计算机科学
COMPUTER SCIENCE

SemFA:基于语义特征与关联注意力的大规模多标签文本分类模型

王振东, 董开坤, 黄俊恒, 王佰玲

引用本文

王振东, 董开坤, 黄俊恒, 王佰玲. SemFA:基于语义特征与关联注意力的大规模多标签文本分类模型[J]. 计算机科学, 2023, 50(12): 270-278.

WANG Zhendong, DONG Kaikun, HUANG Junheng, WANG Bailing. SemFA:Extreme Multi-label Text Classification Model Based on Semantic Features and Association Attention [J]. Computer Science, 2023, 50(12): 270-278.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300239

SemFA:基于语义特征与关联注意力的大规模多标签文本分类模型

王振东　董开坤　黄俊恒　王佰玲
哈尔滨工业大学(威海)计算机科学与技术学院　山东 威海 264209
哈尔滨工业大学(威海)网络空间安全研究院　山东 威海 264209
(wangzhendong20@163.com)

摘　要　大规模多标签文本分类(XMTC)是从一个庞大且复杂的标签集合中查找与文本样本最相关标签的一项具有挑战性的任务。目前,基于 Transformer 模型的深度学习方法在 XMTC 上取得了巨大的成功。然而,现有方法都没能充分利用 Transformer 模型的优势,忽略了文本不同粒度下细微的局部语义信息,同时标签与文本之间的潜在关联尚未得到稳健的建立与利用。对此,提出了一种基于语义特征与关联注意力的大规模多标签文本分类模型 SemFA(An Extreme Multi-Label Text Classification Model Based on Semantic Features and Association-Attention)。在 SemFA 中,首先拼接多层编码器顶层输出作为全局特征。其次,结合卷积神经网络从多层编码器浅层向量中获取局部特征。综合丰富的全局信息和不同粒度下细微的局部信息获得更丰富、更准确的语义特征。最后,通过关联注意力机制建立标签特征与文本特征之间的潜在关联,引入关联损失作为潜在关联不断优化模型。在 Eurlex-4K 和 Wiki10-31K 两个公开数据集上的实验结果表明,SemFA 优于大多数现有的 XMTC 模型,能有效地融合语义特征与关联注意力,提升整体的分类性能。

关键词:自然语言处理;大规模多标签文本分类;语义特征;预训练模型;注意力机制
中图法分类号　TP391

SemFA:Extreme Multi-label Text Classification Model Based on Semantic Features and Association Attention

WANG Zhendong,DONG Kaikun,HUANG Junheng and WANG Bailing
School of Computer Science and Technology,Harbin Institute of Technology(Weihai),Weihai,Shandong 264209,China
Research Institute of Cyberspace Security,Harbin Institute of Technology(Weihai),Weihai,Shandong 264209,China

Abstract　Extreme multi-label text classification(XMTC) is a challenging task that involves finding the most relevant labels from a large and complex label set for a given text sample. Currently,deep learning methods based on the Transformer model have achieved great success in XMTC. However,existing methods have not fully utilized the advantages of the Transformer model,ignoring the subtle local semantic information of texts at different granularities,and failing to establish and utilize the potential associations between labels and texts robustly. To address this issue,this paper proposes SemFA model—an extreme multi-label text classification model based on semantic features and association-attentionthat leverages semantic features and association attention for XMTC. In SemFA,the top-level outputs of multiple encoders are firstly concatenated as global features. Then,a convolutional neural network is used to extract local features from shallow vectors of multiple encoders. By combining the rich global information and subtle local information at different granularities,more accurate and comprehensive semantic features are obtained. Finally,the potential association is established between label features and text features using an association-attention mechanism,and an association loss is introduced to continuously optimize the model. Experimental results on the Eurlex-4K and Wiki10-31K public datasets show that SemFA outperforms most existing XMTC models,effectively integrating semantic features and association attention to improve overall classification performance.

Keywords　Natural Language Processing,Extreme multi-label text classification,Semantic features,Pre-trained models,Attention mechanisms

到稿日期:2023-03-31　返修日期:2023-08-25
基金项目:国家自然科学基金(62272129);中央高校基本科研业务费专项资金(HIT.NSRIF.2020098);国家重点研发计划(2020YFB2009502)
This work was supported by the National Natural Science Foundation of China(62272129),Fundamental Research Funds for the Central Universities of Ministry of Education of China(HIT.NSRIF.2020098) and National Key R&D Program of China(2020YFB2009502).
通信作者:王佰玲(wbl@hit.edu.cn)

1 引言

大规模多标签文本分类(Extreme Multi-label Text Classification，XMTC)是一个从庞大且复杂的标签集合中查找每个文本最相关标签的任务。近年来，XMTC 任务受到了极大的关注，被广泛地应用在主题识别、搜索系统、推荐系统、智能系统[1]和医疗诊断[2]等诸多工业领域，例如维基百科文章的主题标注[3]、社交网络的用户分析[4-5]、诈骗案件的关系分析、电子商务的关键词搜索[6]和产品推荐[7]等。

与经典的多标签文本分类(MLTC)任务不同，XMTC 任务中通常有大量的文本和标签数据，因此，如何从每个文本中提取丰富且准确的语义特征以进行标签预测是 XMTC 任务面临的一个核心挑战。近年来，不少学者陆续提出了多种卓有成效的方法。从文本表示学习的角度来看，这些方法可以分为两类：一类是传统的机器学习方法，通常使用稀疏向量表示文本信息，并直接输入到分类器；另一类是深度学习方法，通常采用基于神经网络的模型，在输入分类器之前将文本用特征向量表示。在传统的机器学习方法中，最典型的是 FastXML[8]和 Dismec[9]模型，它们通常采用 BoW(Bag-of-Words)和 TF-IDF 来表示文本语义信息。然而，这种方法忽略了原始文本语义和语句顺序等因素。此外，较为经典的基于嵌入的方法，例如 SLEEC[10]和 AnnexML[11]等在标签压缩与标签解压的过程中也会丢失一部分语义信息，导致获取的语义特征不够丰富。

最近，深度学习方法受到了广泛的关注，它基本解决了传统方法无法获取丰富语义特征的问题，特别是基于 Transformer[12]架构的模型，如 BERT[13]，Roberta[14]和 XLNet[15]等。一些利用 Transformer 模型的方法，如 X-Transformer[16]和 LightXML[17]等在 XMTC 任务上取得了巨大的成功。这些方法通常关注 Transformer 模型顶层的全局语义特征，虽然这些特征是通过多个自注意力机制获得的丰富语义特征，但并没有充分利用 Transformer 模型的优势，忽略了浅层的局部语义特征，这些局部语义特征同样包含十分重要的信息。与顶层的全局语义特征相比，浅层的局部语义特征可能保留了原始文本更细微的细节，但是这些细节没有被充分考虑，而且原始文本的语义信息也没有被完全挖掘。

