



(12)发明专利申请

(10)申请公布号 CN 111492437 A

(43)申请公布日 2020.08.04

(21)申请号 201780097565.8

(51)Int.Cl.

(22)申请日 2017.12.29

G16H 10/60(2006.01)

(85)PCT国际申请进入国家阶段日
2020.06.08

G16H 50/00(2006.01)

(86)PCT国际申请的申请数据
PCT/RU2017/000819 2017.12.29

(87)PCT国际申请的公布数据
W02019/132685 RU 2019.07.04

(71)申请人 英泰曲尔德直线有限公司
地址 俄罗斯莫斯科

(72)发明人 I·S·德罗金 O·L·布赫瓦洛夫
S·Y·索罗金

(74)专利代理机构 上海一平知识产权代理有限公司
31266

代理人 徐迅 祝莲君

权利要求书1页 说明书19页 附图8页

(54)发明名称

支持医学决策的方法和系统

(57)摘要

一种在服务器上实现的使用患者数学模型支持医学决策的方法,包括:生成包含按患者分组的患者电子医学记录的训练数据集;预处理从训练数据集中选择的患者的医学记录中包含的数据;使用医学本体将处理后的数据转换为针对每个患者的医学事实序列;使用从患者的医学记录中提取的感兴趣的事实,自动标记产生的每个患者的医学事实序列;为每种模态分别训练初始表示;培训联合表示;训练最终模型和集合参数;获取未包含在训练数据集中的患者的医学记录;预处理获得的患者医学记录中包含的数据;使用医学本体将预处理数据转换为医学事实序列;将产生的事实集发送给生成的模型以供输入;根据所呈现的事实集,确定诊断,并对患者指导分析和预测最可能的疾病发展。

1. 一种使用服务器上执行的患者表示数学模型来支持医学决策的方法,其特征在于,包括以下步骤:

- 形成包括按每个患者分组的患者电子健康记录的训练数据集;
- 对从所述训练数据集中选择的所述电子健康记录中包含的数据进行初步处理;
- 使用医学本体将处理后的数据转换为每位患者的医学事实序列;
- 使用从所述健康记录中提取的诊断或其他感兴趣的事实,对每位患者获得的医学事实序列进行自动布局;
- 对每种模态分别进行主要表示的训练;
- 进行联合表示的训练;
- 进行最终模型和聚合参数的训练;
- 获取未包含在训练数据集中的患者的健康记录;
- 对获取的患者健康记录中包含的数据进行初步处理;
- 使用医学本体将经过初步处理的数据转换为医学事实序列;
- 将获得的医学事实序列提交给最终模型的输入;
- 为患者做出诊断以及对患者疾病病程做出分析和预后,最大可能性与获得的医学事实序列相对应。

2. 根据权利要求1所述的方法,其特征在于,电子健康记录至少包括以下数据:患者的状况、患者的治疗方法、用于治疗患者的手段、测试结果。

3. 一种用于使用患者表示数学模型支持医学决策的系统,其特征在于,包括至少一个处理器、随机存取存储器、包含被下载到随机存取存储器中并由至少一个处理器执行的指令的存储设备,所述指令包括以下步骤:

- 形成包括按每个患者成组的患者电子健康记录的训练数据集;
- 对从所述训练数据集中选择的所述电子健康记录中包含的数据执行初步处理;
- 使用医学本体将处理后的数据转换为每个患者的医学事实序列;
- 使用从所述健康记录中提取的诊断或其他感兴趣的事实,对每个患者的获得的医学事实序列执行自动布局;
- 对每种模态分别执行主要表示的训练;
- 执行联合表示的训练;
- 执行最终模型和聚合参数的训练;
- 获取未包含在训练数据集中的患者的健康记录;
- 对获取的患者健康记录中包含的数据执行初步处理;
- 使用医学本体将经过初步处理的数据转换为医学事实序列;
- 将获得的医学事实序列提交给最终模型的输入;
- 为患者做出诊断以及对患者疾病病程做出分析和预后,最大可能性与获得的医学事实序列相对应。

支持医学决策的方法和系统

技术领域

[0001] 该技术方案涉及人工智能领域,即医学决策支持系统。

背景技术

[0002] 对患者疾病的检测和诊断是一个复杂的问题,其通常由医生解决。影响医生诊断结果的因素很多,包括医生的经验、专心程度、当前病例的复杂性。使用不同的决策支持系统来消除这些缺陷。

[0003] 在大多数方法中,决策支持系统能够检测某些病理(疾病)的存在与否,但是,它无法对特定患者进行疾病过程的分析和预后。

发明内容

[0004] 该技术解决方案能够创建患者数学模型,从而可以提高诊断的准确性并针对特定患者进行疾病过程的分析和预后。

[0005] 在服务器上执行的使用患者表示数学模型支持医学决策的方法包括以下步骤:

[0006] 形成包括按每个患者分组的患者电子健康记录的训练数据集;

[0007] 对从训练数据集中选择的电子健康记录中包含的数据进行初步处理;

[0008] 使用医学本体将处理后的数据转换为每个患者的医学事实序列;

[0009] 使用从健康记录中提取的诊断或其他感兴趣的事实,对每个患者获得的医学事实序列进行自动布局;

[0010] 对每种模态分别进行主要表示的培训;

[0011] 进行联合表示的培训;

[0012] 进行最终模型和聚合参数的训练;

[0013] 获取未包含在训练数据集中的患者的健康记录;

[0014] 对获取的患者健康记录中包含的数据进行初步处理;

[0015] 使用医学本体将经过初步处理后的数据转换为医学事实序列;

[0016] 将获得的医学事实序列提交给最终模型的输入;

[0017] 对患者做出诊断并对患者疾病病程做出分析和预后,最大可能性与获得的医学事实序列相对应。

[0018] 使用患者表示数学模型的医学决策支持系统包括至少一个处理器,随机存取存储器,包括下载到随机存取存储器中并由至少一个处理器执行的指令的存储设备,该指令包括以下步骤:

[0019] 形成包含按患者分组的电子健康记录的训练数据集;

[0020] 初步处理从培训数据集中选择的电子健康记录中包含的数据;

[0021] 使用医学本体将处理后的数据转换为每个患者的医学事实序列;

[0022] 使用从健康记录中提取的诊断或其他感兴趣的事实,对获得的每个患者的医学事实序列进行自动布局;

- [0023] 对每种模态分别培训主要表示；
- [0024] 联合表示的培训；
- [0025] 最终模型和集合参数的训练；
- [0026] 获取未包含在训练数据集中的健康记录；
- [0027] 初步处理获得的健康记录数据；
- [0028] 使用医学本体将经过初步处理后的数据转换为医学事实序列；
- [0029] 将获得的事实集提交给形成的模型输入；
- [0030] 对患者做出诊断并对患者疾病病程做出分析和预后，最大可能性与提交的事实集相对应。
- [0031] 在其广泛的方面之一中，该技术解决方案能够对患者体内的过程和趋势进行建模，定义药物和指定疗法的作用，确定手术或指定疗法后的患者死亡率。

具体实施方式

- [0032] 在此技术解决方案中使用以下术语和定义。
- [0033] 矢量表示-在自然语言处理中表示用于语言建模和训练的不同方法的通用名称，旨在从 R_n 的矢量字典中匹配 n 的单词(可能还有短语)匹配，其数量要比字典中的单词的数量少得多。分布语义是矢量表示的理论基础。有多种创建此类匹配的方法，例如，使用神经网络，应用于单词共现矩阵以及使用上下文出现单词训练的显式表示的降维方法。
- [0034] 患者矢量表示-基于患者生理数据、既往病历，病史及其治疗(方法和治疗过程，处方药等)等的患者数学模型，从而能够针对特定患者预测疾病过程，进行诊断，提出建议、治疗策略等。
- [0035] 组织学-涉及活生物体组织的结构、活性和发育的生物学分支。
- [0036] 人类组织学-处理人体组织结构的医学分支。
- [0037] 分布语义学-语言学的一个分支，基于语言单元在大语言数据集(文本语料库)中的分布，用于处理语言单元之间语义相似度的计算。分布语义是基于分布假设的：在相似上下文中出现的语言单元具有紧密的含义。
- [0038] 元数据-有关其他信息的信息，或关于内容或对象的其他信息的数据。元数据会公开有关特征和属性的信息，这些特征表征某些基本要素，从而使它们能够在大型信息流中自动搜索和控制它们。
- [0039] 数据模态-对定义所述数据结构，其格式的某些数据源的数据所有权，并且还使得能够将所述结构与任何器官和/或疾病学和/或程序系统相关联。作为患者自身的患者数据收集装置可以是数据源。
- [0040] 本体论-通过概念方案全面而详细地将某些知识领域形式化。通常，这种方案由分层数据结构组成，该分层数据结构包括在该领域中识别的所有相关类的对象、它们的链接和规则(定理、约束)。
- [0041] 正则化(用于统计、机器学习、逆问题理论)-一种为语句添加一些附加信息的方法，目的是解决不适当地的问题或防止重新训练。该信息通常以对模型复杂度的惩罚形式给出，例如，它可能是结果函数平滑度的约束或矢量空间范数的约束。
- [0042] 词干-查找给定源词的词干的过程。词干无需与形态学的词根相同。

[0043] 事实(医学事实)-描述患者的数据,包括其治疗方法以及所述数据与其他医学事实的链接。

[0044] 电子健康记录(电子病历-EMR)-包含患者数据的数据库:患者生理数据、病历,疾病史及其治疗方法(治疗方法和过程、处方药等)等。电子健康记录还包括至少包括以下数据的患者记录:记录添加日期、诊断代码、症状、程序以及药物,以自然语言显示的病历文本描述、与病历相关的生物医学图像、病患研究和分析结果。

[0045] 医学信息系统的医务人员或其他用户通过该系统界面将包括健康记录、患者生理数据以及其他信息的患者数据下载到服务器。

[0046] 在另一个实施例中,数据可以自动下载到服务器,而无需人工参与,例如在分析收集和分析、处理程序等。

[0047] 在病人去看医生,进行检查、提供医学检查或其他医学程序期间,有关每个此类程序的数据都在医学信息系统中形成(填写和存储)。数据可以包括但不限于患者检查记录、诊断代码、症状、患者规定和/或服用的程序以及药物、以自然语言描述的病史、生物医学图像、分析和研究的结果、观察结果和生理数据测量结果、ECG、EEG、MRT、超声检查、活检、细胞学筛查、X射线、乳房X线照相。

[0048] 所述数据可以以时间序列、图像、视频、基因组数据、信号的形式表示,但不限于文本、表格格式。数据也可以结构化和非结构化形式呈现。此外,上述数据之间的链接可以用作数据。

[0049] 分析可以包括但不限于血液分析、脑脊液分析、尿液分析、粪便分析、基因检测分析等。在技术解决方案的范围内,分析类型不受限制。

[0050] 让我们考虑图1中所示的示例:

[0051] 一名患者105拜访了一位专科医生进行初步评估。医生进行必要的医学活动,然后形成患者症状的描述,开具检查处方。然后,医生将这些信息通过医学信息系统110的界面输入计算机中,随后将这些数据存储在电子健康记录中。另一位患者106再次拜访了治疗学家。该治疗师为患者开药,并将这些数据输入至电子健康记录。通过此过程,将在每个患者的电子健康记录中形成所需的记录集,并且其他医生或决策支持系统可以使用此记录集。

[0052] 为了创建医学决策支持系统,需要收集一定的数据量,从而能够使训练系统根据可用的医学数据识别诊断、诊断组或事实。当在医学信息系统中收集了所需的数据量时,可以将其用作训练数据集。大多数现有的决策支持系统都以不同的临床表现使用机器学习。

