# 基于机器学习的在线问诊平台智能分诊研究

王若佳[1,2]　张　璐[1]　王继民[1]
[1](北京大学信息管理系　北京 100871)
[2](北京大学海洋研究院　北京 100871)

**摘要:**【目的】比较不同机器学习算法在智能分诊任务中的准确率, 针对性地分析在线问诊平台的类目设置问题, 尝试从数据中提取新特征提升分类器效果。【方法】基于“春雨医生”13 个科室 33 073 条实际问诊数据, 比较两种文本向量化方式在支持向量机、多项式贝叶斯、Logistic 回归、随机森林、k 近邻以及集成分类模型这 6 种分类器上实现智能分诊的准确率; 通过高频词分析及词语共现对不同科室的错分数据进一步分析。【结果】文本向量化方法为 TF-IDF、分类算法为支持向量机的分类器在智能分诊中的总体效果最优, 增加年龄和性别特征后分类准确率可达 76.3%。该分类器对外科数据分诊准确率仅为 40.9%, 原因在于问诊平台类目设置的混淆。【局限】假设现有数据中患者选择的科室是正确的。【结论】机器学习可用于在线问诊平台的智能分诊任务, 根据医疗数据特点增加输入特征是分类器提高准确率的一个方向。部分疾病及症状的跨科室性影响了分类器的效果, 在线问诊平台可通过推荐多个科室的方式来提升患者问诊体验。

**关键词:** 在线问诊　智能分诊　机器学习　支持向量机
**分类号:** TP393　G35
**DOI:** 10.11925/infotech.2096-3467.2019.0147

## 1　引　言

在“互联网+”背景下, 传统医疗健康模式正在向“互联网+医疗健康”转型, 网络预约、远程治疗、在线问诊等医疗服务也逐渐普及。其中, 在线问诊平台作为医生与患者线上沟通咨询的重要媒介, 整合了全国各地、各领域的专家医生。利用此类平台, 患者可随时随地根据具体病情选择真实的医生及专家进行咨询, 通过文字、图片、语音等多种方式与医生互动交流; 医生通过与患者的多次即时问答, 了解其病情, 并向患者提供医疗知识、诊断信息及诊疗建议等。在问诊咨询前, 患者首先需要根据自己对病情的认知, 自行选择某一科室的医生。然而, 由于大多数患者对自身病情缺乏全面了解, 且医学专业知识存在严重不足, 因此往往在自主选择医生时产生困惑, 从而出现挂错号、找错医生的情况。线下医院通常设置专门的分诊人员指导患者选择科室门诊, 但该方法对于在线问诊平台并不适用, 不仅会增加平台运营成本, 更会增加患者在线问诊的时间成本, 降低平台的便捷性和及时性。

机器学习技术是人工智能的一个分支, 在医疗健康领域已得到了多种应用, 如流行疾病预警[1]、辅助临床诊断[2]、药物不良反应识别[3]、医疗费用管理[4]等, 并取得了较好的应用效果。本研究将机器学习技术应用到在线问诊平台的智能分诊中, 目的是根据患者输入的问题, 为其推荐合适的科室。本研究基于“春雨医生”平台中 3 万余条真实数据, 构建多种分类模型, 比较其在智能分诊场景中的准确率, 选择最优

通讯作者: 王继民, ORCID: 0000-0002-3573-7788, E-mail: wjm@pku.edu.cn。

分类模型，然后深入分析分诊错误数据，提出分类器的改进方向，为在线问诊平台的分类体系建设提供参考。

## 2　文献综述

### 2.1　在线问诊平台使用的影响因素

李嘉等[5]基于好大夫在线①的医生数据分析医生价格溢价的影响因素，包括地位、声誉、医生在线服务年限、疾病隐私程度和严重程度等。刘笑笑[6]和薛书峰[7]分别基于好大夫在线问诊数据发现，医生的在线努力和声誉对其咨询量有显著影响，服务价格在其中发挥中介作用。邓朝华等[8]采用问卷调查的方法探究在线医疗健康网站医患信任的影响因素，发现网站、医院和医生的可信度对在线患者信任有显著影响。范晓妞等[9]结合医患问诊数据，发现医患双方的知识交换量、信任关系、患者收益、沟通成本等均会对在线医疗咨询的效果产生影响。

### 2.2　在线问诊平台中的医患角色与行为特征

从医生角度来看，Björk 等[10]通过对内科医生的访谈发现，以文本为基础的在线咨询平台下，医生扮演着医疗信息提供者和问询者与传统医疗服务之间中介者的角色，许多内科医生在回答问题过程中能够提升自己的沟通能力。

从患者角度来看，Umefjord 等[11-12]先后通过调查和分析实际问诊数据，针对患者使用在线问诊平台的动机和特征开展研究，其中方便性、匿名、线下医生过于忙碌、没有时间等是患者选择在线问诊的重要原因；使用群体集中在年轻和中等年龄的女性，且经常发生在晚上。Ma 等[13]进一步分析“分答”平台在线问诊对话内容，梳理患者的问诊内容和动机，问诊内容包括核实、确认、问询、推荐等，问诊动机包括没有其他选择、减少不确定性、消除疑虑、提前准备、建立联系等。

此外，吴江等[14-16]着重关注在线医疗社区的用户行为。他们采用社会网络分析方法发现在线医疗社区存在小世界效应，还分析了不同用户之间的交互行为与活跃度、不同用户群体的知识共享行为与活跃时长的差异以及社区内朋友关系的影响因素。

