(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 10 月 29 日 (29.10.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/215671 A1

(51) 国际专利分类号:
G16H 50/80 (2018.01)

(21) 国际申请号: PCT/CN2019/116942

(22) 国际申请日: 2019 年 11 月 11 日 (11.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910763137.5 2019年8月19日 (19.08.2019) 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 518000 (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 518000 (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 518000 (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 518000 (CN)。

(74) 代理人: 深圳众鼎专利商标代理事务所（普通合伙) (SHENZHEN ZHONGDING INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区笋岗东路3012号中民时代广场B座7楼701室, Guangdong 518000 (CN)。

(54) **Title:** METHOD AND DEVICE FOR SMART ANALYSIS OF DATA, AND COMPUTER DEVICE AND STORAGE MEDIUM

(54) **发明名称:** 数据智能分析方法、装置、计算机设备及存储介质

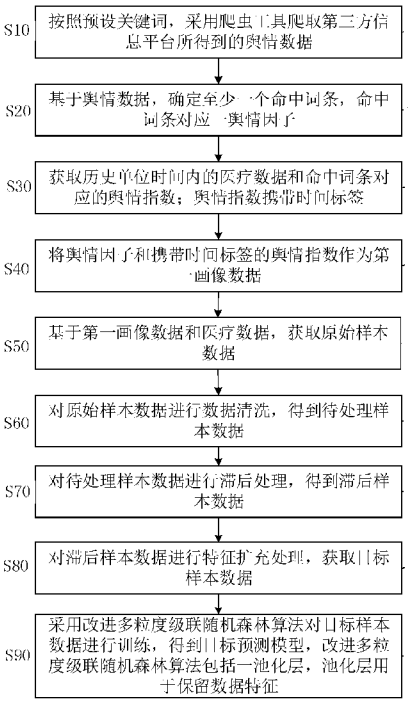


图 2

S10 On basis of predetermined keyword, use crawling tool to grab public opinion data obtained by third party information platform
S20 On basis of public opinion data, determine at least one command line, said command line corresponding to public opinion factor
S30 For past period of time, obtain medical data and public opinion index corresponding to command line, said public opinion index having a time marker
S40 Treat public opinion factor and public opinion index having time marker as first representation data
S50 On basis of first representation data and medical data, obtain original sample data
S60 Perform data cleaning on original sample data to obtain sample data to be processed
S70 Perform time-lag processing on sample data to be processed, so as to obtain time-lag sample data
S80 Perform feature expansion processing on time-lag sample data to obtain target sample data
S90 Perform training on target sample data by using improved multi-grained cascade random forest algorithm, so as to obtain a target prediction model, improved multi-grained cascade forest algorithm comprising pooling layer configured to retain data features

(57) **Abstract:** A method and device for the smart analysis of data, a computer device and a storage medium. The method for the smart analysis of data comprises: using an obtained public opinion factor and a public opinion index having a time marker as first representation data (S40); acquiring original sample data on the basis of the first representation data and medical data (S50); performing data cleaning on the original sample data to obtain sample data to be processed (S60); performing time-lag processing on the sample data to be processed, so as to obtain time-lag sample data (S70); performing feature expansion processing on the time-lag sample data

WO 2020/215671 A1

[见续页]

(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, 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))。
- 在修改权利要求的期限届满之前进行，在收到该修改后将重新公布(细则48.2(h))。
- 根据申请人的请求，在条约第21条(2)(a)所规定的期限届满之前进行。

to obtain target sample data (S80); performing training on the target sample data by using an improved multi-grained cascade random forest algorithm, so as to obtain a target prediction model, the improved multi-grained cascade random forest algorithm comprising a pooling layer configured to retain data features (S90). The method for the smart analysis of data can effectively increase model prediction accuracy and model training efficiency.

(57) 摘要：一种数据智能分析方法、装置、计算机设备及存储介质，该数据智能分析方法包括：将获取到的舆情因子和携带时间标签的舆情指数作为第一画像数据（S40）；基于所述第一画像数据和医疗数据，获取原始样本数据（S50）；对所述原始样本数据进行数据清洗，得到待处理样本数据（S60）；对所述待处理样本数据进行滞后处理，得到滞后样本数据（S70）；对所述滞后样本数据进行特征扩充处理，获取目标样本数据（S80）；采用改进多粒度级联随机森林算法对所述目标样本数据进行训练，得到目标预测模型；所述改进多粒度级联随机森林算法包括一池化层，所述池化层用于保留数据特征（S90），该数据智能分析方法可有效提高模型预测准确率和模型训练效率。

数据智能分析方法、装置、计算机设备及存储介质

本申请以 2019 年 8 月 19 日提交的申请号为 201910763137.5，名称为“数据智能分析方法、装置、计算机设备及存储介质”的中国发明专利申请为基础，并要求其优先权。

技术领域

本申请涉及数据预测技术领域，尤其涉及一种数据智能分析方法、装置、计算机设备及存储介质。

背景技术

随着信息时代的飞速发展，数据预测技术也在不断发展。目前各大科研机构针对医疗数据进行预测时，由于部分医疗数据具有滞后性，导致模型预测准确率较低，例如对于具有一定潜伏期的传染性疾病（如水痘），在满足疾病爆发的条件（如气温、湿度等）时，可能会在下一时段爆发，这就导致模型预测准确率较低，使公民不能及时预防，无法对疾病爆发的严重程度起到控制作用。

发明内容

本申请实施例提供一种数据智能分析方法、装置、计算机设备及存储介质，以解决目前对滞后性的数据进行数据预测时，模型预测准确率较低的问题。

一种数据智能分析方法，包括：

按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据；

基于所述舆情数据，确定至少一个命中词条；所述命中词条对应一舆情因子；

获取历史单位时间内的医疗数据和所述命中词条对应的舆情指数；所述舆情指数携带时间标签；

将所述舆情因子和所述携带时间标签的舆情指数作为第一画像数据；

基于所述第一画像数据和所述医疗数据，获取原始样本数据；

对所述原始样本数据进行数据清洗，得到待处理样本数据；

对所述待处理样本数据进行滞后处理，得到滞后样本数据；

对所述滞后样本数据进行特征扩充处理，获取目标样本数据；

采用改进多粒度级联随机森林算法对所述目标样本数据进行训练，得到目标预测模型；所述改进多粒度级联随机森林算法包括一池化层，所述池化层用于保留数据特征。

一种数据智能分析装置，包括：

舆情数据获取模块，用于按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据。

命中词条确定模块，用于基于舆情数据，确定至少一个命中词条；命中词条对应一舆情因子。

舆情指数获取模块，用于获取历史单位时间内的医疗数据和所述命中词条对应的舆情指数；所述舆情指数携带时间标签。

第一画像数据获取模块，用于将所述舆情因子和所述携带时间标签的舆情指数作为第一画像数据。

原始样本数据获取模块，用于基于所述第一画像数据和所述医疗数据，获取原始样本数据。

待处理样本数据获取模块，用于对所述原始样本数据进行数据清洗，得到待处理样本数据；

滞后样本数据获取模块，用于对所述待处理样本数据进行滞后处理，得到滞后样本数据；

目标样本数据获取模块，用于对所述滞后样本数据进行特征扩充处理，获取目标样本数据；

目标预测模型获取模块，用于采用改进多粒度级联随机森林算法对所述目标样本数据进行训练，得到目标预测模型；所述改进多粒度级联随机森林算法包括一池化层，所述池化层用于保留数据特征。

一种计算机设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令，所述处理器执行所述计算机可读指令时实现上述数据智能分析方法的步骤。

一种可读存储介质，所述可读存储介质存储有计算机可读指令，所述计算机可读指令被处理器执行时实现上述数据智能分析方法的步骤。

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，本申请的其他特征和优点将从说明书、附图以及权利要求变得明显。

附图说明

为了更清楚地说明本申请实施例的技术方案，下面将对本申请实施例的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性的前提下，还可以根据这些附图获得其他的附图。

图 1 是本申请一实施例中数据智能分析方法的一应用环境示意图；

图 2 是本申请一实施例中数据智能分析方法的一流程图；

图 3 是图 2 中步骤 S60 的一具体流程图；

图 4 是图 2 中步骤 S80 的一具体流程图；

图 5 是本申请一实施例中数据智能分析方法的一流程图；

图 6 是图 2 中步骤 S90 的一具体流程图；

图 7 是图 6 中步骤 S92 的一具体流程图；

图 8 是本申请一实施例中数据智能分析装置的一示意图；

图 9 是本申请一实施例中计算机设备的一示意图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

本申请实施例提供的数据智能分析方法可应用在本方法可应用在一种数据智能分析工具中，该数据智能分析工具可根据不同的主题（例如水痘、流感等）对应的样本数据训练不同的预测模型，尤其针对具有滞后性的样本数据，可有效保证模型预测的准确率。该数据智能分析方法可应用在如图 1 的应用环境中，其中，计算机设备通过网络与服务器进行通信。计算机设备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。服务器可以用独立的服务器来实现。