[0053] 医学决策支持系统中的重要组成部分之一是患者矢量表示(患者数学模型),它可以为特定患者提供疾病病程的预后、诊断、提出建议、治疗策略等。

[0054] 在一个实施例中,能够形成患者矢量表示的功能位于单独的服务器中。

[0055] 在一个实施例中,使得能够形成患者矢量表示的功能位于医学决策支持系统所位于的同一服务器中。

[0056] 在一个实施例中,使得能够形成患者矢量表示的功能是使用分布式服务器系统中的云计算的云服务。

[0057] 在第一步中,为了形成患者矢量表示,形成了训练数据集210,图2,其随后将用于算法机器学习,包括深度学习算法。

[0058] 在一个实施例中,训练数据集由医学信息系统的用户通过选择患者记录来形成。

[0059] 在一个实施例中,可以根据指定的标准来执行记录的选择。至少可以将以下内容用作标准:

[0060] • 纳入/排除训练数据集的规则:

[0061] ◦ 具有既往病历的指定组中的既往病历的患者(例如,仅具有肿瘤既往病历的患者);

[0062] ◦ 符合指定性别或年龄参数(例如,年龄在30至45岁之间的男性)的患者;

[0063] ◦ 与已经包含在训练数据集中的患者相关的患者,并且该关系至少由既往病历的相似性、治疗方法等来定义。

[0064] • 包括先前形成的训练数据集的规则。

[0065] 在此技术解决方案范围内,培训数据集包括按患者分组的电子健康记录。用于训练数据集形成的电子健康记录包括至少包含以下数据的患者记录:记录添加日期、诊断代码、症状、程序以及药物、自然语言的病历文本描述、与病历相关的生物医学图像、患者研究以及分析的结果。

[0066] 以下是从健康记录中摘录的片段作为示例:

[0067]

01.01.2001

病人抱怨发烧和湿咳。该患者接受了胸部 X 光检查,并服用了<剂量>剂量的退热药<退热药>。

02.01.2001

确诊为肺炎。指定了以下治疗方法<自然语言治疗描述>

[0068] 所使用的数据表示格式可能会根据所应用的技术而变化或改变。所描述的格式不是唯一的可能,并且为了最好地理解此技术解决方案的概念而进行了描述。

[0069] 电子健康记录可以以下格式显示:开放式电子健康记录(openEHR)、HL7等。格式和标准的选择不会影响技术解决方案的本质。

[0070] 在一个实施例中,健康记录是一组字段,其包括至少描述以下内容的参数:

[0071] • 患者状态;

[0072] • 患者的治疗方法(技术、应用方法、特征);

[0073] • 用于患者治疗的手段(药物、剂量等);

[0074] • 分析结果等;

[0075] 元数据将描述的参数与其他记录中的参数链接起来。

[0076] 如果训练数据集中的记录未按每个病人进行分组,则在获取数据后使用已知的算法或功能(例如,现有技术中已知的任何排序和采样,包括在从数据库中数据采样的SQL查询中使用的GROUP BY,ORDER BY命令)对它们进行分组)。

[0077] 健康记录数据可以以时间序列、图像、影像、基因组数据、信号的形式表示,但不限于文本、表格格式。该数据也可以结构化和非结构化形式表示。

[0078] 记录添加日期可以仅存储日期、日期和时间、时间戳,并且所述记录可以包括绝对形式的所述时间对象、相对形式(相对于其他记录的时间对象)的所述时间对象。

[0079] 诊断、症状、程序和药物的代码可以以MKB格式(例如MKB-10)、SNOMED-CT、CCS(临床分类软件)等表示。格式的选择不影响技术方案的实质。

[0080] 分析结果可以表格形式呈现。

[0081] 结构化和非结构化形式的案例历史记录的文本文描述(自然语言描述)。

[0082] 生物医学图像可以以图像(jpg、png、tiff和其他图形格式)、视频(avi、mpg、mov、mkv和其他视频格式)、3D照片、3D视频、3D模型(obj、max、dwg等)呈现。心电图(ECG)、脑电图(EEG)、MRT、超声检查、活检、细胞学筛查、X射线、乳腺X线照相等的结果可以以生物医学图像的形式呈现。

[0083] RNA测序数据可以以TDF格式(平铺数据格式)、未索引格式(例如GFF、BED和WIG)、索引格式(例如BAM和Goby)以及bigWig和bigBed格式呈现。

[0084] 以上格式表示为使用以上数据(创建、修改等)运行而设计的最低软件。

[0085] 在服务器上形成并接收培训数据集时,服务器会对数据进行初步处理220,该数据包含在患者的健康记录中,从培训数据集中被选择。

[0086] 数据初步处理是特定于域的,并且取决于数据类型和数据源。

[0087] 为每种数据类型和/或数据源分配了特殊的数据处理程序。如果没有为数据类型和/或数据源分配处理程序,或数据类型和/或数据源不需要处理程序,则使用空处理程序,或者忽略该数据类型的处理程序。

[0088] 在实施例之一中,基于至少为电子健康记录中的一种数据字段指定的元数据来定义数据类型。

[0089] 例如,在医学数字影像和通讯(dicom)中,元数据以外显式指示数据模态,并且根据dicom内部定义解释数据模态。

[0090] 在实施例之一中,可以借助于鲜明特征来定义数据类型。前提是服务器或外部源具有鲜明特征的数据库,通过该数据库可以在记录中定义数据类型。

[0091] 例如,数据(字段或文件)开头的GIF89a字节序列的可用性意味着它是GIF格式的位图图像,而BM字节的可用性意味着它是BMP格式的位图图像。

[0092] 在实施例之一中,可以使用预先指定的规则基于记录信息来定义数据类型。

[0093] 例如,基于可执行文件(.rdata)结构分析来定义存储在可执行文件资源(PE-file,.exe)中的图像数据(位图、图标)、多媒体数据(视频、声音)的类型。

[0094] 在一个实施例中,一种类型的数据可以被转换成另一种类型的数据(视频-被转换为一组图像,反之亦然,3D对象-被转换为所述对象映射图像,反之亦然,等等)。

[0095] 对于一个示例,如果用于拍摄该图像的设备参数是已知的,则可以为CT图像分配一个处理程序,以将其转换为一系列可能进行归一化的位图图像。

[0096] 对于另一个示例,对于文本,可以分配一个处理程序,该处理程序执行标准文本NLP转换(通常为小写字母、数字取代、删除停用词和介词、词干)。

[0097] 对于另一个示例,对于自然语言的文本,可以分配一个处理程序,该处理程序通过将文本映射到医学本体的术语和/或医学术语的词典,从文本形成医学事实的序列。

[0098] 在一个用于以自然语言分析文本的实施例中,可以使用从现有技术中已知的至少词法分析和句法分析的算法,基于这些算法,从文本中提取词法单元并组合到表示一连串的医学事实的对象中。

[0099] 在一个实施例中,当映射文本时,用健康记录中的日期和/或与日期相对应的时间和/或当前记录的时间,对每个医学事实进行注释(标记)。

[0100] 例如,如果处理程序处理的字段包含来自日期为2017年1月20日的患者记录中的自然语言文本,则所有医学事实都将以日期2017年1月20日进行注释(标记)。

[0101] 在一个用于以自然语言分析文本的实施例中,可以使用从现有技术中已知的至少词法分析和句法分析的算法,基于这些算法,从文本中提取词法单元并组合到表示一连串的医学事实的对象中。

[0102] 在用于分析文本的另一实施例中,使用了预先训练的(通过一种机器学习方法)文本识别模型,并且由于该模型操作而形成了一组医学事实。此外,如果所形成的医学事实不符合先前指定的标准(例如,当结果由专家分析时),则可以(使用监督学习方法)对所述模型进行再训练。

[0103] 在自然语言处理程序的一个实施例中,处理程序(在预处理之后)从本体论或词典中的文本中搜索每个词。如果在本体论或词典中找到单词,则处理程序将本体概念或与之相对应的词典中的单词保存下来,并拒绝在本体论或词典中找不到的单词。

[0104] 在实施例之一中,可以使用更复杂的文本映射规则(过程)将自然语言的文本映射到事实序列。

[0105] 例如,对于某些概念,可以使用能够提取相关概念和/或值的附加模板(正则表达式)。

[0106] 在实施例之一中,本体论和/或字典位于服务器中。

[0107] 在另一个实施例中,服务器可以从外部源接收本体论和字典。

[0108] 例如,通过本体论查找服务,其提供网络服务接口,以统一的数据输出格式从一个地方查询许多本体论。

[0109] 在一个实施例中,医学数据的任何源可以代替本体论和字典用作知识源,从所述医学数据可以形成知识森林(各种面向非循环的知识图)。尤其是医学计划指南等都属于此类源。

[0110] 在实施例之一中,专用医学文章和/或培训手册可用作知识源。前提是,通过现有技术已知的文本识别方法(通过上述词汇和句法分析,使用经过训练的文本识别模型等)来处理初步找到的文章。

[0111] 在另一个实施例中,开放式生物学本体论被用作知识源。

[0112] 在实施例之一中,处理程序对数据进行归一化(每个数据类型可以使用单独的归一化规则)。

[0113] 例如,处理程序可以将表格形式的某些特定血液测量值(特征缩放)归一化,并在训练数据集上计算此类转换的参数。尤其是计算样本均值 μ 和方差 σ^2 ,以及

$$[0114] \quad \tilde{x} = \frac{x - \mu}{\sigma}$$

[0115] 在图像的另一个示例中,每个像素值对应于根据豪恩斯菲尔德(Hounsfield)强度测得的介质密度的值,处理程序可以将豪恩斯菲尔德比例映射到 $[-1, 1]$ 范围内(对于黑白图像,将在 $[0..255]$ 范围内的所有整数值归一化为在 $[-1.0..1.0]$ 范围内的实际值)。具体而言,可以用以下公式描述归一化:

$$[0116] \quad x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

[0117] $y = y_{\min} + (y_{\max} - y_{\min}) \times x'$

[0118] 其中

[0119] x -值空间中的可归一化值 $\{X\}$;

[0120] $x_{\min}$ -值空间中的最小值 $\{X\}$;

[0121] $x_{\max}$ -值空间中的最大值 $\{X\}$;

[0122] x' -值空间中的归一化值 $\{X\}$;

[0123] $y_{\min}$ -值空间中的最小值 $\{Y\}$;

[0124] $y_{\max}$ -值空间中的最大值 $\{Y\}$;

[0125] y' -值空间中的归一化值 $\{Y\}$;

[0126] 在另一个示例中,来自表的包含特定血液测量值的数据经过预处理-数据归一化(将每个参数减少到零平均单位方差,在训练数据集上计算这种变换的参数)。

[0127] 在另一个示例中,将从显微镜获得的RGB图像形式提供的数据进行后处理-从概率到分类二值化。如果从模型的角度来看,病理概率超过指定的阈值,则将图像标记为包含病理而不是不包含病理。

[0128] 在一个实施例中,处理程序过滤噪声或减少噪声(从有用信号中去除噪声的过程,改善其主观质量或减少数字数据传输通道和存储系统中的错误级别),该噪声为正在分析的数据的噪声。特别是,在图像处理中,可以使用一种用于视频的空间噪声抑制方法(自适应滤波,中值滤波,数学形态学,离散小波的方法等)-时间噪声抑制方法、空间噪声抑制方法或时空噪声抑制方法之一。

[0129] 让我们考虑图3中所示的示例:

[0130] 让我们假设有一个训练数据集310,其由患者记录组成。服务器对从每个样本记录的记录字段中选择的数据执行预处理。例如,服务器从训练数据集310提取患者健康记录中的一个记录301,定义字段内容和/或记录中包含的数据类型。在示例中,给定的记录301包括日期、自然语言描述、CT图像、血液分析。然后,服务器通过处理程序池320(处理程序3201..32N)中的相应处理程序来处理每个记录字段的数据。例如,日期可以由空处理程序3201处理,以自然语言描述的文本-由处理程序3202执行标准文本NLP处理,CT图像-由CT处理程序3203处理,血液分析-由处理程序3204进行数据归一化处理。处理完患者记录301后,包含处理后的数据301*,其中字段旁边的“*”符号表示该字段包含更改后的记录(与原始记录不同)。

[0131] 在实施例之一中,处理程序以使用脚本语言之一的形式形成,例如以插件、库(包括可执行的PE文件,例如.dll)的形式。

[0132] 在一个实施例中,服务器具有一组内置的程序,用于对数据类型的原始动作。以适合于用户的顺序组合这些程序,确保在没有外部帮助的情况下创建处理程序。在这种情况下,通过内置支持脚本语言或通过接口平台创建处理程序,以创建此类处理程序。

[0133] 当服务器已经执行所需的数据预处理后,服务器使用医学本体论将处理过的数据映射为每个患者的医学事实序列230。

[0134] 服务器将整个健康记录映射到有关患者的医学事实序列中。事实可能包含其他信息,例如生物医学图像、心电图(ECG)、分析结果等

[0135] 让我们考虑电子健康记录中的两个记录,以及将它们转换成医学术语序列的结

果,作为一个示例:

[0136] 转换前记录:

[0137]

2001 年 1 月 1 日

病人抱怨发烧和湿咳。该患者接受了胸部 X 光检查,并服用了<剂量>剂量的退热药<退热药>。

2001 年 1 月 1 日

<胸部 X 射线图像>

2001 年 1 月 2 日

确诊为肺炎。指定了以下治疗方法<自然语言治疗描述>

转换后的医学事实。

2001 年 1 月 1 日

<发烧>

<咳嗽>

<胸部 X 射线图像>

<退热药, 剂量>

[0138]

2001 年 1 月 2 日

<肺炎>

<药物 1, 剂量>

<药物 2, 剂量>

<程序 1>

[0139] 在将健康记录转换为一组医学事实之后,服务器会使用从健康记录中提取的诊断信息或其他感兴趣的事实,为每位患者自动排列所获得的医学事实序列240。如果数据被排列好,则服务器将跳过此步骤。

[0140] 在实施例之一中,感兴趣的事实由服务器用户指定或累积在该技术解决方案的用户(例如医生)中。

[0141] 例如,临床试验的包容性标准列表和排他性标准列表,即,一个人为了被包括在诊所中而必须遵守的标准列表(包容性列表),否则将被排除在诊所之外或不被允许进入诊所(排他性列表)。

[0142] 在另一个示例中,肿瘤最大5毫米的肝癌可能是一个包容性标准。

[0143] 在另一个示例中,吸烟或患者年龄超过65岁可能是排他性标准。

[0144] 在实施例之一中,感兴趣的事实是从外部源提取的。

[0145] 例如,感兴趣的事实可以从医学信息系统中提取。

[0146] 然后,服务器通过按时检查对事实进行整理和分组。这种分组对于在一次检查中同时考虑一组事实是必要的。

[0147] 在实施例之一中,分析可以与他们被命令时的出勤有关,或者可以分为个体本质(个体检查)。

[0148] 在实施例之一中,CT、MRT和组织学涉及个体检查。

[0149] 在实施例之一中,至少包含未处理的数据(不是医生的报告,而是图像、视频、时间标记形式的即时结果)的所有检查方法都与单个检查有关。如果仅提供医生的报告或仅检查的事实,则将此类数据视为检查的一部分。

[0150] 对于这样的检查分组,服务器使用与每个事实有关的时间和/或日期的信息。

[0151] 然后,服务器根据检查分组形成{事实集,诊断}对或{事实集,感兴趣的事实}对。

[0152] 在实施例之一中,通过简单搜索来形成配对。

[0153] 然后,服务器为每个数据模态准备训练数据集。如前所述,组织学、X射线、CT、乳腺X线照相等数据可以用作数据模态。

[0154] 例如,为了形成用于CT模态的训练数据集,服务器选择包含CT的记录:

```
{<CT, 图像>, <诊断_i>}, {<CT, 图像>, <诊断_j>},  
{<CT, 图像>, <诊断_k>}, ...
```

例如,对于概念(术语),服务器选择以下记录:

[0155]

```
{[<发烧>], <肺炎>}  
{[<发烧>, <咳嗽>], <肺炎>}  
{[<发烧>, <咳嗽>, <胸部 X 射线>], <肺炎>}  
{[<发烧>, <咳嗽>, <胸部 X 射线>, <退热药>], <肺炎>}
```

[0156] 然后,在为每个模态形成培训数据集之后,服务器分别针对每种模态执行主要代表的培训250。

[0157] 在服务器上为每种模式设置一个模型(一组模型),以预测在此训练数据集中显示并在这些模式中可用的诊断的训练。

[0158] 在一个实施例中,以下机器学习算法可以用作模型:

- [0159] • 线性回归;
- [0160] • 逻辑回归;
- [0161] • 最近邻k算法;
- [0162] • 随机森林;
- [0163] • 树上梯度增强;
- [0164] • 贝叶斯分类法;
- [0165] • 神经网络(完全连接、卷积、递归及其组合)。

[0166] 在一个实施例中,所述模型符合要求-对应于该模型将使用的模态。

[0167] 例如,卷积网络用于图像。

[0168] 在实施例之一中,为每个模态指定几个模型。

[0169] 在实施例之一中,模态被分组为具有相似的模型架构(相同的n参数族)并且可以

一起训练的群集(例如,所有组织学、X射线,乳房X线照相等),并且每个群集模型具有不同的权重集。

[0170] 在一个实施例中,为每个模态形成一组模型n参数族。N参数族意味着存在具有一组参数的通用模型类型,这些参数明确地定义指定模型。

[0171] 例如,如果将神经网络用作模型,则n参数族的示例之一是多层感知器,并且层数和层中神经元数将成为参数。另一个示例-具有固定结构的任何神经网络,其生成n参数族(例如,用于图像分割、图像分类等的族)。

[0172] 在此技术解决方案的范围内,应该使用以下主要的n参数族:

[0173] • 卷积神经网络,用于处理图像、视频、信号;

[0174] • 递归神经网络,用于处理患者健康记录中的事实序列并建立预测模型,用于处理非结构化文本信息;

[0175] • 贝叶斯方法和决策树,用于处理表数据。

[0176] 让我们仔细看一下处理图像的示例。

[0177] 在处理图像时,存在以下主要问题:图像分类(向每个图像分配一个或多个类别或标记)、图像分割(向每个图像像素分配一个或多个标记),感兴趣对象的定位(构建封闭矩形,对于图像中的每个对象,在其内部具有一个感兴趣的对象)。为这些问题中的每一个分配了一个解决问题的体系结构。通常,模态在输入图像的大小和感兴趣的目标标记/对象的数量上不同。密集网络是每个此类模型的基础,其架构概念如图5所示。

[0178] 该体系结构的概念使用模型中信息移动的其他方式,从而可以有效地训练甚至具有大量卷积层的大型模型。在指定的模态下,输入图像的大小和此类模型的类数构成n参数族,而神经网络权重只是该族的参数。它们是在训练过程中定义的,在此过程中将图像和目标标记呈现给模型,并且神经网络会更改其权重,以便其响应与训练数据集布局中的内容相匹配(所谓的目标值或目标响应)。

[0179] 然后,对于每种模态,服务器在该训练数据集中搜索提供最佳结果的族参数。

[0180] 在实施例之一中,至少使用以下方法来搜索族参数,给出最佳结果:

[0181] • 蒙特卡罗方法;

[0182] • 贝叶斯优化。

[0183] 在实施例之一中,将交叉验证用于模型评估,基于该模型评估,为此模态的该训练数据集选择最佳参数模型。交叉验证执行如下。通过不同的方法将数据集 X^L 划分为两N个不相交的子集:

$$[0184] \quad X^L = X_n^m \cup X_n^k,$$

[0185] 其中

[0186] X_n^m -训练子集的长度为m,

[0187] X_n^k -验证子集的长度为 $k=L-m$,

[0188] $n=1..N$ -分区号。

[0189] 为每个分区n构建算法

$$[0190] \quad a_n = \mu(X_n^m)$$

[0191] 并计算质量功能值

[0192] $Q_n = Q(a_n, X_n^k)$

[0193] 所有分区的算术平均值 Q_n 称为交叉验证评估:

[0194] $CV(\mu, X^L) = \frac{1}{N} \sum_{n=1}^N Q(\mu(X_n^m), X_n^k).$

[0195] 交叉验证评估用于选择最佳模型。

[0196] 在实施例之一中,在收集数据时添加模型,并添加新的患者数据源。

[0197] 例如,新型检查(例如ECG)已在医学信息系统中可用。然后,针对ECG模态形成训练数据集,然后在该训练数据集上训练神经网络(图4)模型(模型组),从中获得该模态的表示形式。

[0198] 训练完模型后,服务器为每个模态形成主要的矢量表示。为此,服务器将预处理的患者数据发送到为此模态训练的每个模型的输入,为每个记录定义模型输出值和该模型的最后一个隐藏层的权重值。服务器会将最后一个隐藏层的权重值用作主要矢量表示,其被模态映射到由模型定义的固定大小的矢量中。结果,服务器形成了新的数据集,这是原始数据集的一种转换。

[0199] 每个模态都有其自己的矢量维,例如如果模态为 m_u ,通用空间维为 n ,则建立映射:

[0200] $R^{m_u} \rightarrow R^n$

[0201] 例如,假设

[0202] $x \in R^{m_u}, y \in R^n,$

[0203] 然后,映射:

[0204] $y = Ax + b,$

[0205] 其中

[0206] $A \in R^{(n, m_u)}, b \in R^n$

[0207] 在另一个例子中

[0208] $y = f(Ax + b),$

[0209] 其中

[0210] f -非线性函数(例如线性整流函数(ReLU),S型等)。

[0211] 即 A 是矩阵的大小 n, m_u , and b is a vector of size n

[0212] 在实施例之一中,文本矢量表示被建立在空间中,所有其他模态的主要矢量表示被映射到该空间中,即,通过身份映射将文本模态的主要矢量表示映射到公共表示的空间。

[0213] 如果没有,则使用神经网络模型,有两种可能的情况:

[0214] • 模型输出本身被视为数据表示;

[0215] • 神经网络模型是将输入数据映射到一组事实的分类器,对于这些事实,已有矢量表示。由于可以保证任何模型都会生成事实可用性的概率矢量,因此将通过事实矢量表示的数量对矢量表示进行加权。

[0216] 例如,构建以下模型:

[0217] $f(x, \xi) : X \rightarrow Y,$

[0218] 其中

[0219] X -特征集,

[0220] $Y = \{y_i, i = 1, \dots, n\}$ -目标事实集

[0221] ξ -模型参数。

[0222] 在不失一般性的前提下,可以将问题重新表述为:

[0223] $P=f(x, \xi): X \rightarrow R^n$,

[0224] 和

[0225] $P=(p_1, p_2, \dots, p_n); \sum_{i=1}^n p_i = 1, p_i \geq 0$ 。

[0226] 对模型 p_i 的此类限制可被解释为患者(或其他变体,取决于所解决的问题,例如具有一年期的事实的出现,等等)的事实可用性概率 y_i 。

[0227] 具有训练数据集 $\{x_j, t_j\}, j=1, \dots, N$ 可以定义模型参数 ξ 。我们将训练中获得的模型参数称为 $\hat{\xi}$ 。而且,由于 y_i 中的每一个都是医学事实,因此矢量表示 V_i 对应于它

[0228] 然后对于一个新案例,我们获得以下内容:建立对应于此案例的输入特征矢量 $\bar{x}$;获得对应的概率矢量 $\bar{p}=f(\bar{x}, \hat{\xi})$;在这种情况下,该模态的矢量表示将如下构建:

[0229] $\bar{V}=\sum_{i=1}^n \bar{p}_i V_i$ 。

[0230] 服务器为每个模态形成主要矢量表示,从而产生一组医学事实和术语(诊断、症状、程序和药物)的矢量表示,以及用于将主要矢量表示映射到联合表示空间的模型。

[0231] 在实施例之一中,服务器另外预训练医学术语(概念)的矢量表示,例如,

[0232] • 以大量医学文献的形式提供额外的数据源;

[0233] • 给出了另一种训练语料库,该语料库与当前的训练语料库无关。

[0234] 服务器使用分布语义和单词矢量表示来预训练医学术语(概念)。

[0235] 各个上下文矢量被分配给每个单词。矢量集形成言语的矢量空间。

[0236] 在实施例之一中,医学术语(概念)的预训练是使用软件工具来执行的,该软件工具使用本体进行正则化来分析自然语言语义Word2vec。

[0237] 统计、机器学习、逆问题理论中的正则化-一种在语句中添加一些附加信息的方法,目的是解决不适当地的问题或防止重新训练。该信息通常以对模型复杂性的惩罚形式给出。

[0238] 例如,可能是:

[0239] • 结果函数平滑度的约束;

[0240] • 矢量空间范数的约束;

[0241] • 体重和神经元激励(firing)的正则化;

[0242] • 现有技术中已知的正则化方法。

[0243] 该技术解决方案使用现有技术中已知的主要和通用机器学习以及深度学习正则化方法。让我们假设E-误差函数,在训练过程中最小化,W-模型权重,A-激励隐藏层的所有神经元(相对于神经网络)。然后,最广泛使用的正则化技术之一称为 L_1 (L_2)正则化描述如下:代替最小化E,解决最小化问题

[0244] $E + \alpha L_1(W) \xrightarrow{W} \min(L_1 \text{重量调整})$,

[0245] $E + \alpha L_2(W) \xrightarrow{W} \min(L_2 \text{重量调整})$,

[0246] $E + \alpha L_1(A) \xrightarrow{W} \min(L_1 \text{激励正则化})$,

[0247] $E + \alpha L_2(A) \xrightarrow{W} \min(L_2 \text{激励正则化})$,

[0248] 其中

[0249] $L_p(x) = (\sum_{i=1}^n |x|^p)^{\frac{1}{p}}$ - L_p -范数。

[0250] 给定情况的不同变体是可能的。给定正则求和(条件)替代其他(轻微)限制。即对可能的模型权重的限制未以明确的方程组集和/或生成有效模型权重的 $\bar{W} \subset R^n$ 集的不等式的形式分配,从而避免了重新训练。以下内容还可以与 L_1/L_2 正则化一起使用:

[0251] • 提前停止:

[0252] 此方法的原理是从训练数据集中选择一个小的测试集,该测试集未明确涉及训练,但用于训练过程中的模型误差测量。一旦该测试集中的错误开始增加,训练就会停止。

[0253] • 数据扩充:

[0254] 该方法的原理是,以一定的概率将变换应用于训练数据集的每个示例。此转换不会更改所需的响应,也不会通过应用类似的转换来确保获取新的正确的所需响应。例如,在对胸部X射线图像是否存在肺炎征象进行分类时,可以对输入图像绕轴应用镜像映射,因为它显然不会改变目标标记。

[0255] • 可以通过对模型权重矢量的范数值的限制来明确地对模型参数施加限制: $L_1(W) < \gamma$ 或 $L_2(W) < \gamma$ 。

[0256] • 也可以应用机器和深度学习中广泛接受的其他正则化方法。

[0257] 让我们对 word2vec 工具的操作进行说明性描述: word2vec 输入接收较大的文本语料库,并将矢量与每个单词输出单词坐标相关联。首先,它创建一个字典,从输入的文本数据中“学习”,然后计算单词矢量表示。矢量表示基于上下文相似性:在文本中出现的,与相同单词接近(因此具有相似含义)的单词在矢量表示中将具有矢量-单词的紧密坐标。获得的矢量词(例如,图9A用于支气管炎,图9B用于风湿病)可用于自然语言处理和机器学习。

[0258] word2vec 中有两种主要的学习算法: CBOW (连续词袋) 和连续跳跃元语法 (Skip-gram)。CBOW-模型体系结构,它根据环境上下文预测当前单词。连续跳跃元语法架构以不同的方式起作用:它使用当前单词来预测环境单词。上下文顺序不影响任何这些算法的结果。

[0259] 在输出处获得的矢量文字的坐标表示使得能够计算文字之间的“语义距离”。基于这些词的上下文相似性, word2vec 技术给出了其预测。

[0260] 在实施例之一中,当使用本体论作为正则化(约束空间结构)时,使用注意。

[0261] 在实施例之一中,当使用本体论进行正则化(约束空间结构)时,基于本体术语之间的关系,使用多级(分层)误差函数。在特定情况下,本体论以知识图的形式使用,它指定了术语及其类别的层次结构。因为很明显,知识图中的相似性应该表示矢量空间中术语和类别之间的相似性,所以它可以预先安排矢量表示空间。使用此方法可以对矢量表示训练施加惩罚。此后,惩罚与主误差函数一起被最小化。让我们将当前术语矢量称为 c 。针对本体论的两个术语 c_1 和 c_2 之间的二进制相似性度量,我们称为 q 。如果 $q(c_1, b_2) = 0$, 则术语可以被认为是相似的,如果 $q(b_1, c_2) = 1$, 则术语可以被认为是关系较远的。然后将本体论错误函数定义如下:

$$[0262] \quad OD(c_i) = \sum_{c_j: q(c_i, c_j) = 0} \|c_i - c_j\| - \sum_{c_j: q(c_i, c_j) = 1} \|c_i - c_j\|$$

[0263] 现在,在矢量表示训练期间,OD(c)可以以类似于 L_1/L_2 正则化的方式使用。通过本体论使用正则化确保在不扩展训练数据集的情况下提高模型质量。由于对正则化所施加的表示空间的合理限制,因此改进了模型质量,尤其确保避免重新训练,从而使算法相对于训练集中的超限运行和错误更为稳定。经典正则化的标准方法也对表示空间施加了限制;但是,与基于本体的正则化相比,它们仅缩小了变体,后者基于域区域的外部信息对表示空间施加了限制。

[0264] 在实施例之一中,用于正则化的本体是相对于系统的外部参数,其可以被预先指定并且可以取决于相应的疾病代码系统(例如Idc9/10、MKB-10等)。

[0265] 对于在此步骤中获得的每个神经网络模型,提取主矢量表示作为隐藏层输出。所述模型使得能够将指定模态的输入数据映射到主要矢量表示中。它需要对训练后的模型进行简单的操作,实际上减少了从模型中删除输出层的工作。

[0266] 在训练并获得主要表示之后,服务器执行联合表示的协调多模态机器学习260(图10中所示)(更多信息:“多模式机器学习:调查和分类法”,Tadas Baltrusaitis,Chaitanya Ahuja和Louis-Philippe Morency)。

[0267] 为了在联合表示的协调的多模态机器学习过程中使用非文本数据,例如医学“图像,有必要训练从此模态空间到通用矢量空间的映射模型。为此,可以使用模态的主要矢量化和从主要矢量表示到通用矢量表示的可训练表示功能,前提是上述模型(训练用于从指定模态空间映射到通用矢量空间的模型)可以用作可训练功能。

[0268] 例如,如果模型仅处理图像分类,则几个隐藏的完全连接层可以跟随最后的卷积层。在这种情况下,仅获取最后一个隐藏卷积层而不是完全连接层的输出。

[0269] 如果在健康记录中出现非文本形式,例如CT扫描,采用多层感知器对其主要矢量表示进行处理,然后将该感知器输出视为此图像的矢量表示,并在其上计算其相邻像素之间的余弦距离。

[0270] 此后,使用连续跳跃元语法(skip-gram),但是,当涉及到非文本模式(例如医学图像)时,此模式的功能输出将用作其矢量表示,前提是要从健康记录或医学文本提取医学事实序列的语料库被传输到连续跳跃元语法输入。

[0271] 然后,在联合表示的协调多模态机器学习之后,服务器将执行最终模型和聚合参数的学习270。

[0272] 聚合是从一组矢量中获取单个矢量,其中每个矢量都是来自所选患者的健康记录的医学事实。

[0273] 学习期间获得的权重分配给每个事实。形成了一组权重,这组权重在特定患者的预后/诊断过程中使用-聚合参数。然后,通过将健康记录中的每个矢量乘以相应的权重来定义加权量,然后将获得的矢量相加。通常,在矢量表示的汇总过程中,将使用矢量的直接总 $c_{ag} = \sum_{i=1}^k c_i$ 和,其中汇总的患者表示是-此患者健康记录中事实的矢量表示。 $c_{ag} c_i$ 但是,聚合的这种变体并不总是最佳的,因为从为每种疾病或对该患者做出决定的角度来看,每个事实的权重都可能不同。因此,建议使用以下方法进行汇总: $c_{ag} = \sum_{i=1}^k a_i c_i$,其中 a_i 是标量, $\sum_{i=1}^k a_i = 1$ 。它们的每个 a_i 可以作为模型显式参数,并且可以在训练过程中定义,如函数 $a_i = f(i, c_1, c_2, \dots, c_k, \psi)$ 一样,其中 ψ 是该函数的参数,它们在训练过程中与其他模型权重

一起定义。

[0274] 构建计算图,其中权重作为参数。然后,通过梯度下降法根据当前数据集优化图形参数。所得的权重集是可训练的,即在训练过程中它与其他模型权重一起被修改。权重定义了n参数族的特定函数,该族从多个输入矢量形成一个输出矢量。

[0275] 上述所有内容可以总结如下:基于图形执行针对一组诊断的分类器训练,训练数据集是基于NLP技术从可用的EHR自动生成的(事实提取+它们在时间上的排列,然后成对的“事实”-“诊断”从中生成)。分类器的选择取决于使用未分类矢量特征的可能性,在这种方法中,这些是带有余数的多层完全连接神经网络。

[0276] 用于ECG和生物学图像的两条管道,例如胸部CT,如图7所示。首先,数据进入预处理块。预处理块是特定于域的,并将输入数据转换为模型希望接收的数据类型。转换后的数据被传输到与所述数据模态相对应的模型-例如,胸部CT被传输到神经网络,该神经网络对该检查进行分析。该模型可以输出两种结果(两个输出):它是所需的模型输出(病理概率、分割图等)和此特定示例的矢量表示。所需的模型输出将传输到与模型连接的后处理模块,并且该块输出将显示给人类专家,例如以报告的形式或以可接受的其他任何形式发送给客户。

[0277] 中心方案描述了医学概念的矢量空间,该矢量空间是基于连续跳跃元语法(skip-gram)+正则化构建的,但是本体和每个概念都映射到该空间的某个点。对于每个模型,也从矢量表示构建到该医学概念空间的映射,该矢量表示由模型通过映射功能形成管道生成。