### 2.3　在线问诊的文本挖掘

一些学者基于自然语言处理方法，对在线问诊平台的内容进行分析，如吴江等[17]整合 LDA 主题模型及机器学习算法，设计了一个中文用户文本挖掘流程，遵循此流程可探究在线问诊平台中的社会支持类型，并通过社会支持理论揭示患者用户行为；刘通等[18]为评估医生线上回答内容的准确性，通过词汇共现网络表示医疗领域知识，然后基于信号传播算法计算实际回答和标准回答的相似度。

### 2.4　在线问诊的文本分类任务

一些研究致力于解决在线问诊的文本分类任务，帮助满足患者或医生的线上需求。Himmel 等[19]为帮助专家医生从大量咨询问题中选择适合自己回答的内容，将在线问诊平台中用户的请求分为两个维度，38 个类别，结合主成分分析、奇异值分解对数据降维，通过训练回归模型完成分类；Abdaoui 等[20]为帮助患者选择合适类别的医生，对比支持向量机、朴素贝叶斯、随机森林、决策树 J48 和 JRip 等 5 种分类算法的准确度，针对每一种类别建立二分类器，最后得到一个推荐列表。医生推荐系统也是文本分类的应用场景之一，刁必颂[21]将该任务转化为三层分类问题，遵循科室–二级科室–大类疾病的层级顺序，对患者的提问逐层分类，然后基于聚类的协同过滤推荐算法，通过查询和病人病情类似的已解决病例，推荐该解决病例的回答医生；王静[22]在此基础上，增加了对医生回答的质量评估，综合病例相似性和答案质量两方面给出最优推荐；刘通[23]选择挖掘医生专业背景信息，对平台上的医生基于专业相似度进行聚类，然后对比患者咨询问题的短文本与不同医生类别的相似程度。

综上，当前在线问诊的相关研究或使用量化方式探讨某变量的影响因素，或使用调查、访谈、内容分析等定性方法深入挖掘医患需求及使用行为，在解决文本分类任务时，虽然一些研究应用了机器学习算法，但更多是从算法本身进行讨论，而忽略了在线问诊数据本身的特征。本研究在建立分类模型的同时，对数据内容进一步分析，以挖掘分类器准确率背后的情况。

①https://www.haodf.com/.

## 3 数据与方法

### 3.1 数据采集与预处理

本研究数据来源为“春雨医生”网站[①]，该在线问诊平台创立于 2011 年 7 月，是中国最大的移动医患交流平台之一。“春雨医生”在其官方网站中提供 13 个科室的经典问答，如图 1 所示。每个问答由多条患者和医生的对话组成，为保护患者隐私，网站会将患者的身份信息自动删除，如图 2 所示。

患者使用春雨医生进行问诊的流程如图 3 所示，可以看到在进入问诊界面之前，有两种可选路径，红色箭头所指路径是指用户先提问，该提问进入问题库后，由医生选择回复哪个问题；蓝色箭头所指路径则是用户自行选择合适的医生。若针对该流程进行智能分诊，在分诊之前能获取的有效信息即为用户的第一条提问以及用户可能填写的健康档案(包括年龄、性别、过敏史等)。假设数据中每位患者实际咨询的科室都与其病情相符，则可通过机器学习方法训练出合适的智能分诊模型。

为避免科室差异所导致的误差，分别爬取 13 个科室经典问答中患者的第一个问题，并为其标上对应科室名称的标签。不同科室的示例如表 1 所示。数据预处理部分，通过导入自定义医学词典，使用 Python 的 Jieba 分词对患者的提问文本进行自动分词，并去掉“的”、“了”等无意义的停用词。以“科室”为分层依据，抽取每个科室 75%的数据作为训练集，25%作为测试集。


图 1 春雨医生网站的经典问答界面


图 2 医生和患者的对话示例

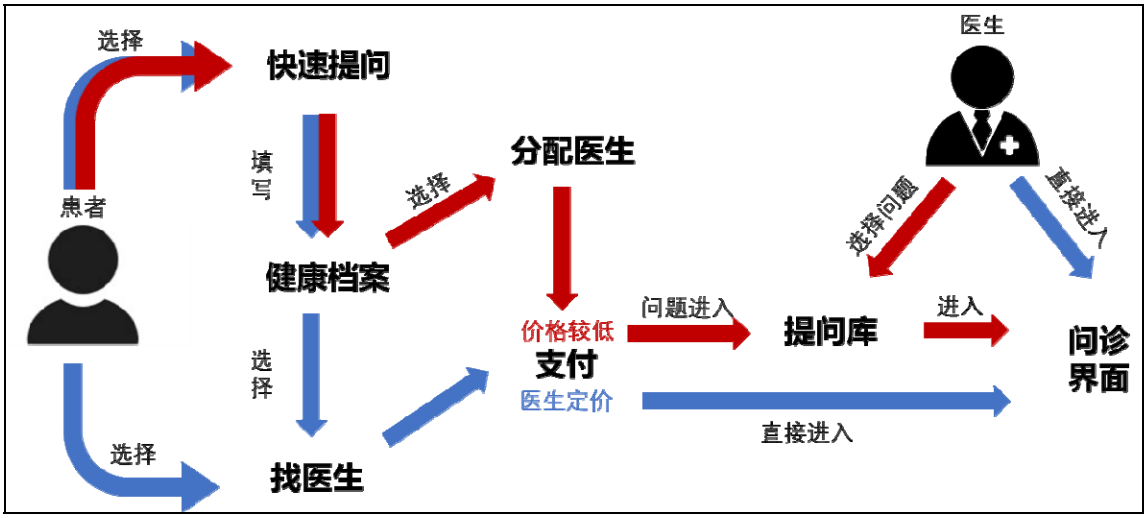


图 3 “春雨医生”平台的问诊流程

①https://www.chunyuyisheng.com/pc/qalist.

