基于大语言模型及电子病历的临床医学NLP系统构建:以患者饮酒信息分析为例

——李春晓 芮欣凯

【摘 要】 以患者饮酒信息分析为切入点,使用多种算法及机器学习模型,构建了一种结合分类规则与大语言模型的临床医学自然语言处理系统。该系统可以识别电子病历中的患者饮酒信息,并对整份病历所包含的饮酒状态进行标记。与已有自然语言处理系统相比,该系统更贴合中文病历语言特点,并在细粒度饮酒状态识别上表现突出,为电子病历文本的智能化分析提供了参考路径。

【关键词】 大语言模型;自然语言处理;特征选择;电子病历
中图分类号:R197.323 文献标识码:B
Construction of a Clinical Medicine NLP System Based on Large Language Models and Electronic Medical Records: a Case Study of Patient Alcohol Consumption Information Analysis/LI Chunxiao, RUI Xinkai. //Chinese Health Quality Management, 2025, 32(12): 18-22

Abstract This study takes the analysis of patient alcohol consumption information as the entry point, employing multiple algorithms and machine learning models to construct a clinical medicine natural language processing system that integrates classification rules with large language models. This system is capable of identifying patient alcohol consumption information within electronic medical records and labeling the alcohol consumption status encompassed in the entire medical record. Compared with existing natural language processing systems, this research is more aligned with the linguistic characteristics of Chinese medical records and demonstrates outstanding performance in fine-grained alcohol consumption status recognition, providing a reference pathway for the intelligent analysis of Chinese electronic medical records texts.

Key words Large Language Model; Natural Language Processing; Feature Selection; Electronic Medical Record

First-author's address Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, 200020, China

电子病历(electronic medical records, EMR)是临床医学信息的重要来源之一[1]。合理、有效地利用EMR在提升医疗服务质量、保障医疗安全及降低医疗成本方面具有重要意义。然而,在国内主流的EMR系统中,大量临床相关信息以非结构化自由文本的形式存储,并呈现出独特的语言特征。这些特征主要包括短文本用词不规范,包含大量专有名词、音译词、专用符号及中英文混杂的表达方式,且涵盖多类临床医学领域等。此外,EMR文本常表现出语义不完整,如大量使用缩略词或简写,并存在同义词、同音词等现象。这些特性使临床医学文本在语言学特征上显著区别于新闻报道、文学作品及生物医学学术文献等文本类型,从而给基于大语言模型的文本处理、信息检索及知识挖掘带来了新的挑战。当前的通用汉语自然语言处理(natural language processing, NLP)工具并非专门针对临床医学领域设计,因而难以准确解析病历文本的数据特征。在实际应用中,临床医生和医学研究人员为了充分挖掘和使用这些临床数据,通常需要依赖人工分析病历文本,其过程不仅成本高、效率低,且容易产生错误,针对临床医学子领域构建专业化的NLP工具已成为迫切需求。

患者的饮酒状态在多种健康问题的研究和诊断中,尤其是对食道癌和肝硬化等疾病的影响方面[2],扮演了重要角色。少量的饮酒信息通常隐藏在海量的EMR非结构化叙述文本中,需要使用复杂技术进行提取和分类。为此,本研究以患者饮酒信息分析为例,提出了一种临床医学NLP系统,将基于分类规则的方法与大语言模型学习技术相