与此同时，DECAF[18]表明标签的相关性和语义信息可以提升分类的效果，然而 X-Transformer 和 LightXML 对于标签信息的使用却欠考虑。因此，如何有效地将标签信息融入 XMTC 任务也受到了研究者们的普遍关注。最典型的方法是 C2AE[19]模型，它通过稀疏线性网络来寻找文本与标签之间的联系。但这种方法依赖的标签与文本之间的相关性并不完全可靠，并且忽略了标签本身所具有的语义信息。随后，研究者们通过图结构表示标签之间的相关性取得了一定的成功，例如 DXML[20]和 LAHA[21]模型。然而，标签图结构却隐藏了标签的原始语义信息。此外，随着图神经网络的发展，LDGN[22]使用标签图神经网络捕获标签与文本之间的关系。但对于大规模数据集而言，通过图神经网络计算的时间和空间的成本开销比较大。从自然语言的角度来看，标签和文本之间的语义信息是相互作用和影响的，标签语义特征和文本语义特征之间也必然存在着一定的潜在关联。然而，这些潜在的关联尚未被稳健地建立和发掘，难以利用潜在的关联来优化模型以提升分类性能。

综上所述，目前 XMTC 任务需要重视以下两个问题：(1)如何充分利用 Transformer 模型的优势，通过丰富的全局信息和不同粒度下细微的局部信息来获取更丰富、更准确的语义特征；(2)如何建立稳健的标签语义特征与文本语义特征之间的潜在关联，并利用这种潜在关联不断优化模型。针对以上问题，本文提出了一种基于语义特征与关联注意力的大规模多标签文本分类模型 SemFA。SemFA 由 4 部分组成：全局特征提取层、局部特征提取层、关联注意力层和预测分类层。对于全局特征提取层，本文拼接 BERT 预训练模型顶层的向量表示来获取丰富的全局语义特征。对于局部特征提取层，本文使用卷积神经网络结合 BERT 模型浅层的向量表示来获取不同粒度下细微的局部语义特征。对于关联注意力层，本文使用关联注意力机制和关联损失，建立和利用了文本与标签之间的潜在关联来不断优化模型。

本文的主要工作如下：

(1)提出了一种综合全局信息与局部信息获取丰富且准确的语义特征的方法，充分发挥 Transformer 模型的优势，在减少计算量的同时提升了模型的文本理解能力。

(2)提出利用关联注意力机制建立稳健的标签特征与文本特征之间的潜在关联，同时引入一种关联损失作为潜在关联来不断优化模型，有效地利用了标签信息来提高模型的分类能力。

(3)在 Eurlex-4K[23]和 Wiki10-31K[24]两个广泛使用的数据集上进行了一系列实验，验证了本文提出的 SemFA 模型在 XMTC 任务中的有效性和优越性。

2 相关工作

自 XMTC 任务被提出以来，研究者们陆续提出了众多成效显著的方法。根据文本输入方式的不同，这些方法通常被分为两类：一类是使用文本稀疏特征表示作为输入的传统机器学习方法；另一类是近年来广为流行的使用原始文本作为输入的深度学习方法。对于传统的机器学习方法，可将其具体分为 3 种：一对多方法、基于树方法和基于嵌入方法。

一对多方法：一对多方法的基本思想是将一个多分类问题转化为多个二分类问题，例如 Babbar 等提出的 DiSMEC 模型和 Yen 等[25]提出的 PDSparse 模型。虽然一对多方法在解决 XMTC 任务上取得了一定的成功，但其计算复杂度高和模型规模大的问题仍没有得到合理的解决，同时，该方法通过 BOW 和 TF-IDF 所表示的语义信息也不够丰富。

基于树方法：基于树方法的目标是解决一对多方法中计算复杂度高的问题，它通过树结构来组织和管理多个二分类器，使得每个分类器只需要处理相对较小的标签集合，从而提高分类的效率和准确性。比较具有代表性的是 Prabhu 等[26]提出的 Parabel 模型，它对标签空间采用树结构进行分层分区，并为每个分区构建局部分类器，这些分类器结合起来对输入实例进行预测。其次，FastXML 使用 PfastreXML 标签损失函数优化决策树的分裂和叶子节点分配。虽然树结构降低

了计算复杂度，但当树的深度增加时，分类性能仍会受到影响。

基于嵌入方法：基于嵌入方法通常将高维的文本数据映射到一个低维嵌入空间中，从而简化了 XMTC 问题。然而，标签压缩与标签解压的过程会丢失一部分语义信息，从而阻碍了分类性能的进一步提升。因此，Bhatia 等提出的 SLEEC 模型通过在嵌入空间中构建稀疏局部分类器，Tagami 等提出的 AnnexML 模型通过近邻近似搜索方法，都试图改进标签压缩和解压部分来避免丢失语义信息，但该问题并没有得到完全解决。

深度学习方法：传统的机器学习方法通常都会输入稀疏的文本特征，但这种特征无法充分利用文本的语义信息。近年来，基于深度学习的方法在 XMTC 任务上取得了巨大的成功，并且开始逐渐主导 XMTC 领域的发展。Liu 等[27]提出的 XML-CNN 是第一个在 XMTC 任务中应用深度学习方法的模型，该模型通过端到端的训练将嵌入的单词输入卷积神经网络中学习文本表示，通过全连接层进行分类。然而，在面对大规模数据集时，XML-CNN 仅使用一个简单的全连接层进行分类是十分困难的。随后，You 等[28]提出的 Attention-XML 模型巧妙地采用概率标签树（PLT）来处理大规模的标签数据，使用循环神经网络和注意力机制来处理原始文本，并针对概率标签树的每一层使用不同的模型。虽然 Attention-XML 模型能够处理大规模的数据，但它在预测阶段的处理速度仍然非常缓慢。