表 1　不同科室的经典问答示例

| 科室 | 示例 | 样本数(个) |
|---|---|---|
| 内科 | 我的心跳最近跳的次数在九十跳左右算正常吗 | 3 405 |
| 外科 | 60 岁老人脚后跟摔了里面有小碎片，怎么治疗 | 2 362 |
| 妇科 | 盆腔炎会肚子隐隐痛吗，没异味，白带特别粘 | 4 205 |
| 产科 | 怀孕四个月喝酒抽烟熬夜对胎儿有影响吗 | 1 937 |
| 儿科 | 7 天新生儿综合评分 36 分踏步反射 0 分是脑瘫吗 | 2 294 |
| 男科 | 睾丸紧缩好像变小了，是怎么回事呢? | 2 553 |
| 骨伤科 | 手肘关节处肿胀，可以不用打石膏固定吗 | 1 914 |
| 营养科 | 为什么有一种人每天暴饮暴食都不会胖的呢 | 3 691 |
| 肿瘤科 | 59 岁老人宫颈癌放化疗后尿失禁带点血怎么回事 | 2 103 |
| 眼科 | 62 岁青光眼晚期如何治疗 | 2 822 |
| 耳鼻咽喉科 | 鼻子塞得很严重，擦了油和通鼻贴完全没有效果，怎么办 | 2 036 |
| 口腔颌面科 | 最近这几天刷牙流血越来越厉害了怎么回事 | 1 926 |
| 皮肤性病科 | 尖锐湿疣有什么特征 | 1 825 |
| 总计 | | 33 073 |

### 3.2　算法原理

基于支持向量机、多项式贝叶斯、Logistic 回归、随机森林及 k 近邻分类算法建立智能分诊模型，并通过投票机制将这 5 种分类器的结果整合为集成分类器。

(1) 支持向量机

支持向量机(Support Vector Machine, SVM)使用非线性映射把原训练数据映射到较高的维上，在新的维上搜索最佳分离超平面，从而使不同类的数据总可以被超平面分开。由于本研究是一个多分类问题，因此基于支持向量机算法调用了 Python Scikit-learn 中的 One-Vs-The-Rest[24]。所谓 One-Vs-The-Rest，是指假设有 n 个类别，建立 n 个二项分类器，每个分类器针对其中一个类别和剩余类别进行分类。预测时利用这 n 个二项分类器进行分类并得到数据属于当前类的概率，选择其中概率最大的一个类别作为最终预测结果。

(2) 多项式朴素贝叶斯

多项式朴素贝叶斯(Multinomial Naïve Bayes, MNB)适合所有特征都是离散型的随机变量，由于文本分类时所使用的词向量就是离散型，因此该算法是文本分类任务中的常用算法之一[25]。该算法的通用公式如公式(1)所示。

$$P(w_k \mid c_i) = \frac{N_{ki} + 1}{\sum_{k=1}^{|V|} N_{ki} + |V|} \tag{1}$$

其中，$N_{ki}$ 是 $w_k$ 类别 $c_i$ 的所有文档中出现的总次数，$|V|$ 是训练数据集的总单词数。

(3) Logistic 回归

Logistic 回归(Logistic Regression)是一种广义线性回归，它将线性回归与 Sigmoid 函数相结合，进而得到一个 0-1 之间的数值，一般来说概率值大于 0.5 的数据被归为 1 代表的类别，而小于 0.5 的则被归为 0 类[26]。Sigmoid 函数及 Logistic 回归常用公式如公式(2)和公式(3)所示。

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

$$h_\theta(x) = f(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n) \tag{3}$$

该算法一般用于处理因变量为二分类变量的回归问题，本研究基于 One-Vs-The-Rest 方法将 Logistic 回归用于处理文本多分类问题。

(4) 随机森林

随机森林(Random Forest)是在以决策树为个体学习器的基础上，加入了随机样本选择和随机特征选择。简单来说，该算法采用有放回抽样策略从原数据集中抽取样本，采用无放回抽样策略抽取不同特征作为输入变量，在每个新数据集上构建决策树，综合多棵决策树的预测结果作为随机森林的预测结果[27]。

(5) k 近邻

k 近邻法(K-Nearest Neighbor, KNN)是一种传统的基于统计的模式识别方法，其分类算法思想较为简单：如果一个样本在特征空间中的 k 个最相似的样本中有大多数属于某一个类别，则该样本也属于这个类别[28]。算法的关键参数设定为 k 值，即最近邻居的个数，由于本研究中的预测类为 13 个科室，因此选择 k 为 13。

(6) 集成分类

集成学习中，最为简单的机制就是分类器投票。本研究整合以上 5 类算法的分类器，使用投票机制，以少数服从多数的原则，得到最终分类结果。

## 4 研究结果

### 4.1 最优分类器的选择

本研究比较了两种文本向量化方式在 6 个分类器上的准确率(分对样本数与所有样本数之比)。两种文本向量化方式分别为 Count 和 TF-IDF, 其中 Count 方法是将文本中的词语直接转换为词频矩阵。但有时候出现频次高的词语并不一定能代表该文本的特征, 而 TF-IDF 在限定词语频次高的同时, 还要求该词语具有一定的区分能力, 一般来说效果会更好。

