

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 9 月 10 日 (10.09.2020)



(10) 国际公布号
WO 2020/177230 A1

- (51) 国际专利分类号:
G16H 50/70 (2018.01)
- (21) 国际申请号: PCT/CN2019/090873
- (22) 国际申请日: 2019 年 6 月 12 日 (12.06.2019)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201910171593.0 2019年3月7日 (07.03.2019) CN
- (71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。
- (72) 发明人: 陈娴娴(CHEN, Xianxian); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号

平安金融中心 23 楼, Guangdong 518033 (CN)。
阮晓雯(RUAN, Xiaowen); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。徐亮(XU, Liang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。

(74) 代理人: 广州华进联合专利商标代理有限公司(ADVANCE CHINA IP LAW OFFICE); 中国广东省广州市天河区珠江东路 6 号 4501 房(部位: 自编 01-03 和 08-12 单元)(仅限办公用途), Guangdong 510623 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS,

(54) Title: MEDICAL DATA CLASSIFICATION METHOD AND APPARATUS BASED ON MACHINE LEARNING, AND COMPUTER DEVICE AND STORAGE MEDIUM

(54) 发明名称: 基于机器学习的医疗数据分类方法、装置、计算机设备及存储介质

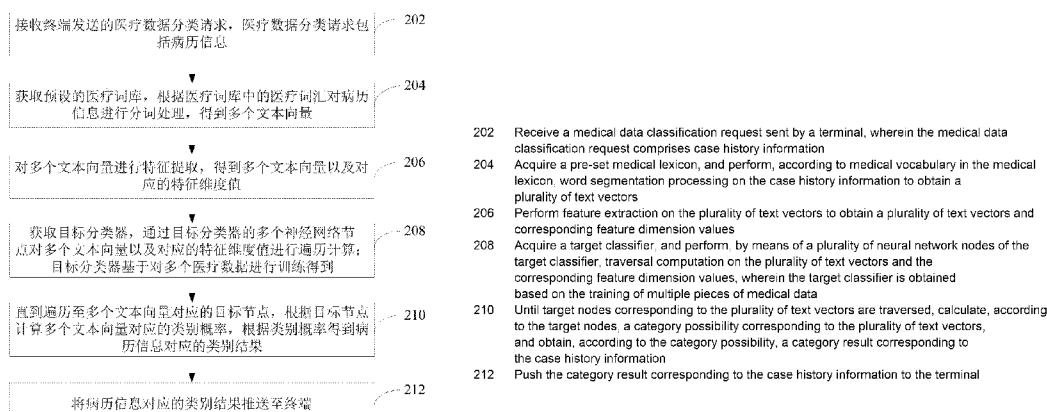


图 2

(57) Abstract: A medical data classification method based on machine learning. The method comprises: receiving a medical data classification request sent by a terminal, wherein the medical data classification request comprises case history information; acquiring a pre-set medical lexicon, and performing, according to medical vocabulary in the medical lexicon, word segmentation processing on the case history information to obtain a plurality of text vectors; performing feature extraction on the plurality of text vectors to obtain a plurality of text vectors and corresponding feature dimension values; acquiring a target classifier, wherein the target classifier is obtained based on the training of multiple pieces of medical data, and performing, by means of a plurality of neural network nodes of the target classifier, traversal computation on the plurality of text vectors and the corresponding feature dimension values; until target nodes corresponding to the plurality of text vectors are traversed, calculating, according to the target nodes, a category possibility corresponding to the plurality of text vectors, and obtaining, according to the category possibility, a category result corresponding to the case history information; and pushing the category result corresponding to the case history information to the terminal.

JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 一种基于机器学习的医疗数据分类方法, 该方法包括: 接收终端发送的医疗数据分类请求, 医疗数据分类请求包括病历信息; 获取预设的医疗词库, 根据医疗词库中的医疗词汇对病历信息进行分词处理, 得到多个文本向量; 对多个文本向量进行特征提取, 得到多个文本向量以及对应的特征维度值; 获取目标分类器, 目标分类器基于对多个医疗数据进行训练得到, 通过目标分类器的多个神经网络节点对多个文本向量以及对应的特征维度值进行遍历计算; 直到遍历至多个文本向量对应的目标节点, 根据目标节点计算多个文本向量对应的类别概率, 根据类别概率得到病历信息对应的类别结果; 将病历信息对应的类别结果推送至终端。

说明书

发明名称：基于机器学习的医疗数据分类方法、装置、计算机设备及存储介质

本申请要求于 2019 年 03 月 07 日提交中国专利局，申请号为 2019101715930，申请名称为“基于机器学习的医疗数据分类方法、装置和计算机设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及计算机技术领域，特别是涉及一种基于机器学习的医疗数据分类方法、装置、计算机设备及存储介质。

背景技术

近年来，癌症患病率不断在增加，癌症作为一个重要的健康问题，对于癌症的早期诊断和治疗能够使得癌症病人的存活率显著升高。随着计算机技术和医疗技术的迅速发展，出现了一些对大量医疗数据进行智能分类的方式，例如从医案书籍中提取单片医案中的结构化词表，并建立医案主题模型，并对医案主题进行训练得到相应的类别。或者利用先验知识对输入样本进行训练，进而对癌症类型进行分类，有助于减轻医务人员的劳动强度。

传统的医疗数据分类方式中，分类分析的数据大多是采用已有的固定的数据，数据来源比较有限，无法对用户实际的病历信息进行分类分析，而病历信息多为较繁杂和具体的病历分析、记录文本，由于医学文本的特殊性，病历信息中的词汇的偏差将会导致语义的完全不一致。

发明内容

一种基于机器学习的医疗数据分类方法，由计算机设备执行，所述方法包括：

接收终端发送的医疗数据分类请求，所述医疗数据分类请求包括病历信息；

获取预设的医疗词库，根据所述医疗词库中的医疗词汇对所述病历信息进行分词处理，得到多个文本向量；

对所述多个文本向量进行特征提取，得到多个文本向量以及对应的特征维度值；

获取目标分类器，通过所述目标分类器的多个神经网络节点对所述多个文本向量以及对应的特征维度值进行遍历计算；所述目标分类器基于对多个医疗数据进行训练得到；

直到遍历至所述多个文本向量对应的目标节点，根据所述目标节点计算所述多个文本向量对应的类别概率，根据所述类别概率得到所述病历信息对应的类别结果；及

将所述病历信息对应的类别结果推送至所述终端。

在其中一个实施例中，所述病历信息中包括多个文本数据，所述对所述病历信息进行分

词处理的步骤包括：获取预设的医疗词库，所述医疗词库中包括多个医疗词汇；将所述病历信息中的多个文本数据与所述医疗词库进行匹配，计算所述病历信息中的文本数据与多个医疗词汇的匹配度，提取达到预设匹配度的文本数据；根据匹配后的文本数据对所述病历信息进行分词，得到分词后的多个文本数据；及对所述分词后的多个文本数据进行向量转换，得到多个文本向量。

在其中一个实施例中，所述对所述多个文本向量进行特征提取，得到多个文本向量以及对应的特征维度值的步骤包括：计算所述多个文本向量的词频和逆向文件频率；根据所述词频和所述逆向文件频率按照预设算法计算多个文本向量的权重；提取出所述权重达到预设阈值的文本向量；及根据预设算法和所述权重计算所述文本向量对应的特征维度值。

在其中一个实施例中，构建所述目标分类器的步骤包括：获取多个医疗数据，根据所述多个医疗数据生成对应的训练集数据和验证集数据；对所述训练集数据中的多个医疗数据进行聚类分析，得到聚类结果；对所述聚类结果进行特征提取，提取出多个特征变量；获取预设的神经网络模型，通过所述神经网络模型对所述训练集数据进行训练，得到多个特征变量对应的特征维度值和权重，根据多个特征变量对应的特征维度值和权重构建初始分类器；及利用所述验证集数据对所述分类器进行进一步训练和验证，直到所述验证集数据中满足预设阈值的数量达到预设比值时，则停止训练，得到所需的目标分类器。