在一实施例中，如图 2 所示，提供一种数据智能分析方法，以该方法应用在图 1 中的服务器为例进行说明，包括如下步骤：

S10：按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据。

其中，预设关键词是预先设置的涉及传播性疾病的一些关键词，如水痘、红肿、瘙痒性疱疹和水疱疹等。舆情数据是指第三方信息平台中不同用户所公开发布的文字数据，用于反映社会事件的发生。具体地，随时信息时代的飞速发展，用户更倾向于采用各种信息平台查询所需的信息，例如根据自身症状查询是否患有疾病等，当某一传播性疾病爆发（如水痘）时，必然会有更大的搜索量或关注度，故本实施例中还按照预设关键词，采用爬虫工具爬取第三方信息平台（如百度、微博或微信）中包含预设关键词的舆情数据。需要说明的是，本实施例中的涉及传播性疾病的一些预设关键词可预先设定部分默认的关键词，再取该默认的关键词对应的近义词，以得到更多的关键词进行爬取，获取更多相关的信息，为后续模型训练提供充足的数据集。

S20：基于舆情数据，确定至少一个命中词条，命中词条对应一舆情因子。

具体地，随时信息时代的飞速发展，用户更倾向于采用各种信息平台查询所需的信息，例如根据自身症状查询是否患有疾病等，当某一传播性疾病爆发（如水痘）时，必然会有更大的搜索量或关注度，故本实施例中选择不同地区历史 20 年每天的舆情因子作为另一部分画像数据。该舆情因子包括但不限于水痘、红肿、瘙痒性疱疹和水疱疹等。

其中，舆情数据包括至少一个原始词条（如百度词条）。具体地，通过专家根据爬取到的每一原始词条中所包含的信息判断是否与水痘有关，以确定至少一个与水痘真实相关的词条作为命中词条。然后，再根据确定的命中词条。每一命中词条对应一舆情因子。该舆情因子是指命中词条中包含的至少一个与预设关键词相关的因子，如水痘、红肿、瘙痒性疱疹和水疱疹。

S30：获取历史单位时间内的医疗数据和命中词条对应的舆情指数舆情指数携带时间标签。

其中，医疗数据是指疾控中心提供的，不同地区哨点医院历史单位时间，如历史 20 年的单位时间内的历史发病人数（即标签数据）。可以理解地，该单位时间即为时间标签，该单位时间可由用户自定义选定，此处不做限定。本实施例中，该单位时间可为一天、一周、一个月、一个季度或者一年等等，在此不一一列举。

本实施例中，以单位时间为一周为例进行说明，具体地，获取单位时间内命中词条对应的舆情指数和医疗数据，每一舆情指数携带时间标签，该时间标签即指命中词条的发布时间。

S40：将舆情因子和携带时间标签的舆情指数作为第一画像数据。

其中，第一画像数据即指将舆情因子和携带时间标签的舆情指数作为模型训练的特征数据。具体地，当需要预测未来一个时间区间内某疾病是否爆发，该时间区间可为一周、一个月、一个季度或者一年，根据预测的时间区间不同，在样本数据的处理上会有所不同，以时间区间为一周进行举例说明，可以舆情因子（如水痘、红肿和疱疹）为列标签，以第 N 周的舆情指数为行标签，建立部分画像数据。其中，第 N 周舆情指数包括但不限于第 N 周平均舆情指数（即一周 7 天的舆情指数取平均）、第 N 周最大舆情指数以及第 N 周最小舆情指数。

需要说明的是，如下表格为本实施例中根据舆情因子建立的画像数据示意图。可以理解地，该示意图仅做示例，在此不做限定。

第 N 周舆情指数	红肿	水痘	疱疹	……
第一周平均舆情指数	x_1	x_2	x_3	……
第一周最大舆情指数	y_1	y_2	y_3	……

第一周最小舆情指数	Z_1	Z_2	Z_3	……
……	……	……	……	……
第 N 周	……	……	……	……

S50：基于第一画像数据和医疗数据，获取原始样本数据

具体地，将第一画像数据作为模型训练的特征数据，将医疗数据作为模型训练的标签数据，以获取原始样本数据。

S60：对原始样本数据进行数据清洗，得到待处理样本数据。

具体地，由于原始样本数据中可能包括缺失值或异常值，为进一步的保证后续模型预测的准确率，需要对原始样本数据进行数据清洗，以保证待处理样本数据的质量。

S70：对待处理样本数据进行滞后处理，得到滞后样本数据。

其中，滞后处理是一种特征工程方法，通过扩充样本数据集，即增大特征画像以收集更多信息的方法。从业务逻辑层面理解就是，延迟特征的效果。具体地，由于部分模型预测的主题不同，其对应的样本数据存在滞后性，如疾病的爆发或者与经济相关的数据。本实施例中，假设预测主题为预测水痘，而水痘的爆发存在滞后性，例如本周的气温突然升高且气候潮湿，可能这一周并不会带来水痘的爆发，但是下一周会迎来爆发期，故需要对待处理样本数据进行滞后处理，以保证后续模型预测的准确率。具体地，对待处理样本数据进行 n 次滞后处理(n一般取 1~3)，假设 n 取 1，则对待处理样本数据进行滞后处理，即将原第一周的数据作为第二周的数据，第二周的数据作为第三周的数据，以此类推，以得到滞后样本数据。若 n 取 2，由于是在第一次时候得到的样本数据的基础上进行再次之后，故对待处理样本数据进行滞后处理，即将原第一周的数据作为第三周的数据，第二周的数据作为第四周的数据，以此类推，得到滞后数据，将每次得到的滞后数据集成，以得到滞后样本数据，实现扩充样本数据集的目的

最后，采用 concat 函数将多次滞后处理得到的滞后样本数据与待处理样本数据合并为一个数据帧（DataFrame）即滞后样本数据。其中，concat 函数是用于连接两个或多个数组的函数。数据帧是二维数据结构,即数据以行和列的表格方式排列。

S80：对滞后样本数据进行特征扩充处理，获取目标样本数据。

具体地，为了扩充样本数据集，进一步提高模型预测的准确率，本实施例中会对滞后样本数据进行特征扩充处理，得到目标样本数据，以达到进一步扩充样本数据集的目的。

S90：采用改进多粒度级联随机森林算法对目标样本数据进行训练，得到目标预测模型，改进多粒度级联随机森林算法包括一池化层，池化层用于保留数据特征。

其中，改进多粒度级联随机森林算法是在多粒度级联随机森林算法中引入卷积神经网络中池化思想的算法。多粒度级联随机森林算法是一种决策树集成方法，通过级联的方式堆叠多层随机森林，以获得更好的特征表示和学习性能，该算法无需过多调节超参数，即可达到良好的性能。

其中，多粒度级联随机森林（Gcforest）中每一层都由多个随机森林组成。通过随机森林学习输入特征向量的特征信息，经过处理后输入到下一层。为了增强模型的泛化能力，每一层选取多种不同类型的随机森林，例如每一层选取两种随机森林结构，分别为 completely-random tree forests（完全随机森林）和 random forests（随机森林）。

本实施例中，先按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据，以便基于舆情数据，确定至少一个与预测主题真实相关的命中词条，以保证后续获取的舆情因子的有效性和准确性。再获取单位时间内命中词条对应的舆情指数和医疗数据。最后，将舆情因子和携带时间标签的舆情指数作为原始样本数据，以使模型通过对历史 20 年，单位时间内的舆情数据进行分析。然后，通过对原始样本数据进行数据清洗，得到待处理样本数据，以保证待处理样本数据的质量。然后，对对待处理样本数据进行滞后处理，

得到滞后样本数据，以扩充样本数据集。并且，针对具有滞后性数据，可实现延迟特征的效果，保证模型预测的准确率。接着，对滞后样本数据进行特征扩充处理，获取目标样本数据，以达到进一步扩充样本数据集的目的，提高模型预测的准确率。最后，采用改进多粒度级联随机森林算法对目标样本数据进行训练，得到目标预测模型，以获得更好的特征表示和学习性能，且算法无需过多调节超参数，即可达到良好的性能，保证模型预测的准确率。并且，改进多粒度级联随机森林算法还包括一池化层，以充分保留数据特征，进一步提高模型预测的准确率。

