}_n$ 经过 Embedding 层得到 $\boldsymbol{X}_S^n$

Fig. 6 Transforming $\boldsymbol{S}_n$ to $\boldsymbol{X}_S^n$ through Embedding layer

我们将上述构建好的三维矩阵 $\boldsymbol{X}_S^n$ 输入到图 7 所示的网络中，以捕获文本不同层次的序列信息。首先将 $\boldsymbol{X}_S^n$ 中每个词嵌入矩阵 $\boldsymbol{X}_S^{n,m}$ 作为一条词向量序列送入第一层 LSTM，捕获每个申报要素内部单词级别的序列语义特征，由此可以得到每个词新的特征表示：

$$\boldsymbol{h}_{n,m}^t = LSTM(\boldsymbol{x}_{n,m}^t, \boldsymbol{h}_{n,m}^{t-1}) \tag{8}$$

接着，通过平均池化和最大值池化来聚合 $\boldsymbol{X}_S^{n,m}$ 中的全部信息和关键信息，从而得到每个申报要素的初始嵌入表示：

$$\boldsymbol{\eta}_n^m = concat(\underset{t\in[1,T]}{avg}(\boldsymbol{h}_{n,m}^t), \underset{t\in[1,T]}{\max}(\boldsymbol{h}_{n,m}^t)) \tag{9}$$

经过第一层 LSTM，我们得到了申报要素嵌入矩阵 $\boldsymbol{H}_n = \{\boldsymbol{\eta}_n^1, \cdots, \boldsymbol{\eta}_n^m, \cdots, \boldsymbol{\eta}_n^M\}$，其中 $\boldsymbol{\eta}_n^m$ 表示第 m 个申报要素的初始嵌入表示。然后将 $\boldsymbol{H}_n$ 作为一条序列送入第二层 LSTM，捕获申报要素之间的序列性，从而得到申报要素的深层嵌入表示：

$$\boldsymbol{\mu}_n^m = LSTM(\boldsymbol{\eta}_n^m, \boldsymbol{h}_n^{m-1}) \tag{10}$$

最后，同样采用平均池化和最大池化聚合所有申报要素的嵌入表示，得到商品描述文本 CD_n 最终的序列特征表示：

$$\boldsymbol{c}_S^n = concat(\underset{m\in[1,M]}{avg}(\boldsymbol{\mu}_n^m), \underset{m\in[1,M]}{\max}(\boldsymbol{\mu}_n^m)) \tag{11}$$

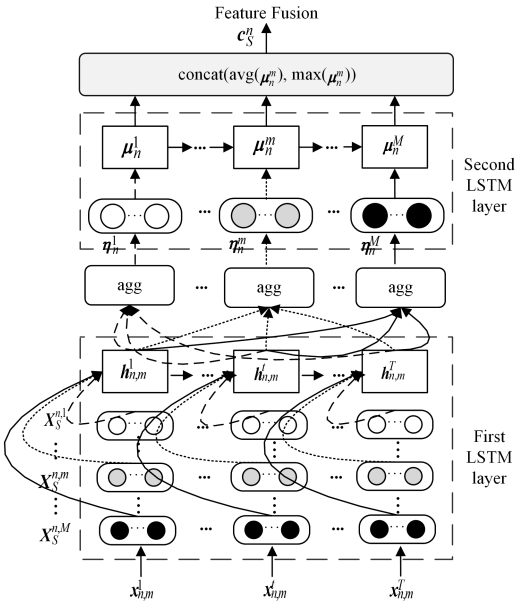

图 7 序列学习模块结构

Fig. 7 Architecture of sequence learning module

4.2.3 融合分类模块

在前两个模块中，我们分别关注了商品文本中词的非连续关联关系和分层的序列信息，这两种特征可能对分类结果产生不同程度的影响，因此我们采用加权融合的方法为这两个部分分配可学习的权重参数，以此来权衡这两部分对分类结果的重要程度。最终我们得到的商品描述文本的特征表示为：

$$\boldsymbol{c}_n = \boldsymbol{W}_G \circ \boldsymbol{c}_G^n + \boldsymbol{W}_S \circ \boldsymbol{c}_S^n \tag{12}$$

其中，$\boldsymbol{W}_G$ 和 $\boldsymbol{W}_S$ 是可学习参数，$\circ$ 为哈达玛积。

最后，将商品描述文本最终的特征表示 $\boldsymbol{c}_n$ 送入一个两层全连接神经网络，并利用 softmax 分类器得到商品 n 在标签 y 上的概率分布预测结果：

$$\boldsymbol{c}_n' = ELU(\boldsymbol{W}_1 \boldsymbol{c}_n + \boldsymbol{b}_1) \tag{13}$$

$$p(y|CD_n) = softmax(\boldsymbol{W}_2 \boldsymbol{c}_n' + \boldsymbol{b}_2) \tag{14}$$