与此同时，研究者们也注意到标签信息在提升分类性能中的重要作用。C2AE 首次假设标签与文本之间存在潜在联系，并进行了初步的探索。随后，Wang 等[29]改进了 C2AE，提出的 RankAE 模型通过将标签排序来捕捉标签之间的相关性。Du 等[30]引入交互机制，利用交互模块捕捉文本与标签之间的复杂关系。然而，这些方法都忽略了标签本身所具有的语义信息。其次，DXML 模型通过考虑标签之间的结构，开创性地使用原始标签信息构建标签结构图来寻找标签之间的相关性。LAHA 模型通过带有标签共存图的混合注意力机制来整合标签和文本语义，其首先通过自注意力机制学习文本内容，再通过交互注意力机制学习标签与文本之间的联系。但是标签结构图却隐藏了标签的原始语义信息。随着图神经网络的兴起，LDGN 模型通过将文本表示和标签共现图输入图卷积神经网络（GCN）中捕获标签与文本之间的关系，并在训练过程中动态重构图。Zong 等[31]提出的 BGNN-XML 模型使用低通图过滤器执行图卷积操作以联合建模标签依赖性和标签特征，从而生成语义标签聚类。虽然通过图神经网络取得了一定的成功，但对于大规模数据集而言，其时间和空间的成本开销比较大。此外，Wang 等[32]提出的 GUDN 模型通过建立指导网络引导模型学习文本与标签之间的潜在关联。Shen 等[33]提出的 Taxo-Class 模型通过计算文本与标签之间的相似度来识别出核心标签以增强分类器。与之类似，Zhang 等[34]提出的 CRG 模型计算标签相关性矩阵，并使用该矩阵指导文本特征的学习。

最近，大规模预训练模型的发展十分迅猛，在大多数的 NLP 任务中都取得了优越的成绩。为了进一步提高模型的准确性和效率，X-Transformer 模型首次将 Transformer 架构应用在 XMTC 领域，该模型通过多层编码器，将最后一层编码器的输出作为文本特征，并通过微调模型输出分类结果。虽然 X-Transformer 模型取得了巨大的成功，但该模型的计算复杂度较高，模型规模也较大，而且没能充分利用文本的语义信息。LightXML 模型对此进行了改进，仅使用一个预训练模型来减少计算量和模型规模，并且通过拼接多层编码器的顶层输出作为文本特征，增强了语义信息，达到了最先进的水平。随后，Xiong 等[35]提出的 XRR 模型和 Zhang 等[36]提出了 XR-Transformer 模型都进一步提高了计算效率和分类精度。然而，这些方法都没有充分利用 Transformer 模型的优势，忽略了多层编码器的浅层所具有的不同粒度下细微的局部语义信息。同时，对于使用标签信息来提升分类性能的巨大作用也欠考虑。

综上所述，本文提出了一种基于语义特征与关联注意力的大规模多标签文本分类模型 SemFA，以提取丰富且准确的语义特征，并通过建立和利用标签特征与文本特征之间的潜在关联来提升整体的分类性能。

3 SemFA 模型

本文提出的 SemFA 模型的整体结构如图 1 所示。

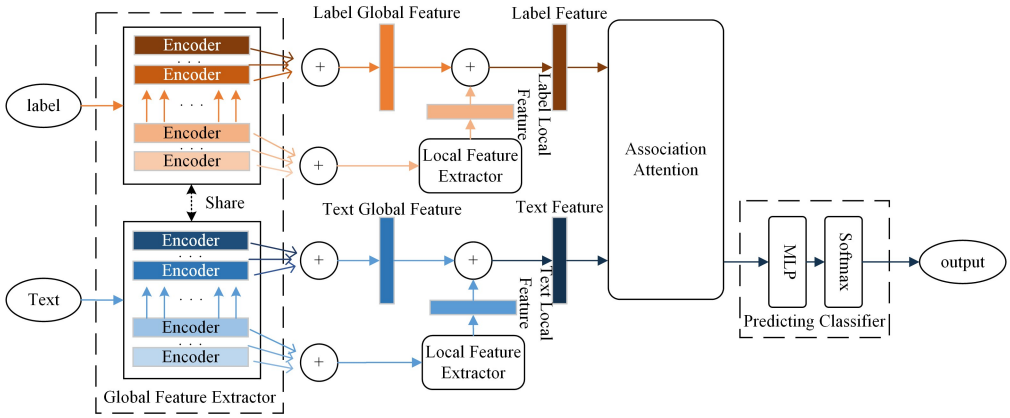


图 1 SemFA 模型

Fig. 1 SemFA model

SemFA 模型主要由 4 个部分组成：全局特征提取层、局部特征提取层、关联注意力层、预测分类层。全局特征提取层以一个参数共享的 BERT 预训练模型作为主干网络，通过拼接模型顶层的向量表示提取出文本和标签的全局特征。在局部特征提取过程中，局部特征提取层以拼接的 BERT 模型浅层的向量表示作为输入，通过卷积神经网络提取文本和标签的局部特征，并将全局特征与局部特征融合为文本特征和标签特征。通过综合丰富的全局信息和不同粒度下细微的局部信息，从而获得了更丰富、更准确的语义信息。其次，将标签特征和文本特征输入到关联注意力层，建立标签特征与文本特征之间的潜在关联，引入关联损失作为建立的潜在关联来不断优化模型。最后，预测分类层通过关键且准确的语义特征得到最终分类结果。

3.1 全局特征提取

BERT 模型在近年来被证明在大多数自然语言处理任务中有着极佳的表现，与 LightXML 相同，本文也使用一个 BERT 模型（$L=12$，$H=768$）进行特征提取以减小计算量。相比仅仅使用 one-hot 编码标签的方法，本文使用原始标签进行特征提取来强化标签的语义信息。

全局特征提取层的结构如图 1 所示。本文使用一个 BERT 模型提取文本的全局特征，为了充分利用 BERT 模型的优势，与以往模型不同，本文拼接最后 5 层隐藏状态的[CLS]向量作为文本的全局特征。高维文本表示包含更充分、更丰富的语义信息。同时，为了降低模型复杂度和计算量，本文使用一个参数共享的 BERT 模型来实现对标签全局特征的提取。与文本全局特征相同，本文拼接最后 5 层隐藏状态的[CLS]向量作为标签的全局特征。文本全局特征和标签全局特征的定义如式(1)所示：

$$\boldsymbol{e}_{\text{global}}=[\boldsymbol{h}_1^{[\text{cls}]},\boldsymbol{h}_2^{[\text{cls}]},\boldsymbol{h}_3^{[\text{cls}]},\boldsymbol{h}_4^{[\text{cls}]},\boldsymbol{h}_5^{[\text{cls}]}] \tag{1}$$

其中，$\boldsymbol{h}_i^{[\text{cls}]}$ 表示最后 i 层隐藏层状态的[CLS]向量，e_{global} 表示文本和标签的全局特征。

