CCL25-Eval 任务9：中医辨证辨病及中药处方生成评测 实验报告

1120232535 汪隽宁 数据科学与大数据技术

摘要

本项目完成的是CCL25-Eval 任务9：中医辨证辨病及中药处方生成评测的子任务1：中医多标签辨证辨病（TCM Mulit-label Syndrome and Disease Differentiation）。本项目使用手工构建特征、神经网络、微调大模型、将手工构建特征和微调大模型结合四种方法，通过在800条数据上的训练，完成了输入为病历和症状，输出为证型多分类标签和疾病单分类标签的任务。经过实验，手工构建特征和微调大模型结合的方法在测试集上的表现最佳，按照官方的正确率评判标准，证型和疾病预测正确率为0.3125，微调大模型、神经网络、手工构建特征的表现则依次下降。由于官方提供的200条公开测试集数据标准答案均为同一种证型和疾病输出结果，实验结果的可靠性有待使用CCL会议结束后公开的500条暂不公开的测试集进行进一步验证。

关键词： 中医诊断；大模型微调；神经网络；特征工程

CCL25-Eval Task 9: Evaluation Report on Syndrome and Disease Differentiation, and Prescription Recommendation in Traditional Chinese Medicine

1120232535 Wang Junning Data Science and Big Data Technology

Abstract

This project addresses Subtask 1 of Task 9 from the CCL25-Eval Challenge: TCM Multi-label Syndrome and Disease Differentiation. We explore four approaches—manual feature engineering, neural networks, fine-tuning large language models (LLMs), and combining LLMs with structured features—to perform syndrome (multi-label) and disease (single-label) classification based on patient records and symptom descriptions. Trained on a dataset of 800 samples, our best-performing method is the combination of manually constructed features with fine-tuned LLMs, achieving a prediction accuracy of 0.3125 on the official test set according to the evaluation metric. Other methods such as standalone LLM fine-tuning, neural networks, and manual features alone perform progressively worse. However, since the 200 public test samples all share the same ground-truth labels for both syndrome and disease, the reliability of this result awaits further validation on the 500 additional samples expected to be released after the CCL conference.

关键词： Traditional Chinese Medicine Diagnosis；Large Language Model Fine-tuning；Neural Networks；Feature Engineering

1 引言

中医作为中国传统医学的重要组成部分，历经数千年的发展，已形成独具特色的理论体系和诊疗方法，对中国乃至全球人民的医疗健康做出了重要贡献。辨证论治是中医认识疾病和治疗疾病的核心原则和方法，其基本思想是通过望、闻、问、切的方法，收集患者症状、舌苔、脉象等临床信息，通过分析、综合，辨清疾病的病因、病机，概括、判断为某种性质的证，进而制定个性化的治疗方案，开具合适的中药处方予以治疗。