其中，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 是可学习的参数矩阵，$\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 是偏置向量。

接着，采用交叉熵损失函数优化模型：

$$\mathcal{L} = -\sum_n \sum_l q_{n,l} \log p(y_{n,l}|CD_n, \theta) \tag{15}$$

其中，θ 表示模型所有的参数，$p(y_{n,l}|CD_n, \theta)$ 表示模型预测商品 n 的标签为 l 的概率。$\boldsymbol{q}_n$ 为一个关于商品 n 标签的 0-1 向量，当商品 n 的真实标签为 l 时，$q_{n,l}$ 为 1，否则为 0。

5 实验与结果分析

为了检验本文模型的性能，我们在两个真实的海关通关数据集上进行了对比实验，同时进行了消融实验。

5.1 数据集

本文使用两个来自中国某省的真实通关数据集，它们分别属于车辆行业和纺织制造业。其中，车辆数据集(VID)包含了车辆、航空器、船舶及相关运输设备等通关商品的进出口贸易信息，共包括 31 600 条数据，分别属于 158 种不同的 HS 编码。纺织数据集(TMD)包含了纺织原料及纺织制品等通关商品的进出口贸易信息，共包括 179 800 条数据，分别属于 899 种不同的 HS 编码。数据集的具体统计信息如表 1 所列。

表 1 数据集的统计信息

Table 1 Statistics of datasets

数据集	VID	TMD
样本数	31 600	179 800
训练集样本数	25 280	143 840
测试集样本数	6 320	35 960
语料库词数	40 282	98 823
类别数	158	899
申报要素种类数	82	228
样本平均词数	30	22

5.2 实验设置

将词向量的维度 d 设置为 300，采用 Word2Vec 进行预训练；子图学习模块中第一层图注意力网络采用 8 头注意力；图池化比例为 0.1；序列学习模块中每个 LSTM 层均采用 2 层单向 LSTM；batch-size 设为 64；学习率初始值为 0.000 4，之后按[2,10,30,50]进行多步长衰减，衰减率为 0.5。

5.3 基准方法

我们将本文方法与 3 种经典的分类模型、一种先进的语言模型 Transformer 以及目前最先进的多种文本分类方法进行对比。基准方法的介绍如下。

TF+LR：基于词频的逻辑回归模型。

TF+DT：基于词频的决策树模型。

TF+XGBoost：基于词频的树集成模型。

Transformer[13]：使用 Transformer 中的 Encoder 结构，并采用多头自注意力为序列中的每个单词生成嵌入表示。

TextCNN[1]：利用多个滤波器捕获文本的局部语义，使用最大池化捕获最重要的特征。

TextRNN[7]：将 LSTM 最后一个时间步的隐藏状态作为文本的表示，捕获长程双向的上下文依赖。

DeepMoji[12]：一种情绪检测模型，使用双向 LSTM 捕获序列性并利用注意力机制学习文档表示。

FastText[5]：一种简单高效的文本分类方法，利用与位置无关的全连接网络学习文档嵌入。我们分别评估了使用和不用 bigram 的情况。

TextGCN[16]：构建全语料单词-文档的大型异构图，并利用GCN学习节点嵌入表示将文本分类转化为节点分类问题。

本文采用正确率(Accuracy)作为评价指标。

5.4 实验结果分析

5.4.1 模型效果分析

表2列出了本文提出的TSGINN模型与上述基准方法的正确率对比。可以看到，本文方法在两个数据集上均达到了最优的性能。我们还观察到，传统的机器学习方法由于建模能力有限，其分类效果并不理想。通过比较可以发现，关注文本局部序列性的方法(如TextCNN，fastText(bigram))和关注文本非连续词关联性的方法(如TextGCN)均取得了比传统方法更好的分类效果。而本文方法同时捕获了商品文本中的非连续词关联性和层次序列性，整合了图信息和序列信息，并且在捕获非连续词关联性时不再使用基于PMI等统计信息的边权重建模方法，而是采用注意力机制自适应地学习节点间的边权值，从而提升网络模型的学习能力。此外，在捕获文本中的序列信息时，本文模型结合了商品描述文本特有的层次结构特点，捕获了两个层面的序列信息。因此，本文模型取得了最优的效果，同时也说明了本文模型在捕获图信息和序列信息上的优势。