为了避免过拟合，本文采用高丢失率的 Dropout 层。然后经过一个 MLP 层进一步增强特征。全局特征提取层 GloF 的定义如式(2)所示：

$$\boldsymbol{F}_{\text{global}}=W_f\sigma(D(\boldsymbol{e}_{\text{global}}))+b_f \tag{2}$$

其中，D 表示 Dropout 层，σ 表示 ReLU 激活函数，W_f 和 b_f 表示 MLP_f 的参数，$\boldsymbol{F}_{\text{global}}$ 表示 $\boldsymbol{e}_{\text{global}}$ 经过全局特征提取层 GLoF 的输出。

3.2 局部特征提取

一段文本不仅包含全局特征，其本身的局部特征也拥有极其丰富的语义信息。BERT 模型的浅层向量信息往往包含着原始文本的不同粒度的语义信息和表面级别的含义，但该部分信息一直未得到充分的利用与挖掘。同时，卷积神经网络可以有效地提取原始文本的局部信息，它能通过不同大小的卷积核提取文本中的 n-gram 特征，从而取得优异的分类效果。在以前的工作中，通常使用卷积神经网络在 BERT 模型的深层特征上进一步提取语义信息，但获得的语义信息仍不全面。因此，局部特征提取层旨在获取 BERT 模型浅层的不同粒度下更细微的局部特征来增强语义信息。

局部特征提取层的结构如图 2 所示。首先，为了充分发挥 BERT 模型的浅层信息，本文拼接 BERT 模型的前 5 层隐藏状态的[CLS]向量作为浅层特征。浅层特征的定义如式(3)所示：

$$\boldsymbol{e}_{\text{shallow}}=[\boldsymbol{h}_1^{[\text{cls}]},\boldsymbol{h}_2^{[\text{cls}]},\boldsymbol{h}_3^{[\text{cls}]},\boldsymbol{h}_4^{[\text{cls}]},\boldsymbol{h}_5^{[\text{cls}]}] \tag{3}$$

其中，$\boldsymbol{h}_i^{[\text{cls}]}$ 表示前 i 层隐藏层状态的[CLS]向量，$\boldsymbol{e}_{\text{shallow}}$ 表示浅层特征。

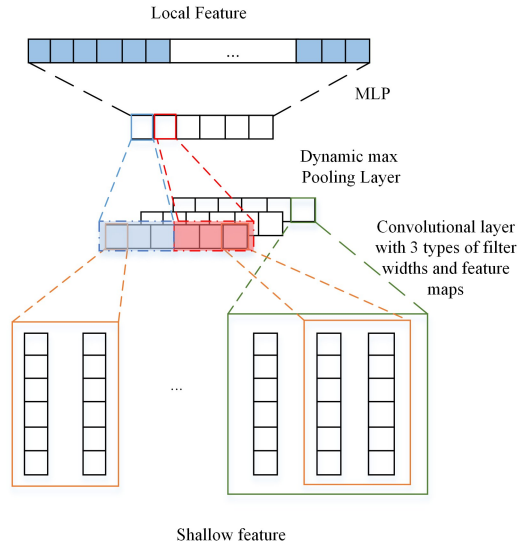


图 2 局部特征提取层

Fig. 2 Local feature extraction layer

将式(3)中输出的浅层特征 $\boldsymbol{e}_{\text{shallow}}$ 作为 CNN 层的输入，本文分别采用卷积核大小为 2，3，4 的二维卷积核对浅层特征 $\boldsymbol{e}_{\text{shallow}}$ 进行卷积操作，以获得不同粒度大小的特征图。特征图可以表示为 $\boldsymbol{C}=[c_1,c_2,\cdots,c_i]$。特征图中的每个元素 c_i 均通过式(4)获得：

$$c_i=f_c(W_c(\boldsymbol{e}^i_{\text{shallow}})+b_c \tag{4}$$

其中，f_c 表示 ReLU 激活函数，W_c 和 b_c 表示卷积核参数。

与 XML-CNN 类似，本文对各个特征图使用动态最大池化操作进行特征筛选。动态最大池化取 top-k 值进行保留，从而避免了传统的最大池化操作仅依赖一个最主要特征而无法区分语义信息特别相近的标签的问题。经过筛选后的特征可以表示为 $\boldsymbol{P}(\boldsymbol{c})$。

其次将局部特征经过一个 MLP 层增强特征。局部特征提取层 LocaF 的定义如式(5)所示：

$$\boldsymbol{F}_{\text{local}}=W_l\sigma(\boldsymbol{P}(\boldsymbol{c}))+b_l \tag{5}$$

其中，σ 表示 ReLU 激活函数，W_l 和 b_l 表示 MLP_l 的参数，$\boldsymbol{F}_{\text{local}}$ 表示 $\boldsymbol{e}_{\text{shallow}}$ 经过局部特征提取层 LocaF 的输出。

最后，将全局特征与局部特征进行融合以获得最终的文本特征 $\boldsymbol{F}_{\text{text}}$ 和标签特征 $\boldsymbol{F}_{\text{label}}$。文本特征 $\boldsymbol{F}_{\text{text}}$ 和标签特征 $\boldsymbol{F}_{\text{label}}$ 分别通过式(6)和式(7)获得：

$$\boldsymbol{F}_{\text{text}}=\alpha*\boldsymbol{F}^t_{\text{global}}+\beta*\boldsymbol{F}^t_{\text{local}} \tag{6}$$

$$\boldsymbol{F}_{\text{label}}=\alpha*\boldsymbol{F}^l_{\text{global}}+\beta*\boldsymbol{F}^l_{\text{local}} \tag{7}$$

其中，α 和 β 是常数，$\boldsymbol{F}_{\text{global}}$ 和 $\boldsymbol{F}_{\text{local}}$ 表示文本的全局特征和局部特征，$\boldsymbol{F}^t_{\text{global}}$ 和 $\boldsymbol{F}^t_{\text{local}}$ 表示标签的全局特征和局部特征。

3.3 关联注意力机制

仅仅依靠一个简单的分类网络建立的文本和标签之间的联系往往是不稳定的，解决这一问题的方法就是在标签特征

与文本特征之间建立潜在的关联与指引。在以前的工作中，建立的潜在关联通常是不稳健的，同时也忽略了标签本身所具有的语义信息。因此，本文使用原始标签来提取标签特征，使其具有标签本身的语义信息。其次，本文使用关联注意力机制建立稳健的标签特征与文本特征之间的潜在关联，将关联损失$L_{\text{association}}$作为这种潜在关联，并辅助整个模型训练。