在其中一个实施例中，所述文本中包括多个文本句，所述多个文本句组成文本块，所述通过所述目标分类器的多个神经网络节点对所述多个文本向量以及对应的特征维度值进行遍历计算多个文本向量对应的类别步骤包括：利用所述目标分类器根据所述特征维度值计算所述多个文本向量之间的相关性，根据所述相关性计算所述文本中成句的文本句，并计算所述文本句的句向量；提取所述句向量的特征，根据所述多个句向量的特征计算出文本块向量；及计算所述文本块向量对应每个类别的概率，提取达到预设概率值的类别，并对所述文本块添加对应的类别标签。

在其中一个实施例中，所述方法还包括：根据预设频率从预设数据库中获取多个历史医疗数据；对多个历史医疗数据进行聚类分析，得到分析结果；根据所述分析结果进行特征选择，得到多个特征变量；根据预设的算法计算多个特征变量的权重；及根据多个特征变量和对应的权重对所述目标分类器进行优化调整。

一种基于机器学习的医疗数据分类装置，所述装置包括：

请求接收模块，用于接收终端发送的医疗数据分类请求，所述医疗数据分类请求包括病历信息；

分词处理模块，用于获取预设的医疗词库，根据所述医疗词库中的医疗词汇对所述病历信息进行分词处理，得到多个文本向量；

特征提取模块，用于对所述多个文本向量进行特征提取，得到多个文本向量以及对应的特征维度值；

数据分类模块，用于获取目标分类器，通过所述目标分类器的多个神经网络节点对所述多个文本向量以及对应的特征维度值进行遍历计算；所述目标分类器基于对多个医疗数据进行训练得到；直到遍历至所述多个文本向量对应的目标节点，根据所述目标节点计算所述多个文本向量对应的类别概率，根据所述类别概率得到所述病历信息对应的类别结果；及

数据推送模块，用于将所述病历信息对应的类别结果推送至所述终端。

在其中一个实施例中，所述分词处理模块还用于获取预设的医疗词库，所述医疗词库中包括多个医疗词汇；将所述病历信息中的多个文本数据与所述医疗词库进行匹配，计算所述病历信息中的文本数据与多个医疗词汇的匹配度，提取达到预设匹配度的文本数据；根据匹配后的文本数据对所述病历信息进行分词，得到分词后的多个文本数据；及对所述分词后的多个文本数据进行向量化，得到多个文本向量。

一种计算机设备，包括存储器和处理器，所述存储器存储有至少一条计算机可读指令，所述计算机可读指令由所述处理器加载并执行以下步骤：

接收终端发送的医疗数据分类请求，所述医疗数据分类请求包括病历信息；

获取预设的医疗词库，根据所述医疗词库中的医疗词汇对所述病历信息进行分词处理，得到多个文本向量；

对所述多个文本向量进行特征提取，得到多个文本向量以及对应的特征维度值；

获取目标分类器，通过所述目标分类器的多个神经网络节点对所述多个文本向量以及对应的特征维度值进行遍历计算；所述目标分类器基于对多个医疗数据进行训练得到；

直到遍历至所述多个文本向量对应的目标节点，根据所述目标节点计算所述多个文本向量对应的类别概率，根据所述类别概率得到所述病历信息对应的类别结果；及

将所述病历信息对应的类别结果推送至所述终端。

一种非易失性的计算机可读存储介质，所述存储介质中存储有至少一条指令，所述计算机可读存储介质中存储有至少一条计算机可读指令，所述计算机可读指令由处理器加载并执行以下步骤：

接收终端发送的医疗数据分类请求，所述医疗数据分类请求包括病历信息；

获取预设的医疗词库，根据所述医疗词库中的医疗词汇对所述病历信息进行分词处理，得到多个文本向量；

对所述多个文本向量进行特征提取，得到多个文本向量以及对应的特征维度值；

获取目标分类器，通过所述目标分类器的多个神经网络节点对所述多个文本向量以及对应的特征维度值进行遍历计算；所述目标分类器基于对多个医疗数据进行训练得到；

直到遍历至所述多个文本向量对应的目标节点，根据所述目标节点计算所述多个文本向

量对应的类别概率，根据所述类别概率得到所述病历信息对应的类别结果；及
将所述病历信息对应的类别结果推送至所述终端。

本申请的一个或多个实施例的细节在下面的附图和描述中提出。本申请的其它特征和优点将从说明书、附图以及权利要求书变得明显。

5

附图说明

下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

10 图 1 为一个实施例中基于机器学习的医疗数据分类方法的应用场景图；

图 2 为一个实施例中基于机器学习的医疗数据分类方法的流程示意图；

图 3 为一个实施例中对病历信息进行分词处理步骤的流程示意图；

图 4 为一个实施例中构建目标分类器步骤的流程示意图；

图 5 为一个实施例中基于机器学习的医疗数据分类装置的结构框图；

15 图 6 为一个实施例中计算机设备的内部结构图。

具体实施方式

为了使本申请的技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本
20 申请。

本申请提供的基于机器学习的医疗数据分类方法，可以应用于如图 1 所示的应用环境中。
终端 102 通过网络与服务器 104 通过网络进行通信。医务人员可以利用对应的终端 102 向服务器 104 发送医疗数据分类请求，医疗数据分类请求中包括了病历信息。服务器 104 接收终端 102 发送的医疗数据分类请求后，对病历信息进行分词处理，得到多个文本向量，服务器
25 104 进而对多个文本向量进行特征提取，得到多个文本向量以及对应的特征维度值。服务器 104 则进一步获取目标分类器，目标分类器基于对多个医疗数据进行训练得到，通过目标分类器的多个神经网络节点对得到多个文本向量以及对应的特征维度值进行分类分析，由此能够有效得到病历信息对应的类别结果，服务器 104 并将病历信息对应的类别结果推送至对应的终端 102。通过对病历信息进行有效地分词和特征提取，并利用预先训练构建的分类器对
30 提取的文本数据进行分类，从而有效地提高了病历信息的分类准确率。终端 102 可以但不限于各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备，服务器 104 可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

在其中一个实施例中，如图 2 所示，提供了一种基于机器学习的医疗数据分类方法，以

该方法应用于图 1 中的服务器为例进行说明，包括以下步骤：

步骤 202，接收终端发送的医疗数据分类请求，医疗数据分类请求包括病历信息。

病历信息可以包括就医人员的身份标识、资本资料、病史记录信息和历史诊断信息等。医务人员在对就医人员进行诊断的时候，可以利用对应的终端获取就医人员的病历信息，病历信息可以包括医务人员输入的信息，也可以包括根据就医人员的身份标识从数据库获取的病历信息。终端获取该就医人员的病历信息后，则根据病历信息向服务器发送医疗数据分类请求，医疗数据分类请求中包括了病历信息和身份标识。

进一步地，服务器还可以根据就医人员的身份标识从第三方数据库获取该就医人员的历史病历信息，例如该就医人员在其他地方就医的病历信息，以有效获取该就医人员对应的完整的病历信息。

步骤 204，获取预设的医疗词库，根据医疗词库中的医疗词汇对病历信息进行分词处理，得到多个文本向量。

服务器对病历信息进行分词处理之前，可以获取大量的医疗数据，并对获取的大量的医疗数据进行语义分析，例如可以通过预设的语义分析模型对大量的医疗数据进行语义分析，得到多个类型的医疗词汇。服务器进而利用分析得到的医疗词汇生成医疗领域内多个类型对应的医疗词库。

服务器接收到终端发送的医疗数据分类请求后，则对病历信息进行分词处理。具体地，服务器获取预设的医疗词库，医疗词库中包括了大量的医疗词汇和对应的向量。服务器则将病历信息中的多个文本数据与医疗词库中多个医疗词汇进行匹配，具体地，服务器可以通过预设的距离算法计算病历信息中文本数据和医疗词汇之间的相似度，进而计算出病历信息中的文本数据与医疗词汇的匹配度。服务器进一步提取出达到预设匹配度的文本数据。服务器则根据匹配后的文本数据对病历信息进行分词，得到分词后的多个文本数据。服务器进一步对分词后的多个文本数据进行向量化，将文本数据转换为对应的量化信息，得到多个文本数据对应的多个文本向量。