具体结果如表 2 所示, 总体来说, 使用 TF-IDF 提取文本特征的支持向量机算法具有相对最高的准确率, 为 75.1%。当向量化方式为词频矩阵时, 集成分类器准确率最高, 为 74.5%。其他分类器中, 逻辑回归分类器的准确率分别为 74.2%和 74.0%, 效果较好, 而 k 近邻分类器准确率最低, 仅为 48.4%和 54.9%。综上, 选择文本向量化方法为 TF-IDF、分类算法为支持向量机的分类器作为最优分类器, 并对其在测试集上的结果作进一步分析。

表 2 分类器的准确率比较

| 分类算法 | Count | TF-IDF |
|---|---|---|
| 支持向量机 | 73.4% | 75.1% |
| 随机森林 | 68.6% | 70.0% |
| 多项式贝叶斯 | 71.8% | 69.1% |
| 逻辑回归 | 74.2% | 74.0% |
| k 近邻 | 48.4% | 54.9% |
| 集成分类 | 74.5% | 74.4% |

### 4.2 科室准确率的比较

为进一步了解支持向量机模型在不同科室智能分诊的有效性, 分别计算不同科室的分诊准确率, 如表 3 所示。13 个科室的数据量与分诊准确率并不显著相关($t$=1.064, $p$=0.310), 不同科室数据量的差异并未直接影响各科室的分诊准确率。具体来看, 在 13 个科室中, 共有 12 个科室准确率超过 60%, 眼科的准确率最高, 接近 95%, 有 5 个科室的准确率介于 80%-90%, 有 3 个科室的准确率介于 70%-80%, 可见, 分类器在不同科室的分诊效率虽存在差异, 但整体上均能保持较好的性能。然而, 外科的智能分诊准确率只有 40%左右。外科是在线问诊平台及线下医院最为常见的科室之一, 患者咨询问题的多样性也会更高, 需要进一步分析错误分诊的特征及原因, 找出提升智能分诊准确率的潜在方法。

表 3 分类器在不同科室中的分诊效果

| 科室 | 数据量 | 分诊准确率 |
|---|---|---|
| 眼科 | 565 | 94.9% |
| 营养科 | 738 | 85.0% |
| 口腔颌面科 | 385 | 84.2% |
| 耳鼻喉科 | 407 | 82.6% |
| 肿瘤科 | 421 | 82.2% |
| 妇科 | 841 | 82.2% |
| 骨伤科 | 383 | 72.6% |
| 内科 | 681 | 72.2% |
| 男科 | 511 | 72.2% |
| 产科 | 387 | 66.1% |
| 儿科 | 459 | 64.3% |
| 皮肤性病科 | 365 | 62.5% |
| 外科 | 472 | 40.9% |

### 4.3 错分数据的分析

提取测试集数据中被错误分诊的样本, 计算两两科室的分诊错误率, 进而找出各科室错误率最高的对应科室, 分析其特征, 不同科室的错分情况如表 4 所示。外科与男科是最容易混淆的两个科室, 有 25%的外科数据被错误预测为男科; 有 10%的男科数据被错误预测为外科。其次是产科与妇科, 有 22%的产科数据被错误预测为妇科, 7%的妇科数据被错误

表 4 不同科室的错分情况

| 原始科室 | 预测科室 | 错误率 |
|---|---|---|
| 外科 | 男科 | 25% |
| 产科 | 妇科 | 22% |
| 儿科 | 内科 | 10% |
| 男科 | 外科 | 10% |
| 妇科 | 产科 | 7% |
| 内科 | 儿科 | 6% |
| 骨伤科 | 皮肤性病科 | 5% |
| 皮肤性病科 | 内科 | 5% |
| 营养科 | 儿科 | 5% |
| 肿瘤科 | 妇科 | 5% |
| 耳鼻喉科 | 内科 | 4% |
| 口腔颌面科 | 内科 | 3% |
| 眼科 | 皮肤性病科 | 1% |

预测为产科。再次是内科与儿科，分别为 10%和 6%。此外，尽管内科与其他单一科室的错误率不算高，但内科与儿科、皮肤性病科、耳鼻喉科、口腔颌面科等多个科室均存在混淆，这也与内科本身的特征密切相关。

为进一步探究两两科室混淆的原因，分别统计不同科室在线问诊数据的高频词，并找出错误率较高的两两科室共同出现的常见高频词。在各科室的前 100 高频词中，外科和男科的高频词有 62%重叠，产科与妇科的高频词有 52%重叠，儿科与内科的高频词有 54%重叠。常见的高频易混词如表 5 所示。可以看到，患者在咨询这些科室的医生时，常提到的症状或疾病具有一定的跨科室性，并不是单一对应于某一科室，比如儿童发烧，不仅是儿科能解决的问题，也是内科诊治的病症。事实上，临床实践中某一科室的医生也很有可能会遇到并处理其他相关科室的病情，比如普通外科医生会解决涉及擦伤、骨折等骨伤科病症。这是外科、儿科、产科等科室智能分诊准确率偏低的重要原因。

表 5　科室常见高频易混词

| 科室 | 常见高频易混词举例 |
|---|---|
| 外科–男科 | 龟头、阴茎、手淫、勃起、早泄、包皮、尿、睾丸、疼、精子、性生活、前列腺炎、痒、手术、龟头炎 |
| 产科–妇科 | 月经、怀孕、流产、检查、子宫、出血、严重、疼、自然流产、分泌物、流血、孩子 |
| 儿科–内科 | 发烧、咳嗽、治疗、感冒、药、反复、症状、大便、吐、检查、拉肚子、痰 |