在一实施例中，步骤 S10 之前中，该数据智能分析方法还包括：

S101：获取气象因子和对应的气象数据。

可以理解地，本实施例可根据预测主题的不同选择不同的画像数据，本实施例中以预测水痘为例进行说明，由于气候情况和水痘病毒存在非常紧密的相关性，故选择不同地区历史 20 年每天的气象因子作为一部分画像数据。该气象因子包括但不限于不同地区的昼夜气温、昼夜气压、昼夜降水量、湿度、光照强度和风力等。

S102：将气象因子和对应的气象数据作为第二画像数据；

其中，第二画像数据即指将气象因子和对应的气象数据作为模型训练的特征数据。具体地，针对气象因子建立画像数据的方式与步骤 S40 一致，即可以气象因子为列标签，以第 N 周的气象情况为行标签，建立第二画像数据。其中，第 N 周的气象情况包括但不限于第 N 周的平均气象情况（如平均降水量）、第 N 周的最大气象情况（如最大降水量）以及第 N 周的最小气象情况（如最小降水量）。

相应地，步骤 S50 中，即基于第一画像数据和医疗数据，获取原始样本数据，包括：

S51：将第一画像数据、第二画像数据和医疗数据作为原始样本数据。

本实施例中，通过气象情况结合舆情数据大量传播的思想，以有效预测疾病爆发时段，提高模型预测的准确率。

在一实施例中，如图 3 所示，步骤 S60 中，即对原始样本数据进行数据清洗，得到待处理样本数据， 具体包括如下步骤：

S61：对原始样本数据进行缺失值填充，得到第一样本数据。

其中，缺失值填充方法包括但不限于均值填充、众数填充、中位数填充、期望值最大化方法、多重填补以及 k-means 聚类方法等。具体地，以 k-means 聚类方法进行填充为例，将缺失值所在的画像数据进行聚类，并将缺失值以所聚类类簇的均值进行填充。

S62：对第一样本数据进行异常值检测，得到至少一个异常值，将异常值标记为空。

S63：对标记为空的异常值进行缺失值填充，得到待处理样本数据。

具体地，异常值检测包括但不限于采用统计变量分析（如箱型图分析、平均值、最大最小值分析以及 3σ 法则）、基于距离的方法、基于密度的离群点检测、基于密度的离群点检测和孤立森林(Isolation Forest)等。本实施例中，以 3σ 法则为例，若数据服从正太分布，在 3σ 原则下，异常值被定义为一组测定值中与平均值的偏差超 3 倍标准差的值，因为在正态分布的假设下，距离平均值 3σ 之外的值出现的概率小于 0.003），即超过 $\mu+3\sigma$ 的数据以及不超过 $\mu-3\sigma$ 的数据作为异常值。

具体地，由于异常值对应的样本数据不一定是不必要的，若直接将该异常值对应的样本数据删除，会导致样本数据中的特征缺失，影响样本数据的质量，进而影响模型预测的准确率，故本实施例中会将异常值删除并标记为空值，再对标记为空值的异常值再次进行缺失值填充，得到待处理样本数据。本实施例中，通过对标记为空值的异常值进行缺失值填充，得到待处理样本数据，以避免直接将异常值对应的样本数据去除，导致样本数据缺少该部分特征，影响模型预测的准确率的问题。

本实施例中，通过对原始样本数据进行缺失值填充，得到第一样本数据，再对对第一

样本数据进行异常值检测，得到至少一个异常值，以通过对样本数据中的异常值和缺失值进行处理，达到数据清洗的目的，保证样本数据的质量。然后，将得到的异常值标记为空，以便对标记为空的异常值再次进行缺失值填充，得到待处理样本，以通过对原始样本数据进行两次缺失值填充，保证样本数据的质量和规范性，提升模型预测的准确率

在一实施例中，如图 4 所示，步骤 S80 中，即对滞后样本数据进行特征扩充处理，获取目标样本数据，具体包括如下步骤：

S81：对滞后样本数据进行特征扩充，得到至少一个统计指标对应的特征值。

S82：将特征值与滞后样本数据进行拼接，获取目标样本数据。

其中，统计指标包括但不限于每一行数据对应的最大值、最小值、均值和标准差，将每一统计指标作为新的列加入滞后样本数据中，以扩充数据集并增大特征画像收集更多特征信息，提高模型预测的准确率。可以理解地，该滞后样本数据为一矩阵，将特征值与滞后样本数据进行拼接，获取目标样本数据，即在样本矩阵中增加 N 个列，N 为统计指标（如每一行对应的数据的最大值、最小值和均值）的个数，每一行对应的数据的最大值、最小值和均值即为特征值。

本实施例中，通过对滞后样本数据进行特征扩充，得到至少一个统计指标对应的特征值，将特征值与滞后样本数据进行拼接，获取目标样本数据以扩充数据集并增大特征画像收集更多特征信息，提高模型预测的准确率。

在一实施例中，如图 5 所示，步骤 S80 之后，该数据智能分析方法还包括如下步骤：

S111：对目标样本数据进行方差分析，去除方差小于预设方差阈值的数据，得到第二样本数据。

S112：对第二样本数据进行奇异值分解，以更新目标样本数据。

具体地，由于数据量有时过犹不及，在数据分析应用中大量的数据反而会产生更坏的性能。故需要对目标样本数据进行筛选，以去除冗余数据，达到减少数据列数的同时保证丢失的数据信息尽可能少。

其中，方差分析是指根据数据列的方差进行分析，以去除方差过于小（即小于预设方差阈值）的序列，得到第二样本数据。具体地，方差的大小描述的是一个变量的信息量，方差过于小的序列则认为其包含的信息量少，故去除所有方差小的数据列，以达到数据降维的效果，降低数据处理量，提高后续模型训练效率。

具体地，在目标样本数据中包含多个特征，但某些特征对于模型的预测精度的影响并不大，或者可认为相关性过大的特征可同等替换，故可将冗余变量去除，以达到数据降维的目的，节约模型训练时间。具体地，采用方差分析时，是将方差小于预设方差阈值的数据列去除，故方差分析的准确性取决于预设方差阈值，因此，为进一步去除冗余数据，且能保证丢失的数据信息尽可能少，本实施例中还需对第二样本数据进行奇异值分解，以除冗余数据，实现数据压缩的目的，保证目标样本数据的质量。

本实施例中，通过对目标样本数据进行方差分析，去除方差小于预设方差阈值的数据，得到第二样本数据，以去除冗余数据，达到减少数据列数的同时保证丢失的数据信息尽可能少，节约模型训练时间。然后，对第二样本数据进行奇异值分解，更新目标样本数据，以进一步去除冗余数据，保证目标样本数据的质量。

在一实施例中，改进多粒度级联随机森林算法包括多粒子扫描算法和级联随机森林算法，多粒子扫描算法对应至少一个滑动窗口，如图 6 所示，步骤 S90 中，具体包括如下步骤：

S91：采用多粒子扫描算法，按照至少一个滑动窗口，对目标样本数据进行多粒子扫描，得到至少一个中间数据。

其中，多粒子扫描是指采用滑动窗口对目标样本数据进行扫描，得到至少一个中间数据。本实施例中，可设置不同维度的滑动窗口，可以理解地，该滑动窗口可为 i*j 的窗口。

例如目标表样本数据行标签为第 i 周，则滑动窗口 window_size 可取 2（每 2 周）、4（每个月）、12（每个季度）等。需要说明的是，该滑动窗口可扫描至少一个特征画像，即可扫描每一列、每两列、每 j 列，以极大化地搜寻特征与标签集、特征与特征之间的内在关联性。

S92：基于池化层，对至少一个中间数据进行池化处理，得到待训练数据。

本实施例中，通过池化层对至少一个中间数据进行池化处理，得到待训练数据，以达到对数据进行降维的目的，减小计算量，提高模型训练效率。

S93：采用级联随机森林算法对待训练数据进行训练，获取目标预测模型。

具体地，多粒度级联随机森林算法基于神经网络集成的思想，将第 i 次 complete-random tree forest 预测得到的标签列 $cforest_i$ 和 random forest 预测得到的标签列 $rforest_i$ 作为不断加入目标样本数据的画像列，以进一步特征扩充，最终得到如下的特征画像 $[orgf_1, orgf_2, \mathrm{K}, orgf_n, cforest_1, rforest_1, \mathrm{K}, cforest_k, rforest_k]$。其中，orgf 是目标样本数据。最后，将该特征画像输入到最后 m 个（m 一般取 3~5，一般数量级取 3，千万数量级取 3~4，超过千万数量级取 4~5）random forest 中进行预测，取最终的 Max 值作为最终的预测概率值。