步骤 206，对多个文本向量进行特征提取，得到多个文本向量以及对应的特征维度值。

服务器对病历信息对应的文本向量进行分词，得到多个文本向量后，进一步对文本数据进行特征提取。服务器根据预设算法计算分词后的多个文本向量的权重。例如，服务器可以通过 TF-IDF 算法计算多个文本向量的 TF 值和 IDF 值，TF 词频(Term Frequency)，表示文本向量在文档中出现的频率。IDF 逆向文件频率(Inverse Document Frequency)，指词语普遍重要性的度量。并根据多个词的 TF 值和 IDF 值计算多个对应的权重，例如通过计算 TF 值和 IDF 值的乘积可以得到文本向量对应的权重，服务器进而根据文本向量的权重对文本向量进行特征提取，则提取出达到预设阈值的文本向量。

服务器提取出达到预设阈值的文本向量后，则根据预设算法和文本向量的权重计算出多

个文本向量的特征维度值，特征维度值可以表示文本向量所属的特征维度。通过计算出文本向量的权重，由此根据权重对文本向量进行过滤，从而能够有效地对文本向量进行特征提取，并得到文本向量对应的特征维度值。

步骤 208，获取目标分类器，通过目标分类器的多个神经网络节点对多个文本向量以及
5 对应的特征维度值进行遍历计算；目标分类器基于对多个医疗数据进行训练得到。

步骤 210，直到遍历至多个文本向量对应的目标节点，根据目标节点计算多个文本向量对应的类别概率，根据类别概率得到病历信息对应的类别结果。

服务器在获取目标分类器之前，还可以预先构建和训练得到目标分类器。具体地，服务器可以预先从本地数据库或第三方数据库中获取大量的医疗数据，根据多个医疗数据生成对
10 应的训练集数据和验证集数据。服务器对医疗数据对应的多个字段数据进行向量化，得到多个文本数据对应的特征向量，并将特征向量转换为对应的特征变量。服务器进而采用预设的聚类算法对训练集数据对应的特征变量进行聚类分析，提取出达到预设阈值的特征变量。服务器则获取预设的神经网络模型，通过神经网络模型对训练集数据进行训练，得到多个特征变量对应的特征维度值和权重，根据多个特征变量对应的特征维度值和权重构建初始分类器。
15 利用验证集数据对分类器进行进一步训练和验证，直到验证集数据中满足预设阈值的数量达到预设比值时，则停止训练，得到所需的目标分类器。

服务器对文本数据进行特征提取，得到多个文本数据对应的多维度向量后，则获取已经训练好的目标分类器，将多个文本向量以及对应的维度特征值输入至目标分类器中，其中，目标分类器中包括多个预设的神经网络层节点和对应的节点权重。通过目标分类器中的多个
20 节点预设损失函数对多个文本向量以及对应的维度特征值进行遍历计算，直到得到多个文本词向量对应的目标节点，根据目标节点计算多个文本向量对应的类别概率，根据类别概率得到文本向量对应的类别结果，进而得到病历信息对应的类别结果。

步骤 212，将病历信息对应的类别结果推送至终端。

服务器通过目标分类器对病历信息进行分类，得到病历信息对应的类别结果后，则将病历
25 信息对应的类别结果推送至对应的终端。通过对病历信息进行有效地分词和特征提取，并利用预先训练构建的目标分类器对提取的文本信息进行分类，能够有效地提高病历信息的分类准确率，由此能够有利于医务人员根据推送的病历信息对应的类别结果进行有效地诊断，进而有效提高了医务人员的诊断效率。

例如，病历信息中包括了就医人员对应的历史病历记录信息，包括多个历史症状描述、
30 历史处方信息、历史诊断信息等数据。通过对病历信息进行多次筛选和文本提取后，利用预先训练的目标分类器对提取出的文本进行分类分析，当对该就医人员的病历信息中所有数据进行分类分析后，得到该病历信息所对应的类别结果，例如当就医人员患病为癌症时，则可以分类得到具体的癌症类别。

上述基于机器学习的医疗数据分类方法中，服务器接收终端发送的医疗数据分类请求后，对医疗数据分类请求中携带的病历信息进行分词处理，由此能够有效地根据医疗领域进行分词得到多个文本向量，服务器进而对多个文本向量进行特征提取，能够有效地提取出得到多个文本向量以及对应的特征维度值。服务器则进一步获取目标分类器，目标分类器基于对多个医疗数据进行训练得到，通过目标分类器的多个神经网络节点对得到多个文本向量以及对应的特征维度值进行遍历计算，直到遍历至多个文本向量对应的目标节点，根据目标节点计算多个文本向量对应的类别概率，根据类别概率得到病历信息对应的类别结果，由此能够有效得到病历信息对应的类别结果，通过预先训练构建的分类器对提取的文本数据进行分类，从而有效地提高了病历信息的分类准确率。服务器则将病历信息对应的类别结果推送至对应的终端。由此能够有利于医务人员根据推送的病历信息对应的类别结果进行有效地决策，通过对病历信息进行准确地分类，进而能够有效提高医疗数据的处理效率。

在其中一个实施例中，如图3所示，病历信息中包括了多个文本数据，对病历信息进行分词处理的步骤，具体包括以下内容：

步骤302，获取预设的医疗词库，医疗词库中包括多个医疗词汇；将病历信息中的多个文本数据与医疗词库进行匹配，计算病历信息中的文本数据与多个医疗词汇的匹配度，提取达到预设匹配度的文本数据。

步骤304，根据匹配后的文本数据对病历信息进行分词，得到分词后的多个文本数据。

步骤306，对分词后的多个文本数据进行向量转换，得到对应的多个文本向量。

服务器对医疗数据进行处理之前，可以预先建立医疗词库。具体地，服务器可以获取大量的医疗数据，并对获取的大量的医疗数据进行语义分析，例如可以通过预设的语义分析模型对大量的医疗数据进行语义分析，得到多个类型的医疗词汇。服务器进而利用分析得到的医疗词汇生成医疗领域内多个类型对应的医疗词库。

医务人员可以利用对应的终端向服务器发送医疗数据分类请求，医疗数据分类请求中包括了病历信息。服务器接收终端发送的医疗数据分类请求后，对医疗数据分类请求中的病历信息进行分词处理。具体地，服务器获取预设的医疗词库，医疗词库中包括了大量的医疗词汇和对应的向量。服务器则将病历信息中的多个文本数据与医疗词库中多个医疗词汇进行匹配，具体地，服务器可以通过预设的距离算法计算病历信息中文本数据和医疗词汇之间的相似度，进而计算出病历信息中的文本数据与医疗词汇的匹配度。服务器进一步提取出达到预设匹配度的文本数据。服务器则根据匹配后的文本数据对病历信息进行分词，得到分词后的多个文本数据。

服务器进一步对分词后的多个文本数据进行向量化，将文本数据转换为对应的量化信息，得到多个文本数据对应的多个文本向量。例如，可以通过 Doc2Vec 与 Word2Vec 算法对分词后的多个文本数据进行词向量化和段落向量化，进而得到对应的文本向量。其中，文本

向量可以包括字向量、词向量和句向量等。

服务器得到多个文本数据对应的文本向量后，根据预设算法计算出文本向量的特征维度值，并对多个文本向量进行特征提取，得到多个文本向量和对应的特征维度值。服务器则进一步获取预设的分类器，通过分类器对多个文本向量以及对应的特征维度值进行分类分析，由此能够有效得到病历信息对应的类别结果，服务器并将病历信息对应的类别结果推送至对应的终端。通过对病历信息进行有效地分词和特征提取，并利用预先训练构建的分类器对提取的文本信息进行分类，能够有效地提高病历信息的分类准确率，由此能够有利于医务人员根据推送的病历信息对应的类别结果进行有效地诊断。

在其中一个实施例中，对多个文本数据进行特征提取，得到多个文本向量对应的多维度向量的步骤包括：计算多个文本向量的词频和逆向文件频率；根据词频和逆向文件频率按照预设算法计算多个文本向量的权重；提取出权重达到预设阈值的文本向量；根据预设算法和权重计算文本向量对应的特征维度值。

医务人员可以利用对应的终端向服务器发送医疗数据分类请求，医疗数据分类请求中包括了病历信息。服务器接收终端发送的医疗数据分类请求后，对医疗数据分类请求中的病历信息进行分词处理，得到多个文本向量。