[0278] 然后从这个通用空间中提取特定患者的矢量并将其传输到最终模型,在此将它们汇总到统一的患者模型中,根据该模型进行诊断和/或推荐治疗等。

[0279] 执行所有必需的操作后,管理员、医生或其他用户将要分析的患者记录添加(发送)到服务器。

[0280] 这样的记录的示例可以是:

[0281] 60岁的男子抱怨走路时胸痛,呼吸困难。健康记录附有心电图检查说明。心电图显示窦性心律,QRS复合体正常,V4-V6导联中明显ST段压低(1毫米)。未分配治疗。无鉴别诊断。

[0282] 然后,服务器对数据进行预处理,选择关键事实并将其转换为医学事实,例如:

[0283] <胸痛>

[0284] <呼吸困难>

[0285] <QRS复合体正常>

[0286] <ST段压低>

[0287] 此后,服务器将获得的事实集发送到现有模型的输入,并以与提交的事实集相对应的最大概率进行诊断。

[0288] 然后服务器接收模型使用的结果。服务器输出以下模型结果作为说明性示例:75%对应于“心绞痛”诊断,建议进行额外检查:自行车肌力测试,每日ECG监测。

[0289] 结果可以以建议的形式呈现,以医学图像的形式选择感兴趣的区域,以报告的形式呈现。

[0290] 疾病过程的分析和预后可能是由于(一个或多个)模型的应用所致。

[0291] 在某些实施例中,可以预测患者的死亡率。

[0292] 在一些实施例中,查询列表进入服务器。例如,查询是需要服务器处理的一系列CT检查。服务器执行处理,例如,它在CT切片中用红色圈出(突出显示)模型指定的感兴趣的潜在区域,并将所有获得的感兴趣体积(包含感兴趣区域)显示为列表。感兴趣的区域位于切片中,感兴趣的体积通过将多个感兴趣的区域聚合为一个而建立。

[0293] 在一些实施例中,可以预测心脏功能不全、肝病和其他疾病(病理)。

[0294] 下面给出了在其一个实施例中使用该技术方案解决方案的实验数据。

[0295] 来自MIMIC-III(重症监护医学信息市场)的数据免费访问,其包括用于研究的匿名健康数据,该数据与2001-2012年期间留在贝思以色列医学中心重症监护病房的约4万患者有关。

[0296] MIMIC-III包含有关人口统计学、实验室测量、程序、处方药、患者死亡率和患者留在医学中心期间的生命体征的信息。

[0297] 为了比较不同的模型,使用以下方法处理了数据。

[0298] 首先,对于每个患者,参照指定的预约,按预约日期排列,从数据库中提取有关诊断的信息(以ICD9代码的形式,其已在原始数据库中使用),处方药(以NDC标识符的形式,或者,如果无法获得,则以医学文本描述的形式),以及分配的过程(以过程ICD9代码的形式)。

[0299] 结果,获得的数据集包含大量患者,这些患者到医学中心的访问历史简短但信息量较少。可以使用有关此类患者的信息来训练体重矩阵,以建立诊断之间的关系。但是,在训练模型时,此类信息将无用:对于仅到诊所就诊一次的患者,无法识别在下次就诊时发生的事件。

[0300] 因此,为准备体重矩阵,使用了所有患者信息,并进行了模型训练,还对数据进行了处理:通过一年的滑动窗口滚动患者的健康记录,并将该年内记录的所有就诊记录视为独立的功能集,而访问次数少于2次的滑动窗口则不在考虑之列。

[0301] 从患者一年内访问医学中心的每个序列中,除了最新的预约,所有预约都被提取出来,并与相关信息一起使用,以提取传输到特定模型输入中的特征。然后,将所有此类序列以4:1的比例分为训练数据集和测试数据集。

[0302] 附加处理可能取决于特定模型输入接收的内容。

[0303] MIMIC-III数据库的设计使每次患者访问医学中心都可以关联多种诊断、药物和程序,并且在预约中的顺序也不是唯一定义的。因此,如果模型接收到有序的事件序列并且未考虑预约时间,则在训练该模型时,将以随机方式重新安排一次预约中的诊断、药物和操作,然后将“预约”统一起来成一个序列。

[0304] 由于这样的事件序列对于不同的患者将具有不同的长度,并且长期事件将对诊断的预测贡献较小,因此在一些实施例中,从最近的N个事件中学习模型(在患者健康记录中事件较少的情况下,事件增加为零)。模型仅考虑每1年长窗口的整个访问序列。

[0305] 权重矩阵的预训练

[0306] 在构建某些分类器时,使用了所谓的嵌入矩阵来获得MCV矢量或医学概念矢量-健康记录的紧凑表示形式,即有限长度的矢量,其元素为实数。

[0307] 图6示出了为患者获取医学概念矢量的方案。

[0308] 健康记录中按时间顺序排列的事件被考虑用于其建立。如果发生某种类型的事件,我们在事件对应的矢量位置中放入“1”。因此,我们获得了一个由1和0构成的高维矢量,

在1位置其对应于健康记录中发生的事件,在0位置其对应于健康记录中未发生事件。

[0309] 在某些模型中,修改了此通用方案:例如,考虑在一个预约(和一个时间标记)内以随机方式重新安排的事件,将固定数量的最新事件记录到事件矢量中,或者另外添加一个预约对应的事件乘以由事件多久之前发生定义的权重。

[0310] 健康记录的这种稀疏表示需要大量内存,基于此的模型训练需要大量时间。为了减少时间和紧凑数据,根据健康记录的稀疏表示绘制患者的医学概念矢量。

[0311] 为了最简单地获得紧凑表示,使用了所谓的嵌入矩阵,通过该矩阵可以将健康记录的稀疏矢量相乘。已经考虑了几种矩阵:

[0312] Word2Vec:作为一个矩阵,我们采用了一个系数矩阵,该系数矩阵是根据一次预约中存在的辅助诊断、方向和药物的分析而获得的。为了进行训练,我们使用具有skipgram机制的word2vec模型,以便获得一定长度的医学概念矢量,其用于任何诊断、医学程序或处方药(对应于嵌入矩阵列)。

[0313] 该体重矩阵是从健康记录中获悉的,其包括诊断代码、症状、处方案程序和处方药,以提取有关诊断之间关系的更多信息。

[0314] 本体论嵌入:使用了本体信息,即,以ICD9代码表示的,位于诊断树的高级节点上的代码用于获取事件的紧凑表示。

[0315] 嵌入到ICD9树中:以获得紧凑表示,可以使用非标准的正则化函数,该函数最大化树中到远处对象的距离并最小化树中到近处对象的距离(同时校正ICD9代码树中父节点的矢量)。与现有方法相反,权重矩阵已进行了预训练,并且已经过训练,将其用于模型训练,在现有方法中,针对特定问题对诊断权重及其家族树中父母进行训练。

[0316] 对于每个集问题,可以考虑几个分类器:

[0317] 基本的1-hot编码:此模型是在逻辑回归的基础上构建的,将健康记录的0和1-稀疏矢量表示的数组输入到该模型中。仅考虑疾病。

[0318] TFI-IDF编码:此模型是在逻辑回归的基础上建立的,输入阵列传送到该逻辑回归,输入阵列的箱位(slots)与患者疾病相关。它的构建与先前的模型类似,除了考虑了健康记录中的疾病发生数量的事实,然后使用TF-IDF算法处理输入特征,以将更大的权重与诊断罕见病例相关联(通常罕见,但在特定患者中经常发生)。

[0319] 以下几种模型使用相似的神经网络架构进行分类,但权重矩阵不同。

[0320] Word2Vec嵌入:该模型使用权重矩阵以患者矢量的形式获取健康记录的紧凑表示形式。基于连续跳跃元语法的Word2vec矩阵用作权重矩阵。将获得的紧凑表示用作逻辑回归的特征。

[0321] Word2Vec嵌入+注意:在模型中使用Word2Vec权重矩阵来获取模型中患者矢量的紧凑表示。此外,使用了具有注意机制的神经网络架构。

[0322] 具有ICD9树的嵌入:具有嵌入矩阵的模型,基于ICD9代码树构建。通过将矩阵乘以患者矢量获得的患者的紧凑表示,用于建立基于逻辑回归的模型。

[0323] 具有ICD9树+注意的嵌入:基于ICD9代码树构建的带有嵌入矩阵的模型,其中使用了具有注意机制的神经网络架构。

[0324] 具有ICD9树+注意+反比文档频数权重评价方法(tfidf)的嵌入:该模型与前一个模型的不同之处在于,将TFI-IDF编码模型为指定患者返回的值另外传输到其输入。

[0325] Choi嵌入+注意:具有嵌入矩阵的模型,该模型基于矢量压缩表示构建,在Choi等所有“GRAM中考虑:使用注意机制,基于图的注意模型用于医学保健表示学习。

[0326] 基于时间的模型:复制了患者mcv矢量的绘制方法。

[0327] 图8示出了通用计算机系统的示例,在该通用计算机系统中可以实现当前技术解决方案,该计算机系统包括计算机20或服务器的形式的多功能计算单元,其包括处理器21、系统存储器22和系统总线23,它链接了不同的系统组件,包括带处理器21的系统内存。

[0328] 系统总线23可以是任何不同的总线结构类型,其包括使用多种总线体系结构中的任何一种的存储器总线或存储器控制器、外围总线和本地总线。系统存储器包括只读存储器(ROM) 24和随机存取存储器(RAM) 25。ROM 24存储基本的输入/输出系统26(BIOS),其由一些程序组成,这些程序用于在启动过程中在计算机20内部的各个元素之间交换信息。

[0329] 计算机20还可以包括用于读取和写入硬盘的硬盘驱动器27,用于读取和写入可移动磁盘29的磁盘驱动器28以及用于读取和写入可移动光盘31(例如,压缩盘、数字视盘和其他光学手段)的光盘驱动器30。硬盘驱动器27、磁盘驱动器28和光盘驱动器30分别通过硬盘驱动器接口32、磁盘驱动器接口33和光盘驱动器接口34连接到系统总线23。磁盘驱动器及其计算机可读的相应方式,确保了计算机20的指令、数据结构、程序模块和其他计算机可读数据的非易失性存储。

[0330] 尽管本文所述的典型配置使用硬盘/可移动磁盘29和可移动光盘31,但对于本领域技术人员显而易见的是,在典型的操作环境中,可以使用其他类型的计算机可读方式来存储从计算机可访问的数据,例如磁带、闪存驱动器、数字视频磁盘、伯努利盒式磁带、随机存取存储器(RAM)、只读存储器(ROM)等。

[0331] 可以将包括操作系统35在内的不同程序模块保存到硬盘、磁盘29、光盘31、ROM 24或RAM 25。计算机20包括链接到操作系统35或集成到操作系统35的文件系统36,一个或多个软件应用程序37,其他程序模块38和程序数据39。用户可以通过键盘40和定点设备42等输入设备向计算机20输入命令和信息。其他输入设备(未示出)可以包括麦克风、操纵杆、游戏手柄、卫星天线、扫描仪或其他任何设备。

[0332] 这些设备和其他设备通常通过链接到系统总线的串行端口接口46连接到处理器21,但是它们可以通过并行接口、游戏端口或通用串行总线(USB)等其他接口连接。监视器47或其他可视化显示单元也通过接口(例如视频适配器48)连接到系统总线23。除监视器47外,个人计算机通常还包括其他外围输出设备(未示出),例如扬声器和打印机。