DOI:10.13912/j.cnki.chqm.2025.32.12.04
上海交通大学医学院附属瑞金医院 上海 200020

结合，以期为临床医学文本的智能化处理提供解决方案。

1　现状分析

目前，国外已有多个具有较大影响的 NLP 系统应用于生物医学和临床医学领域。美国医学图书馆[3]开发了 MetaMap，该系统能够将非结构化文本映射到统一医学语言系统(unified medical language system，UMLS)概念[4]；Jain 等人[5]设计的医学语言提取与编码系统最初用于从放射报告中提取诊断信息，后进一步扩展至其他临床专业；Hua 等人[6]开发的 MedEx 系统最初应用于出院总结，随后扩展到门诊病历，该系统结合了查找算法、正则表达式和业务规则等多种方法，显著提高了临床病历中药物名称、剂量、用法和频率等信息的提取和效率精度；此外，Nadkarni 等人[7]开发的 UMLS 概念定位器是一个带有用户图形界面的系统，支持通过词组、概念及词组与概念的组合对临床文本进行精确查询。相较于国外将 NLP 技术应用于临床信息处理，特别是在 EMR 信息的利用方面取得的显著进展，国内在这一领域起步较晚，目前的研究主要集中在 EMR 的基本数据集定义[8]、互操作性[9]、临床路径[10]等方面，较少涉及 EMR 中叙述文本的处理、查找及检索。因此，本研究提出的临床医学 NLP 系统应用于汉语书写的入院病历，是自动化汉语处理技术在 EMR 应用中的一次尝试，扩展了医学语言处理的范围。

2　研究设计

2.1　数据采集及预处理

作为研究基础，需构建实验所需的患者饮酒状态标记标准分类数据集。为此，从某三级甲等医院 EMR 系统中随机抽取 350 份住院患者入院病历，采用本地化 HL7 CDA R2[11]兼容格式存储。所有病历数据均经过脱敏处理，移除了患者身份证号、姓名、性别、年龄、所在病区及床位号等隐私信息，每份病历平均长度为 742 字。在此基础上，依据 Ogren 等人[12]提出的数据开发指南，由领域专家为每份病历手工标注一个饮酒状态标签。标签类别包括“未确认”、“从未饮酒”、“饮酒”和“曾饮酒”四种。上述四类饮酒状态标签构成了一个细粒度的患者饮酒状态分类集，其定义如下：(1)未确认。入院病历中未包含任何明确描述患者饮酒状况的信息。(2)从未饮酒。入院病历中明确指出患者从未饮酒。(3)饮酒。入院病历中包含患者当前的饮酒信息，且未明确提及停止饮酒或戒酒。(4)曾饮酒。入院病历中明确提及患者存在饮酒历史或习惯，但已在入院前停止饮酒。

在数据集的标签分布方面，“未确认”标签占比最高，共 250 份；“从未饮酒”标签有 48 份；“饮酒”标签有 20 份；“曾饮酒”标签有 32 份。为支持后续的模型训练与评估，将数据集按 8∶2 的比例随机划分，其中 80%(280 份)作为训练集，20%(70 份)作为测试集。

2.2　系统架构

首先，假设一份病历的分类状态基于病历中患者饮酒状态的描述，可以将复杂的文档分类任务简化为句子分类任务。为此，需构建一个高性能分类器。结合基于规则的方法与基于大语言模型的学习方法，该分类器能够识别与患者饮酒状态相关的句子，并为这些句子分配相应的饮酒状态标签。其次，通过一系列权重计算规则，这些句子级的饮酒状态信息将被归一化，从而最终形成对应病历的分类标签。例如，一份入院病历中所有与饮酒状态相关的句子均被归类为“未饮酒”，则该病历也将被标记为“未饮酒”。为实现上述目标，本系统共分为五个模块，旨在将一个复杂的四分类问题分解为多个二元分类任务(图 1)。

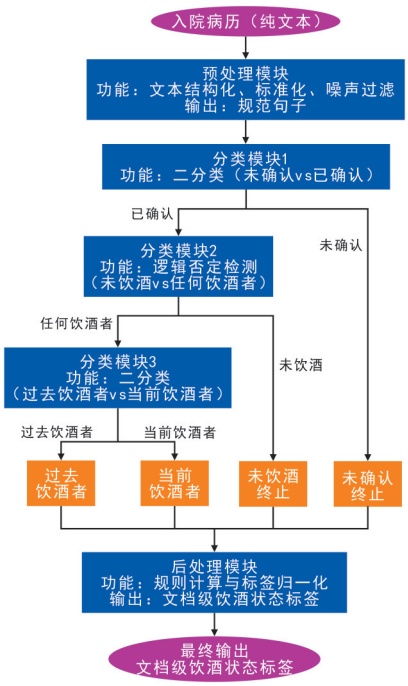

图 1　临床医学 NLP 系统处理患者饮酒状态信息流程

(1)预处理模块。该模块的核心任务是将自由录入的纯文本信息转换为结构化、标准化的计算机可识别格式。利用病历文档的结构特征与临床文档的元数据(如入院病程章节标题)，对信息进行粗粒度筛选，并过滤掉可能妨碍识别患者饮酒状态的错误提示与噪声。经过处理的相关句子将被传递给分类模块 1。(2)分类模块 1。此模块的目标是将预处理后的句子分类为“未确认”或“已确认”。其中，“已确认”作为临时通用标签，用以提取与患者饮酒状态相关的详细信息，为后续处理阶段提供支持。所有标记为“已确认”的句子将传递至分类模块 2。(3)分类