服务器得到病历信息对应的多个文本向量后，根据预设算法计算分词后的多个文本向量的权重。例如，服务器可以通过 TF-IDF 算法计算多个文本向量的 TF 值和 IDF 值，TF 词频(Term Frequency)，表示文本向量出现的频率。IDF 逆向文件频率(Inverse Document Frequency)，可以表示词语普遍重要性的度量。并根据多个词的 TF 值和 IDF 值计算多个对应的权重，例如通过计算 TF 值和 IDF 值的乘积可以得到文本数据对应的权重。

例如，可以用如下公式计算多个文本向量的 TF 值：

$$tf_{i,j} = \frac{t_{i,j}}{\sum_i n_i}$$

计算文本向量的 IDF 值的公式可以如下：

$$idf_i = \lg \frac{1}{\frac{1}{n_i}}$$

计算文本向量权重的公式可以如下：

$$tfidf_{i,j} = tf_{i,j} \times idf_{i,j}$$

如果包含文本向量 t 的文档越少，也就是 n 越小，IDF 越大，则说明文本向量 t 具有很好的类别区分能力。如果某一类文档 C 中包含词条 t 的文档数为 m，而其它类包含 t 的文档总数为 k，显然所有包含 t 的文档数 n=m+k，当 m 大的时候，n 也大，按照 IDF 公式得到的 IDF 的值会小，就说明该词条 t 类别区分能力不强。如果一个词条在一个类的文档中频繁出现，则说明该词条能够很好代表这个类的文本的特征，则该词条具有较高的权重。通过计算 TF 与 IDF 的乘积，进而计算出文本向量的权重，服务器进而根据文本向量的权重对文本向量进行特征提取，则提取出达到预设阈值的文本向量。

服务器提取出达到预设阈值的文本向量后，则根据预设算法和文本向量的权重计算出多个文本向量的特征维度值，特征维度值可以表示文本向量所属的特征维度。文本向量可以包括多个特征维度，服务器计算出文本向量的权重后，则可以利用权重对文本向量的特征维度的重要程度进行计算，进而得到文本向量对应的特征维度值。通过计算出文本向量的权重，
5 由此根据权重对文本向量进行过滤，从而能够有效地对文本向量进行特征提取，并得到文本向量对应的特征维度值。

在其中一个实施例中，如图 4 所示，获取目标分类器之前，还包括构建目标分类器的步骤，该步骤具体包括以下内容：

步骤 402，获取多个医疗数据，根据多个医疗数据生成对应的训练集数据和验证集数据。

10 服务器在获取目标分类器之前，还需要构建和训练出目标分类器。具体地，服务器可以预先从本地数据库或第三方数据库中获取大量的医疗数据，医疗数据可以包括医疗诊断信息、临床数据以及调研数据等。服务器则将大量的医疗数据生成训练集数据和验证集数据，其中，训练集数据可以是经过人工标注后的数据。

步骤 404，对训练集数据中的多个医疗数据进行聚类分析，得到聚类结果。

15 步骤 406，对聚类结果进行特征提取，提取出多个特征变量。

步骤 408，获取预设的神经网络模型，通过神经网络模型对训练集数据进行训练，得到多个特征变量对应的特征维度值和权重，根据多个特征变量对应的特征维度值和权重构建初始分类器。

20 步骤 410，利用验证集数据对分类器进行进一步训练和验证，直到验证集数据中满足预设阈值的数量达到预设比值时，则停止训练，得到所需的目标分类器。

服务器首先对训练集数据中的医疗数据进行数据清洗和数据预处理，具体地，服务器对医疗数据对应的多个字段数据进行向量化，得到多个文本数据对应的特征向量，并将特征向量转换为对应的特征变量。服务器进一步对特征变量进行衍生处理，得到处理后的多个特征变量。如对特征变量进行缺失值填充、异常值抽取更替等。

25 服务器进而采用预设的聚类算法对训练集数据对应的特征变量进行聚类分析。例如，预设的聚类算法可以为 k-means (k-均值算法) 聚类的方法。服务器通过对特征变量进行多次聚类后得到多个聚类结果。服务器并根据预设算法计算多个特征变量之间的相似度，提取出相似度达到预设阈值的特征变量。

例如，服务器可以对多个聚类结果内的特征变量分别进行组合，得到多个组合特征变量。
30 获取目标变量，利用目标变量对多个组合特征变量进行相关性检验。检验通过时，对组合特征变量添加交互标签。利用添加交互标签后的组合特征变量解析对应的特征变量。添加交互标签后的组合特征变量可以为达到预设阈值的特征变量，服务器则提取出达到预设阈值特征变量。通过对特征变量进行特征处理和特征提取，能够有效地提取出有价值的特征变量。

服务器则获取预设的机器学习模型，例如可以是基于决策树的 Xgboot 机器学习模型。例如，机器学习模型中包括多个神经网络模型，神经网络模型可以包括预设的输入层、多个 LSTM 层、dropout 层和输出层。神经网络模型中包括多个网络节点，其中每一层网络节点的舍弃率可以为 0.2。神经网络模型的 LSTM 层包括激活函数以及损失函数，通过 LSTM 层输出的完全连接人工神经网络也包括对应的激活函数。神经网络模型中还包括确定误差的计算方式，例如可以采用均方误差算法；还包括确定权重参数的迭代更新方式，例如可以采用 RMSprop 算法。神经网络模型中还可以包括一层普通的神经网络层，用于输出结果的降维。

服务器获取预设的神经网络模型后，进一步将训练集数据中的医疗数据输入至神经网络模型中进行学习和训练。服务器通过对训练集中的大量医疗数据进行训练后，可以得到多个特征变量对应的特征维度值和权重，进而根据多个特征变量对应的特征维度值和权重构建初始分类器。

服务器得到初始分类器后，则获取验证集数据，通过验证集数据中的大量医疗数据对构建的初始分类器进行训练和验证。直到验证集数据中满足预设阈值的数量达到预设比值时，则停止训练，进而得到训练完成的目标分类器。通过对大量的医疗数据进行训练和学习，进而能够有效构建出预测准确率较高的分类器，从而有效提高了医疗数据的分类准确率。

在其中一个实施例中，文本中包括多个文本句，多个文本句组成文本块，通过分类器的多个神经网络节点对多个文本向量以及对应的特征维度值进行遍历计算多个文本向量对应的类别的步骤包括：利用目标分类器根据特征维度值计算多个文本向量之间的相关性，根据相关性计算文本中成句的文本句，并计算文本句的句向量；提取句向量的特征，根据多个句向量的特征计算出文本块向量；计算文本块向量对应每个类别的概率，提取达到预设概率值的类别，并对文本块添加对应的类别标签。

医务人员可以利用对应的终端向服务器发送医疗数据分类请求，医疗数据分类请求中包括了病历信息。服务器接收终端发送的医疗数据分类请求后，对医疗数据分类请求中的病历信息进行分词处理，得到多个文本数据对应的文本向量。服务器进一步对文本向量进行特征提取，得到多个文本向量以及对应的特征维度值。

服务器提取出多个文本向量以及对应的特征维度值后，则获取目标分类器，并将多个文本向量以及对应的特征维度值作为目标分类器的输入。其中，目标分类器中包括多个预设的神经网络层节点和对应的节点权重，通过目标分类器中的多个神经网络层节点对多个文本向量以及对应的特征维度值进行遍历计算。具体地，文本中可以包括多个词语和短句，即文本句。文本向量可以包括词向量和短语向量。服务器可以首先根据文本向量和对应的维度特征值计算出文本中多个文本向量之间的相关性，进而根据相关性计算文本中成句的文本句，并计算出文本句对应的句向量。服务器则提取句向量的特征，并根据多个句向量的特征计算出文本块向量。其中，文本块包括多个文本句，文本块向量可以由多个句向量组成。服务器根