[0333] 通过与一台或多台远程计算机49的逻辑连接,计算机20可以在网络附近工作。远程计算机(或计算机)49可以是另一台计算机、服务器、路由器、联网的P、对等设备或单个网络的其他节点,并且通常也包含有关计算机20的大多数或所有上述元件,尽管仅示出了存储设备50。逻辑连接包括局域网(LAN) 51和广域网(WAN) 52。这样的网络邻居通常用于办公室、企业内部网、互联网。

[0334] LAN邻域中使用的计算机20通过网络接口或适配器53连接到本地网络51。WAN邻域中使用的计算机20通常使用调制解调器54或其他设备与全球计算机网络52(例如Internet)进行通信。

[0335] 调制解调器54(可以是内部调制解调器或外部调制解调器)通过串行端口接口46连接到系统总线23。在网络附近,可以将有关计算机20所述的程序模块或其部分存储在远

程存储设备中。有必要考虑到图示的网络连接是典型的,其他方式也可用于计算机之间的通信。

[0336] 最后的评论是,说明书中给出的数据是示例,其不限制由权利要求书限定的该技术方案的范围。对于本领域技术人员而言显而易见的是,可以存在符合该技术方案的实质和范围的该技术方案的其他实施例。

[0337] 文献:

[0338] 1.<https://hackemoon.com/attention-mechanism-in-neural-network-30aaf5e39512>

[0339] 2.<https://medium.com/@Synced/a-brief-overview-of-attention-mechanism-13c578ba9129>

[0340] 3.“从电子健康记录中学习医学概念表示及其在心力衰竭预测中的应用”,爱德华·崔(Andrew Choi),安迪·舒茨(Andy Schuetz),沃尔特·斯图尔特(Walter F.Stewart),孙继孟,2016年2月11日。