具体地，将得到的待训练数据输入到级联森林中进行训练。例如，本实施例中采用三种维度的滑动窗口，首先使用第一维度的滑动窗口进行扫描得到一特征向量，再将该原始特征向量输入到 complete-random tree forest 和 random forest 中，分别得到两个预测序列（即 $cforest_i$ 和 $rforest_i$），再将这两个预测序列拼接，得到第一特征向量，将原始特征向量输入到第一层级联森林中进行训练，得到第一预测序列。然后将得到的第一预测序列与第一特征向量进行拼接，得到第二特征向量，作为第二层的级联森林的输入数据；第二层级联森林训练得到的第二预测序列再与第二维度的滑动窗得到的第三特征向量（与第一特征向量的获取方法相同）进行拼接，作为第三层级联森林的输入数据；第三层级联森林训练得到的第三预测序列再与第三维度的滑动窗得到的第四特征向量进行拼接，作为下一层的输入，不断重复上述过程，直至收敛，得到目标预测模型。

本实施例中，通过采用多粒子扫描算法，按照至少一个滑动窗口，对目标样本数据进行多粒子扫描，得到至少一个中间数据，以极大化地搜寻特征与标签集、特征与特征之间的内在关联性。然后，通过结合池化层，对至少一个中间数据进行池化处理，得到待训练数据，以将机器学习和神经网络思想相结合，获取更多直观无法获取的信息，来丰富模型，进一步提高模型预测准确率。

在一实施例中，如图 7 所示，步骤 S92 中，即基于池化层，对至少一个中间数据进行池化处理，得到待训练数据，具体包括如下步骤：

S921：选取相邻的两个中间数据作为一组待处理数据组，以得到中间数据对应的至少一组待处理数据组。

S922：对每组待处理数据组进行取平均运算，得到第一数据序列。

S923：对每组待处理数据组进行最小值运算，得到第二数据序列，第二数据列中包括每组待处理数据组的两个中间数据中的最小值。

S924：对每组待处理数据组进行最大值运算，得到第三数据序列，第三数据列中包括每组待处理数据组的两个中间数据中的最大值。

S925：将第一数据序列、第二数据序列和第三数据序列进行拼接，得到待训练数据。

具体地，从业务逻辑层面上来说，模型预测需要更多线性、或者非线性的方法来对数

据进行空间扭曲，从而获取更多直观无法获取的信息，来丰富模型，故本实施例中，采用三种池化方式对至少一个中间数据进行池化，再对每种方式进行池化所得到的结果进行集成，得到待训练数据，以获取更多直观无法获取的信息，来丰富模型，并可充分保留数据特征。假设中间是中间数据中某一列画像数据为 $Feature$：$f_1, f_2, f_3, f_4, f_5, \mathrm{K}\ f_n$，则采用如下三种池化方式对至少一个中间数据进行池化。

$$Feature_new_1: (f_1+f_2)/2, (f_2+f_3)/2, \mathrm{K}, (f_{n-1}+f_n)/2$$

$$Feature_new_2: \max(f_1, f_2), \max(f_2, f_3), \mathrm{K}, \max(f_{n-1}, f_n)$$

$$Feature_new_3: \min(f_1, f_2), \min(f_2, f_3), \mathrm{K}, \min(f_{n-1}, f_n)$$

本实施例中，通过采用三种池化方式对至少一个中间数据进行池化，再对每种方式进行池化所得到的结果进行集成，得到待训练数据，以充分保留数据特征，保证样本数据质量，提高模型预测准确率。

应理解，上述实施例中各步骤的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本申请实施例的实施过程构成任何限定。

在一实施例中，提供一种数据智能分析装置，该数据智能分析装置与上述实施例中数据智能分析方法一一对应。如图 8 所示，该数据智能分析装置包括舆情数据获取模块 10、命中词条确定模块 20、舆情指数获取模块 30、第一画像数据获取模块 40、原始样本数据获取模块 50、待处理样本数据获取模块 60、滞后样本数据获取模块 70、目标样本数据获取模块 80、和目标预测模型获取模块 90。各功能模块详细说明如下：

舆情数据获取模块 10，用于按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据。

命中词条确定模块 20，用于基于舆情数据，确定至少一个命中词条；命中词条对应一舆情因子。

舆情指数获取模块 30，用于获取历史单位时间内的医疗数据和命中词条对应的舆情指数；舆情指数携带时间标签。

第一画像数据获取模块 40，用于将舆情因子和携带时间标签的舆情指数作为第一画像数据。

原始样本数据获取模块 50，用于基于第一画像数据和医疗数据，获取原始样本数据。

待处理样本数据获取模块 60，用于对原始样本数据进行数据清洗，得到待处理样本数据。

滞后样本数据获取模块 70，用于对待处理样本数据进行滞后处理，得到滞后样本数据。

目标样本数据获取模块 80，用于对滞后样本数据进行特征扩充处理，获取目标样本数据。

目标预测模型获取模块 90，用于采用改进多粒度级联随机森林算法对目标样本数据进行训练，得到目标预测模型；改进多粒度级联随机森林算法包括一池化层，池化层用于保留数据特征。

具体地，待处理样本数据获取模块包括第一样本数据获取单元、异常值获取单元和待处理样本数据获取单元。

第一样本数据获取单元，用于对原始样本数据进行缺失值填充，得到第一样本数据。

异常值获取单元，用于对第一样本数据进行异常值检测，得到至少一个异常值，将异常值标记为空。

待处理样本数据获取单元，用于对标记为空的异常值进行缺失值填充，得到待处理样本数据。

具体地，目标样本数据获取模块包括特征值获取单元和目标样本数据获取单元。

特征值获取单元，用于对滞后样本数据进行特征扩充，得到至少一个统计指标对应的特征值。

目标样本数据获取单元，用于将特征值与滞后样本数据进行拼接，获取目标样本数据。

具体地，该数据智能分析装置包括第二样本数据获取单元和目标样本数据更新单元。

第二样本数据获取单元，用于对目标样本数据进行方差分析，去除方差小于预设方差阈值的数据，得到第二样本数据。

目标样本数据更新单元，用于对第二样本数据进行奇异值分解，以更新目标样本数据。

具体地，改进多粒度级联随机森林算法包括多粒子扫描算法和级联随机森林算法，多粒子扫描算法对应至少一个滑动窗口；目标预测模型获取模块包括目标预测模型、待训练数据获取单元和目标预测模型获取单元。

中间数据获取单元，用于采用多粒子扫描算法，按照至少一个滑动窗口，对目标样本数据进行多粒子扫描，得到至少一个中间数据。

待训练数据获取单元，用于基于池化层，对至少一个中间数据进行池化处理，得到待训练数据。

目标预测模型获取单元，用于采用级联随机森林算法对待训练数据进行训练，获取目标预测模型。

具体地，待训练数据获取单元包括待处理数据组获取子单元、第一数据序列获取子单元、第二数据序列获取子单元、第三数据序列获取子单元、和待训练数据获取子单元。

待处理数据组获取子单元，用于选取相邻的两个中间数据作为一组待处理数据组，以得到中间数据对应的至少一组待处理数据组。

第一数据序列获取子单元，用于对每组待处理数据组进行取平均运算，得到第一数据序列。

第二数据序列获取子单元，用于对每组待处理数据组进行最小值运算，得到第二数据序列，第二数据列中包括每组待处理数据组的两个中间数据中的最小值。

第三数据序列获取子单元，用于对每组待处理数据组进行最大值运算，得到第三数据序列，第三数据列中包括每组待处理数据组的两个中间数据中的最大值。

待训练数据获取子单元，用于将第一数据序列、第二数据序列和第三数据序列进行拼接，得到待训练数据。

关于数据智能分析装置的具体限定可以参见上文中对于数据智能分析方法的限定，在此不再赘述。上述数据智能分析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 10 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可读存储介质、内存储器。该可读存储介质存储有操作系统、计算机可读指令和数据库。该内存储器为可读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执行数据智能分析方法过程中生成或获取的数据，如目标样本数据。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一种数据智能分析方法。

在一个实施例中，提供了一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机可读指令，处理器执行计算机可读指令时实现上述实施例中的数据智能分析方法的步骤，例如图 2 所示的步骤 S10-S90，或者图 3 至图 7 中所示的步骤。