据多个神经网络层节点中的预设损失函数计算文本块向量属于每个类别的概率，并根据类别概率将多个文本块向量输入至下一个神经网络层节点进行计算，直到得到多个文本块向量对应的目标节点，进而根据目标节点计算得到多个文本块向量对应的类别概率，获取出类别概率最高的类别结果，由此得到多个文本块向量所属的类别结果。通过利用大量数据训练得到的目标分类器对病历信息中的文本向量进行分类，从而能够有效及准确地得到病历信息所属的类别，由此能够有效地提高病历信息的分类准确率。

在其中一个实施例中，该方法还包括：根据预设频率从预设数据库中获取多个历史医疗数据；对多个历史医疗数据进行聚类分析，得到分析结果；根据分析结果进行特征选择，得到多个特征变量；根据预设的算法计算多个特征变量的权重；根据多个特征变量和对应的权重对分类器进行优化调整。

服务器训练得到目标分类器后，还可以根据预设的频率对分类器进行调参优化。具体地，服务器可以根据预设频率从本地数据库或第三方数据库中获取大量的历史医疗数据，例如预设频率可以为一个月、三个月、六个月等，服务器则可以获取过去一个月、三个月、六个月内的历史医疗数据，历史医疗数据可以包括医疗诊断信息、临床数据以及调研数据等。

服务器首先获取的大量历史医疗数据进行数据清洗和数据预处理，具体地，服务器对历史医疗数据对应的多个字段数据进行向量化，得到多个字段数据对应的特征变量，并对特征变量进行衍生处理，得到处理后的多个特征变量。如对特征变量进行缺失值填充、异常值抽取更替等。

服务器进而采用预设的聚类算法对训练集数据对应的特征变量进行聚类分析。例如，预设的聚类算法可以为 k-means (k-均值算法) 聚类的方法。服务器通过对特征变量进行多次聚类后得到多个聚类结果。服务器并根据预设算法计算多个特征变量之间的相似度，提取出相似度达到预设阈值的特征变量。

例如，服务器可以对多个聚类结果内的特征变量分别进行组合，得到多个组合特征变量。获取目标变量，利用目标变量对多个组合特征变量进行相关性检验。检验通过时，对组合特征变量添加交互标签。利用添加交互标签后的组合特征变量解析对应的特征变量。添加交互标签后的组合特征变量可以为达到预设阈值的特征变量，服务器则提取出达到预设阈值特征变量。通过对特征变量进行特征处理和特征提取，能够有效地提取出有价值的特征变量。

服务器进一步根据预设的算法计算多个特征变量的权重，进而根据多个特征变量和对应的权重对目标分类器进行优化调整。具体地，服务器可以根据多个特征变量和对应的权重对目标分类器中的参数进行调整，由此能够有效地对目标分类器进行调参优化。

应该理解的是，虽然图 2-4 的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，图 2-4 中的至少一部分步

骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

5 在其中一个实施例中，如图 5 所示，提供了一种基于机器学习的医疗数据分类装置，包括：请求接收模块 502、分词处理模块 504、特征提取模块 506、数据分类模块 508 和数据推送模块 510，其中：

请求接收模块 502，用于接收终端发送的医疗数据分类请求，医疗数据分类请求包括病历信息；

10 分词处理模块 504，用于获取预设的医疗词库，根据医疗词库中的医疗词汇对病历信息进行分词处理，得到多个文本向量；

特征提取模块 506，用于对多个文本向量进行特征提取，得到多个文本向量以及对应的特征维度值；

15 数据分类模块 508，用于获取目标分类器，通过目标分类器的多个神经网络节点对多个文本向量以及对应的特征维度值进行遍历计算；目标分类器基于对多个医疗数据进行训练得到；直到遍历至多个文本向量对应的目标节点，根据目标节点计算多个文本向量对应的类别概率，根据类别概率得到病历信息对应的类别结果；

数据推送模块 510，用于将病历信息对应的类别结果推送至终端。

20 在其中一个实施例中，病历信息中包括多个文本数据，分词处理模块 504 还用于获取预设的医疗词库，医疗词库中包括多个医疗词汇；将病历信息中的多个文本数据与医疗词库进行匹配，计算病历信息中的文本数据与多个医疗词汇的匹配度，提取达到预设匹配度的文本数据；根据匹配后的文本数据对病历信息进行分词，得到分词后的多个文本数据；对分词后的多个文本数据进行向量化，得到多个文本向量。

25 在其中一个实施例中，特征提取模块 506 还用于计算多个文本向量的词频和逆向文件频率；根据词频和逆向文件频率按照预设算法计算多个文本向量的权重；提取出权重达到预设阈值的文本向量；根据预设算法和权重计算文本向量对应的特征维度值。

30 在其中一个实施例中，该装置还包括目标分类器构建模块，用于获取多个医疗数据，根据多个医疗数据生成对应的训练集数据和验证集数据；对训练集数据中的多个医疗数据进行聚类分析，得到聚类结果；对聚类结果进行特征提取，提取出多个特征变量；获取预设的神经网络模型，通过神经网络模型对训练集数据进行训练，得到多个特征变量对应的特征维度值和权重，根据多个特征变量对应的特征维度值和权重构建初始分类器；利用验证集数据对分类器进行进一步训练和验证，直到验证集数据中满足预设阈值的数量达到预设比值时，则停止训练，得到所需的目标分类器。

在其中一个实施例中，文本中包括多个文本句，多个文本句组成文本块，数据分类模块 508 还用于利用目标分类器根据特征维度值计算多个文本向量之间的相关性，根据相关性计算文本中成句的文本句，并计算文本句的句向量；提取句向量的特征，根据多个句向量的特征计算出文本块向量；计算文本块向量对应每个类别的概率，提取达到预设概率值的类别，
5 并对文本块添加对应的类别标签。

在其中一个实施例中，该装置还包括目标分类器优化模块，用于根据预设频率从预设数据库中获取多个历史医疗数据；对多个历史医疗数据进行聚类分析，得到分析结果；根据分析结果进行特征选择，得到多个特征变量；根据预设的算法计算多个特征变量的权重；根据多个特征变量和对应的权重对目标分类器进行优化调整。

关于基于机器学习的医疗数据分类装置的具体限定可以参见上文中对于基于机器学习的医疗数据分类方法的限定，在此不再赘述。上述基于机器学习的医疗数据分类装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处
10 理器调用执行以上各个模块对应的操作。

在一个实施例中，提供了一种计算机设备，该计算机设备可以是服务器，其内部结构图可以如图 6 所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机可读指令和数据库。该内存储器为易失性存储介质中的操作系统和计算机可读指令的运行提供环境。该计算机设备的数据库用于存储医疗数据、病历信息等数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机可读指令被处理器执行时以实现本申请任意一个实施例中提供的基于机器学习的医疗数据分类方法的步骤。

本领域技术人员可以理解，图 6 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机可读指令来指令相关的硬件来完成，所述的计算机可读指令可存储于一非易失性计算机可读存储介质中，该计算机可读指令在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器（ROM）、可编程 ROM（PROM）、电可编程 ROM（EPROM）、电可擦除可编程 ROM（EEPROM）或闪存。

易失性存储器可包括随机存取存储器 (RAM) 或者外部高速缓冲存储器。作为说明而非局限, RAM 以多种形式可得, 诸如静态 RAM (SRAM)、动态 RAM (DRAM)、同步 DRAM (SDRAM)、双数据率 SDRAM (DDRSDRAM)、增强型 SDRAM (ESDRAM)、同步链路 (Synchlink) DRAM (SLDRAM)、存储器总线 (Rambus) 直接 RAM (RDRAM)、直接存储器总线动态 RAM (DRDRAM)、以及存储器总线动态 RAM (RDRAM) 等。

以上实施例的各技术特征可以进行任意的组合, 为使描述简洁, 未对上述实施例中的各个技术特征所有可能的组合都进行描述, 然而, 只要这些技术特征的组合不存在矛盾, 都应当认为是本说明书记载的范围。

以上所述实施例仅表达了本申请的几种实施方式, 其描述较为具体和详细, 但并不能因此而理解为对发明专利范围的限制。应当指出的是, 对于本领域的普通技术人员来说, 在不脱离本申请构思的前提下, 还可以做出若干变形和改进, 这些都属于本申请的保护范围。因此, 本申请专利的保护范围应以所附权利要求为准。

