

AI大语言模型辅助医学影像信息系统数据清洗与整合技术研究

豆 圆

杭州融御科技有限公司 浙江 杭州 310000

摘 要: 在医疗数字化转型浪潮下, 医学影像信息系统积累海量数据, 但数据质量欠佳与整合困难问题突出。本文聚焦 AI 大语言模型, 深入探究其辅助医学影像信息系统进行数据清洗与整合的关键技术。通过剖析医学影像数据特性与问题, 基于模型原理及优势, 实现数据去噪、错误纠正、语义对齐等功能。实际应用表明, 该技术提升了数据质量与医疗效率, 为临床诊断、医学科研提供有力支撑, 对推动医疗信息化高质量发展意义重大。

关键词: AI大语言模型; 医学影像信息系统; 数据清洗; 数据整合

引言

近年来, 医疗信息化飞速发展, 医学影像信息系统成为医疗数据重要来源, X光、CT、MRI等影像数据对疾病诊疗意义重大。但当前医学影像数据存在诸多问题, 采集时因设备、患者等因素导致数据模糊、缺失, 元数据易遗漏, 人工标注存在误差, 数据格式多样导致兼容性差, 且数据分散存储缺乏整合机制。AI大语言模型凭借强大自然语言处理能力, 有望通过精准理解医学文本, 实现数据高效清洗与整合, 推动医疗信息化发展。

1 医学影像信息系统数据特性与问题剖析

1.1 数据多样性与复杂性

医学影像数据堪称种类繁多, 其来源基于多种不同的成像原理。X光影像利用X射线穿透人体, 根据不同组织对X射线的吸收差异来成像, 主要用于骨骼、胸部等部位的初步检查, 图像相对简单, 以黑白影像为主, 能清晰显示骨骼的大致形态和轮廓。而CT影像则通过对人体进行断层扫描, 将人体内部结构以断层图像的形式呈现出来, 数据量极为庞大, 每个断层图像都包含了丰富的解剖细节, 如人体的各个脏器、血管等结构都能清晰展现, 分辨率通常在毫米级别, 能够提供高精度的解剖信息。MRI影像则是基于核磁共振原理, 对人体软组织的分辨能力极强, 在神经系统、关节等部位的检查中具有独特优势, 图像能够呈现出不同组织的细微差异, 为医生提供更详细的病变信息。

除了影像本身, 这些数据还伴随着大量的元数据。患者的基本信息, 如姓名、年龄、性别等, 是对患者个体的基本标识; 检查时间记录了影像采集的具体时刻, 对于疾病的动态观察和诊断具有重要参考价值; 诊断报告则是医生对影像的解读和分析, 包含了对病变的

判断、初步诊断结果等关键信息。然而, 这些元数据的格式和存储方式在不同的医疗机构和设备中各不相同, 有的以结构化表格形式存储, 有的则以自由文本形式记录, 这无疑极大地增加了数据处理的难度。

1.2 数据质量问题

数据缺失: 在影像采集的复杂过程中, 设备故障是导致数据缺失的常见原因之一。例如, CT设备的探测器出现故障, 可能会使部分扫描层面的图像无法正常获取, 呈现出模糊或空白的状态。患者的配合度也至关重要, 当患者在检查过程中无法保持静止, 如儿童在进行MRI检查时难以长时间安静配合, 就容易导致图像出现运动伪影甚至部分数据丢失^[1]。在元数据方面, 患者自身对病史的记忆模糊、医护人员记录疏忽等因素, 都可能造成患者关键信息的缺失, 如重要的既往病史、过敏药物史等, 这些信息的缺失可能会影响医生对病情的全面判断, 甚至导致误诊或漏诊。

错误标注: 人工标注医学影像存在难以避免的主观性和误差。不同医生对医学影像的理解和判断标准存在差异, 经验丰富的医生和新手医生在标注时可能会得出不同的结论。此外, 医学影像中的病变形态和特征往往具有多样性和复杂性, 一些早期病变或不典型病变容易被误判。例如, 在肺部X光影像中, 一些小结节可能被误诊为炎症, 而实际上可能是早期肺癌的表现。这种错误标注不仅会误导医生的诊断思路, 还会影响后续的治疗方案选择, 降低数据在临床应用和医学研究中的可靠性。