或者，处理器执行计算机可读指令时实现数据智能分析装置这一实施例中的各模块/单元的功能，例如图 8 所示的各模块/单元的功能，为避免重复，这里不再赘述。

在一实施例中，提供一个或多个存储有计算机可读指令的可读存储介质，所述计算机可读存储介质存储有计算机可读指令，其特征在于，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行时实现上述实施例中数据智能分析方法的步骤，例如图 2 所示的步骤 S10-S90，或者图 3 至图 7 中所示的步骤，为避免重复，这里不再赘述。或者，该计算机可读指令被处理器执行时实现上述数据智能分析装置这一实施例中的各模块/单元的功能，例如图 8 所示的各模块/单元的功能，为避免重复，这里不再赘述。本实施例中的可读存储介质包括非易失性可读存储介质和易失性可读存储介质。

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取存储介质中，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。

所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，仅以上述各功能单元、模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能单元、模块完成，即将所述装置的内部结构划分成不同的功能单元或模块，以完成以上描述的全部或者部分功能。

以上所述实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围，均应包含在本申请的保护范围之内。

权 利 要 求 书

1. 一种数据智能分析方法，其特征在于，包括：

按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据；

基于所述舆情数据，确定至少一个命中词条；所述命中词条对应一舆情因子；

获取历史单位时间内的医疗数据和所述命中词条对应的舆情指数；所述舆情指数携带时间标签；

将所述舆情因子和所述携带时间标签的舆情指数作为第一画像数据；

基于所述第一画像数据和所述医疗数据，获取原始样本数据；

对所述原始样本数据进行数据清洗，得到待处理样本数据；

对所述待处理样本数据进行滞后处理，得到滞后样本数据；

对所述滞后样本数据进行特征扩充处理，获取目标样本数据；

采用改进多粒度级联随机森林算法对所述目标样本数据进行训练，得到目标预测模型；所述改进多粒度级联随机森林算法包括一池化层，所述池化层用于保留数据特征。

2. 如权利要求 1 所述数据智能分析方法，其特征在于，所述按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据之前，所述数据智能分析方法还包括：

获取气象因子和对应的气象数据；

将所述气象因子和对应的气象数据作为第二画像数据；

所述基于所述第一画像数据和所述医疗数据，获取原始样本数据，包括：

将所述第一画像数据、所述第二画像数据和所述医疗数据作为原始样本数据。

3. 如权利要求 1 所述数据智能分析方法，其特征在于，所述对所述原始样本数据进行数据清洗，得到待处理样本数据，包括；

对所述原始样本数据进行缺失值填充，得到第一样本数据；

对所述第一样本数据进行异常值检测，得到至少一个异常值，将所述异常值标记为空；

对所述标记为空的异常值进行缺失值填充，得到所述待处理样本数据。

4. 如权利要求 1 所述数据智能分析方法，其特征在于，所述对所述滞后样本数据进行特征扩充处理，获取目标样本数据，包括：

对所述滞后样本数据进行特征扩充，得到至少一个统计指标对应的特征值；

将所述特征值与所述滞后样本数据进行拼接，获取所述目标样本数据。

5. 如权利要求 1 或 4 所述数据智能分析方法，其特征在于，在所述获取目标样本数据之后，所述数据智能分析方法包括：

对所述目标样本数据进行方差分析，去除方差小于预设方差阈值的数据，得到第二样本数据；

对所述第二样本数据进行奇异值分解，以更新所述目标样本数据。

6. 如权利要求 1 所述数据智能分析方法，其特征在于，所述改进多粒度级联随机森林算法包括多粒子扫描算法和级联随机森林算法，所述多粒子扫描算法对应至少一个滑动窗口；

所述采用改进多粒度级联随机森林算法对所述目标样本数据进行训练，得到目标预测模型，包括：

采用所述多粒子扫描算法，按照至少一个所述滑动窗口，对所述目标样本数据进行多粒子扫描，得到至少一个中间数据；

基于所述池化层，对至少一个所述中间数据进行池化处理，得到待训练数据；

采用级联随机森林算法对所述待训练数据进行训练，获取目标预测模型。

7. 如权利要求 6 所述数据智能分析方法，其特征在于，所述对至少一个所述中间数据进行池化处理，得到待训练数据，包括：

选取相邻的两个中间数据作为一组待处理数据组，以得到所述中间数据对应的至少一组所述待处理数据组；

对每组所述待处理数据组进行取平均运算，得到第一数据序列；

对每组所述待处理数据组进行最小值运算，得到第二数据序列，所述第二数据列中包括每组所述待处理数据组的两个所述中间数据中的最小值；

对每组所述待处理数据组进行最大值运算，得到第三数据序列，所述第三数据列中包括每组所述待处理数据组的两个所述中间数据中的最大值；

将所述第一数据序列、所述第二数据序列和所述第三数据序列进行拼接，得到所述待训练数据。

8. 一种数据智能分析装置，其特征在于，包括：

舆情数据获取模块，用于按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据；

命中词条确定模块，基于所述舆情数据，确定至少一个命中词条；所述命中词条对应一舆情因子；

舆情指数获取模块，用于获取历史单位时间内的医疗数据和所述命中词条对应的舆情指数；所述舆情指数携带时间标签；

第一画像数据获取模块，用于将所述舆情因子和所述携带时间标签的舆情指数作为第一画像数据；

原始样本数据获取模块，用于基于所述第一画像数据和所述医疗数据，获取原始样本数据；

待处理样本数据获取模块，用于对所述原始样本数据进行数据清洗，得到待处理样本数据；

滞后样本数据获取模块，用于对所述待处理样本数据进行滞后处理，得到滞后样本数据；

目标样本数据获取模块，用于对所述滞后样本数据进行特征扩充处理，获取目标样本数据；

目标预测模型获取模块，用于采用改进多粒度级联随机森林算法对所述目标样本数据进行训练，得到目标预测模型；所述改进多粒度级联随机森林算法包括一池化层，所述池化层用于保留数据特征。

9. 一种计算机设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令，其特征在于，所述处理器执行所述计算机可读指令时实现如下步骤：

按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据；

基于所述舆情数据，确定至少一个命中词条；所述命中词条对应一舆情因子；

获取历史单位时间内的医疗数据和所述命中词条对应的舆情指数；所述舆情指数携带时间标签；

将所述舆情因子和所述携带时间标签的舆情指数作为第一画像数据；

基于所述第一画像数据和所述医疗数据，获取原始样本数据；

对所述原始样本数据进行数据清洗，得到待处理样本数据；

对所述待处理样本数据进行滞后处理，得到滞后样本数据；

对所述滞后样本数据进行特征扩充处理，获取目标样本数据；

采用改进多粒度级联随机森林算法对所述目标样本数据进行训练，得到目标预测模型；所述改进多粒度级联随机森林算法包括一池化层，所述池化层用于保留数据特征。

10. 如权利要求 9 所述的计算机设备，其特征在于，所述按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据之前，所述数据智能分析方法还包括：

获取气象因子和对应的气象数据；

将所述气象因子和对应的气象数据作为第二画像数据；

所述基于所述第一画像数据和所述医疗数据，获取原始样本数据，包括：

将所述第一画像数据、所述第二画像数据和所述医疗数据作为原始样本数据。

11. 如权利要求 9 所述的计算机设备，其特征在于，所述对所述原始样本数据进行数据清洗，得到待处理样本数据，包括；

对所述原始样本数据进行缺失值填充，得到第一样本数据；

对所述第一样本数据进行异常值检测，得到至少一个异常值，将所述异常值标记为空；

对所述标记为空的异常值进行缺失值填充，得到所述待处理样本数据。

12. 如权利要求 9 所述的计算机设备，其特征在于，所述对所述滞后样本数据进行特征扩充处理，获取目标样本数据，包括：

对所述滞后样本数据进行特征扩充，得到至少一个统计指标对应的特征值；

将所述特征值与所述滞后样本数据进行拼接，获取所述目标样本数据。

13. 如权利要求 9 所述的计算机设备，其特征在于，所述改进多粒度级联随机森林算法包括多粒子扫描算法和级联随机森林算法，所述多粒子扫描算法对应至少一个滑动窗口；

