实验操作-中文临床试验筛选标准分类

**一、背景简介**

临床试验是指通过人体志愿者也称为受试者进行的科学研究，目的是确定一种药物或一项治疗方法的疗效、安全性以及存在的副作用，对于促进医学发展和提高人类健康都起到关键的作用。根据试验目的等不同，受试者可能是患者或健康志愿者。筛选标准是临床试验负责人拟定的鉴定受试者是否满足某项临床试验的主要指标，分为入组标准和排出标准，一般为无规则的自由文本形式。临床试验的受试者招募一般是通过人工比较病历记录表和临床试验筛选标准完成，这种方式费时费力且效率低下。因此，临床试验面临诸多困境，比如受试者招募困难导致临床试验难以按期完成，入组患者流失影响试验的有效性等。近年来，随着临床试验设计越来越复杂，数目越来越多，基于自然语言处理和信息抽取的系统也开始在临床试验受试者招募中崭露头角并表现出不错的效果，且具有很大的实际应用前景和医学临床价值。目前这类研究大多集中在英文临床试验筛选标准及英文电子健康记录数据，针对中文电子健康数据的研究也以及取得了很多进展，然而与中文临床试验筛选标准的自然语言处理研究很少。

此实验希望能通过自然语言处理和机器学习的方法探索对临床试验筛选标准的自动解析，主要集中在筛选标准分类任务。

**二、实验说明**

在本次实验中，我们给定事先定义好的15种筛选标准类别和一系列中文临床试验筛选标准的描述句子，要求返回每一条筛选标准的具体类别。筛选标准类别及其示例如下。

**类别描述及示例：**

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **category** | **description** | **example** |
| 1 | Disease | 疾病描述，包括各种疾病名称，现患病、疾病史等 | 1.胰腺炎病史 2.高血压 3.无严重精神疾病者 4.有严重、未控制的器质性病变或感染者 5.存在慢性免疫系统疾病史 |
| 2 | Addictive Behavior | 成瘾行为，如嗜烟、吸毒、酗酒、药物滥用、 | 1.有药物成瘾的证据 2.既往有酗酒史 3.嗜烟者 4.有吸毒史或酗酒史者 5.有药物依赖或药物滥用史 |
| 3 | Age | 年龄相关描述 | 1.年龄65~75 岁 2.成人（14-60岁） |
| 4 | Allergy Intolerance | 任何过敏相关描述，包括过敏史、过敏源、过敏反应等等。 | 1.既往有药物过敏史者 2.过敏体质且既往有痰热清过敏史者 3.对外源性蛋白有过敏反应 4.不能耐受手术或麻醉者 5.无过敏史 |
| 5 | Compliance with Protocol | 病人依存性描述，包括随访 | 1.依从性差的患者 2.患者依从性差，不能进行随访 3.无法按期随访的患者 |
| 6 | Consent | 知情同意，包括知情同意书、知情同意的意愿 | 1.签署知情同意书 2.患者拒绝 3.知情同意参加本研究，并签署知情同意书 |
| 7 | Diagnostic | 诊断相关描述 | 1.符合肩颈、腰腿痛诊断标准 2.B超、妇检发现继发性痛经的患者 |
| 8 | Enrollment in other studies | 正在或之前参加其他临床试验 | 1.正在参加影响本研究结果评价的其它临床试验者 2.当前正参与另一项可能会干扰本研究结果的研究 |
| 9 | Laboratory Examinations | 在实验室进行的各项检查 | 1.左室射血分数（LVEF）≥50% 2.HIV检测阳性 3.腋下体温≥38.5℃ |
| 10 | Life Expectancy | 生存期描述 | 1.预期复发后30天内可能会出现早期死亡的患者 2.预计生存期≥12周 |
| 11 | Pharmaceutical Substance or Drug | 药物相关描述 | 1.有精神或神经科药物服用史者 2.正在服用影响维生素D代谢的药物  3.未用过糖皮质激素  4.既往注射抗VEGF药物史 |
| 12 | Organ or Tissue Status | 器官、组织状态相关描述 | 1.肾功能正常  2.合并有严重肝功能、肺功能不全的患者 3.心、肾功能在正常范围 |
| 13 | Risk Assessment | 风险评估描述 | 1.ASA分级Ⅰ～Ⅱ级 2.GCS≥8分 |
| 14 | Smoking Status | 吸烟状态相关描述 | 1.吸烟史：大于20包年 2.吸烟量≧10支/天 |
| 15 | Therapy or Surgery | 治疗或手术相关描述 | 1.脊柱外科手术史 2.盆腔手术病史 |

**输入输出示例：**

|  |  |  |
| --- | --- | --- |
| ID | 输入（筛选标准） | 输出（分类） |
| s1 | 年龄>80 岁 | Age |
| s2 | 近期颅内或椎管内手术史 | Therapy or Surgery |
| s3 | 血糖<2.7mmol/L | Laboratory Examinations |