模块 2。该模块对标记为“已确认”的句子进行显式否定检测,以识别那些表明患者为“未饮酒”状态的句子。剩余的句子将被标记为“任何饮酒者”并传递至分类模块 3,这一标签同样为临时通用标签。(4)分类模块 3。这是最终的分类器,用于为句子分配“过去饮酒者”或“当前饮酒者”标签。该模块基于时空特征,利用支持向量机(support vector machine,SVM)进行分类。(5)后处理模块。此模块采用一系列权重计算规则,将所有句子的标签综合考虑,从而为整个病历分配一个最终的饮酒状态标签。

2.2.1 预处理模块

该模块的任务是基于病历文档元数据进行粗粒度信息过滤,对文本进行断句和分词处理,并针对病历文档中每个词组的语言学特性进行特征提取和处理。这些特性包括标点符号、虚词和连字符等。具体做法为利用病历文档的元数据搜索方法,剔除与目标任务无关的文本段落,以避免此类错误提示可能干扰患者饮酒状态的正确分类。此外,由于临床病历书写的习惯和特点,文本中常出现数字与中外文混杂的现象,例如病历中的数量词通常表现为单个数字或数字后紧跟单位标识(如“cm”、“ml”等)。为有效处理此类数据,模块设计了一种映射关系:在分词处理前,若检测到数量词,其中的小数点将被替换为字符“e”,运算符号则分别用“a”、“b”、“c”、“d”替代;分词完成后,再将这些临时字符转换回原始符号。为了提升分词的准确性,采用优化后的 ICTCLAS 分词系统[13]对文本进行断句和分词处理。经过调整后,该系统的分词准确率可达 94.2%,基本满足后续文本分析需求。

2.2.2 分类模块 1

假设对于分类任务而言,入院病历中不同位置出现的词组具有不同的重要性,位于热点词周边的词组对分类任务最具影响力。基于此,研究构建了一个热点词集,使用预处理模块对数据集完成分词后,对分词结果中的词组进行归一化处理,将书写相同的词组合并,并统计其原型词组的出现频率。依据 Joachims 提出的特征选择标准[14],筛选出直接频度大于 3 的词组,构成初始特征集,共计 4 822 个特征。随后,采用信息增益技术[15]对这些特征进行排序。信息增益用于衡量特征与文档类别之间的相关性,其计算过程首先评估类别本身的熵,即类别分布的不确定性;其次计算给定特征条件下的条件熵,表示特征出现或不出现时类别的不确定性;最后通过两者的差值,量化特征对减少分类不确定性所贡献的信息量。使用此方法从初始特征集中提取排名前 500 的特征,组成新的特征子集,并邀请专家对其进行评估。确定热点词需同时满足以下两个条件:该词必须属于特征子集中的特征词;该词必须与分类任务(即患者饮酒状态的区分)具有直接相关性。基于此标准,将与饮酒状态分类任务无直接关联的词组从特征子集中剔除。筛选后剩余特征词构成了最终的热点词集:酒、酒精、罐、瓶、乙醛、乙醇、饮酒、喝酒、酗酒。

热点词集构建完成后,使用 TF-IDF(term frequency - inverse document frequency)算法对每个训练样本进行特征权重计算。该算法计算每个词组在样本中的出现频率以反映其在局部文档中的重要性,计算逆文档频率以衡量词组在整个数据集中的稀有性,最后将两者相乘,得出每个词组的 TF-IDF 权重值。通过此方法,热点词列表被转换为二元向量,向量中的每个元素表示训练样本中对应热点词的状态及其特征权重值。