所述采用改进多粒度级联随机森林算法对所述目标样本数据进行训练，得到目标预测模型，包括：

采用所述多粒子扫描算法，按照至少一个所述滑动窗口，对所述目标样本数据进行多粒子扫描，得到至少一个中间数据；

基于所述池化层，对至少一个所述中间数据进行池化处理，得到待训练数据；

采用级联随机森林算法对所述待训练数据进行训练，获取目标预测模型。

14. 如权利要求 13 所述的计算机设备，其特征在于，所述对至少一个所述中间数据进行池化处理，得到待训练数据，包括：

选取相邻的两个中间数据作为一组待处理数据组，以得到所述中间数据对应的至少一组所述待处理数据组；

对每组所述待处理数据组进行取平均运算，得到第一数据序列；

对每组所述待处理数据组进行最小值运算，得到第二数据序列，所述第二数据列中包括每组所述待处理数据组的两个所述中间数据中的最小值；

对每组所述待处理数据组进行最大值运算，得到第三数据序列，所述第三数据列中包括每组所述待处理数据组的两个所述中间数据中的最大值；

将所述第一数据序列、所述第二数据序列和所述第三数据序列进行拼接，得到所述待训练数据。

15. 一个或多个存储有计算机可读指令的可读存储介质，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行如下步骤：

按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据；

基于所述舆情数据，确定至少一个命中词条；所述命中词条对应一舆情因子；

获取历史单位时间内的医疗数据和所述命中词条对应的舆情指数；所述舆情指数携带时间标签；

将所述舆情因子和所述携带时间标签的舆情指数作为第一画像数据；

基于所述第一画像数据和所述医疗数据，获取原始样本数据；

对所述原始样本数据进行数据清洗，得到待处理样本数据；

对所述待处理样本数据进行滞后处理，得到滞后样本数据；

对所述滞后样本数据进行特征扩充处理，获取目标样本数据；

采用改进多粒度级联随机森林算法对所述目标样本数据进行训练，得到目标预测模

型；所述改进多粒度级联随机森林算法包括一池化层，所述池化层用于保留数据特征。

16. 如权利要求 15 所述的可读存储介质，其特征在于，所述按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据之前，所述数据智能分析方法还包括：

获取气象因子和对应的气象数据；

将所述气象因子和对应的气象数据作为第二画像数据；

所述基于所述第一画像数据和所述医疗数据，获取原始样本数据，包括：

将所述第一画像数据、所述第二画像数据和所述医疗数据作为原始样本数据。

17. 如权利要求 15 所述的可读存储介质，其特征在于，所述对所述原始样本数据进行数据清洗，得到待处理样本数据，包括；

对所述原始样本数据进行缺失值填充，得到第一样本数据；

对所述第一样本数据进行异常值检测，得到至少一个异常值，将所述异常值标记为空；

对所述标记为空的异常值进行缺失值填充，得到所述待处理样本数据。

18. 如权利要求 15 所述的可读存储介质，其特征在于，所述对所述滞后样本数据进行特征扩充处理，获取目标样本数据，包括：

对所述滞后样本数据进行特征扩充，得到至少一个统计指标对应的特征值；

将所述特征值与所述滞后样本数据进行拼接，获取所述目标样本数据。

19. 如权利要求 15 所述的可读存储介质，其特征在于，所述改进多粒度级联随机森林算法包括多粒子扫描算法和级联随机森林算法，所述多粒子扫描算法对应至少一个滑动窗口；