表2 不同方法的正确率对比结果

Table 2 Accuracy comparison of different methods

Model	VID	TMD
TF+LR	0.7364	0.2771
TF+DT	0.7983	0.6033
TF+XGBoost	0.8579	0.6601
Transformer	0.8968	0.7093
TextCNN	0.8941	0.7068
TextRNN	0.8489	0.6826
DeepMoji	0.8600	0.6863
fastText(1-gram)	0.8952	0.6770
fastText(bigram)	0.8988	0.7115
TextGCN	0.8880	0.7087
ours	**0.9193**	**0.7407**

5.4.2 消融实验

我们分别只使用子图学习模块和序列学习模块进行消融实验，并与完整模型的效果进行对比。图8(a)和图8(b)分别给出了消融实验在数据集VID和TMD上的实验结果。

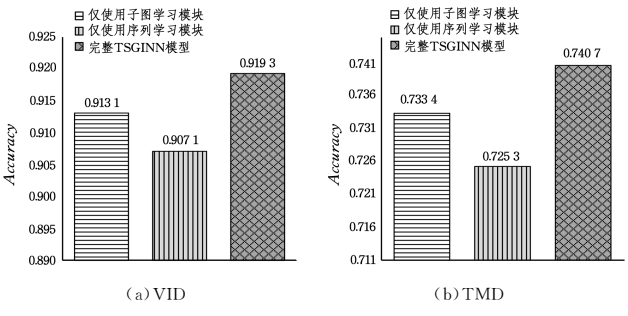

图8 消融实验正确率的对比

Fig. 8 Accuracy comparison of ablation experiment

可以看出，仅使用子图学习模块或序列学习模块时，分类效果仍然优于基准方法并取得了较优的效果，这表明了所提模型在捕获文本序列信息和图信息时的有效性。同时我们还看到，仅使用子图学习模块时的分类效果优于仅使用序列学习模块时的分类效果，这表明了通过建模非连续词关联关系来捕获文本中的关键信息对HS编码分类的重要性。完整模型TSGINN同时使用子图学习模块和序列学习模块，其分类效果相比单独使用任一模块都有较为明显的提升，这也说明了捕获融合了文本图信息和序列信息的语义特征对HS编码分类的重要意义。

5.4.3 文本图构建方法比较

前文提到了文本本身并不存在图的拓扑结构，因此我们尝试以下几种不同的构建文本图的方法。

(1)PMI-DOC。每条数据中的词构建成一个全连通图，用衡量两个词之间关联程度的PMI值作为词节点间的边权值，即节点A与节点B连边的权值ew为：

$$ew=\lg(p(A,B)/(p(A),p(B))) \tag{16}$$

(2)CP-DOC。将每条数据中的词构建成一个全连通图，利用条件概率计算节点边权值，即节点A与节点B之间连边的权值ew为：

$$ew=\left(\frac{p(A,B)}{p(A)}+\frac{p(A,B)}{p(B)}\right)\Big/2 \tag{17}$$

(3)CP-SW。使用一个固定大小为10的滑动窗口作为共现窗口，窗口内部节点两两相连。边权值的计算方式与CP-DOC相同。

(4)CP-DE。以申报要素为窗口，窗口内部节点两两相连。节点边权值计算方式与CP-DOC相同。

(5)DE-GAT(本文模型使用的方法)。以申报要素为窗口，窗口内部节点两两相连，节点边权值由GAT自适应学习得到。

以上前4种方法均为采用词共现统计信息作为边权值的文本图构建方式，同时使用两层图卷积[20]按照边权值聚合邻居节点信息，最终使用全局平均池化得到子图嵌入表示。而本文采用的DE-GAT方法使用图注意力自适应地学习词节点的边权值，为了便于对比，我们使用两层单头图注意力用于实验，并采用同样的全局平均池化得到子图嵌入表示。图9给出了这几种方法在VID数据集上的实验结果。不难看出，文本图的构建对图神经网络的学习效果有着很大的影响，不同方法之间正确率的差距甚至可以达到0.039(3.9%)。从CP-DOC，CP-SW，CP-DE这3种方法的结果可以看出，共现窗口的选择对实验效果有一定的影响，其中将申报要素作为词共现窗口符合商品文本结构特点，因此取得了较好的性能。此外，我们还发现，相比基于统计信息的边权值计算方法，本文提出的DE-GAT方法能够自适应地学习节点间的边权值，从而提升网络模型的学习能力，取得了最优的效果。