关联注意力层的结构如图 3 所示。关联注意力层由一个 Attention 层和一个 MLP 层构成。本文将标签特征 $\boldsymbol{p}_{\text{label}}$ 和文本特征 $\boldsymbol{F}_{\text{text}}$ 输入到 Attention 层。注意力机制可以计算标签特征与文本特征之间的关联度，进而有效地捕捉标签特征与文本特征之间的关键联系，从而建立起标签特征与文本特征之间的潜在关联。然后通过 MLP 层和残差连接增强关联特征。

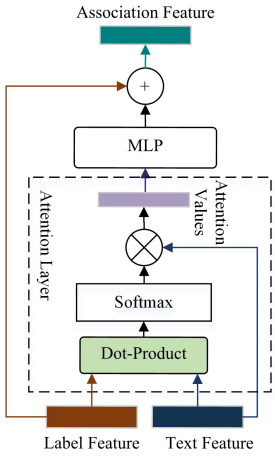


图 3 关联注意力层

Fig. 3 Association-attention layer

最后，本文引入损失函数$L_{\text{association}}$作为标签特征与文本特征之间潜在的关联。关联损失函数$L_{\text{association}}$选用一个交叉熵(BCE)损失函数。损失函数$L_{\text{association}}$的定义如式(8)所示：

$$L_{\text{association}}(y,\tilde{y})=\sum_{i=1}^{n}\sum_{j=1}^{L}-y_{ij}\log(\tilde{y_{ij}})-(1-y_{ij})\log(1-\tilde{y_{ij}}) \tag{8}$$

其中，y 表示真实标签，$\tilde{y}$表示关联特征通过预测分类层的预测标签。

本文仅在训练阶段使用关联损失函数$L_{\text{association}}$，在预测阶段并不使用标签信息。

3.4 预测分类层

预测分类层进行预测并得到最终分类结果。预测分类层的结构如图 1 所示，其包括一个 MLP 层和一个 Sotfmax 层。损失函数$L_{\text{prediction}}$选用 BCE 损失函数。预测分类层的定义如式(9)所示：

$$\hat{y}=\tau(W_p(\boldsymbol{F}_{\text{text}})+b_p) \tag{9}$$

其中，τ 表示 Softmax 层，W_p和b_p表示MLP_p的参数。

损失函数$L_{\text{prediction}}$的定义如式(10)所示：

$$L_{\text{prediction}}(y,\hat{y})=\sum_{i=1}^{n}\sum_{j=1}^{L}-y_{ij}\log(\hat{y}_{ij})-(1-y_{ij})\log(1-\hat{y}_{ij}) \tag{10}$$

其中，y 表示真实标签，$\hat{y}$ 表示文本特征通过预测分类层的预测标签。

3.5 训练阶段与预测阶段

SemFA 模型最终由全局特征提取层、局部特征提取层、关联注意力层和预测分类层构成。SemFA 的目标是最小化损失函数L_{all}。损失函数L_{all}的定义如式(11)所示：

$$L_{\text{all}}=L_{\text{prediction}}+L_{\text{association}} \tag{11}$$

SemFA 模型在训练阶段充分利用所有可用的语义信息。首先通过全局特征提取层获得文本和标签的全局特征，然后通过局部特征提取层获得文本和标签的局部特征，并将全局特征与局部特征融合为丰富且准确的文本语义特征和标签语义特征。其次将标签语义特征与文本语义特征输入到关联注意力层，通过建立稳健的标签特征与文本特征之间的潜在关联，将关联损失作为建立的潜在关联不断优化整个模型。

在预测阶段，SemFA 模型并不使用标签信息，只保留全局特征提取层、局部特征提取层和预测分类层，使模型能在短时间内预测分类结果。

4 实验

4.1 实验数据集

本文选用 XMTC 领域广泛使用的两个公开数据集来评估 SemFA 模型的有效性，包括一个多语言基准数据集 EUR-Lex 和一个英语基准数据集 Wiki10-31K。EUR-Lex 数据集是关于欧盟法律的数据集，有近 2 万份样本和近 4 千个类别。Wiki10-31K 数据集是一个维基百科文章的数据集，包含 3 万多的标签。两个数据集的统计信息如表 1 所列。其中，L 表示标签数量，$Train$ 表示训练样本数量，$Test$ 表示测试样本数量，$\bar{L}$表示每个样本的平均标签数量，$\bar{\bar{L}}$表示每个标签的平均样本数量。

表 1 数据集信息统计

Table 1 Information statistics of datasets

Datasets	L	$Train$	$Test$	$\bar{L}$	$\bar{\bar{L}}$
Eurlex-4K	3956	15449	3865	5.30	20.79
Wiki10-31K	30938	14146	6616	18.64	8.52

4.2 评价指标

本文选用在 XMTC 中广泛使用的 P@k 作为评价指标，P@k 的定义如式(12)所示：

$$P@k=\frac{1}{k}\sum_{i\in rank_k(\hat{y})}y_i \tag{12}$$

其中，k 是一个固定的常数，$rank_k(\hat{y})$表示预测排名前 k 个得分最高的标签。

4.3 实验设置

本文实验的硬件环境为 NVIDIA Tesla P40 单个 GPU，软件环境为 Pytorch 1.11.0。本文使用“bert-base”版本的 BERT($L=12$，$H=768$)，模型训练的最小批次大小为 8，通过 AdamW 优化方法进行训练，学习率大小为 0.00005，平滑因子 $eps=1\times10^{-8}$。为了避免过拟合，本文设置 Dropout 随机失活率为 0.5。

4.4 实验结果及分析

4.4.1 对比模型

为了验证本文提出的 SemFA 模型的有效性，本文将其

与近年来的基准模型进行比较。

DiSMEC：基于分布式稀疏学习的大规模多标签文本分类算法。

Parabel：基于分区标签树的大规模多标签分类算法。

Bonsai[37]：基于多样化与浅层的标签树的大规模多标签分类算法。

DXML：基于标签结构图的大规模多标签分类算法。

XML-CNN：基于卷积神经网络的大规模多标签文本分类算法。

AttentionXML：基于标签树与注意力机制结合的大规模多标签文本分类算法。

X-Transformer：基于 Transformer 的大规模多标签文本分类算法。

LightXML：基于动态负采样的大规模多标签文本分类算法。

DEPL[38]：基于生成伪标签描述的长短尾大规模多标签文本分类算法。

GUDN：基于具有标签强化策略指导网络的大规模多标签文本分类算法。

A-XML+TailMix[39]：基于长尾混合机制的大规模多标签文本分类算法。

值得注意的是，AttentionXML，X-Transformer 和 LightXML 在原论文中采用 BERT，Roberta 和 XLNet 这 3 个模型，最后整合输出结果。为了实验的公平性，本文均只使用 BERT 模型的输出结果作为本文的对比模型实验结果。