[0341] 4.“用于医学保健表示学习的基于图的注意模型”,爱德华·崔,穆罕默德·塔哈·巴哈多里,松乐,沃尔特·斯图尔特,孙继孟,2017年。

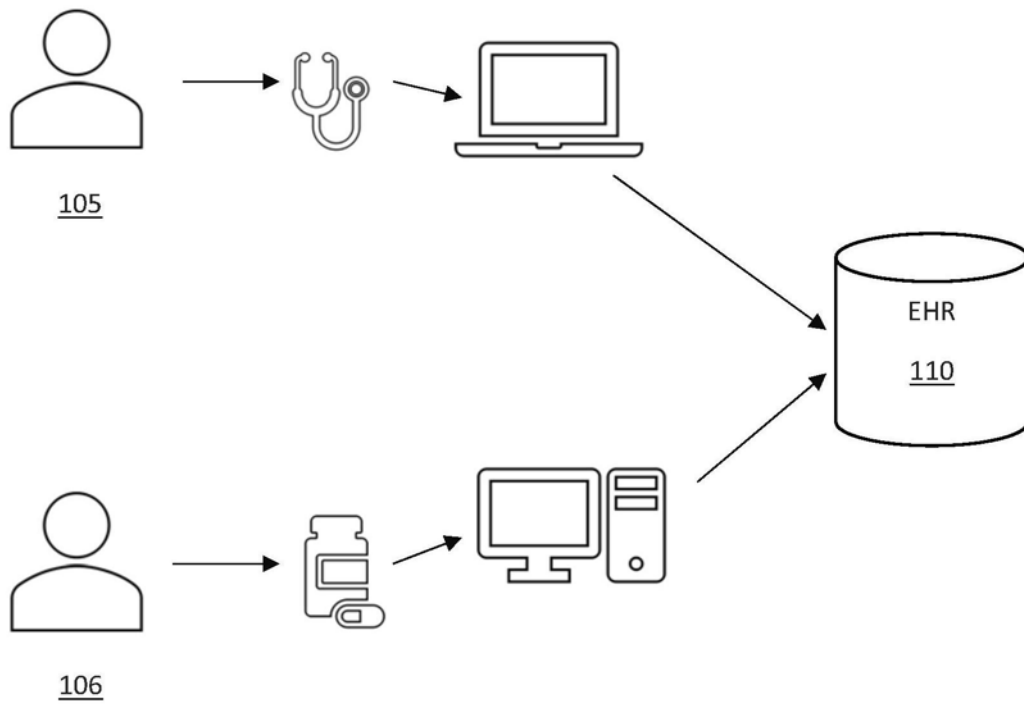


图1

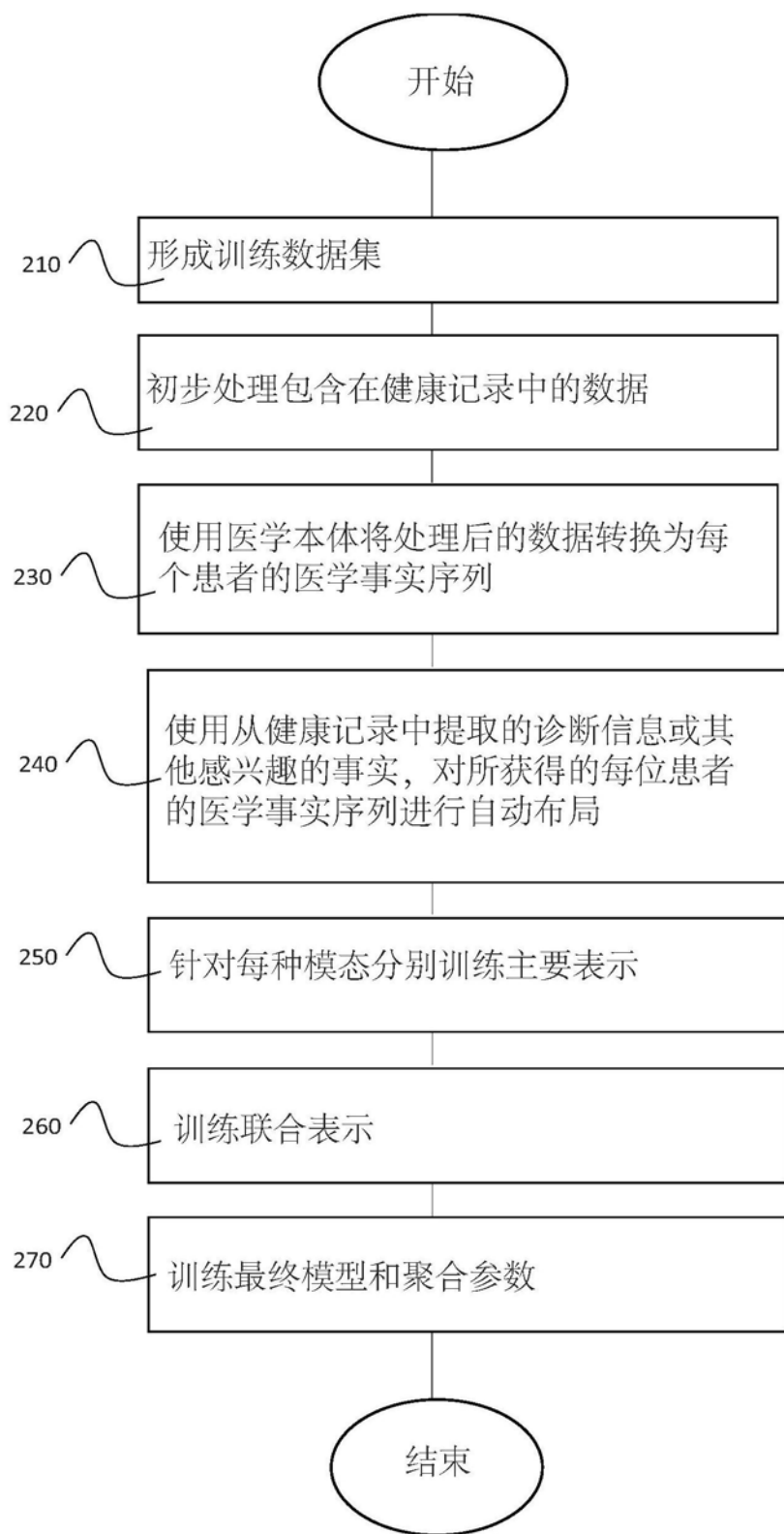


图2

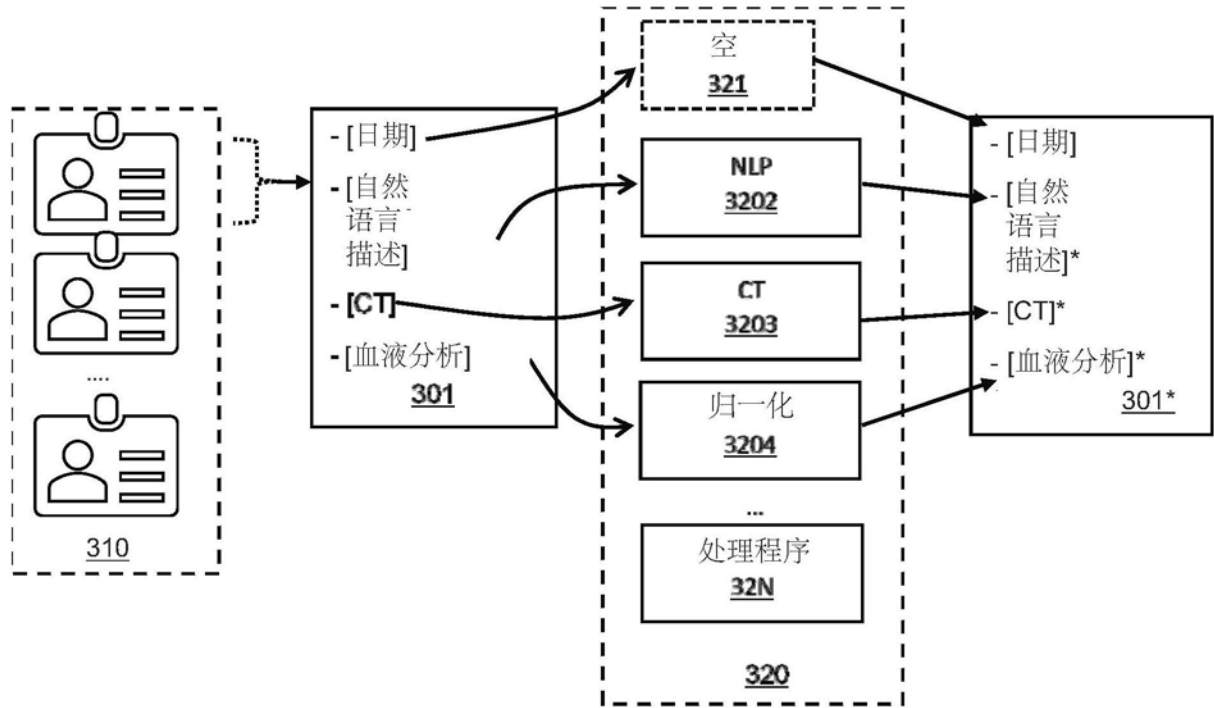


图3

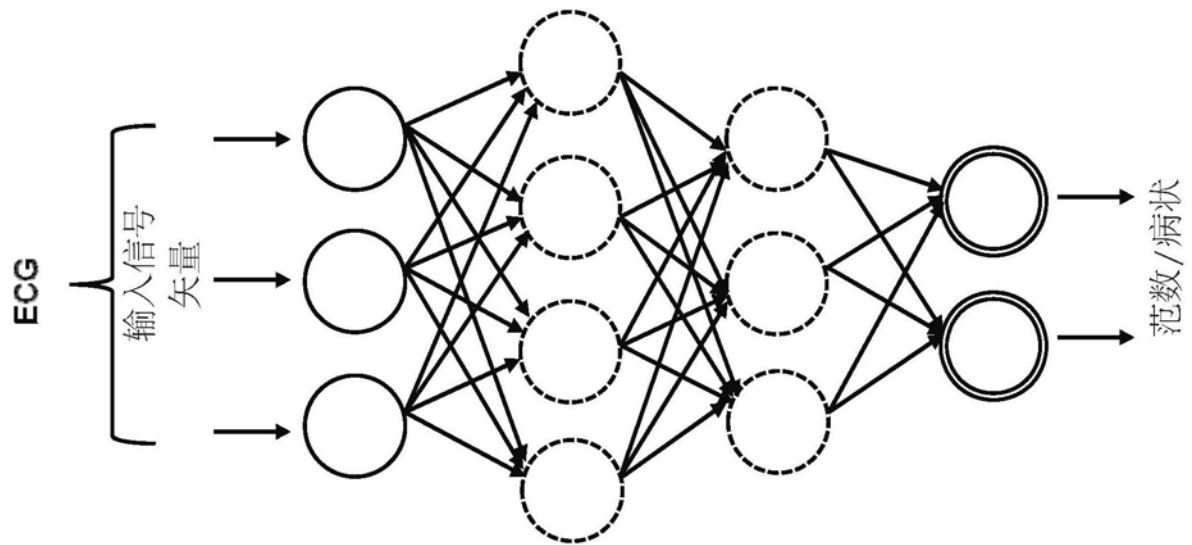


图4

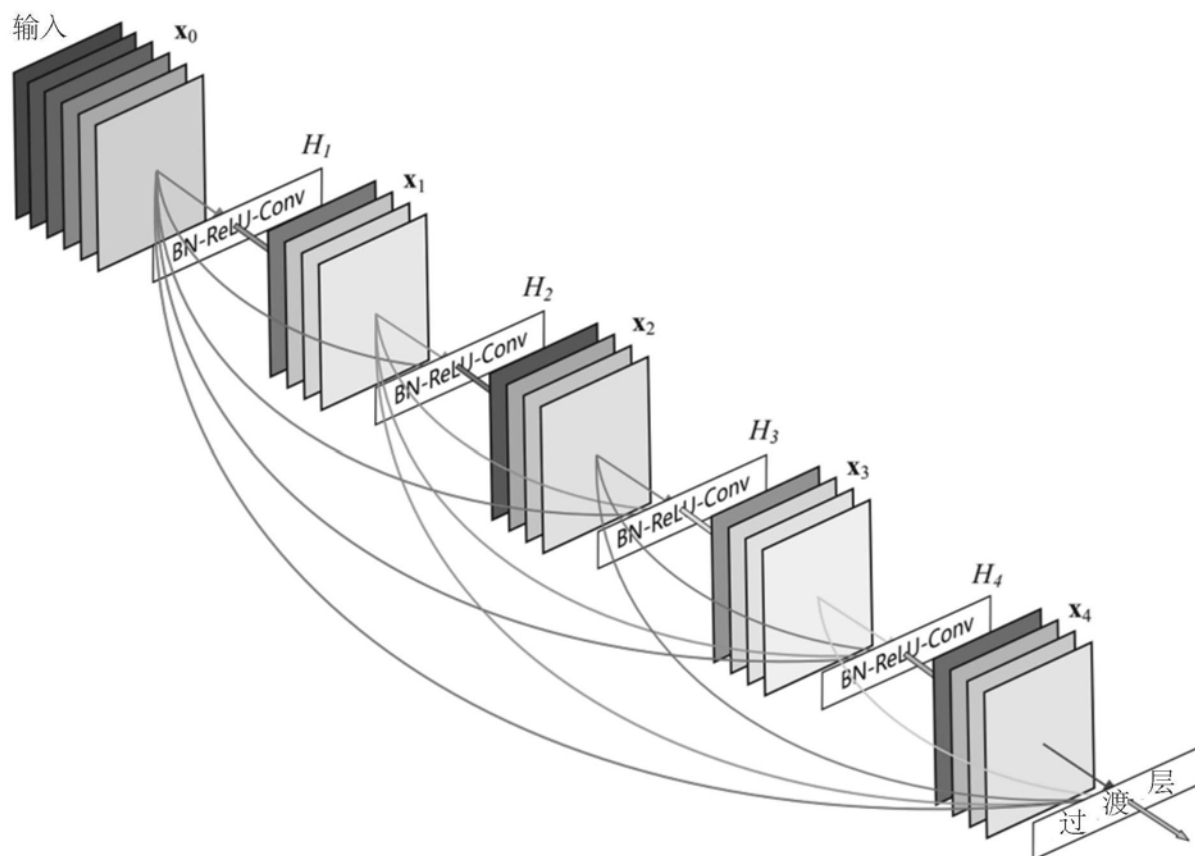


图5

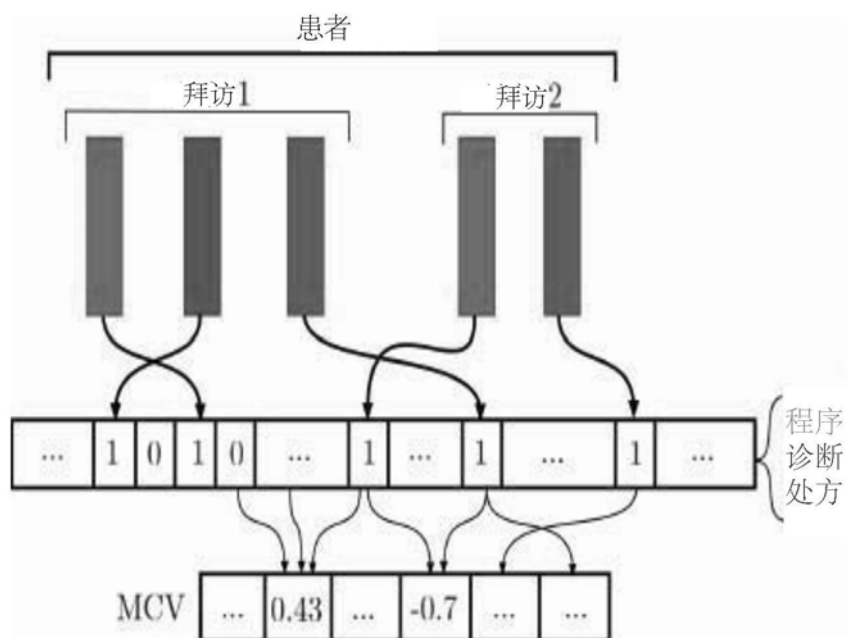


图6