本项目基于官方构建的用于评估中医辨证辨病及处方生成的数据集，对于数据集中给定的患者临床文档，对患者所患的证型和疾病进行判断，

2 数据预处理与手工特征构建

训练数据共计800条，来源为官方构建的用于评估中医辨证辨病及处方生成的数据集，格式如：

{
"ID": "35",
"性别": "女",
"职业": "退休",
"年龄": "66岁",
"婚姻": "已婚",
"病史陈述者": "本人",
"发病节气": "立夏",
"主诉": "发作性胸闷20年，加重伴胸痛3月余",
"症状": "胸部疼痛，呈针刺样，胸闷不舒，心慌不安，气短乏力，眼干眼涩，口干口苦，纳可，食后反酸烧心，眠可，二便调。",
"中医望闻切诊": "中医望闻切诊：表情自然，面色暗红，形体正常,动静姿态，语气低,气息平；无异常气味,舌暗红、苔黄腻,舌下脉络曲张,脉弦。",
"病史": "现病史：患者20年前因劳累后出现胸闷，无胸痛，于***"就诊，行心电图、心脏彩超等检查，诊断为\冠心病"，具体治疗不详，好转后出院。院外规律服用\拜阿司匹林"等，症状控制可。3月前劳累后出现上述症状加重，自服\复方丹参滴丸"，效不佳。后患者于***就诊"，行冠脉CT示：\LM轻度狭窄，LAD中度狭窄，LCX轻度狭窄，RCA轻度狭窄，PDA中度狭窄"，予\吲哚布芬"、\喜格迈"等，效不佳。现为求进一步中西医结合系统治疗，入住我病区。入院症见：既往史：否认慢性支气管炎等慢性疾病病史。否认肝炎、否认结核等传染病史。预防接种史不详。否认手术史、否认重大外伤史。否认输血史。否认药物过敏史、否认其他接触物过敏史。个人史：生于******，久居本地，无疫水、疫源接触史，无嗜酒史，无吸烟史，无放射线物质接触史，否认麻醉毒品等嗜好，否认冶游史，否认食物过敏史，否认传染病史。婚育史：已婚，适龄婚育。月经史：已绝经，既往月经规律。家族史：父母已故，死因不详。兄弟姐妹6人，均体健。育有1子，儿子及配偶均体健，家人体健，否认家族性遗传病史。",
"体格检查": "体温：36.5℃ 脉搏：61次/分 呼吸：18次/分 血压：158/71mmHg（R）、152/71mmHg（L） Padua评分：3分（低危）生命体征一般情况：患者老年，女,发育正常，营养良好，神志清楚，步入病房,查体合作，皮肤黏膜：全身皮肤及粘膜无黄染,未见皮下出血,淋巴结浅表淋巴结未及肿大。标题定位符头颅五官无畸形，眼睑无水肿,巩膜无黄染,双侧瞳孔等大等圆，对光反射灵敏，外耳道无异常分泌物，鼻外观无畸形，口唇红润，伸舌居中，双侧扁桃体正常，表面未见脓性分泌物，标题定位符颈软，无抵抗感，双侧颈静脉正常，气管居中，甲状腺未及肿大，未闻及血管杂音。标题定位符胸廓正常,双肺呼吸音清晰，未闻及干、湿罗音，未闻及胸膜摩擦音。心脏心界不大，心率61次/分，心律齐整，心音低,各瓣膜听诊区未闻及杂音，未闻及心包摩擦音。脉搏规整，无水冲脉、枪击音、毛细血管搏动征。腹部腹部平坦，无腹壁静脉显露，无胃肠型和蠕动波，腹部柔软，无压痛、反跳痛，肝脏未触及，脾脏未触及，未触及腹部包块，麦氏点无压痛及反跳痛，Murphy's征－，肾脏未触及，肝浊音界正常，肝肾区无明显肾区叩击痛，肝脾区无明显叩击痛，腹部叩诊鼓音，移动性浊音-，肠鸣音正常，无过水声，直肠肛门、生殖器肛门及外生殖器未查。生理反射存在，病理反射未引出，双下肢无水肿。",

"辅助检查": " 2020-4-29 冠脉CT示：LM轻度狭窄，LAD中度狭窄，LCX轻度狭窄，RCA轻度狭窄，PDA中度狭窄。（于***）2020-5-12 心电图示：窦性心律，ST-T改变。",
"疾病": "胸痹心痛病",
"证型": "气虚血瘀证|痰热蕴结证",
"处方": " ['丁香', '广藿香', '黄芪', '檀香', '砂仁', '木香', '草豆蔻', '附片 ', '花椒', '制川乌', '细辛', '桔梗', '麸炒枳壳', '葛根']"
}

在数据预处理过程中，我们选择将'性别', '年龄', '发病节气', '主诉', '症状', '中医望闻切诊', '病史'等字段进行拼接，构造用于语言模型输入的统一文本字段'text'。而体格检查、辅助检查等字段噪声较多且信息冗余，因此未纳入最终输入。这些字段既包含了患者的基本信息，又蕴含了较丰富的中医症状学描述。为了增强结构与可读性，我们在每一段信息前均添加了"【字段名】"的标签，例如"【主诉】心慌乏力7天。"。

在手工特征构建的过程中，我们对"主诉"和"中医望闻切诊"中可能出现的症状、舌象、脉象、面色特征进行了总结，对每条数据中的"主诉"和"中医望闻切诊"进行了这四类特征的提取，构建了如下结构化特征：在手工特征构建方面，我们设计并归纳了主诉关键词、舌象、脉象及面色四类结构化特征，这些特征大多出现在主诉与中医望闻切诊字段中。我们手动构建了关键词表，并通过关键词匹配进行特征提取，对每条样本进行多热编码（multi-hot encoding）。最终得到如下结构化特征:

- 主诉关键词（33维）
- 舌象特征（8维）
- 脉象特征（16维）
- 面色特征（6维）

经过数据预处理和手工特征构建的数据格式如下:

{
"ID": "1508",
"text": "【性别】男\n【年龄】62\n【发病季节】冬天\n【主诉】心慌乏力7天。\n【症状】阵发性心慌、乏力，无胸闷胸痛，无头晕头痛，纳眠可，二便调。\n【中医望闻切诊】表情自然，面色红润，形体正常，动静姿态，语气清，气息平，无异常气味，舌暗红、舌苔白，脉沉数。\n【病史】现病史，患者于7天前因无明显诱因出现阵发性心慌症状，乏力明显，为明确进一步中西医结合诊断治疗，遂入住我病区，入院症见，阵发性心慌、乏力，无胸闷胸痛，无头晕头痛，纳眠可，二便调，既往史，既往否认糖尿病等慢性疾病病史，否认肝炎、否认结核等传染病史，预防接种史不详，否认手术史、否认重大外伤史，否认输血史，否认药物过敏史、否认其他接触物过敏史，个人史，生于久居本地，无疫水、疫源接触史，无嗜酒史，无吸烟史，无放射线物质接触史，否认麻醉毒品等嗜好，否认冶游史，否认食物过敏史，否认传染病史，婚育史，适龄婚育，家族史，兄弟姐妹数人，均体健，否认家族性遗传病史。",
"syndrome": [0, 5],
"disease": 0,
"face_vec": [1, 0, 0, 0, 0, 0],
"tongue_vec": [0, 0, 0, 0, 0, 1, 0, 0],
"pulse_vec": [0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0],
"main_symptoms": [1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
}