4.4.2 对比实验

本文将 SemFA 与上述对比模型分别在 Eurlex-4K 和 Wiki10-31K 数据集上进行实验，使用 4.2 节中描述的评价指标，对比实验的实验结果如表 2 所列，每列中最好的结果加粗标注。

表 2 模型在两个数据集上的实验结果对比

Table 2 Comparison of model experimental results on two datasets

（单位：%）

Model	Eurlex-4K			Wiki10-31K		
	P@1	*P*@3	*P*@5	*P*@1	*P*@3	*P*@5
DiSMEC	83.21	70.39	58.73	84.13	74.72	65.94
Parabel	82.12	68.91	57.89	84.19	72.46	63.37
Bonsai	82.30	69.55	58.35	84.52	73.76	64.69
DXML	75.63	60.13	48.65	86.45	70.88	61.31
XML-CNN	75.32	60.14	49.21	81.41	66.23	56.11
AttentionXML	85.43	73.30	60.99	86.46	77.22	67.98
X-Transformer	85.46	72.87	60.79	87.12	76.51	66.69
LightXML	86.12	73.87	61.67	87.39	77.02	68.21
DEPL	86.43	73.77	62.19	87.32	77.05	67.39
GUDN	**86.73**	74.35	62.27	88.80	78.55	69.64
A-XML+TailMix	85.80	73.70	61.99	85.17	78.12	68.66
SemFA	86.37	**74.87**	**62.74**	**88.93**	**78.82**	**69.82**

由表 2 可知，本文提出的 SemFA 模型在两个数据集上均取得了优异的结果，分别达到了 86.37%，74.87%，62.74% 和 88.93%，78.82%，69.82% 的准确率，基本优于其他基准模型。在 Eurlex-4K 数据集上，本文提出的 SemFA 模型与各指标表现最好的模型相比，虽然在 P@1 值上降低了 0.36%，但是在 P@3 和 P@5 值上分别提高了 0.52% 和 0.47%。在 3 个评价指标上取得了最好或次好的结果。在 Wiki10-31K 数据集上，SemFA 模型相较于各指标表现最好的模型在 3 个评价指标上均有提升，在 P@1，P@3 和 P@5 值上分别提高了 0.13%，0.27% 和 0.18%。

根据实验结果分析，DiSMEC，Parabel 和 Bonsai 模型使用稀疏矩阵来表示文本特征是不够丰富的，从而影响了模型的分类精度。DXML 模型未能充分利用标签原始信息，影响了模型的性能。XML-CNN 和 AttentionXML 模型分别使用卷积神经网络和注意力机制，在处理长文本时会丢失一部分重要信息，导致模型性能受到影响。X-Transformer 和 LightXML 模型与本文模型一样都使用了 Transformer 架构，但它们并未考虑到标签与文本之间的潜在关联，同时，GUDN 通过建立指导网络提高了分类精度，但没能充分利用文本信息，忽略了细微的局部特征，因此，这 3 种模型在整体性能上表现均不如本文提出的 SemFA 模型。

综上所述，本文提出的 SemFA 模型能够有效地融合语义特征和关联注意力来提升整体的分类效果，在两个广泛使用的 XMTC 数据集上均有优异的表现，具有充分的竞争力。

4.4.3 消融实验

为了验证 SemFA 模型中的不同组件对模型的有效增益，本文设计了相应的消融实验。

本文以 BERT 模型作为主干网络，SemFA-G 为仅加入全局特征提取的模型，SemFA-GL 为仅加入全局特征提取和局部特征提取的模型，SemFA-AA 为仅加入关联注意力机制的模型。消融实验的实验结果如表 3 所列，每列中最好的结果加粗标注。

表 3 模型在两个数据集上的消融实验结果

Table 3 Ablation experimental results of each model on two datasets

（单位：%）

Model	Eurlex-4K			Wiki10-31K		
	P@1	*P*@3	*P*@5	*P*@1	*P*@3	*P*@5
BERT	84.72	71.66	59.12	87.60	76.74	67.03
SemFA-G	85.41	73.59	61.78	88.05	77.08	67.69
SemFA-GL	85.62	74.15	62.35	88.39	**78.34**	**69.07**
SemFA-AA	**85.82**	**74.18**	**62.35**	**88.56**	77.56	67.80

由表 3 可以得出：

(1)全局特征提取的有效性

与基线模型 BERT 相比，SemFA-G 在两个数据集上准确率分别提升了 0.69%，1.93%，2.66% 和 0.45%，0.34%，0.66%。相较于 BERT 模型的单层编码，拼接多层编码输出能够充分利用上下文语义信息获得更充分、更丰富的语义特征，提高模型的分类性能。本文将在 4.4.4 节中深入讨论多层编码对模型性能的影响。

(2)局部特征提取的有效性

与 SemFA-G 模型相比，SemFA-GL 模型在两个数据集上准确率分别提升 0.21%，0.56%，0.57% 和 0.34%，1.26%，1.38%。与丰富的全局特征相比，局部特征中所包含的不同粒度下的语义信息能够作为辅助信息以提高分类精度。本文发现，局部特征提取层可以提取样本中关键词的 n-

gram 信息来增强语义特征。以 Eurlex-4K 为例，如图 4 所示，Eurlex-4K 数据集中的样本并不完全是完整的句子，而是由多个关键词组成。因此，局部特征提取层使用包含原始信息的多层编码器浅层输出，通过卷积神经网络能够提取到局部特征作为辅助信息，与丰富的全局特征融合，提升模型的分类能力。

Texts	Labels
protocol articl decis eea joint committe articl affect case law court justic european commun	european_economic_area protocol_to_an_agreement eea_joint_committee ec_court_of_justice settlement_of_disputes
commiss direct novemb amend annex ii council direct eec market cereal seed eec commiss european commun regard treati establish european econom commun regard council direct eec june market cereal seed amend commiss direct eec present scientif technic knowledg appear ……	cereals seed marketing_standard
decis ec european parliament council establish multiannu commun programm promot safer internet onlin technolog text eea relev european parliament council european union regard treati establish european commun articl thereof regard propos ……	information_policy child_protection internet action_programme public_awareness_campaign community_financing