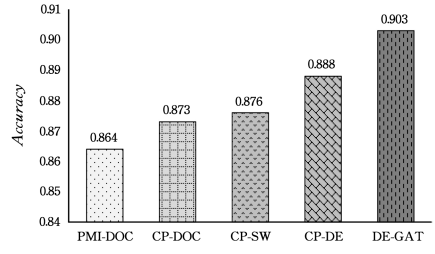

图9 不同文本图构建方法的正确率对比

Fig. 9 Accuracy comparison of different text graph construction methods

结束语 本文针对海关商品 HS 编码分类问题提出了一种新颖的 TSGINN 模型,该模型结合商品数据特有的层次结构,同时捕获商品文本中的非连续词关联关系和多层次序列信息来解决海关商品 HS 编码分类问题。在真实数据集上的实验表明,本文模型的分类效果优于其他已有的分类方法,验证了该模型在捕获非连续词关联关系和多层次序列信息方面的有效性。事实上,除了海关商品 HS 编码分类任务外,所提模型也适用于与此类似的电商平台商品分类任务,同时也为基于图结构的文本分类方法在文本图建模方面提供了新的思路。

参 考 文 献

[1] KIM Y. Convolutional neural networks for sentence classification[C]// Empirical Methods in Natural Language Processing. 2014:1746-1751.

[2] ZHANG X,ZHAO J,LECUN Y,et al. Character-level convolutional networks for text classification[C]// Neural Information Processing Systems. 2015:649-657.

[3] CONNEAU A,SCHWENK H,BARRAULT L,et al. Very deep convolutional networks for text classification[C]// Conference of the European Chapter of the Association for Computational Linguistics. 2017:1107-1116.

[4] JOHNSON R,ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]// Meeting of the Association for Computational Linguistics. 2017:562-570.

[5] JOULIN A,GRAVE E,BOJANOWSHI P,et al. Bag of tricks for efficient text classification[C]// Conference of the European Chapter of the Association for Computational Linguistics. 2017:427-431.

[6] TANG D,QIN B,LIU T,et al. Document modeling with gated recurrent neural network for sentiment classification[C]// Empirical Methods in Natural Language Processing. 2015:1422-1432.

[7] LIU P,QIU X,HUANG X. Recurrent neural network for text classification with multi-task learning[C]// International Joint Conference on Artificial Intelligence. 2016:2873-2879.

[8] LUO Y. Recurrent neural networks for classifying relations in clinical notes[J]. Journal of Biomedical Informatics,2017,72:85-95.

[9] ZHANG Y,LIU Q,SONG L. Sentence-state LSTM for text representation[C]// Meeting of the Association for Computational Linguistics. 2018:317-327.

[10] YANG Z,YANG D,DYER C,et al. Hierarchical attention networks for document classification[C]// Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. 2016:1480-1489.

[11] PAPPAS N,POPESCUBELIS A. Multilingual hierarchical attention networks for document classification[C]// International Joint Conference on Natural Language Processing. 2017:1015-1025.

[12] FELBO B,MISLOVE A,SOGAARD A,et al. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment,emotion and sarcasm[C]// Empirical Methods in Natural Language Processing. 2017:1615-1625.

[13] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017:5998-6008.

[14] ZHAO W,YE J,YANG M,et al. Investigating capsule networks with dynamic routing for text classification[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[15] WANG Y,SUN A,HAN J,et al. Sentiment analysis by capsules[C]// International World Wide Web Conference. 2018:1165-1174.

[16] YAO L,MAO C,LUO Y. Graph convolutional networks for text classification[C]// AAAI Conference on Artificial Intelligence. 2019:7370-7377.

[17] LIU X,YOU X,ZHANG X,et al. Tensor graph convolutional networks for text classification[C]// AAAI Conference on Artificial Intelligence. 2020.

[18] VELICKOVIC P,CUCURULL G,CASANOVA A,et al. Graph attention networks[C]// International Conference on Learning Representations. 2018.

[19] GAO H,JI S. Graph u-nets[C]// International Conference on Machine Learning. 2019:2083-2092.

[20] KIPF T N,WELLING M. Semi-supervised classification with graph convolutional networks[C]// International Conference on Learning Representations. 2016.

DU Shao-hua, born in 1996, postgraduate. Her main research interests include text mining and so on.

WAN Huai-yu, born in 1981, Ph.D, associate professor, Ph. D supervisor, is a member of China Computer Federation. His main research interests include social network mining, text mining, user behavior analysis and spatial-temporal data mining.