在特征选择和权重计算后,部分文档表现为“零特征向量”,即不包含任何热点词,这些样本病历将被直接标记为“未确认”。针对剩余样本,本研究作出进一步假设:并非所有句子内容均与饮酒状态相关,而标记为“未确认”的句子仅缺少“已确认”句子所具备的特征。基于此假设,研究设计了一个文本过滤窗口,仅保留热点词所在位置前后各 20 个字范围内的完整句子。这些句子被提取并作为下一模块的输入,而文档的其余部分则被忽略。此方法明显减少了样本规模,并降低了标记为“已确认”句子的向量描述稀疏性。

2.2.3 分类模块 2

该模块的目标是将标记为“已确认”的句子进一步划分为“任何饮酒者”和“未饮酒”两个类别。在汉语语言学中,否定意义的表达除了通过显式否定标志词(如“不”、“没”、“未”、“否”)实现外,还可以通过反问句或疑问代词等隐式否定方式表达。然而,研究人员分析了 241 条含有否定意义的病历句子后发现,由于病历记录的特性,当前样本中均为直述形式的显式否定句,未见隐式否定句。

为有效处理否定表达,本模块通过从病历文档提取否定词,并依据语言学对否定词的定义,构建了一份医学场景汉语否定词表,如无、未、没、不、否、没有、无法、否认、不会等。同时,统计这些否定句中否定标志词与对象词之间间隔字数的结果显示,间隔 0 字(即否定词紧邻对象词)的句子有 210 条,占比最高;间隔 1 字的有 3 条,间隔 2 字的有 13 条,间隔 3 字的有 8 条,间隔 4 字的有 5 条,间隔 5 字的有 2 条,间隔 6 字的为 0 条。基于此分布,设定否定词与对象词的最大间隔为 5 个字,作为后续处理的阈值。

在医学领域中,医务人员通常会记录“未出现”的症状为显式阴性症状,从本质上来看,“未饮酒”可视

为一种显式阴性症状。为实现显式阴性检测，研究采用扩展后的 NegEx 算法[16]，并结合上述否定词表对患者的饮酒状态进行分析。NegEx 是一种依赖正则表达式规则和否定词表的简单算法，通过输入包含锚点词的句子，判断其中是否存在显式否定意义，并输出检测结果。在本研究中定义了两类正则表达式模式：一类描述否定词位于锚点词之前的情况，另一类描述否定词位于锚点词之后的情况。两种模式均限定否定词与锚点词之间的间隔不超过 5 个词，以确保检测的准确性。对于每个输入句子，NegEx 首先检查是否包含热点词，若存在，则进一步进行显式阴性检测；若句中热点词被判定具有显式否定意义，则该句子被标记为“未饮酒”；剩余未被标的句子则被传递至“分类模块 3”进行后续处理。

2.2.4 分类模块 3

该模块基于“时空特征集”，采用 SVM[17] 对剩余训练集进行分类，以区分“当前饮酒者”和“过去饮酒者”的句子标记。SVM 是一种基于结构风险最小化原则和 VC 维理论构建的判别式机器学习方法，被广泛应用于分类任务。其评估算法对未知数据的分类风险时，将总风险分解为两部分：经验风险(即训练数据上的误差)与置信风险(与样本数量和模型复杂度相关的不确定性估计)。通过优化 VC 维(反映模型复杂度的参数)，SVM 可以确保经验风险和置信风险达到均衡，进而最大化各类别数据之间的分隔间隔，最终降低整体分类风险。例如，在一个二维空间中，SVM 的目标是找到一个超平面，将不同类别(如圆形和方形)的数据点分隔开，并使该超平面与各类别边界之间的间隔尽可能大。最靠近超平面的数据点被称为支持向量，它们对分类边界的确定具有关键作用。在更高维度的空间中，分类器表现为一个 n−1 维的超平面，必要时还需对原始数据进行数学变换以实现线性可分。

为构建适用于本任务的 SVM 分类器，首先要对训练子集进行分词处理，并筛选出直接词频大于 3 的词组，构成初始特征集。其次，采用与前文一致的特征排序方法对这些词组进行筛选，并邀请专家对特征子集进行评估，仅保留与汉语时空量化观念密切相关的特征，包括表示时间顺序的词组(如“前”、“过去”、“一直”、“持续”)与表示时间量度的词组(如“年”、“月”、“日”、“天”)，其余无关词组则被剔除。将筛选后的特征定义为时空特征子集，该子集被进一步转换为二元向量，其中每个元素表示训练样本中对应词组的存在状态及其特征值。最后，利用 WEKA 工具包[18] 中的 SVM 实现判别。模块基于训练子集和时空特征子集，训练出一个线性 SVM 句子分类模型，用于区分“当前饮酒者”和“过去饮酒者”。