权利要求书

1、一种基于机器学习的医疗数据分类方法，所述方法包括：

接收终端发送的医疗数据分类请求，所述医疗数据分类请求包括病历信息；

获取预设的医疗词库，根据所述医疗词库中的医疗词汇对所述病历信息进行分词处理，

5 得到多个文本向量；

对所述多个文本向量进行特征提取，得到多个文本向量以及对应的特征维度值；

获取目标分类器，通过所述目标分类器的多个神经网络节点对所述多个文本向量以及对应的特征维度值进行遍历计算；所述目标分类器基于对多个医疗数据进行训练得到；

直到遍历至所述多个文本向量对应的目标节点，根据所述目标节点计算所述多个文本向
10 量对应的类别概率，根据所述类别概率得到所述病历信息对应的类别结果；及
将所述病历信息对应的类别结果推送至所述终端。

2、根据权利要求1所述的方法，其特征在于，所述病历信息中包括多个文本数据，所述对所述病历信息进行分词处理的步骤包括：

获取预设的医疗词库，所述医疗词库中包括多个医疗词汇；将所述病历信息中的多个文
15 本数据与所述医疗词库进行匹配，计算所述病历信息中的文本数据与多个医疗词汇的匹配度，
提取达到预设匹配度的文本数据；

根据匹配后的文本数据对所述病历信息进行分词，得到分词后的多个文本数据；及
对所述分词后的多个文本数据进行向量转换，得到多个文本向量。

3、根据权利要求1所述的方法，其特征在于，所述对所述多个文本向量进行特征提取，
20 得到多个文本向量以及对应的特征维度值的步骤包括：

计算所述多个文本向量的词频和逆向文件频率；

根据所述词频和所述逆向文件频率按照预设算法计算多个文本向量的权重；

提取出所述权重达到预设阈值的文本向量；及

根据预设算法和所述权重计算所述文本向量对应的特征维度值。

4、根据权利要求1所述的方法，其特征在于，构建所述目标分类器的步骤包括：

获取多个医疗数据，根据所述多个医疗数据生成对应的训练集数据和验证集数据；

对所述训练集数据中的多个医疗数据进行聚类分析，得到聚类结果；

对所述聚类结果进行特征提取，提取出多个特征变量；

获取预设的神经网络模型，通过所述神经网络模型对所述训练集数据进行训练，得到多
30 个特征变量对应的特征维度值和权重，根据多个特征变量对应的特征维度值和权重构建初始
分类器；及

利用所述验证集数据对所述分类器进行进一步训练和验证，直到所述验证集数据中满足
预设阈值的数量达到预设比值时，则停止训练，得到所需的目标分类器。

5、根据权利要求1至4任意一项所述的方法，其特征在于，所述文本中包括多个文本句，所述多个文本组成文本块，所述通过所述目标分类器的多个神经网络节点对所述多个文本向量以及对应的特征维度值进行遍历计算多个文本向量对应的类别的步骤包括：

利用所述目标分类器根据所述特征维度值计算所述多个文本向量之间的相关性，根据所述相关性计算所述文本中成句的文本句，并计算所述文本句的句向量；

提取所述句向量的特征，根据所述多个句向量的特征计算出文本块向量；及

计算所述文本块向量对应每个类别的概率，提取达到预设概率值的类别，并对所述文本块添加对应的类别标签。

6、根据权利要求1所述的方法，其特征在于，所述方法还包括：

根据预设频率从预设数据库中获取多个历史医疗数据；

对多个历史医疗数据进行聚类分析，得到分析结果；

根据所述分析结果进行特征选择，得到多个特征变量；

根据预设的算法计算多个特征变量的权重；及

根据多个特征变量和对应的权重对所述目标分类器进行优化调整。

7、一种基于机器学习的医疗数据分类装置，所述装置包括：

请求接收模块，用于接收终端发送的医疗数据分类请求，所述医疗数据分类请求包括病历信息；

分词处理模块，用于获取预设的医疗词库，根据所述医疗词库中的医疗词汇对所述病历信息进行分词处理，得到多个文本向量；

特征提取模块，用于对所述多个文本向量进行特征提取，得到多个文本向量以及对应的特征维度值；

数据分类模块，用于获取目标分类器，通过所述目标分类器的多个神经网络节点对所述多个文本向量以及对应的特征维度值进行遍历计算；所述目标分类器基于对多个医疗数据进行训练得到；直到遍历至所述多个文本向量对应的目标节点，根据所述目标节点计算所述多个文本向量对应的类别概率，根据所述类别概率得到所述病历信息对应的类别结果；及

数据推送模块，用于将所述病历信息对应的类别结果推送至所述终端。

8、根据权利要求7所述的装置，其特征在于，所述分词处理模块还用于获取预设的医疗词库，所述医疗词库中包括多个医疗词汇；将所述病历信息中的多个文本数据与所述医疗词库进行匹配，计算所述病历信息中的文本数据与多个医疗词汇的匹配度，提取达到预设匹配度的文本数据；根据匹配后的文本数据对所述病历信息进行分词，得到分词后的多个文本数据；及对所述分词后的多个文本数据进行向量化，得到多个文本向量。

9、根据权利要求7所述的装置，其特征在于，所述特征提取模块还用于计算所述多个文

本向量的词频和逆向文件频率；根据所述词频和所述逆向文件频率按照预设算法计算多个文本向量的权重；提取出所述权重达到预设阈值的文本向量；及根据预设算法和所述权重计算所述文本向量对应的特征维度值。

10、根据权利要求 7 所述的装置，其特征在于，所述装置还包括分类器构建模块，用于
5 获取多个医疗数据，根据所述多个医疗数据生成对应的训练集数据和验证集数据；对所述训练集数据中的多个医疗数据进行聚类分析，得到聚类结果；对所述聚类结果进行特征提取，提取出多个特征变量；获取预设的神经网络模型，通过所述神经网络模型对所述训练集数据进行训练，得到多个特征变量对应的特征维度值和权重，根据多个特征变量对应的特征维度值和权重构建初始分类器；及利用所述验证集数据对所述分类器进行进一步训练和验证，直
10 到所述验证集数据中满足预设阈值的数量达到预设比值时，则停止训练，得到所需的目标分类器。

11、根据权利要求 7 所述的装置，其特征在于，所述文本中包括多个文本句，所述多个文本句组成文本块，所述数据分类模块还用于利用所述目标分类器根据所述特征维度值计算所述多个文本向量之间的相关性，根据所述相关性计算所述文本中成句的文本句，并计算所
15 述文本句的句向量；提取所述句向量的特征，根据所述多个句向量的特征计算出文本块向量；及计算所述文本块向量对应每个类别的概率，提取达到预设概率值的类别，并对所述文本块添加对应的类别标签。

12、根据权利要求 7 所述的装置，其特征在于，所述装置还包括模型优化模块，用于根据预设频率从预设数据库中获取多个历史医疗数据；对多个历史医疗数据进行聚类分析，得
20 到分析结果；根据所述分析结果进行特征选择，得到多个特征变量；根据预设的算法计算多个特征变量的权重；及根据多个特征变量和对应的权重对所述目标分类器进行优化调整。

13、一种计算机设备，包括存储器和处理器，所述存储器存储有至少一条计算机可读指令，所述计算机可读指令由所述处理器加载并执行以下步骤：

接收终端发送的医疗数据分类请求，所述医疗数据分类请求包括病历信息；

25 获取预设的医疗词库，根据所述医疗词库中的医疗词汇对所述病历信息进行分词处理，得到多个文本向量；

对所述多个文本向量进行特征提取，得到多个文本向量以及对应的特征维度值；

获取目标分类器，通过所述目标分类器的多个神经网络节点对所述多个文本向量以及对应的特征维度值进行遍历计算；所述目标分类器基于对多个医疗数据进行训练得到；

30 直到遍历至所述多个文本向量对应的目标节点，根据所述目标节点计算所述多个文本向量对应的类别概率，根据所述类别概率得到所述病历信息对应的类别结果；及
将所述病历信息对应的类别结果推送至所述终端。