图 4 Eurlex-4K 数据集示例图

Fig. 4 Illustration of Eurlex-4K dataset

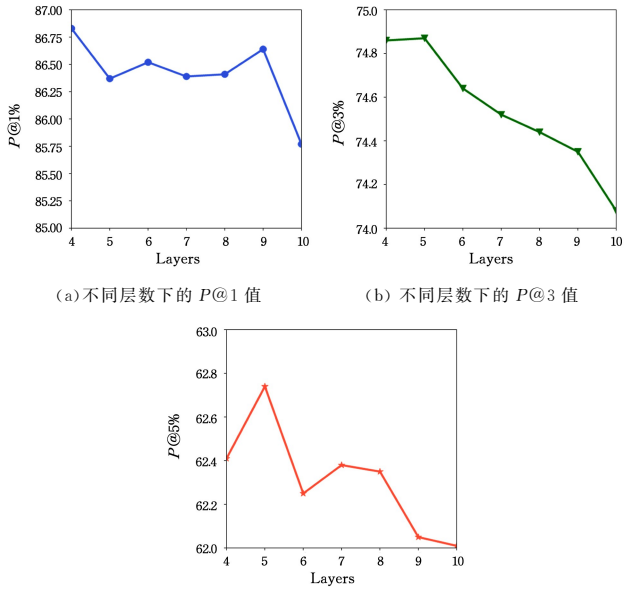
(3)关联注意力机制的有效性

与基准模型 BERT 相比，SemFA-AA 模型在两个数据集上准确率分别提升了 1.1%，2.52%，3.23% 和 0.96%，1.97%，0.77%。依据实践经验，文本和标签之间存在着巨大的潜在空间，相似的文本和标签之间不仅向量表示相似，而且在语义空间中的距离也相近，其中必然隐藏着潜在的语义关联。实验结果表明，本文使用的关联注意力机制可以建立和利用标签和文本之间的潜在关联，并不断优化微调模型，提高模型的分类精度，验证了关联注意力机制的有效性。

综上所述，全局特征提取、局部特征提取和关联注意力机制均有助于提升模型分类精度，当三者协同作用时，模型性能达到最优。

4.4.4 多层编码对比实验

为了验证本文的全局特征提取层与局部特征提取层中使用的 BERT 多层编码器[CLS]向量的拼接输出能够提高分类精度与提取更丰富的语义特征，本文采用不同层数的多层编码拼接输出进行一组对比实验。因受局部特征提取层卷积核大小的影响，仅对 4 层及以上的多层编码进行对比实验。实验结果如图 5 所示。由图 5 可知，多层编码拼接的层数会影响模型的分类性能。根据实验结果分析，当多层编码拼接层数为 5 时，模型的分类效果最佳，但当多层编码拼接层数逐渐增加时，模型的分类效果却开始呈现逐渐下降的趋势。由于多层编码中所包含的语义信息并不是完全有益和准确的，其中必然包含一定的冗杂信息，当层数逐渐增加时，冗杂信息就会逐渐累积，最终影响模型的分类精度。基于综合考虑，本文选用的多层编码拼接层数为 5 层。



(a)不同层数下的 $P@1$ 值 (b) 不同层数下的 $P@3$ 值

(c)不同层数下的 $P@5$ 值

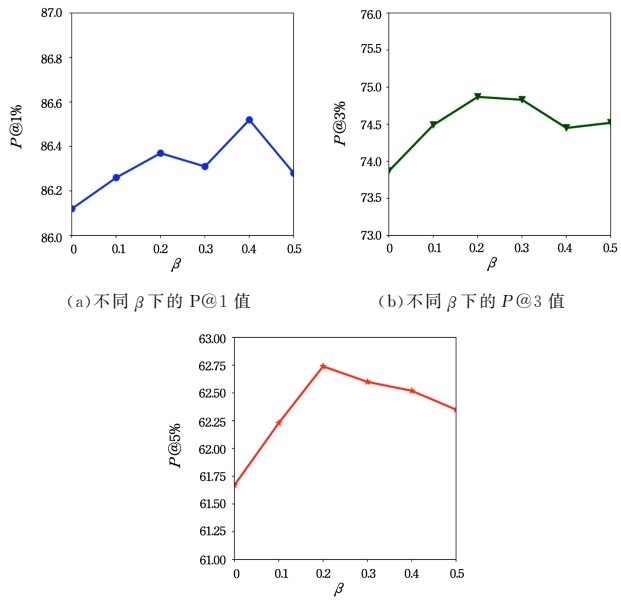
图 5 多层编码层数的影响

Fig. 5 Impact of the number of multi-layer coding layer

4.4.5 参数分析

为了进一步检验模型的鲁棒性，本文对模型中涉及的相关参数进行了讨论，主要为全局特征和局部特征的融合参数 α 和 β。

本文通过融合全局特征和局部特征可以获取更丰富、更充分的语义特征。为探究融合参数 α 和 β 对模型性能的影响，本文通过设置不同参数值对其进行讨论。为方便讨论，将 α 设置为 1，通过调节 β 进行实验。实验结果如图 6 所示。



(a)不同 β 下的 P@1 值 (b)不同 β 下的 $P@3$ 值

(c)不同 β 下的 $P@5$ 值

图 6 融合参数的影响

Fig. 6 Impact of fusion parameters

由图 6 可知，全局特征和局部特征的融合参数在一定程度上会影响模型的分类精度。根据实验结果分析，当 β 为 0.2 时，模型的分类效果最优。这表明局部特征作为全局

特征的辅助信息，若过少则不能被有效地利用，而过多也不一定能提高分类性能，因为其中包含累积的冗杂信息会给模型引入更多的噪声，从而干扰模型分类效果。通过综合对比分析，本文统一使用 β 为 0.2 的融合参数实现全局特征与局部特征的融合。

结束语 针对现存的 XMTC 方法不能充分利用 Transformer 模型的优势而忽略了局部语义信息和不能稳健地建立与利用文本与标签之间的潜在关联的问题，本文提出了一种基于语义特征与关联注意力的大规模多标签文本分类模型 SemFA。具体来说，该模型在提取丰富的全局特征的基础上，通过结合卷积神经网络与预训练模型浅层的向量表示提取不同粒度下细微的局部特征。综合丰富的全局信息和不同粒度下细微的局部信息得到更充分、更准确的语义特征。其次，通过关联注意力机制和关联损失稳健地建立和利用标签特征与文本特征之间的潜在关联不断优化模型，提高分类性能。大量的实验证明了本文提出的 SemFA 模型的有效性和优越性。近年来，不少学者通过图神经网络引入外部知识来增强语义信息，未来工作将结合现有模型和图神经网络进行研究。

参 考 文 献

[1] MIRI M, DOWLATSHAHI M B, HASHEMI A, et al. Ensemble feature selection for multi-label text classification: An intelligent order statistics approach[J]. International Journal of Intelligent Systems, 2022, 37(12): 11319-11341.