### 4.4　分类器效果的提升

增加类别特征是提高分类器准确率的方法之一。在“春雨医生”的平台上，用户咨询医生前可选填健康档案表格，包括年龄、性别、过敏史等内容。对 33 073 条数据进行抽取，发现其中 29 133 条数据具有年龄和性别特征。以这 29 133 条带有特征的数据作为样本进行下一步分析。

对 13 个科室的平均年龄及男女比例做基本统计，结果如表 6 所示。13 个科室的总体平均年龄为 28.6 岁，其中儿科的年龄平均值明显偏低，仅为 10.9 岁，肿瘤科则明显偏大，为 47.3 岁。男女比例方面，男科、妇科、产科这三个科室由于科室的特殊性，男女比例偏差极大，虽然有部分女性/男性替自己的配偶咨询医生，但这种情况还是相对较少；此外，外科患者中有 64%为男性，女性则更倾向于问以保健、减肥为主的营养科。

表 6　不同科室患者的平均年龄和性别比例

| 科室 | 年龄平均值 | 男性比例 | 女性比例 |
|---|---|---|---|
| 妇科 | 27.2 | 3.5% | 96.5% |
| 产科 | 27.0 | 4.1% | 95.9% |
| 营养科 | 23.6 | 35.1% | 64.9% |
| 儿科 | 10.9 | 43.1% | 56.9% |
| 口腔颌面科 | 26.6 | 43.5% | 56.5% |
| 皮肤性病科 | 25.8 | 44.1% | 55.9% |
| 眼科 | 28.3 | 45.0% | 55.0% |
| 肿瘤科 | 47.3 | 45.5% | 54.5% |
| 耳鼻喉科 | 27.8 | 47.8% | 52.2% |
| 内科 | 34.8 | 48.9% | 51.1% |
| 骨伤科 | 34.0 | 51.3% | 48.7% |
| 外科 | 31.3 | 64.0% | 36.0% |
| 男科 | 26.9 | 94.4% | 5.6% |
| 总体 | 28.6 | 43.9% | 56.1% |

综上，部分科室在年龄和性别两个属性上有自身特点。因此将这两个特征和问题文本一起作为分类器的输入变量进行训练，并对比增加特征前和增加特征后不同科室的准确率，如表 7 所示。总体来看，增加特征前的分类器准确率为 75.5%，增加后上升到 76.3%。具体到科室上，年龄平均值显著较低的儿科和男性比例显著较高的男科分诊预测准确率均有 3.5%的提高，平均年龄较高的肿瘤科准确率也提升了 1.0%。

表 7　不同科室特征增加前后的分诊准确率比较

| 科室 | 增加特征前准确率 | 增加特征后准确率 | 提升率 |
|---|---|---|---|
| 妇科 | 82.7% | 83.2% | 0.5% |
| 产科 | 67.4% | 67.9% | 0.5% |
| 营养科 | 86.7% | 87.5% | 0.8% |
| 儿科 | 58.3% | 61.8% | 3.5% |
| 口腔颌面科 | 81.6% | 82.1% | 0.4% |
| 皮肤性病科 | 60.6% | 60.6% | 0.0% |
| 眼科 | 99.4% | 99.4% | 0.0% |
| 肿瘤科 | 75.5% | 76.6% | 1.0% |
| 耳鼻喉科 | 85.8% | 84.7% | –1.1% |
| 内科 | 73.2% | 73.8% | 0.5% |
| 骨伤科 | 70.4% | 71.1% | 0.7% |
| 外科 | 45.8% | 46.4% | 0.6% |
| 男科 | 70.0% | 73.5% | 3.5% |
| 总体 | 75.5% | 76.3% | 0.8% |

## 5 结 语

本研究基于真实的在线问诊数据，使用5种机器学习算法建立5个分类器，并基于投票机制将这5个分类器整合，得到集成分类模型。对比这6个分类器在在线问诊平台智能分诊任务中的效果，发现：

(1) 总体来说，支持向量机、Logistic回归和集成分类器更适合分诊任务。这也验证了支持向量机处理高维、稀疏环境下文本分类的优势[29]。

(2) 不同分类算法适合的文本向量化方式并不相同，多项式贝叶斯、Logistic 回归和集成分类器直接使用统计词频的词袋模型准确率更高。而支持向量机、随机森林和k-近邻更适合处理以TF-IDF方法进行文本向量化的预测。

(3) 不同科室的预测效果差异较大，准确率最高的眼科可达94.9%，而外科仅有40.9%。通过对错分数据做进一步分析，发现一些科室之间存在重叠部分，例如怀孕、流产等问题，妇科和产科的医生都可对此提供咨询服务。所以从根本上，并不能说分类器的结果错误，而是数据所代表的疾病本身存在跨科室性。刘浏等[30]探究使用机器学习方法量化研究社科类论文跨学科性的可行性，发现使用期刊所在学科作为论文初始类别会影响自动分类的效果，而这种分类效果不好并不完全由算法、特征及语料导致，还跟论文本身存在的学科交叉性息息相关。对于智能分诊任务，这种问题或许可以通过为患者推荐多个科室来解决。

