报告题目

基于迁移学习和集成学习的医学短文本分类

张博 孙逸 李孟颖 郑馥琦 张益嘉 王健 林鸿飞 杨志豪
（ 大连理工大学 计算机科学与技术学院 辽宁省 大连市 116024）

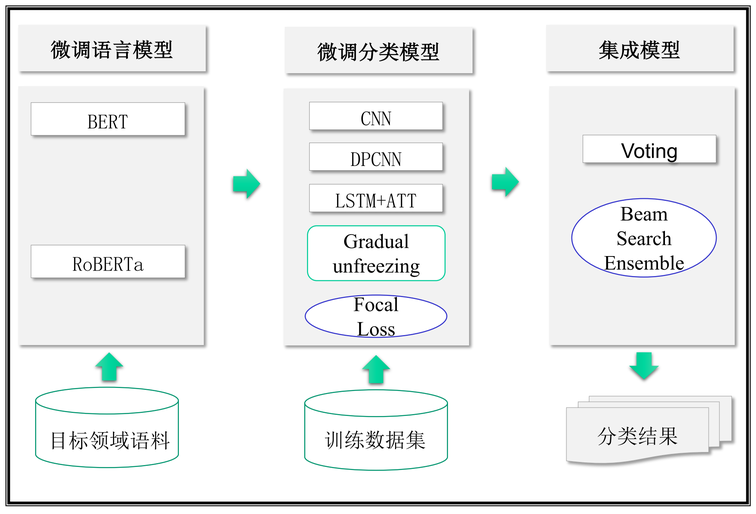
论文摘要

医学文本分类对于辅助医疗、构建医学文本结构化数据具有重要的价值和意义。该文提出一种基于迁移学习和集成学习的临床试验筛选标准短文本分类技术。首先，利用目标领域数据集对预训练语言模型进行微调来实现迁移学习得到在目标领域的语义增强语言模型；其次，将上述含有丰富目标领域语义信息语言模型与主流的神经网络模型结合得到医学文本分类器，再针对医学文本分类任务进行模型分类器的微调；最后，通过模型集成并采用 **beam search ensemble** 算法提高整个文本分类系统的性能，最终在 **CHIP2019** 评测三测试集上 **F1** 值达到了 **0.8111**。

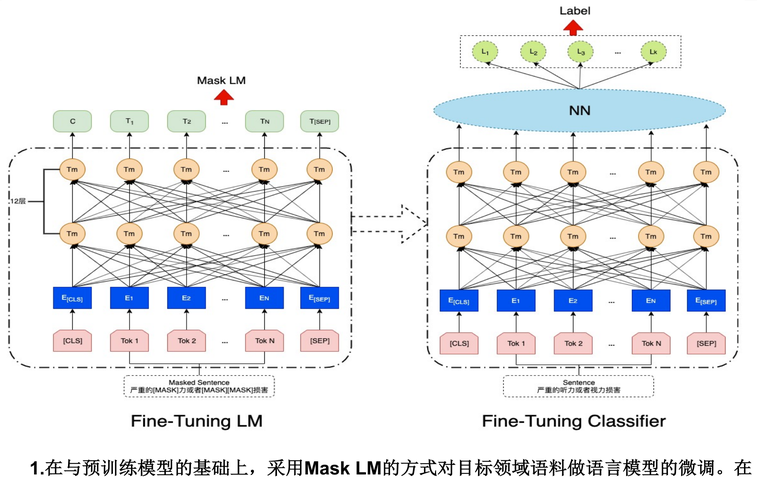
论文简介

本文提出的临床试验筛选标准短文本分类方法主要分为三个阶段，一是微调预训练语言模型，在此阶段使用的语言模型是 **BERT[5]**和 **Roberta[6]**；二是微调分类模型，面向医学短文本分类任务将第一阶段得到的语言模型与神经网络模型进行结合构建医学分类模型，微调过程是针对结合后的整体模型即包含语言模型和神经网络模型，上述两个阶段的微调使用的数据集均为无监督领域相关的外部资源数据集；三是利用集成学习的知识来实现最终的医学文本分类，将第二阶段得到的多个分类器进行集成学习，通过投票方式和 **beam search ensemble** 算法选择出最佳的模型组合作为医学文本分类系统的最终分类模型。