2.2.5 后处理模块

经过之前四个模块的处理，训练集中除了与分类任务无关的章节外，所有句子均被分配了 1 个饮酒状态标记。在本模块中，研究使用基于权重的优先规则聚合每份文档中的句子标记，计算出最终的饮酒状态。

3　结果

3.1　评价指标

采用 NLP 领域和医学统计学领域通用的评价指标[19] 对系统性能进行评估，包括精确率、召回率、特异度、F 值、宏平均 F 值和微平均 F 值等，并与基线方法的性能－频率分布法进行了对比。验证数据来源于测试集中的 70 份病例数据。

3.2　实验结果

临床医学 NLP 系统在测试集上的混淆矩阵结果如表 1 所示，其中，缩写 U、N、P 和 C 分别代表四类患者饮酒状态标签：“未确认(Unconfirmed)”、“未饮酒(Not drinker)”、“过去饮酒者(Past drinker)”和“当前饮酒者(Current drinker)”。图 2 则给出了系统在测试集上的性能评估。

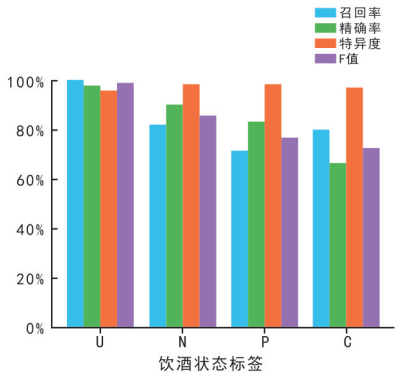

图 2　不同饮酒标签系统效能对比

从表 1 可以看出，系统对“未确认”标签的识别准确度极高，几乎没有误判，说明规则与模型结合的多层分类框架在降低无关信息干扰方面具有显著优势。在“未饮酒”与“过去饮酒者”标签的识别中，仍存在少量交叉错误，但总体保持了较高的分类稳定性。特别是在“当前饮酒者”类别中，系统能够在有限样

表 1　临床医学 NLP 系统在测试集上的混淆矩阵结果　　单位：例

饮酒状态标记	饮酒状态分类				合计
	U	N	P	C	
U	47	0	0	0	47
N	1	9	0	1	11
P	0	1	5	1	7
C	0	0	1	4	5
合计	48	10	6	6	70

本下取得较高的精确率，显示了其在小样本条件下的鲁棒性。根据图 2 数据，本系统在测试集上的微平均 F 值为 0.927，显著高于基线系统的 0.671。结合表 1 和图 2 的对比结果，系统在总体性能上明显优于基线方法，不仅提升了整体 F 值，还在召回率与特异度上取得了平衡，证明了该方法在实际应用中的有效性。整体来看，本研究所构建的系统能够较为全面地覆盖不同饮酒状态的识别需求，并显著提高中文病历饮酒信息的提取效率和准确性，展示了良好的临床应用前景。

4　讨论

本研究提出了一种基于 NLP 技术的患者饮酒状态信息提取和分类系统，将复杂的病历级分类任务分解为多个句子级子任务，通过综合分析句子标记实现整份病历的饮酒状态识别。临床医学 NLP 系统在中文临床文本处理上实现了创新性自动化方案，并结合大语言模型的学习能力，克服了现有 NLP 系统处理中文病历中各类独特语言特征的困难，更贴合中文病历语言特点，并提升了处理性能。

分析系统在测试集上的分类结果，发现有以下改进空间：首先，系统在处理病历中的指代消解问题时存在局限，例如“患者无不良嗜好”容易被误判为“未确认”；其次，对于通过社会行为间接提示的饮酒状态或复合否定结构（如“病人无心脏病、糖尿病、吸烟、酗酒史”）的句子存在识别困难；最后，时空分类组件在区分“过去饮酒者”和“当前饮酒者”时仍依赖特征集和已标注数据，未充分利用病历中的显式时间信息。这些问题为未来优化提供了方向，例如引入高阶 SVM 或深度学习模型，增强时空特征建模能力等。