(4) 探讨年龄、性别特征是否能提高分类器效率时，在删除近4 000条缺失年龄、性别字段数据的基础上，分类器的准确率仍从75.1%上升至75.5%。可见即使样本数量的增加会为机器学习算法提供更多可学习的特征，但数据的质量问题同样重要。

(5) 加入年龄、性别特征后，分类器对平均年龄较低的儿科、平均年龄较高的肿瘤科以及男性比例较高的男科预测准确率都有所提高，而对年龄、性别比例并无太大差异的耳鼻喉科来说，准确率反而有所下降。这说明虽然提取更多的数据特征是分类器提高准确率的一个方向，但该特征需要具体问题具体分析，盲目增加特征有可能得不偿失。

此外，本研究通过机器学习算法建立起患者问题与科室类目之间的映射关系，还发现了在线问诊平台本身的一些问题。

(1) 从科室类目来看，与传统医疗服务模式不同，在线问诊平台的科室分类采用扁平化模式，根据患者的咨询需求设置相应科室类目。以“春雨医生”为代表的科室设置并不是平级并列关系，而是存在部分包含关系，比如骨科，在传统综合医院中隶属于外科，但由于在线问诊需求较多，将其独立出来作为一级类目。因此，某一疾病与科室并不一定是一对一的关系，很有可能是一对多的关系，比如6岁男孩咳嗽对应儿科和内科。但不同患者已有的医疗知识基础不同，对不同科室的定位和诊治范围不够了解，在线问诊选择科室过程中可能会产生困惑。智能分诊任务的实现使平台可以针对患者的具体问题向其推荐1-2个科室，减少患者在寻找科室和医生过程中的负担，提高患者问诊效率。

(2) 从就诊流程来看，在线问诊平台打破传统医疗服务中“科室→医生→问诊”的线性流程，逐渐从以疾病为中心的服务向以患者为中心的问题导向服务转型[31]。为进一步提升患者的问诊体验，平台应能直接面向患者的具体问题提供服务。智能分诊任务能够直接针对患者的就诊问题智能匹配相关科室，平台也可据此提高匹配医生的准确性，从而优化问题导向的在线问诊流程，提高患者满意度。

(3) 从在线健康档案来看，在线问诊平台在问诊环节需要患者填写年龄和性别等个人信息，以辅助医生进行诊断。但数据分析过程中发现，在非本人问诊情境下，问诊人经常会填写本人的年龄性别信息，医生需要在对话过程中进一步询问真正患者的年龄和性别，影响问诊效率。因此，在线问诊平台可进一步优化设计界面，增加患者问题描述的相关提示和说明，以帮助患者更好地描述病情并填写相关信息。

本研究的局限在于假设现有数据中患者选择的科室是正确的，但事实上数据质量并没有那么理想。未来，对于智能分诊任务，除了把控数据质量，还将尝试使用Word2Vec等文本向量化方式和LSTM等深度学习算法提升模型准确率[32]；此外，还可以将智能分诊抽象为文本多标签分类任务，为某些跨科室的患者提问推荐多个科室类别。

## 参考文献:


[1] Pineda A L, Ye Y, Visweswaran S, et al. Comparison of Machine Learning Classifiers for Influenza Detection from Emergency Department Free-text Reports[J]. Journal of Biomedical Informatics, 2015, 58: 60-69.

[2] 孔倩, 王杜娟, 王延章, 等. 基于多目标神经网络的前列腺癌诊断方法[J]. 系统工程理论与实践, 2018, 38(2): 532-544. (Kong Qian, Wang Dujuan, Wang Yanzhang, et al. Multi-Objective Neural Network-Based Diagnostic Model of Prostatic Cancer[J]. Systems Engineering - Theory & Practice, 2018, 38(2): 532-544.)

[3] Nikfarjam A, Sarker A, O'connor K, et al. Pharmacovigilance from Social Media: Mining Adverse Drug Reaction Mentions Using Sequence Labeling with Word Embedding Cluster Features[J]. Journal of the American Medical Informatics Association, 2015, 22(3): 671-681.

[4] Kose I, Gokturk M, Kilic K. An Interactive Machine-Learning-Based Electronic Fraud and Abuse Detection System in Healthcare Insurance[J]. Applied Soft Computing, 2015, 36: 283-299.

[5] 李嘉, 唐洁, 蒋玲, 等. 在线健康咨询市场中的价格溢价研究[J]. 管理科学, 2018, 31(1): 15-32. (Li Jia, Tang Jie, Jiang Ling, et al. Price Premiums in the Online Health Consultation Market[J]. Journal of Management Science, 2018, 31(1): 15-32.)

[6] 刘笑笑. 在线医生信誉和医生努力对咨询量的影响研究[D]. 哈尔滨: 哈尔滨工业大学, 2014. (Liu Xiaoxiao. The Impact of Online Doctor Reputation and Doctor Effort on Consultation Amount[D]. Harbin: Harbin Institute of Technology, 2014.)

[7] 薛书峰. 互联网医疗的定价影响因素研究[D]. 南京: 南京大学, 2016. (Xue Shufeng. Research on the Factors Affecting the Pricing of Online Healthcare Community[D]. Nanjing: Nanjing University, 2016.)

[8] 邓朝华, 洪紫映. 在线医疗健康服务医患信任影响因素实证研究[J]. 管理科学, 2017, 30(1): 43-52. (Deng Zhaohua, Hong Ziying. An Empirical Study of Patient-physician Trust Impact Factors in Online Healthcare Services[J]. Journal of Management Science, 2017, 30(1): 43-52.)