2 AI大语言模型原理及在数据处理中的优势

2.1 模型基本原理

AI大语言模型的核心架构是Transformer, 它摒弃了

传统循环神经网络（RNN）的顺序处理方式，采用了多头注意力机制，能够同时关注输入文本的不同部分，从而更好地捕捉文本中的语义依赖关系。以GPT系列模型为例，在训练过程中，它会对海量的互联网文本进行无监督学习。这些文本涵盖了新闻、学术论文、小说、博客等各种领域和体裁，模型通过对这些文本的学习，不断调整自身的参数，构建起一个庞大而复杂的语言知识图谱。在这个图谱中，包含了丰富的语言知识，如词汇的语义、语法规则、上下文关系以及不同领域的专业知识等。当输入医学相关文本时，模型会基于之前学习到的知识，对文本进行深入分析和理解。它能够识别出文本中的医学术语、疾病名称、症状描述等关键信息，并根据这些信息生成准确的回复或完成特定的任务，如医学问题解答、影像报告分析等。

2.2 在医学影像数据处理中的优势

自然语言理解能力：医学影像报告通常以自然语言的形式记录，其中包含了医生对影像的详细描述和诊断意见。然而，由于医生的表达方式各异，报告中可能存在模糊、隐晦的表述，这给数据提取带来了很大困难。AI大语言模型凭借其强大的自然语言理解能力，能够准确解析这些报告。它可以识别出病变部位的精确描述，如“左肺下叶背段可见一大小约2cm×3cm的结节影”，并提取出关键信息“左肺下叶背段”和“2cm×3cm的结节影”。同时，对于病变性质的描述，如“考虑为恶性肿瘤可能性大”，模型也能准确理解其含义，从而为后续的诊断和分析提供准确的数据支持，有效解决了因语言表达差异导致的数据提取难题。

知识推理与补全：AI大语言模型拥有丰富的医学知识储备，这使得它在面对缺失或不完整的医学影像数据时，能够进行合理的推理和补全。例如，当患者的部分影像数据缺失，但已知其其他检查结果和详细病史时，模型可以根据这些已知信息，结合自身的医学知识，推测出缺失影像可能呈现的特征^[2]。如果已知患者有长期吸烟史且近期咳嗽加重，在胸部CT影像部分缺失的情况下，模型可以推测缺失部分可能存在与肺部疾病相关的病变，如肺气肿、肺部炎症等，从而为医生提供有价值的参考，提高数据的完整性和可用性。

3 AI大语言模型辅助数据清洗与整合技术实现

3.1 数据清洗技术

数据去噪：医学影像报告中常常夹杂着各种噪声数据，这些噪声会干扰对有效信息的提取和分析。AI大语言模型通过对大量正常医学影像报告的学习，能够掌握正常报告的语言模式和特征。当面对一份新的报告时，

模型可以根据学习到的模式判断其中是否存在异常数据。如果报告中出现重复的语句、无意义的乱码或者与医学领域无关的词汇，模型能够准确识别并将其剔除。在一份关于心脏超声检查的报告中，如果出现了一段与检查无关的广告信息，模型可以迅速检测到并将其去除，从而提高报告数据的纯净度。

错误纠正：针对医学影像标注错误的问题，AI大语言模型可以利用其丰富的医学知识和上下文理解能力进行纠正。当模型发现标注信息与影像中的实际情况不符时，它会根据医学知识和影像的其他相关信息进行分析和判断。如果标注中将肝脏的病变位置错误标注为胆囊，模型可以结合影像中肝脏和胆囊的形态、位置以及周围组织的关系，判断出标注错误，并将其修正为正确的肝脏病变位置。同时，对于病变类型的错误标注，如将良性肿瘤误标为恶性肿瘤，模型也能通过对病变特征的分析 and 医学知识的运用，进行准确的纠正，提高标注的准确性。

3.2 数据整合技术

语义对齐：不同来源的医学影像数据在描述和术语上往往存在差异，这给数据整合带来了很大障碍。AI大语言模型通过强大的语义理解能力，能够将不同表述统一为标准医学术语。在医学领域，对于同一种疾病可能存在多种不同的称呼，如“心肌梗塞”和“心肌梗死”实际上指的是同一种疾病。模型可以识别出这些同义词，并将其统一为标准术语，方便数据的整合和分析。在整合不同医院的心脏病患者影像数据时，模型可以将各种关于心脏病的不同表述转化为统一的医学术语，使数据在语义层面上保持一致，提高数据的通用性和可比性。

数据融合算法：为了将不同格式、不同来源的医学影像数据融合为统一的数据集，AI大语言模型结合机器学习算法构建了数据融合模型。在数据融合过程中，首先对不同格式的数据进行预处理，将其转化为统一的特征表示形式。对于DICOM格式的影像数据和JPEG格式的影像数据，通过特定的算法提取它们的关键特征，如影像的纹理特征、形状特征等^[3]。然后，利用机器学习算法，如神经网络、支持向量机等，对这些特征进行融合处理。通过训练数据融合模型，使其能够学习到不同数据之间的内在联系和规律，从而将不同来源的数据有效地融合在一起，形成一个完整、统一的数据集，大大提高了数据的可用性和价值。