所述采用改进多粒度级联随机森林算法对所述目标样本数据进行训练，得到目标预测模型，包括：

采用所述多粒子扫描算法，按照至少一个所述滑动窗口，对所述目标样本数据进行多粒子扫描，得到至少一个中间数据；

基于所述池化层，对至少一个所述中间数据进行池化处理，得到待训练数据；

采用级联随机森林算法对所述待训练数据进行训练，获取目标预测模型。

20. 如权利要求 19 所述的可读存储介质，其特征在于，所述对至少一个所述中间数据进行池化处理，得到待训练数据，包括：

选取相邻的两个中间数据作为一组待处理数据组，以得到所述中间数据对应的至少一组所述待处理数据组；

对每组所述待处理数据组进行取平均运算，得到第一数据序列；

对每组所述待处理数据组进行最小值运算，得到第二数据序列，所述第二数据列中包括每组所述待处理数据组的两个所述中间数据中的最小值；

对每组所述待处理数据组进行最大值运算，得到第三数据序列，所述第三数据列中包括每组所述待处理数据组的两个所述中间数据中的最大值；

将所述第一数据序列、所述第二数据序列和所述第三数据序列进行拼接，得到所述待训练数据。

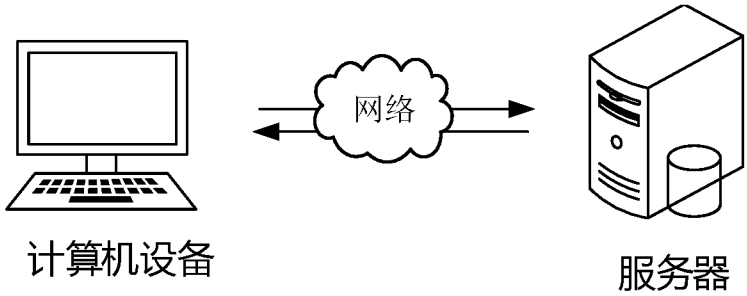


图 1

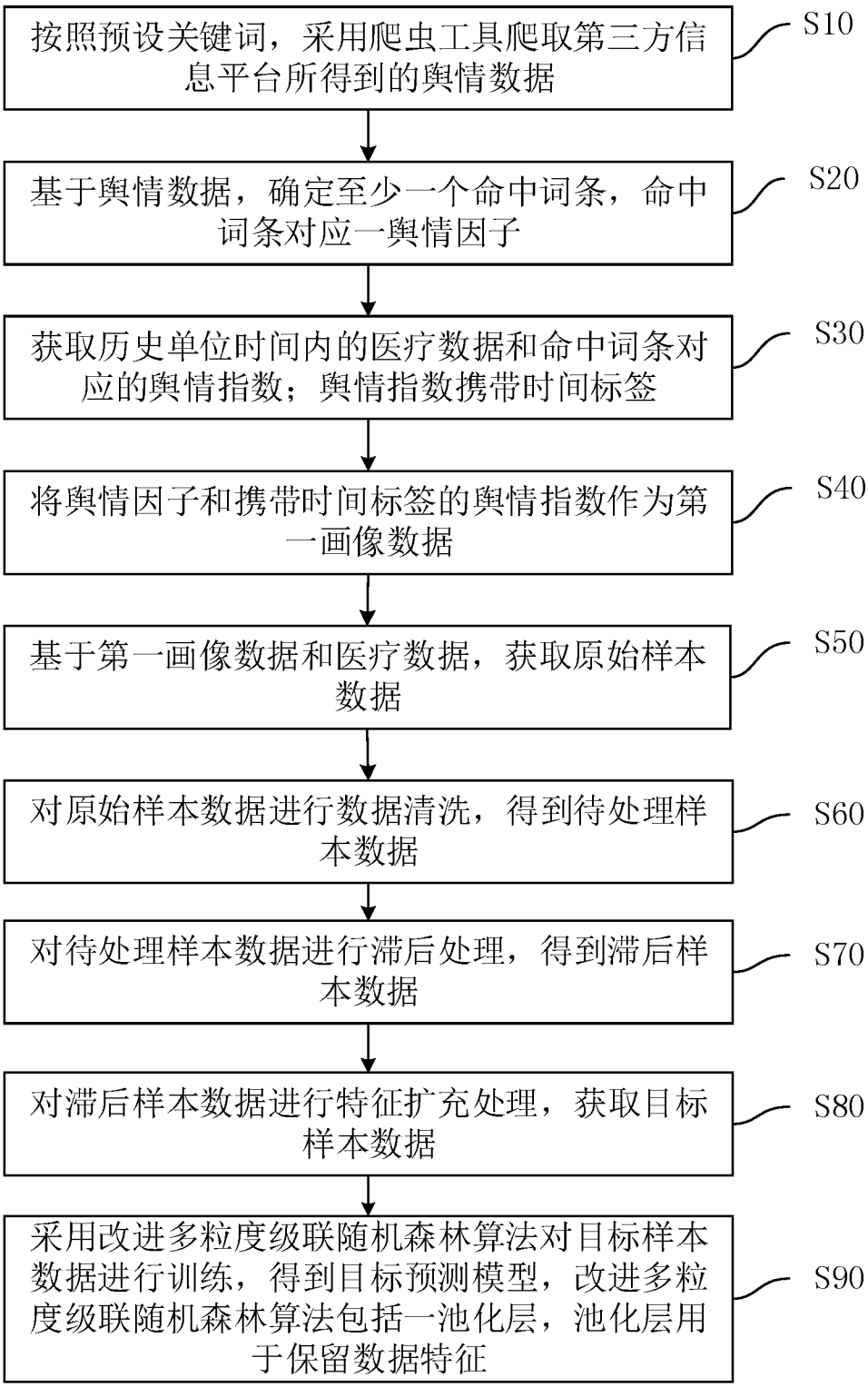

按照预设关键词，采用爬虫工具爬取第三方信息平台所得到的舆情数据
S10
基于舆情数据，确定至少一个命中词条，命中词条对应一舆情因子
S20
获取历史单位时间内的医疗数据和命中词条对应的舆情指数；舆情指数携带时间标签
S30
将舆情因子和携带时间标签的舆情指数作为第一画像数据
S40
基于第一画像数据和医疗数据，获取原始样本数据
S50
对原始样本数据进行数据清洗，得到待处理样本数据
S60
对待处理样本数据进行滞后处理，得到滞后样本数据
S70
对滞后样本数据进行特征扩充处理，获取目标样本数据
S80
采用改进多粒度级联随机森林算法对目标样本数据进行训练，得到目标预测模型，改进多粒度级联随机森林算法包括一池化层，池化层用于保留数据特征
S90

图 2

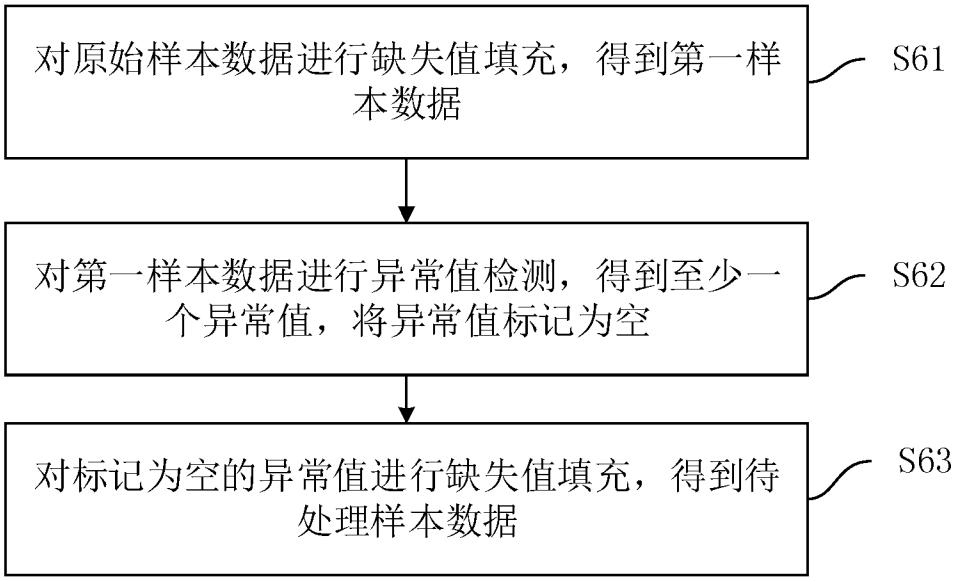

对原始样本数据进行缺失值填充，得到第一样本数据
S61
对第一样本数据进行异常值检测，得到至少一个异常值，将异常值标记为空
S62
对标记为空的异常值进行缺失值填充，得到待处理样本数据
S63

图 3

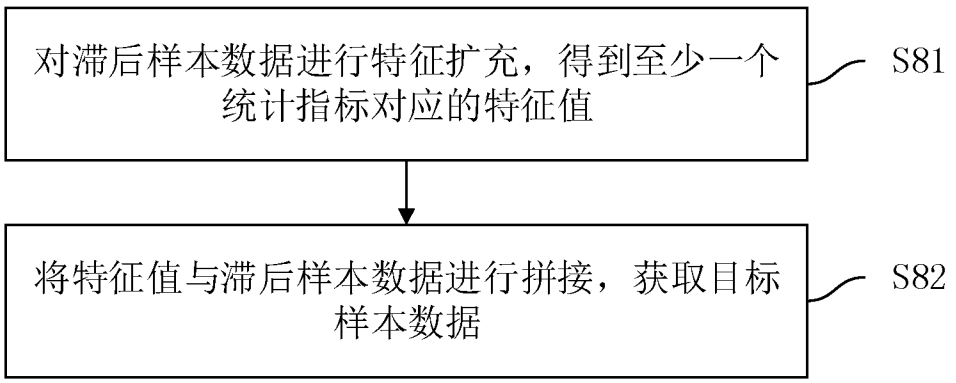

对滞后样本数据进行特征扩充，得到至少一个统计指标对应的特征值
S81
将特征值与滞后样本数据进行拼接，获取目标样本数据
S82

图 4

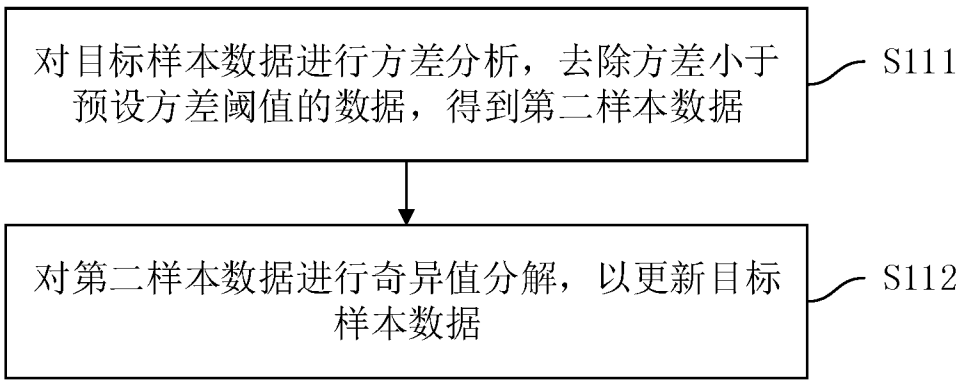

对目标样本数据进行方差分析，去除方差小于预设方差阈值的数据，得到第二样本数据
S111
对第二样本数据进行奇异值分解，以更新目标样本数据
S112

图 5

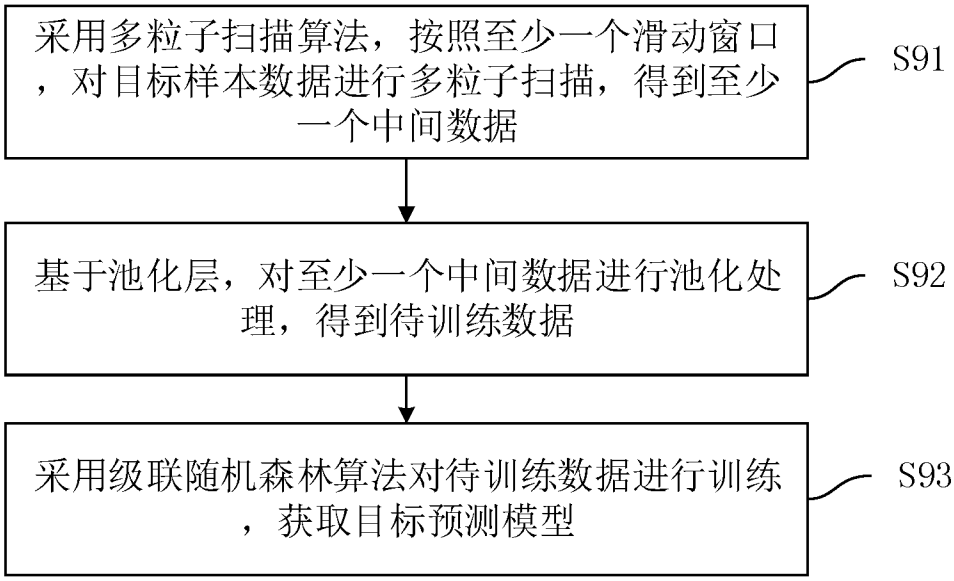

采用多粒子扫描算法，按照至少一个滑动窗口，对目标样本数据进行多粒子扫描，得到至少一个中间数据
S91
基于池化层，对至少一个中间数据进行池化处理，得到待训练数据
S92
采用级联随机森林算法对待训练数据进行训练，获取目标预测模型
S93