3 实验方法

3.1 基于手工构建特征的机器学习方法

在完成结构化特征构建后，我们首先尝试基于这些手工特征直接进行分类。我们选择了两种经典的机器学习模型:

- 逻辑回归
- 随机森林

在逻辑回归中，我们使用所有手工构建的多热向量作为输入，采用对数损失函数进行训练。模型在验证集上的准确率为**0.1050**。

随后，我们采用随机森林模型进行训练，得到了更优的初步结果，其在验证集上的准确率为**0.1325**。

为进一步提升模型性能，我们对随机森林模型进行了参数网格搜索优化，主要调节的参数包括决策树数量、最大树深度、内部节点再划分所需最小样本数等。经过多组参数组合评估后，最终优化后的随机森林模型在验证集上取得了**0.1875** 的准确率，优于前两种方法，但仍低于随机猜测。

3.2 基于手工构建特征的神经网络分类器

我们进一步尝试了使用神经网络对手工构建的结构化特征进行建模，我们设计的神经网络结构如下所示：

- 输入层：维度为结构化特征拼接后的向量（共63维）
- 隐藏层一：线性变换后升维为2 ×hidden_dimffiffi$ReLU$;ffi$DropoutB$æM :hidden_dimffiffi$ReLU$;ffi$Dropout$
- 输出层：
  - 疾病预测头：使用线性层输出4维logits，采用CrossEntropyLoss（单标签分类）
  - 证型预测头：使用线性层输出10维logits，采用BCEWithLogitsLoss（多标签分类）

上述网络结构以结构化特征作为唯一输入，未引入文本语义信息。训练时，我们联合优化两个任务的损失函数。模型最终在验证集上的疾病预测准确率为**0.2300**，仍低于随机猜测。

3.3 预训练模型BERT微调

接下来我们引入了预训练语言模型BERT 进行微调。

我们选用了`google-bert/bert-base-chinese` 作为基础模型(Devlin et al., 2018)，其包含12 层Transformer 编码器，总参数量约为110M。模型输入为经过格式化拼接后的患者信息文本（字段包括性别、年龄、主诉、中医望闻切诊等），输入格式如下所示：

【性别】男 【年龄】62 【发病季节】冬天 【主诉】心慌乏力7天。. . . . . .

最终，微调后的BERT 模型在验证集上获得了**0.3075** 的准确率，显著优于传统模型与神经网络。

3.4 大语言模型Qwen2.5微调

然后，我们尝试对**Qwen2.5** 系列模型进行微调，本实验中选用`Qwen/Qwen2.5-0.5B` 作为基础模型(Team, 2024)。

在微调过程中，我们未采用自然语言生成式prompt 格式，而是将原始病例信息格式化为纯文本输入，并将模型的输出token embedding（即`last hidden state[:, 0, :]`）送入自定义的全连接分类头进行预测，病历输入格式仍与BERT 相同。

我们对Qwen2.5-0.5B模型进行了全参数微调，但是最后在测试集的正确率仅为**0.2150**，并不理想。

3.5 BERT +手工构建特征拼接

在基于`BERT` 微调基础上，我们进一步将手工构建的特征向量与BERT 编码器的输出表示进行拼接，共同作为分类器的输入。

BERT 使用`bert-base-chinese`，输入为标准病例拼接文本，结构特征在Dataset 预处理阶段提取，作为额外向量输入拼接至pooled output（768维）之后，形成总计831 维的特征。

该融合模型在验证集上达到了**0.3125** 的疾病分类准确率，略高于仅使用文本输入的BERT模型（0.3075）。

4 实验设置

4.1 实验环境

本实验在如下环境中完成：

- 操作系统：Ubuntu 22.04
- GPU：NVIDIA L20 GPU 48GB 显存
- 编程语言：Python 3.12
- 深度学习框架：PyTorch 2.7.1，CUDA 12.4，Transformers 4.41.2

4.2 数据集划分

我们使用的数据集共包含约800 条带标注的中医病例数据和200 条不能用于训练的测试集数据，每条数据包含原始病例文本与结构化标签信息。我们将训练数据中的640 条作为训练集，160 条作为验证集。所有实验均在测试集上进行评估。