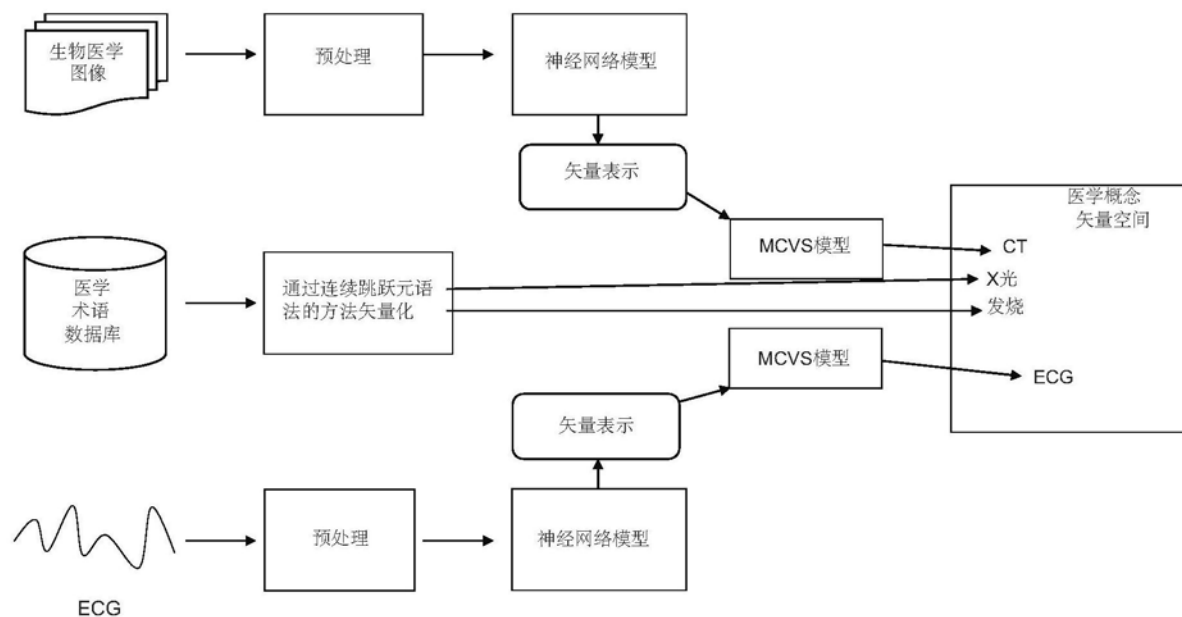


图7

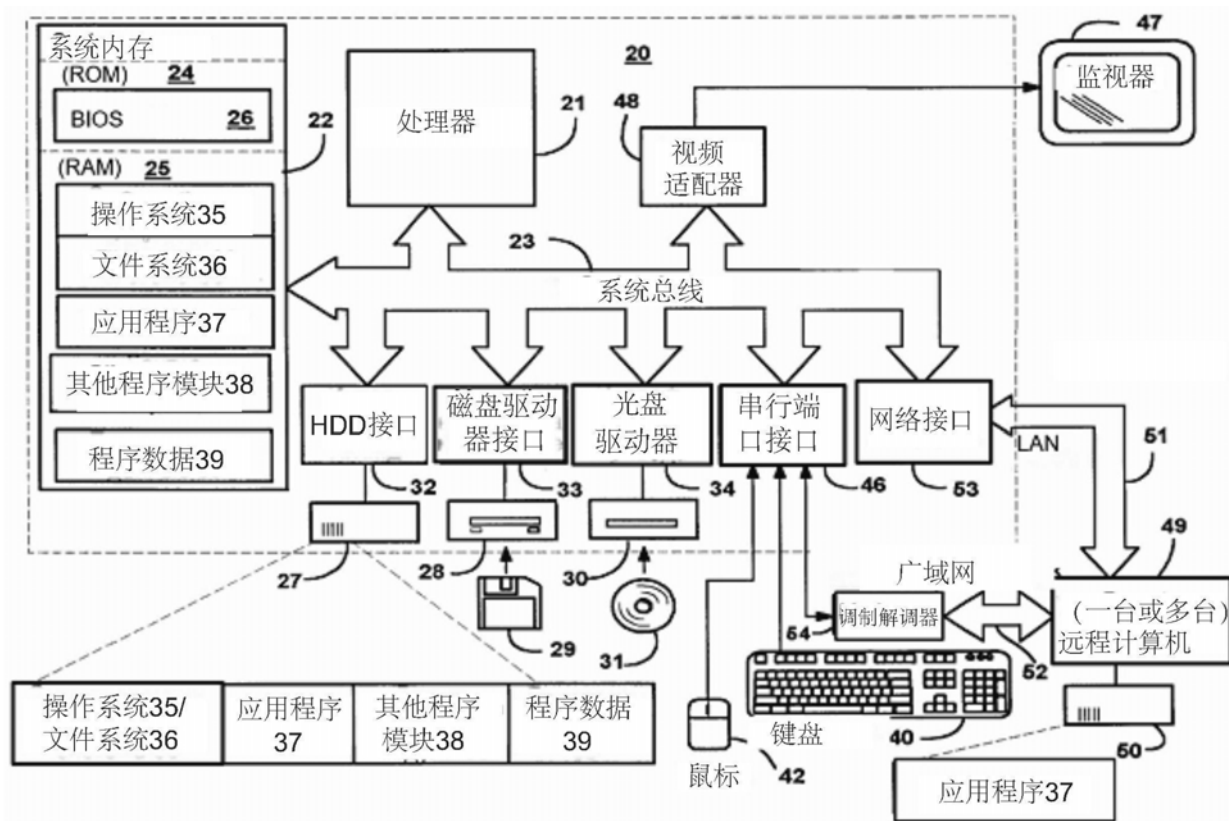
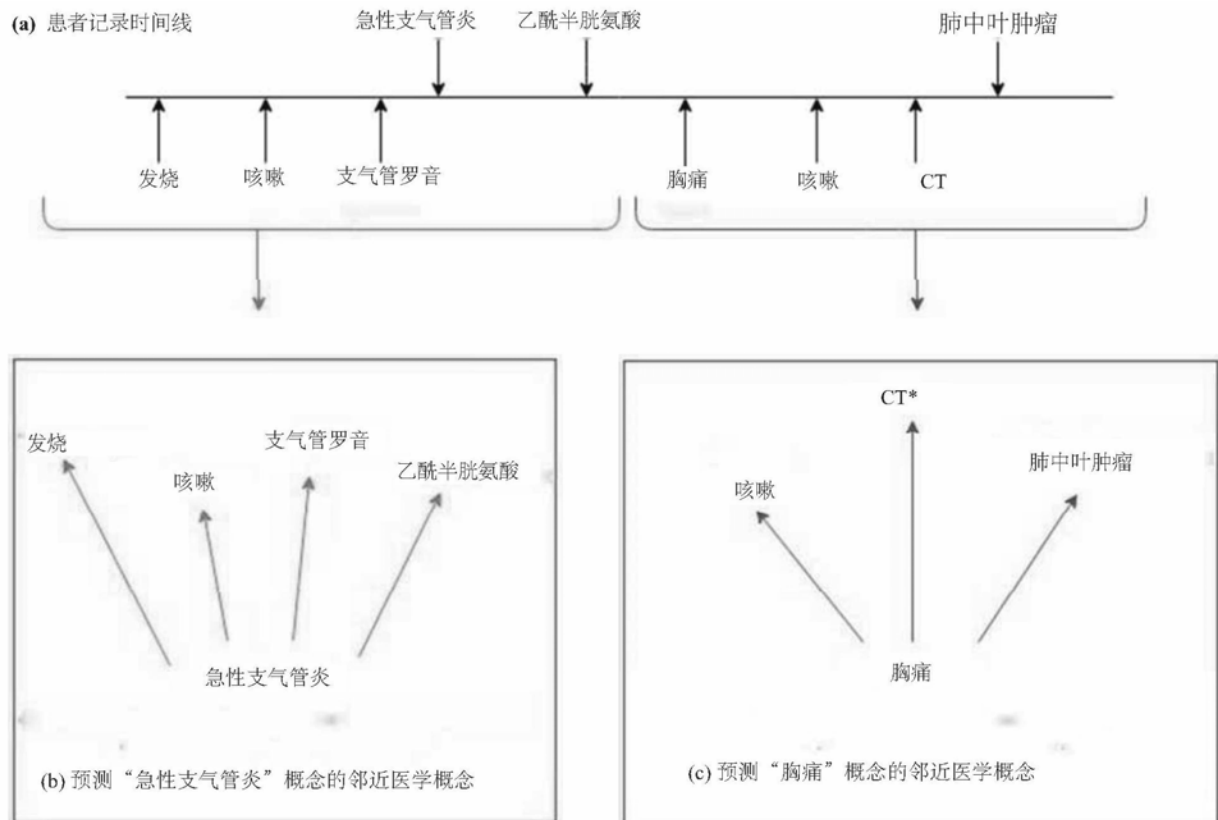


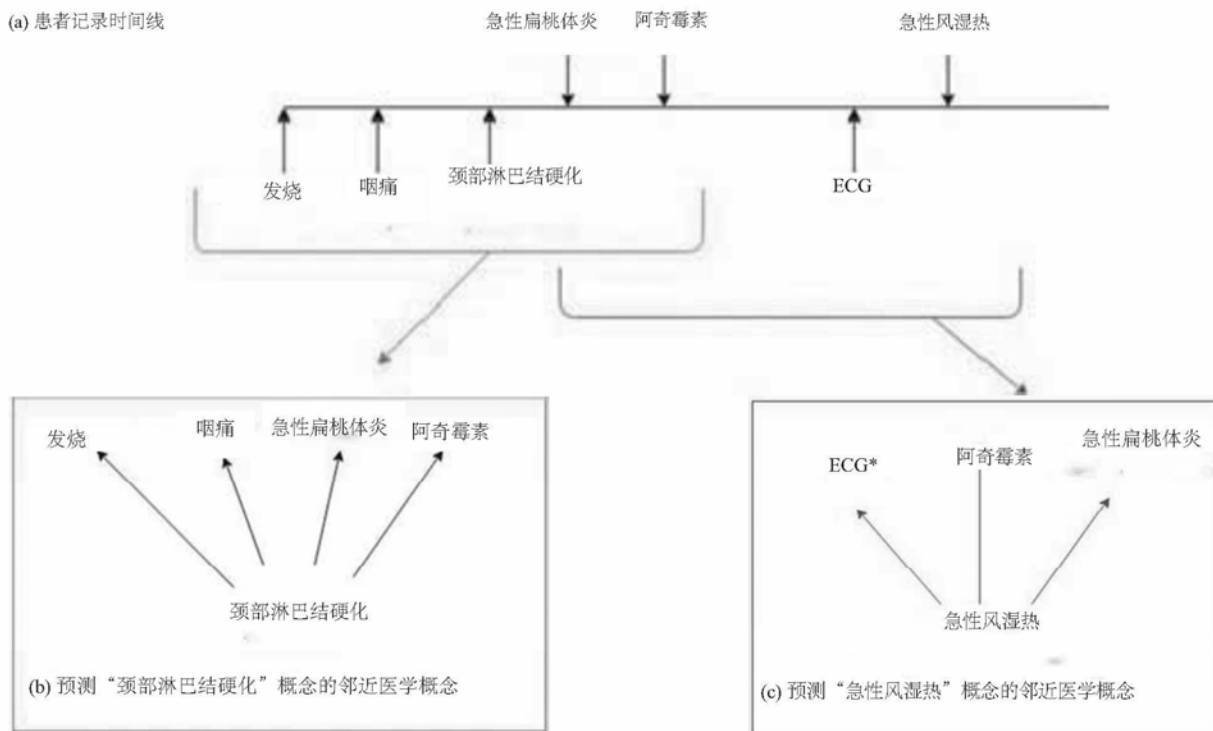
图8



*CT - 由神经网络从原始图像生成的矢量

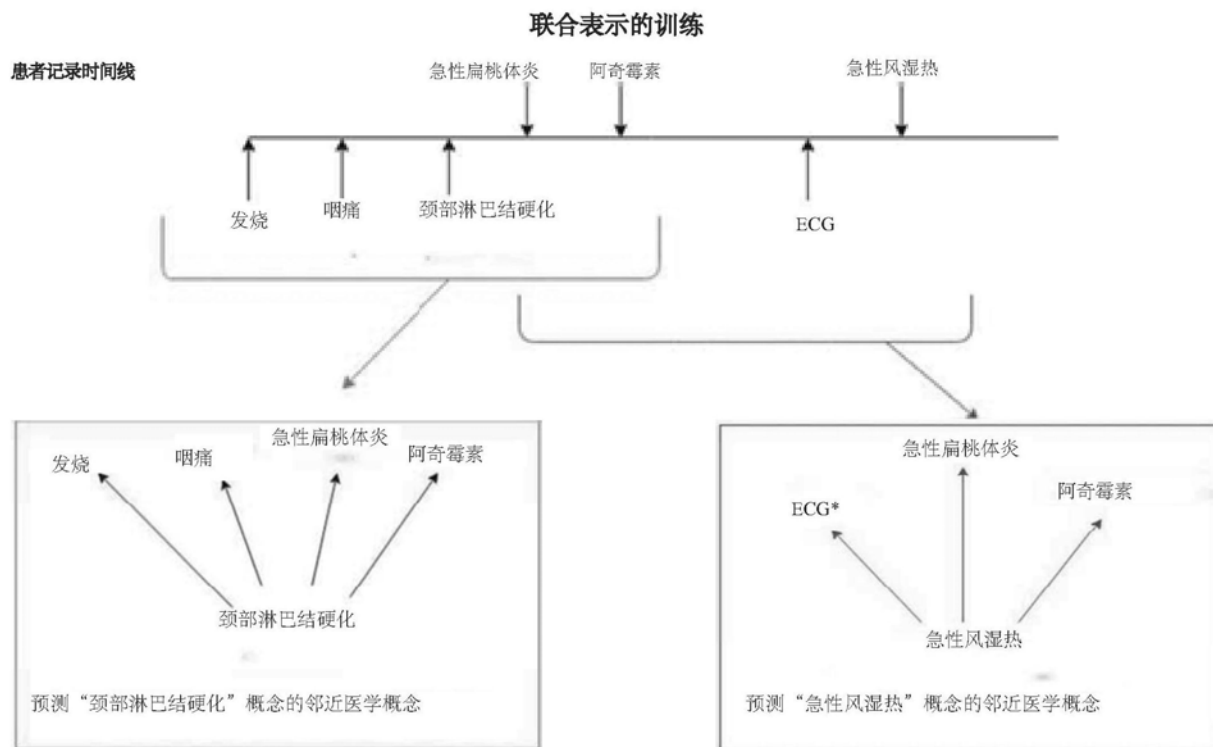
图9A

(a) 患者记录时间线



*ECG - 由神经网络从原始信号生成的矢量

图9B



*ECG - 由原始信号生成的矢量，由神经网络根据原始表示进行处理

图10A

神经网络训练

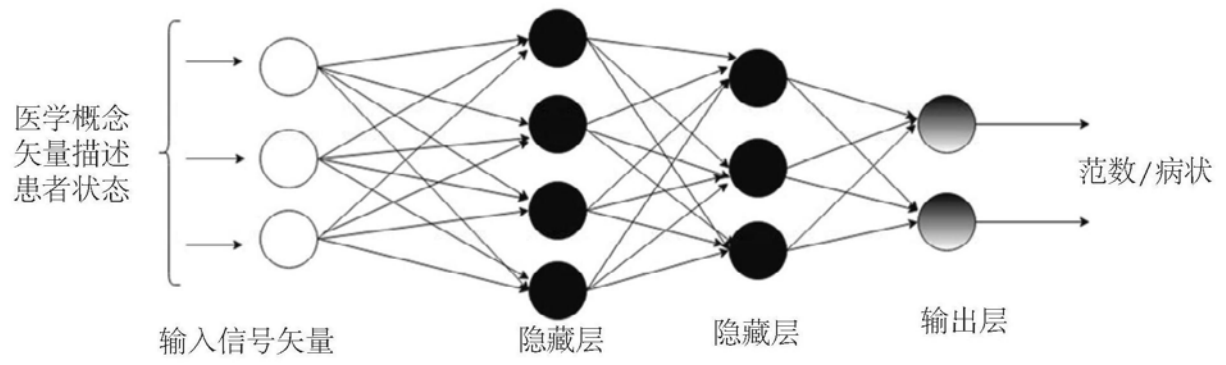


图10B