本研究已经在肝硬化相关研究中实现了初步应用，系统能够快速、准确地识别患者饮酒状态，为该领域流行病学分析和健康风险评估提供了数据支持。未来应进一步拓展系统应用范围，包括放射影像报告、病理报告及多中心病历，以提升数据覆盖面和成果代表性。同时，将探索多类型临床数据的联合分析方法，优化分类算法，为不同疾病和研究场景提供可复用的自动化信息提取工具，增强系统在多中心、多类型临床数据中的通用性和推广价值。

参考文献

[1] 刘丹红，罗小楠，徐勇勇. 电子病历及其应用概述[J]. 中国卫生质量管理，2010，17(4)：2－5，1.

[2] BAGNARDI V，ROTA M，BOTTERI E，et al. Light alcohol drinking and cancer：a meta－analysis[J]. Ann Oncol，2013，24(2)：301－308.

[3] ARONSON AR，LANG FM. An overview of MetaMap：historical perspective and recent advances[J]. J Am Med Inform Assoc，2010，17(3)：229－236.

[4] LINDBERG DAB，HUMPHREYS BL，MCCRAY AT. The united medical language system[J]. Method Inform Med，1993，32(4)：281－291.

[5] JAIN NL，KNIRSCH CA，FRIEDMAN C，et al. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports.[J]. Proc Amia Annu Fall Symp，1996(1)：542－546.

[6] HUA X，STENNER SP，SOU D，et al. MedEx：a medication information extraction system for clinical narratives[J]. J Am Med Inform Assoc，2010，17(1)：19－24.

[7] NADKARNI P，CHEN R，BRANDT C，et al. UMLS concept indexing for production databases：a feasibility study[J]. J Am Med Inform Assoc，2001，8(1)：512.

[8] 娄苗苗，张　浩，刘丹红. 医疗质量测量指标基础数据的标准化方法[J]. 中国卫生质量管理，2013，20(2)：53－56.

[9] 舒　婷，赵　韡，刘海一. 2020 年我国医院电子病历系统应用水平分析[J]. 中国卫生质量管理，2022，29(1)：8－10，20.

[10] 盛文佳，金可可，费宏玮，等. 基于临床路径的电子病历改进思路[J]. 中国卫生质量管理，2011，18(5)：18－20.

[11] 孟晓阳. 电子病历互操作性的实现技术[J]. 中国卫生质量管理，2010，17(4)：19－21.

[12] OGREN P，SAVOVA GK，CHUTE CG. Constructing evaluation corpora for automated clinical named entity recognition[C]. In 6th International Conference on Language Resources & Evaluation. DBLP，2010.

[13] 张华平，商建云. NLPIR－Parser：大数据语义智能分析平台[J]. 语料库语言学，2019，6(1)：87－104.

[14] JOACHIMS T. Text categorization with support vector machines：learning with many relevant features[C]. In 10th European Conference on Machine Learning，1998：137－142.

[15] 要　芳. 基于本体的电子病历知识库研究[D]. 西安：西安电子科技大学，2009.

[16] CHAPMAN WW，BRIDEWELL W，HANBURY P，et al. A simple algorithm for identifying negated findings and diseases in discharge summaries[J]. J Biomed Inform，2001，34(5)：301－310.

[17] 吉旭瑞，魏德健，张俊忠，等. 中文电子病历信息提取方法研究综述[J]. 计算机工程与科学，2024，46(2)：325－337.

[18] HALL M，FRANK E，HOLMES G，et al. The WEKA data mining software：an update[J]. ACM SIGKDD Explor Newsl，2008，11(1)：10－18.

[19] YANG Y. An evaluation of statistical approaches to text categorization[J]. Proc Amia Annu Fall Symp，1999，1(1－2)：358－362.

通信作者：
芮欣凯：上海交通大学医学院附属瑞金医院信息中心副主任，高级工程师
E－mail：rxk@rjh.com.cn

收稿日期：2024－10－14
修回日期：2025－08－21
本文编辑：姚　涛、刘斯妤