5 实验结果

我们尝试了多种方法完成分类任务，其结果如下所示：

模型方法	分类任务准确率
逻辑回归（手工特征）	0.1050
随机森林（手工特征）	0.1325
随机森林+ 网格搜索	0.1875
MLP（结构特征输入）	0.2300
BERT 微调	0.3075
Qwen2.5-0.5B 微调	0.2150
BERT + 结构特征拼接	**0.3125**

Table 1: 不同方法在疾病分类任务上的准确率比较

从表1 中可以看出，BERT相比传统浅层模型具有显著优势；而手工构建特征的加入可能有助于进一步提升分类效果。

6 结论与展望

本实验围绕“中医辨证辨病”任务，探索了从浅层模型到大语言模型微调的表现差异：

- 传统机器学习模型（如逻辑回归与随机森林）在仅使用结构特征的条件下效果有限，表现不佳；
- 神经网络能够一定程度建模结构特征间的非线性关系，准确率达到0.23；
- 微调BERT 模型能够充分利用中医病例文本中的上下文语义信息，取得0.3075 的准确率；
- 在BERT 基础上引入结构化特征拼接，可进一步提升性能至0.3125；
- 微调大语言模型Qwen 却表现不佳，或许更适合使用prompt作为输入的任务形式。

未来展望：

- **优化特征构建**：如从信息更丰富的“症状”中提取特征；
- **改进结构特征融合方式：** 当前使用简单拼接的方式将结构化特征引入BERT，未来可尝试设计门控机制等方法，更有效地融合先验医学特征与文本上下文语义。
- **生成式大语言模型微调策略优化**：选用更适合大预言模型的输入输出形式进行微调，可能有更好的效果
- **引入知识图谱与中医知识库：** 融合中医理论知识，如《中医内科学》中的辨证规律、中药-证型-疾病之间的三元组等结构化知识，帮助模型结合专家知识，提升泛化能力。

7 一些迷惑

官方给出的baseline代码是对bert和qwen2.5-0.5B-Instruct进行了最简单的微调，在子任务1的正确率能够达到45%，我的微调却只能有30%的正确率。我不知道这是因为我一次性训练两个分类头的原因，还是我用的是公开测试集（200条，输入不一但输出证型和疾病类型200条全部一样）但官方用的是封闭测试集（500条），总之至少10%的差距还是很令人困惑的。

参考文献

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Qwen Team. 2024. Qwen2.5: A party of foundation models, September.

A 项目文件夹目录树

```
├ code
│  ├ Bert
│  │  ├ src
│  │  │  ├ data_insight.py
│  │  │  ├ data_preprocess.py
│  │  │  ├ dataset.py
│  │  │  ├ feature_eg.py
│  │  │  ├ model.py
│  │  │  ├ notes.txt
│  │  │  ├ test_and_eval.py
│  │  │  ├ train.py
│  │  │  └ utils.py
│  │  └ config.yaml
│  ├ data
│  │  ├ TCM-TBOSD-test-B.json
│  │  ├ test_input.json
│  │  ├ test_output.json
│  │  ├ test_raw_input.json
│  │  ├ test_raw_output.json
│  │  ├ train_raw.json
│  │  ├ train_val.json
│  │  ├ train.json
│  │  └ val.json
│  ├ manual_feature
│  │  ├ train.py
│  │  └ utils_mf.py
│  ├ MLP
│  │  ├ config_nn.yaml
│  │  ├ dataset_nn.py
│  │  ├ model_nn.py
│  │  ├ pred_and_eval.py
│  │  └ train.py
│  ├ Qwen
│  │  ├ scripts
│  │  │  ├ dataset.py
│  │  │  ├ eval.py
│  │  │  ├ model.py
│  │  │  ├ predict.py
│  │  │  ├ train.py
│  │  │  └ utils.py
│  │  ├ config.yaml
│  │  ├ prepare_env.sh
│  │  └ requirements.txt
│  └ results.md
├ data
│  ├ TCM-TBOSD-test-B.json
│  ├ test_input.json
│  ├ test_output.json
│  ├ test_raw_input.json
│  ├ test_raw_output.json
```

```
|  ├ train_raw.json
|  ├ train_val.json
|  ├ train.json
|  └ val.json
├ 实验报告.pdf
└ 跑测试集.mp4
```

B 评估指标计算公式

$$syndrome_{acc} = \frac{NUM(y \cap \hat{y})}{NUM(y)} \tag{1}$$

$$disease_{acc} = \frac{NUM(y \cap \hat{y})}{NUM(y)} \tag{2}$$

$$task1_{acc} = \frac{1}{2}(syndrome_{acc} + disease_{acc}) \tag{3}$$