系统框架



核心模型



1.在与预训练模型的基础上，采用**Mask LM**的方式对目标领域语料做语言模型的微调。在训练过程中，采用倾斜三角学习率的方法，以使得模型在训练开始时快速收敛到参数空间的合适区域，再细化其参数。

2.对微调过后的语言模型与其他下游分类任务进行衔接，一起做分类任务微调。为使不同模型学到不同的信息用于最后的模型融合，其中下游分类模型包括**CNN**，**DPCNN**，**LSTM+Attention**等。采用**Focal Loss**损失函数，有助于改善不均匀样本和难分类样本的问题，并在训练过程中，对不同模型分别使用了倾斜三角学习率和逐层解冻的技巧。

3.最后采用投票机制进行模型集成。因所有模型一起融合可能带来许多冗余信息，使模型的性能下降，又选取在验证集上表现最好的模型组合有着过拟合、时间过长、模型无法重复投票等问题。对**Beam Search**算法进行改进以应用到模型融合中，有效处理上述问题。

实验仿真

➢ 对比实验结果：

模型	P	R	F1
bert_base	0.7937	0.8105	0.7964
bert_dpcnn	0.8166	0.7913	0.7976
bert_lstm_att	0.8199	0.8006	0.8049
bert_cnn	0.8001	0.8135	0.8012
roberta_base	0.7886	0.8217	0.7990
roberta_att	0.7824	0.8158	0.7948
roberta_dpcnn	0.8117	0.8029	0.8030
roberta_lstm_att	0.7967	0.8188	0.8028
roberta_cnn	0.7943	0.8311	0.8063
all_voting	0.8061	0.8247	0.8099
beam_search_voting	0.8078	0.8249	**0.8111**

➢ 消融实验结果：

模型	P	R	F1
roberta	0.7722	0.8088	0.7850
roberta_fine	0.7919	0.8119	0.7977
roberta_fine_gu	0.7886	0.8217	0.7990
beam_search_voting	0.8078	0.8249	**0.8111**

➢ 在**BERT**预训练语言模型中，拼接**LSTM**与注意力机制模型效果最优，在**Roberta**预训练语言模型中，拼接**CNN**模型效果最好。在评测时，使用了**9**个模型**(**其中**roberta_cnn**的投票权重为**2)**全部集成学习得到分类结果的**F1**值为**0.8099**，比单个模型最优结果高出**0.36%**，比单个**BERT**模型结果高出**1.35%**，证明模型集成的有效性。用我们提出的**beam search ensemble**算法进行模型集成得到的结果比所有模型共同集成结果高出**0.12%**，充分论证了**beam search ensemble**算法的有效性。

➢ 消融实验中，**roberta_fine**表示只对预训练语言模型进行微调，对比预训练**roberta**结果提高**1.27%**，充分说明了微调模型的有效性。**roberta_fine_gu**在微调分类器模型时采用逐层解冻方案来实现模型的快速收敛。通过与不进行任何微调的单个**roberta**模型相比提高了**1.40%**，与只进行语言模型微调的**roberta_fine**相比提高了**0.13%**，证明了逐层解冻策略的有效性。

论文结论

•我们提出一种基于迁移学习和集成学习的医学短文本分类方案，利用相关领域外部资源数据对语言模型和分类模型进行微调，在微调过程中利用了斜三角学习率和逐层解冻的微调方法，最后用模型集成学习来提高医学文本分类系统的性能，在模型集成过程中提出了改进的 **beam search ensemble** 算法，该算法可以选出最佳分类模型组合，进一步提高了分类结果，最终 **F1** 值达到**0.8111**，是目前该文本分类任务的最佳结果。

•未来工作继续利用迁移学习的一些知识，并尝试在神经网络分类模型算法上有所改进，进一步提升医学文本分类系统的性能。

西安电子科技大学 XIDIAN UNIVERSITY