[2] WU H X, HAN M, CHEN Z Q, et al. A review of multi-label classification under supervised and semi-supervised learning[J]. Computer Science, 2022, 49(8): 12-25.

[3] DEKEL O, SHAMIR O. Multiclass-multilabel classification with more classes than examples[C]// Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, 2010: 137-144.

[4] SUN Y, LV H, LIU X, et al. Personalized recommendation for Weibo comic users[C]// 2018 Wireless Telecommunications Symposium. 2018: 1-6.

[5] LI K Y, CHEN Y, NIU S Z. BERT-based text classification algorithm for social e-commerce [J]. Computer Science, 2021, 48(2): 87-92.

[6] CHANG W C, YU H F, ZHONG K, et al. Taming pretrained transformers for extreme multi-label text classification[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 3163-3171.

[7] AGRAWAL R, GUPTA A, PRABHU Y, et al. Multi-label learning with millions of labels: Recommending advertiser bid-phrases for web pages[C]// Proceedings of the 22nd International Conference on World Wide Web. 2013: 13-24.

[8] PRABHU Y, VARMA M. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 263-272.

[9] BABBAR R, SCHOLKOPF B. Dismec: Distributed sparse machines for extreme multi-label classification[C]// Proceedings of the tenth ACM International Conference on Web Search and Data Mining. 2017: 721-729.

[10] BHATIA K, JAIN H, KAR P, et al. Sparse Local Embeddings for Extreme Multi-label Classification[C]// Annual Conference on Neural Information Processing Systems. 2015: 730-738.

[11] TAGAMI Y. Annexml: Approximate nearest neighbor search for extreme multi-label classification[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 455-464.

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

[13] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[14] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv: 1907.11692, 2019.

[15] YANG Z, DAI Z, YANG Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[C]// Advances in Neural Information Processing Systems. 2019: 5753-5763.

[16] CHANG W C, YU H F, ZHONG K, et al. Taming pretrained transformers for extreme multi-label text classification[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 3163-3171.

[17] JIANG T, WANG D, SUN L, et al. Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 7987-7994.

[18] MITTAL A, DAHIYA K, AGRAWAL S, et al. Decaf: Deep extreme classification with label features[C]// Proceedings of the 14th ACM International Conference on WebSearch and Data Mining. 2021: 49-57.

[19] YEH C K, WU W C, KO W J, et al. Learning deep latent space for multi-label classification[J]. arXiv: 1707.00418, 2017.

[20] ZHANG W, YAN J, WANG X, et al. Deep extreme multi-label learning[C]// Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. 2018: 100-107.

[21] HUANG X, CHEN B, XIAO L, et al. Label-aware document representation via hybrid attention for extreme multi-label text classification[J]. Neural Processing Letters, 2021, 54(5): 3601-3617.

[22] MA Q, YUAN C, ZHOU W, et al. Label-specific dual graph neural network for multi-label text classification[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021: 3855-3864.

[23] LOZAMENCIA E, FURNKRANZ J. Efficient pairwise multilabelclassification for large-scale problems in the legal domain [C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2008: 50-65.

[24] ZUBIAGA A. Enhancing navigation on wikipedia with socialtags [J]. arXiv: 1202.5469, 2012.

[25] YEN I E H, HUANG X, RAVIKUMAR P, et al. Pd-sparse: A primal and dual sparse approach to extreme multiclass and mul-

tilabel classification[C]// International Conference on Machine-Learning. 2016:3069-3077.
[26] PRABHU Y, KAG A, HARSOLA S, et al. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising[C]// Proceedings of the 2018 World Wide Web Conference. 2018:993-1002.
[27] LIU J, CHANG W C, WU Y, et al. Deep learning for extreme multi-label text classification[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017:115-124.
[28] YOU R, ZHANG Z, WANG Z, et al. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification[C]// Advances in Neural Information Processing Systems. 2019:5571-6362.
[29] WANG B, CHEN L, SUN W, et al. Ranking-based autoencoder for extreme multi-label classification[J]. arXiv: 1904. 05937, 2019.
[30] DU C, CHEN Z, FENG F, et al. Explicit interaction model towards text classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:6359-6366.
[31] ZONG D M, SUN S L. BGNN-XML: Bilateral Graph Neural Networks for Extreme Multi-label Text Classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(7): 6698-6709.
[32] WANG Q, SHU H, ZHU J. GUDN a novel guide network for extreme multi-label text classification[J]. arXiv: 2201. 11582, 2022.
[33] SHEN J, QIU W, MENG Y, et al. TaxoClass: Hierarchical multi-label text classification using only class names[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. 2021: 4239-4249.
[34] ZHANG Q W, ZHANG X, YAN Z, et al. Correlation-Guided Representation for Multi-Label Text Classification[C]// IJCAI. 2021:3363-3369.
[35] XIONG J, YU L, NIU X, et al. XRR: Extreme multi-label text classification with candidate retrieving and deep ranking[J]. Information Sciences, 2023, 622:115-132.
[36] ZHANG J, CHANG W C, YU H F, et al. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification[C]// Advances in Neural Information Processing Systems. 2021:7267-7280.
[37] KHANDAGALE S, XIAO H, BABBAR R. Bonsai: diverse and shallow trees for extreme multi-label classification[J]. Machine Learning, 2020, 109(11):2099-2119.
[38] ZHANG R, WANG Y S, YANG Y, et al. Long-tailed Extreme Multi-label Text Classification with Generated Pseudo Label Descriptions[J]. arXiv: 2204. 00958, 2022.
[39] HAN S, CHOI E, LIM C, et al. Long-tail Mixup for Extreme Multi-label Classification[C]// Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022:3998-4002.

WANG Zhendong, born in 2000, postgraduate, is a member of China Computer Federation. His main research interests include artificial intelligence, natural language processing and financial security.

WANG Bailing, born in 1978, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include industrial Internet security, information security and financial security.

(责任编辑:何杨)