[9] 范晓妞, 艾时钟. 在线医疗社区参与双方行为对知识交换效果影响的实证研究[J]. 情报杂志, 2016, 35(7): 173-178. (Fan Xiaoniu, Ai Shizhong. An Empirical Study on the Relationship Between Online Medical Community Participants' Behaviors and Knowledge Exchange Effect[J]. Journal of Intelligence, 2016, 35(7): 173-178.)

[10] Björk A B, Hillborg H, Augutis M, et al. Evolving Techniques in Text-Based Medical Consultation–Physicians' Long-Term Experiences at an Ask the Doctor Service[J]. International Journal of Medical Informatics, 2017, 105: 83-88.

[11] Umefjord G, Petersson G, Hamberg K. Reasons for Consulting a Doctor on the Internet: Web Survey of Users of an Ask the Doctor Service[J]. Journal of Medical Internet Research, 2003, 5(4): e26.

[12] Umefjord G, Sandström H, Malker H, et al. Medical Text-Based Consultations on the Internet: A 4-Year Study[J]. International Journal of Medical Informatics, 2008, 77(2): 114-121.

[13] Ma X, Gui X, Fan J, et al. Professional Medical Advice at Your Fingertips: An Empirical Study of an Online[J]. Proceedings of the ACM on Human-Computer Interaction, 2018, 2: Article No. 116.

[14] 吴江, 周露莎. 在线医疗社区中知识共享网络及知识互动行为研究[J]. 情报科学, 2017, 35(3): 144-151. (Wu Jiang, Zhou Lusha. The Study of Knowledge Sharing Network and Users' Knowledge Interaction in Online Health Community[J]. Information Science, 2017, 35(3): 144-151.)

[15] 吴江, 施立. 基于社会网络分析的在线医疗社区用户交互行为研究[J]. 情报科学, 2017, 35(7): 120-125. (Wu Jiang, Shi Li. Study of the User Interaction Behavior in Online Health Community Based on Social Network Analysis[J]. Information Science, 2017, 35(7): 120-125.)

[16] 吴江, 李姗姗, 周露莎, 等. 基于随机行动者模型的在线医疗社区用户关系网络动态演化研究[J]. 情报学报, 2017, 36(2): 213-220. (Wu Jiang, Li Shanshan, Zhou Lusha, et al. Research on Dynamic Evolution of Users' Relationship Network in Online Health Community Based on Stochastic Actor-oriented Model[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(2): 213-220.)

[17] 吴江, 侯绍新, 靳萌萌, 等. 基于 LDA 模型特征选择的在线医疗社区文本分类及用户聚类研究[J]. 情报学报, 2017, 36(11): 1183-1191. (Wu Jiang, Hou Shaoxin, Jin Mengmeng, et al. LDA Feature Selection Based Text Classification and User Clustering in Chinese Online Health Community[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(11): 1183-1191.)

[18] 刘通, 杨敬成. 基于信号传播算法的在线医疗咨询反馈内容评估方法[J]. 数据分析与知识发现, 2017, 1(11): 29-36. (Liu Tong, Yang Jingcheng. Evaluating Online Healthcare Consultation Feedbacks Based on Signal Transmission Algorithm[J]. Data Analysis and Knowledge Discovery, 2017,

1(11): 29-36.)

[19] Himmel W, Reincke U, Michelmann H W. Text Mining and Natural Language Processing Approaches for Automatic Categorization of Lay Requests to Web-Based Expert Forums[J]. Journal of Medical Internet Research, 2009, 11(3): e25.

[20] Abdaoui A, Azé J, Bringay S, et al. Assisting E-patients in an Ask the Doctor Service[J]. Studies in Health Technology and Informatics, 2015, 210: 572-576.

[21] 刁必颂. 基于在线患者咨询数据的在线医生推荐系统研究[D]. 北京: 北京理工大学, 2016. (Diao Bisong. Online Patient Counseling Data Based Online Doctor Recommend System Research[D]. Beijing: Beijing Institute of Technology, 2016.)

[22] 王静. 在线问诊平台相似病例推荐[D]. 哈尔滨: 哈尔滨理工大学, 2017. (Wang Jing. Similar Cases Recommendation on Online Medical Diagnose Platform[D]. Harbin: Harbin University of Science and Technology, 2017.)

[23] 刘通. 基于在线咨询记录的医生自动匹配算法应用研究[J]. 情报理论与实践, 2018, 41(6): 147-152. (Liu Tong. An Application Research of Automatic Physician Matching Algorithm Based on Online Healthcare Consultation Records[J]. Information Studies: Theory & Application, 2018, 41(6): 147-152.)

[24] Scikit-learn. One-Vs-The-Rest[EB/OL]. [2018-02-02]. https://scikit-learn.org/stable/modules/multiclass.html#one-vs-the-rest.

[25] Kibriya A M, Frank E, Pfahringer B, et al. Multinomial Naive Bayes for Text Categorization Revisited[C]// Proceedings of the Australasian Joint Conference on Artificial Intelligence. 2004: 488-499.

[26] Scikit-learn. Logistic Regression[EB/OL]. [2018-02-02]. https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.

[27] Scikit-learn. Random Forests[EB/OL]. [2018-02-02]. https://scikit-learn.org/stable/modules/ensemble.html#random-forests.

[28] Scikit-learn. Nearest Neighbors Classification[EB/OL]. [2018-02-02]. https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification.