4 技术应用案例与效果评估

4.1 医院应用实例

知名三甲医院在其医学影像信息系统中引入了AI大语言模型技术,旨在解决长期以来困扰医院的数据质量和整合问题。在数据清洗阶段,模型对医院历年来积累的海量医学影像报告进行了全面处理。通过数据去噪功能,成功剔除了大量包含重复内容、乱码以及无关信息的噪声数据,使报告数据更加简洁明了。在错误纠正方面,模型对标注错误的影像数据进行了细致的分析和修正,共纠正了数千条错误标注,大大提高了影像数据的准确性。同时,针对缺失的患者信息,模型利用自身的推理能力和知识储备,结合患者的其他检查结果和已有病史,填补了大量缺失的关键信息,如患者的过往手术史、家族病史等,为医生提供了更完整的患者资料。

在数据整合过程中,医院面临着不同科室、不同设备产生的影像数据格式多样、术语不统一的难题。AI大语言模型通过语义对齐功能,将各种不同的医学术语统一为标准术语,消除了数据在语义层面的差异^[4]。例如,将不同科室对“糖尿病视网膜病变”的不同表述统一为标准术语,方便了数据的整合和共享。同时,利用数据融合算法,将来自CT、MRI、超声等不同设备的影像数据以及相关的文本信息进行了深度融合,构建了一个统一的医学影像数据库。

经过一年的实际应用,该医院的医学影像信息系统得到了显著优化。医生在检索影像数据时,效率提高了50%,以往需要花费大量时间在不同系统和数据库中查找患者的完整影像资料,现在通过统一的数据库和智能检索功能,能够快速准确地获取所需信息。诊断准确性也得到了大幅提升,提升了15%。医生在综合分析患者的影像数据和相关信息时,由于数据的准确性和完整性得到了保障,能够做出更准确的诊断,减少了误诊和漏诊的发生,为患者的治疗提供了更可靠的依据。

4.2 效果评估

数据质量提升:通过AI大语言模型的清洗和整合,医学影像数据的准确性、完整性和一致性得到了显著提高。错误标注得到纠正,缺失数据得到补充,不同格式和来源的数据实现了统一和融合,为临床诊断和科研提供了高质量的数据支持。在临床诊断中,医生可以根据

更准确的数据做出更精准的诊断决策,提高治疗效果;在医学科研中,高质量的数据能够为研究人员提供更可靠的研究基础,有助于发现新的医学规律和疾病治疗方法。

医疗效率提升:医生能够更快速、准确地获取所需的影像数据,大大缩短了诊断时间。在紧急情况下,如急性心梗、脑卒中等疾病的诊断和治疗中,快速获取准确的影像数据至关重要。AI大语言模型辅助的医学影像信息系统能够帮助医生在最短时间内了解患者的病情,制定合理的治疗方案,提高医疗服务效率,挽救患者生命。同时,数据的高效整合也使得医院各科室之间的协作更加顺畅,提高了整体医疗服务水平。

结语

AI大语言模型在医学影像信息系统的数据处理中展现出卓越成效,有效攻克了数据质量欠佳与整合艰难的难题。通过精准的数据清洗与整合,不仅提升了数据的准确性、完整性和一致性,为临床诊断和医学科研筑牢了坚实的数据根基,还显著提高了医疗效率,助力医生做出更精准的诊断决策。然而,该技术在实际应用中仍面临数据安全、模型可解释性等挑战。未来,需进一步完善技术,强化安全保障,增进模型的可解释性,推动AI大语言模型在医疗领域的深度应用,为医疗事业的发展注入强大动力。

参考文献

- [1]都娟娟.景泰川电力提灌泵站机电设备的运行与管理[J].湖北农机化,2020,(16):159-160.
- [2]裴丰华.试论电力提灌站机电设备的运行与管理[J].科技风,2019,(04):169.DOI:10.19392/j.cnki.1671-7341.201904148.
- [3]张涛涛.分析电力提灌站机电设备的运行管理[J].甘肃农业,2017,(21):58-59.DOI:10.15979/j.cnki.cn62-1104/f.2017.21.018.
- [4]闫斌.高扬程电力提灌工程机电设备管理探析[J].科技与创新,2023,(05):136-138.DOI:10.15913/j.cnki.kjycx.2023.05.044.