14、根据权利要求 13 所述的计算机设备，其特征在于，所述病历信息中包括多个文本数

据,所述处理器执行计算机可读指令时还执行以下步骤:获取预设的医疗词库,所述医疗词库中包括多个医疗词汇;将所述病历信息中的多个文本数据与所述医疗词库进行匹配,计算所述病历信息中的文本数据与多个医疗词汇的匹配度,提取达到预设匹配度的文本数据;根据匹配后的文本数据对所述病历信息进行分词,得到分词后的多个文本数据;及对所述分词后的多个文本数据进行向量转换,得到多个文本向量。

15、根据权利要求 13 所述的计算机设备,其特征在于,所述处理器执行计算机可读指令时还执行以下步骤:计算所述多个文本向量的词频和逆向文件频率;根据所述词频和所述逆向文件频率按照预设算法计算多个文本向量的权重;提取出所述权重达到预设阈值的文本向量;及根据预设算法和所述权重计算所述文本向量对应的特征维度值。

16、根据权利要求 13 所述的计算机设备,其特征在于,所述文本中包括多个文本句,所述多个文本组成文本块,所述处理器执行计算机可读指令时还执行以下步骤:利用所述目标分类器根据所述特征维度值计算所述多个文本向量之间的相关性,根据所述相关性计算所述文本中成句的文本句,并计算所述文本句的句向量;提取所述句向量的特征,根据所述多个句向量的特征计算出文本块向量;及计算所述文本块向量对应每个类别的概率,提取达到预设概率值的类别,并对所述文本块添加对应的类别标签。

17、一种非易失性的计算机可读存储介质,所述计算机可读存储介质中存储有至少一条计算机可读指令,所述计算机可读指令由处理器加载并执行以下步骤:

接收终端发送的医疗数据分类请求,所述医疗数据分类请求包括病历信息;

获取预设的医疗词库,根据所述医疗词库中的医疗词汇对所述病历信息进行分词处理,得到多个文本向量;

对所述多个文本向量进行特征提取,得到多个文本向量以及对应的特征维度值;

获取目标分类器,通过所述目标分类器的多个神经网络节点对所述多个文本向量以及对应的特征维度值进行遍历计算;所述目标分类器基于对多个医疗数据进行训练得到;

直到遍历至所述多个文本向量对应的目标节点,根据所述目标节点计算所述多个文本向量对应的类别概率,根据所述类别概率得到所述病历信息对应的类别结果;及

将所述病历信息对应的类别结果推送至所述终端。

18、根据权利要求 17 所述的存储介质,其特征在于,所述病历信息中包括多个文本数据,所述计算机可读指令被所述处理器执行时还执行以下步骤:获取预设的医疗词库,所述医疗词库中包括多个医疗词汇;将所述病历信息中的多个文本数据与所述医疗词库进行匹配,计算所述病历信息中的文本数据与多个医疗词汇的匹配度,提取达到预设匹配度的文本数据;根据匹配后的文本数据对所述病历信息进行分词,得到分词后的多个文本数据;及对所述分词后的多个文本数据进行向量转换,得到多个文本向量。

19、根据权利要求 17 所述的存储介质,其特征在于,所述计算机可读指令被所述处理器

执行时还执行以下步骤：计算所述多个文本向量的词频和逆向文件频率；根据所述词频和所述逆向文件频率按照预设算法计算多个文本向量的权重；提取出所述权重达到预设阈值的文本向量；及根据预设算法和所述权重计算所述文本向量对应的特征维度值。

- 20、根据权利要求 17 所述的存储介质，其特征在于，所述文本中包括多个文本句，所述多个文本组成文本块，所述计算机可读指令被所述处理器执行时还执行以下步骤：利用所述目标分类器根据所述特征维度值计算所述多个文本向量之间的相关性，根据所述相关性计算所述文本中成句的文本句，并计算所述文本句的句向量；提取所述句向量的特征，根据所述多个句向量的特征计算出文本块向量；及计算所述文本块向量对应每个类别的概率，提取达到预设概率值的类别，并对所述文本块添加对应的类别标签。
- 5