[29] 王昊, 叶鹏, 邓三鸿. 机器学习在中文期刊论文自动分类研究中的应用[J]. 现代图书情报技术, 2014(3): 80-87. (Wang Hao, Ye Peng, Deng Sanhong. The Application of Machine-Learning in the Research on Automatic Categorization of Chinese Periodical Articles[J]. New Technology of Library and Information Service, 2014(3): 80-87.)

[30] 刘浏, 王东波. 基于论文自动分类的社科类学科跨学科性研究[J]. 数据分析与知识发现, 2018, 2(3): 30-38. (Liu Liu, Wang Dongbo. Identifying Interdisciplinary Social Science Research Based on Article Classification[J]. Data Analysis and Knowledge Discovery, 2018, 2(3): 30-38.)

[31] Ishikawa H, Hashimoto H, Kiuchi T. The Evolving Concept of "Patient-Centeredness" in Patient–Physician Communication Research[J]. Social Science & Medicine, 2013, 96: 147-153.

[32] 赵明, 杜会芳, 董翠翠, 等. 基于 Word2Vec 和 LSTM 的饮食健康文本分类研究[J]. 农业机械学报, 2017, 48(10): 202-208. (Zhao Ming, Du Huifang, Dong Cuicui, et al. Diet Health Text Classification Based on Word2Vec and LSTM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(10): 202-208.)

## 作者贡献声明:

王若佳: 提出研究思路, 设计研究方案, 进行实验, 论文写作;
张璐: 清洗和分析数据, 论文写作;
王继民: 修改研究思路, 修改论文。

## 利益冲突声明:

所有作者声明不存在利益冲突关系。

## 支撑数据:

支撑数据由作者自存储, E-mail: Wangruojia@pku.edu.cn。

[1] 王若佳. Chunyu_doctor.csv. 春雨医生平台中的患者第一条提问及所选科室.

[2] 王若佳. Chunyu_doctor2.csv. 春雨医生平台中患者第一条提问、性别、年龄及所选科室.

[3] 王若佳. Classifier.py. 分类器的 Python 程序.

[4] 王若佳. Precision_specialties.xlsx. 不同科室的分类准确率.

[5] 王若佳. Error rate.xlsx. 两两科室的分诊错误率.

[6] 王若佳. Hot_words.py. 提取高频词的 Python 程序.

[7] 王若佳. Classifier_feature.py. 将文本特征和性别、年龄特征结合在一起的分类器的 Python 程序.

收稿日期: 2019-02-11
收修改稿日期: 2019-03-21

# Automatic Triage of Online Doctor Services Based on Machine Learning

Wang Ruojia[1,2]　Zhang Lu[1]　Wang Jimin[1]
[1](Department of Information Management, Peking University, Beijing 100871, China)
[2](Institute of Ocean Research, Peking University, Beijing 100871, China)

**Abstract: [Objective]** This paper compares the performance of various machine learning algorithms for automatic triage, aiming to improve their effectiveness through analyzing mis-classification data. **[Methods]** First, we retrieved 33,073 real patients' questions from a website named "chunyu doctor". Then, we compared the accuracy of two text vectorization methods and six classification models. Finally, we analyzed the mis-classification data and extracted new features to improve the performance of models. **[Results]** The best automatic triage model used TF-IDF as text vectorization method and support vector machine as classification algorithm. After adding age and gender characteristics, the classification accuracy rate reached 76.3%. The classifier had the lowest accuracy rate for surgery department due to the setting of this platform's categories. **[Limitations]** We assumed that the department selection of the patient was correct. **[Conclusions]** Machine learning techniques could improve the performance of automatic triage services of the online health consulting platforms.

**Keywords:** Ask the Doctor Service　Automatic Triage　Machine Learning　Support Vector Machine

## 新算法识别 Twitter 网络欺凌的准确度高达 90%

近日，某研究团队开发了一个机器学习算法，识别 Twitter 上的霸凌和侵略者的准确度高达 90%。目前，缺少能够有效地检测社交媒体上有害行为的工具，因为这种行为在本质上通常是模棱两可的，并且通常通过看似肤浅的评论和批评表现出来。为了解决这一问题，研究团队分析了有滥用行为的 Twitter 用户所表现出的行为模式以及他们与其他 Twitter 用户之间的差异。

"我们编写爬虫程序，通过各种机制从 Twitter 收集数据，包括用户的推文内容、个人资料以及与社交网络相关的信息(如关注的人和粉丝)。"然后，研究人员对推文本身进行自然语言处理和情感分析，并对用户之间的联系进行各种社交网络分析。研究人员开发了自动分类攻击性网络行为的两种特定类型的算法，即网络欺凌和网络侵略。该算法能够以 90%的准确性识别 Twitter 上有滥用行为的用户，即从事骚扰行为的用户，例如发送死亡威胁或向其他用户发表种族主义言论。"简而言之，算法通过权衡某些特征来学习如何分辨欺凌者和典型用户。"

研究人员认为，尽管这项研究可以帮助减轻网络欺凌，但这只是第一步。"最关键的问题是对人类的伤害，而且很难撤消。研究结果发现，机器学习可以用于自动检测网络欺凌者，从而帮助 Twitter 和其他社交媒体平台剔除有问题的用户。但是，这样的系统从根本上来说是被动的，它并不能从根本上防止欺凌，即使删除了欺凌账户以及所有先前的攻击记录，受害者仍然能看到并受到欺凌者的影响。"该团队目前继续探索积极的缓解技术，以应对骚扰活动。

(编译自:https://www.sciencedaily.com/releases/2019/09/190916092101.htm)

(本刊讯)