**三、数据说明**

来源：中国临床试验注册网站 （<http://chictr.org.cn/>）

特点：文本格式，非结构化，中文，医学相关

数量：10000条

适用于：分类问题

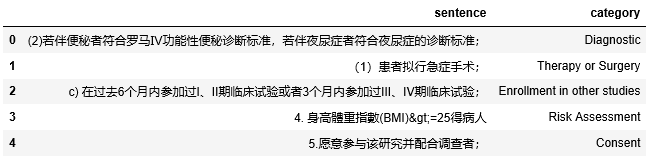
**四、实验步骤**

实验平台：Anaconda3版本的Jupyter Notebook

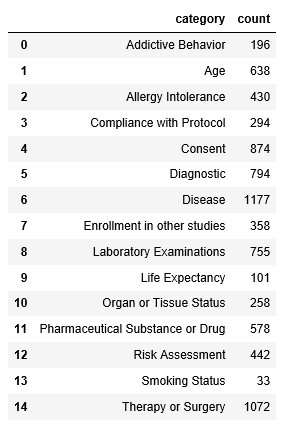
实现思路：首先观察数据，将数据按照4:1的比例分为训练集和测试集，然后通过观察数据特征，构建特征向量(如TF-IDF)，使用Sklearn包的分类算法（如SVM）进行训练，调用预测函数对测试集进行预测，将结果与原始值进行比较，并评估准确度。

1. **获得数据**

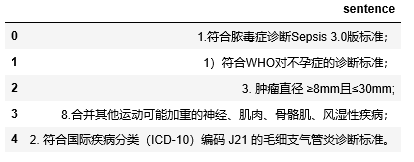
**1.1训练数据示例**



**1.2训练数据类别&数目**



**1.3 测试数据示例**



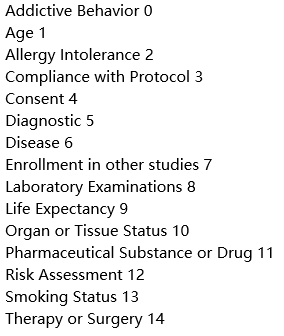
1. **数据格式转化**

**2.1通过TF-IDF算法将文本转换为词向量**

**2.2输入转化为维度为1000的词向量**

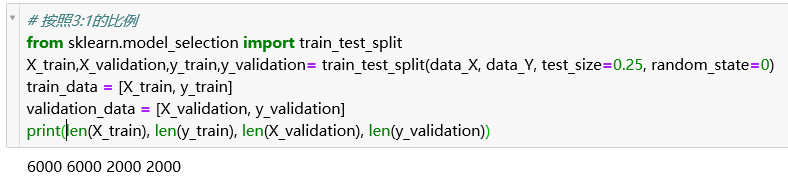


**2.3输出转化为0-14的数字表示**

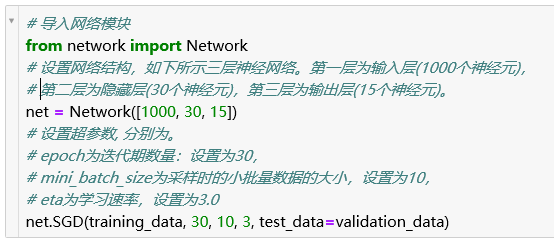


1. **训练神经网络**

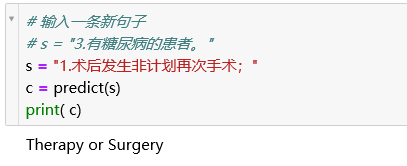
**3.1分割出一部分验证集**

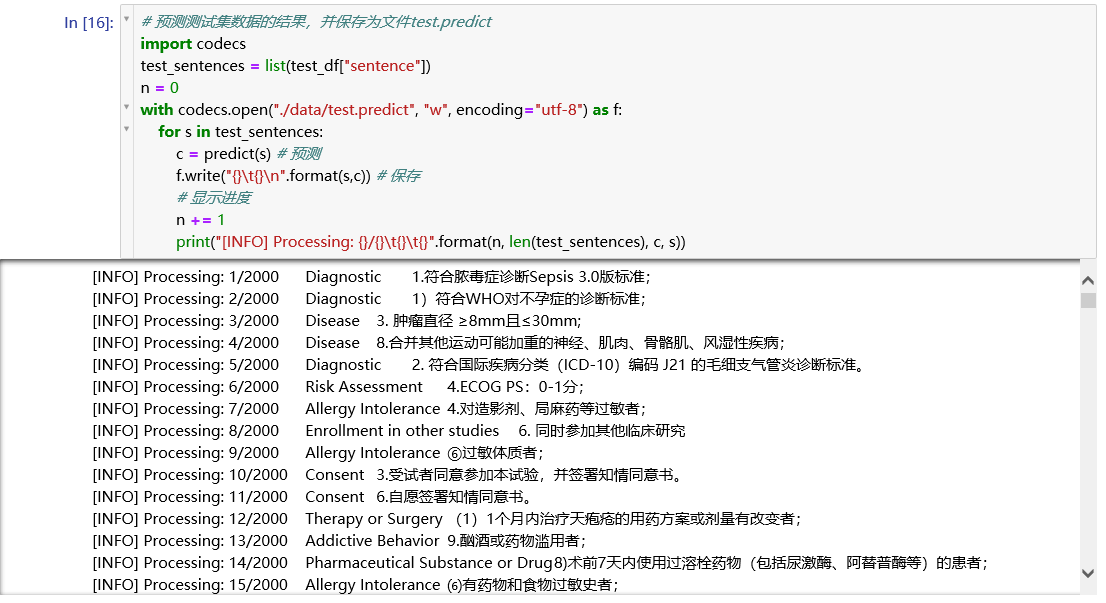


**3.2设置超参数进行训练**



1. **进行预测**





1. **评价**