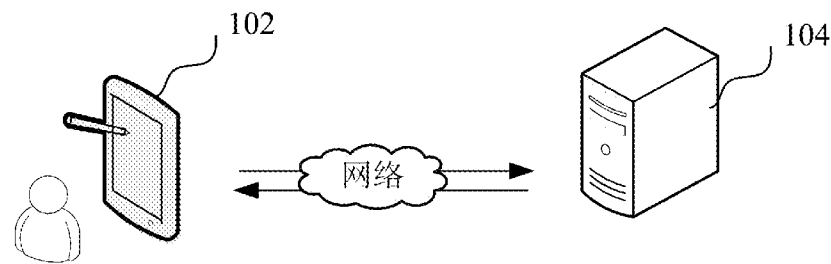


图 1

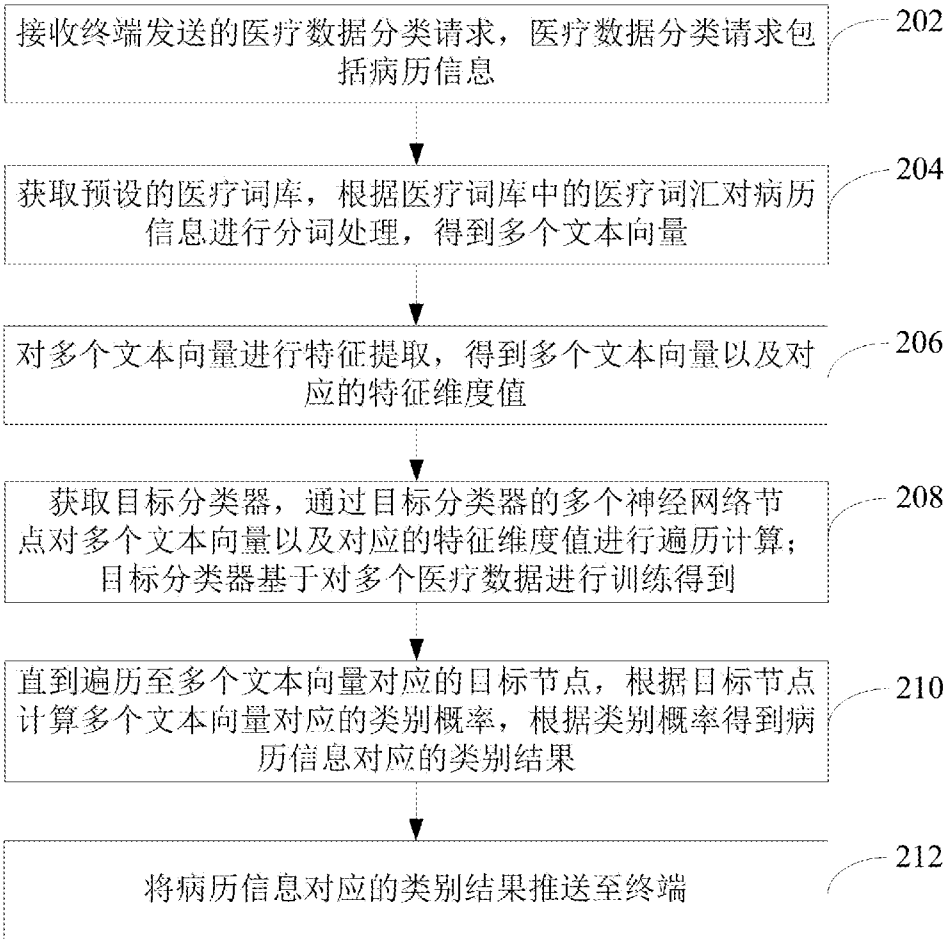


图 2

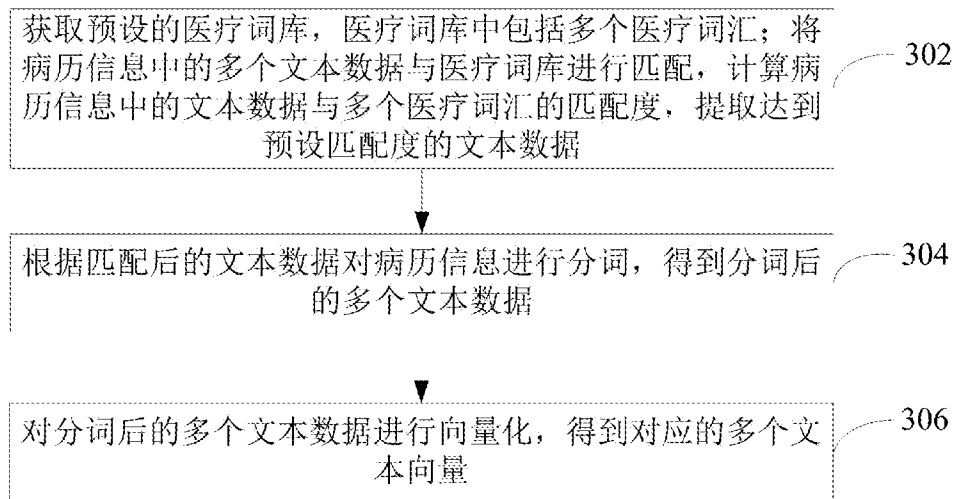


图 3

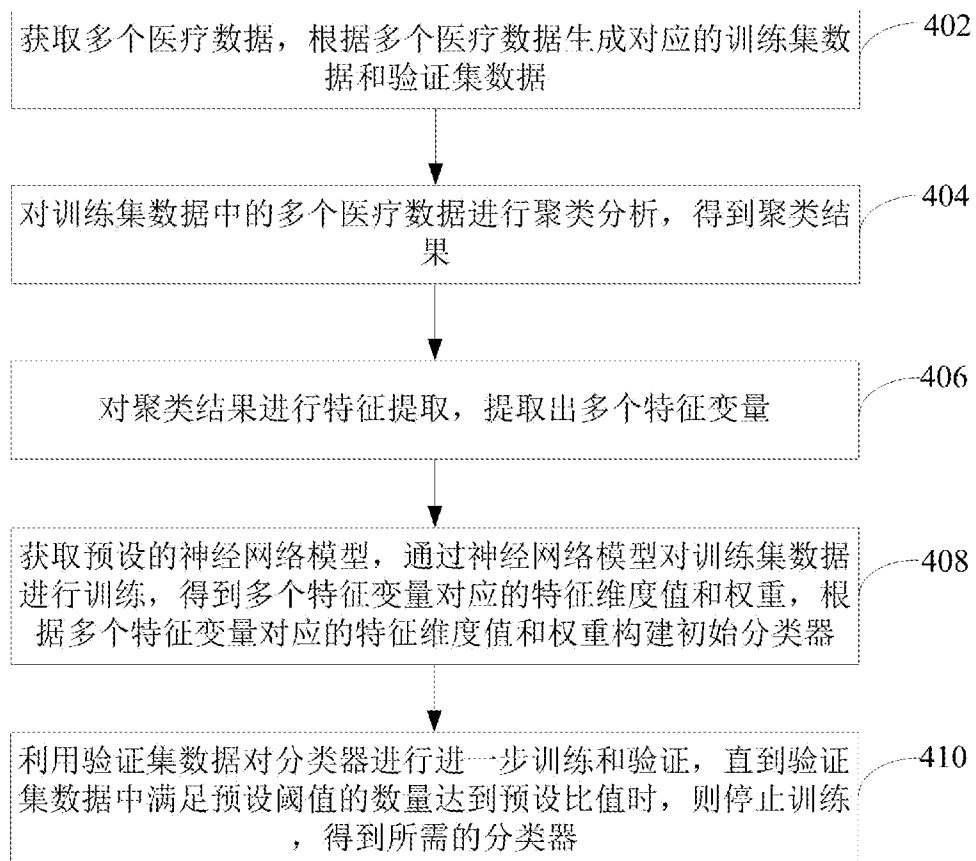


图 4

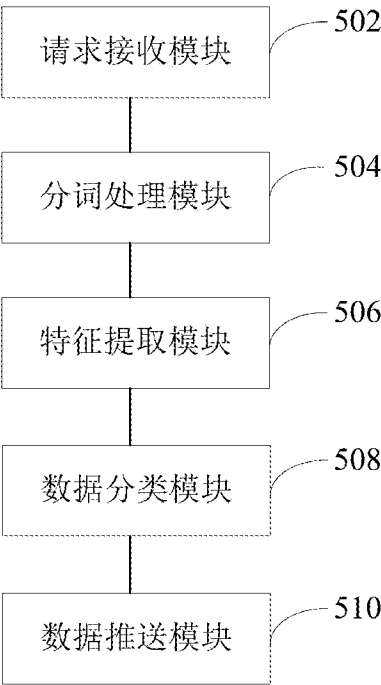


图 5

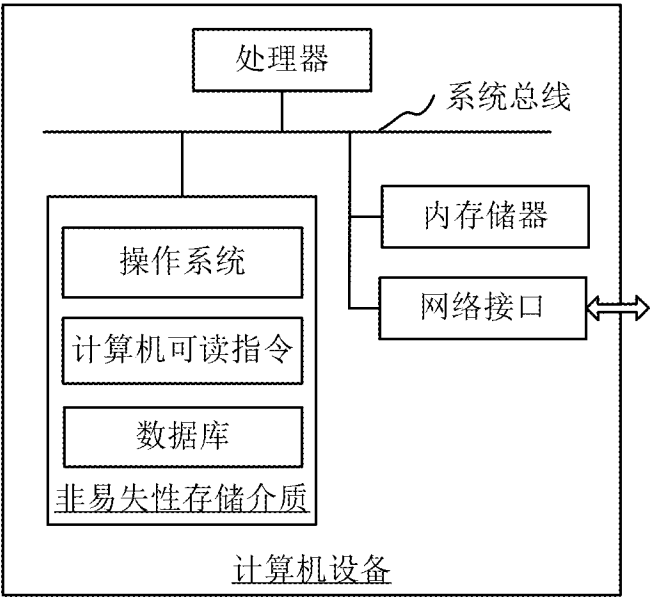


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/090873

A. CLASSIFICATION OF SUBJECT MATTER

G16H 50/70(2018.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G16H G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, EPODOC, CNPAT, CNKI, IEEE: 机器学习, 医疗, 分类, 病历, 文本, 向量, 维度, 神经网络, 训练, machine, learning, classify, medical, record, text, vector, neural, network, dimension, train

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 107680689 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 09 February 2018 (2018-02-09) description, paragraphs [0043]-[0048], [0066] and [0093]-[0095], and figure 3	1-20
A	CN 108447534 A (LINGJOIN SOFTWARE (BEIJING) CO., LTD.) 24 August 2018 (2018-08-24) entire document	1-20
A	US 2017228500 A1 (MASSENGALE, Justin) 10 August 2017 (2017-08-10) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

25 November 2019

Date of mailing of the international search report

17 December 2019

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/090873

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	107680689	A	09 February 2018	WO	2018201772	A1	08 November 2018
CN	108447534	A	24 August 2018	None			
US	2017228500	A1	10 August 2017	None			

国际检索报告

国际申请号

PCT/CN2019/090873

A. 主题的分类

G16H 50/70 (2018.01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

G16H G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

WPI, EPDOC, CNPAT, CNKI, IEEE: 机器学习, 医疗, 分类, 病历, 文本, 向量, 维度, 神经网络, 训练, machine, learning, classify, medical, record, text, vector, neural, network, dimension, train

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 107680689 A (平安科技深圳有限公司) 2018年 2月 9日 (2018 - 02 - 09) 说明书第[0043]-[0048]、[0066]、[0093]-[0095]段、附图3	1-20
A	CN 108447534 A (灵玖中科软件北京有限公司) 2018年 8月 24日 (2018 - 08 - 24) 全文	1-20
A	US 2017228500 A1 (MASSENGALE, Justin) 2017年 8月 10日 (2017 - 08 - 10) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2019年 11月 25日

国际检索报告邮寄日期

2019年 12月 17日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

授权官员

鞠博

电话号码 86-(10)-53961534

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/090873

检索报告引用的专利文件			公布日 (年/月/日)	同族专利		公布日 (年/月/日)
CN	107680689	A	2018年 2月 9日	W0	2018201772	A1 2018年 11月 8日
CN	108447534	A	2018年 8月 24日	无		
US	2017228500	A1	2017年 8月 10日	无		