图 6

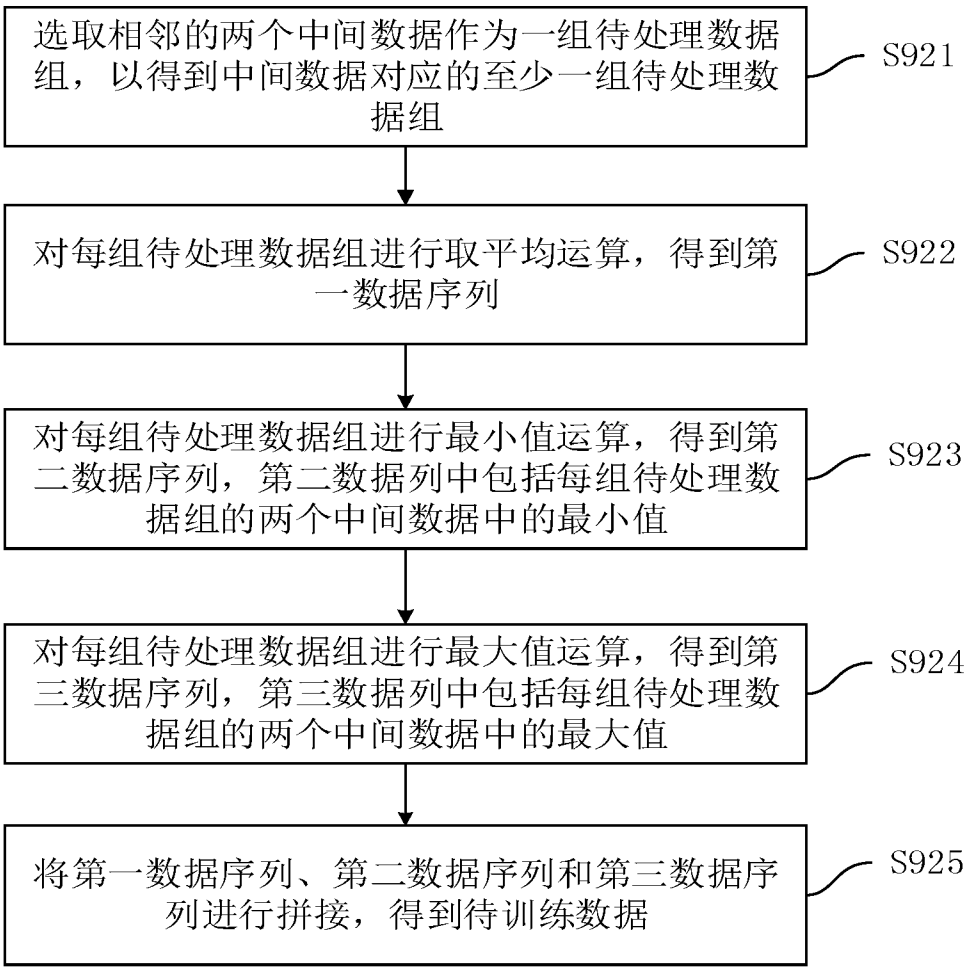

选取相邻的两个中间数据作为一组待处理数据组，以得到中间数据对应的至少一组待处理数据组
S921
对每组待处理数据组进行取平均运算，得到第一数据序列
S922
对每组待处理数据组进行最小值运算，得到第二数据序列，第二数据列中包括每组待处理数据组的两个中间数据中的最小值
S923
对每组待处理数据组进行最大值运算，得到第三数据序列，第三数据列中包括每组待处理数据组的两个中间数据中的最大值
S924
将第一数据序列、第二数据序列和第三数据序列进行拼接，得到待训练数据
S925

图 7

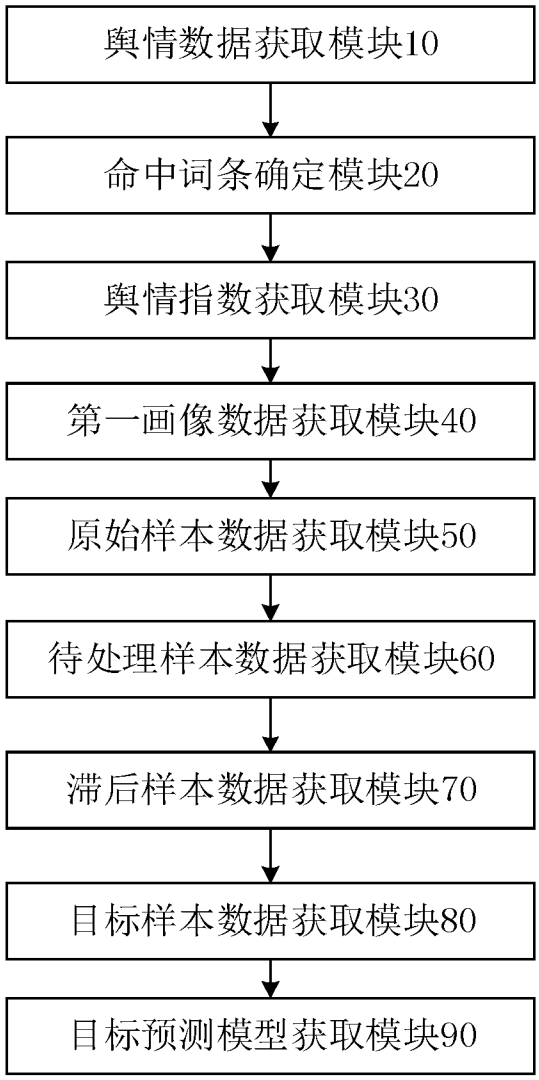

舆情数据获取模块10
命中词条确定模块20
舆情指数获取模块30
第一画像数据获取模块40
原始样本数据获取模块50
待处理样本数据获取模块60
滞后样本数据获取模块70
目标样本数据获取模块80
目标预测模型获取模块90

图 8

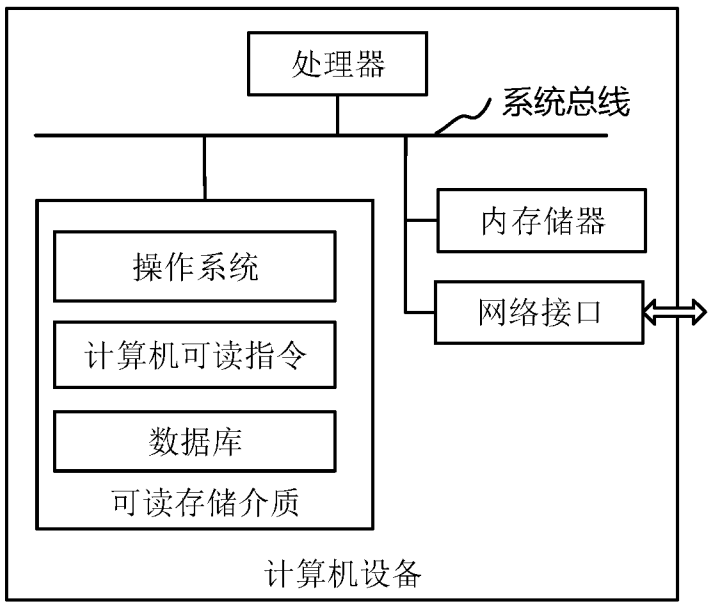

处理器
系统总线
内存储器
网络接口
操作系统
计算机可读指令
数据库
可读存储介质
计算机设备

图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/116942

A. CLASSIFICATION OF SUBJECT MATTER

G16H 50/80(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; G06N; G06Q; G06K

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)

CNPAT, EPODOC, WPI, CNKI: 疾控, 疾病, 疫病, 疫情, 传染病, 流行病, 舆情, 预测, 样本, 填充, 清洗, 滞后, 时间, 随机森林, 深度森林, 粒度, 粒子, 池化, 卷积, 下采样, disease, epidemic, illness, flu, public sentiment, forecast, predict, sample, data cleaning, lagging, time, forest, gcForest, deep, pooling, convolution, down sampling, subsampling, decimation

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 108417274 A (SOUTHEAST UNIVERSITY et al.) 17 August 2018 (2018-08-17) description, paragraphs [0039]-[0079], [0099]-[0102]	1-20
A	CN 108647249 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 12 October 2018 (2018-10-12) entire document	1-20
A	CN 109241987 A (NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 18 January 2019 (2019-01-18) entire document	1-20
A	CN 109656918 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 19 April 2019 (2019-04-19) entire document	1-20
A	WO 2017186048 A1 (4PARADIGM (BEIJING) TECHNOLOGY CO., LTD.) 02 November 2017 (2017-11-02) 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	Date of mailing of the international search report
30 April 2020	**18 May 2020**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/116942

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	108417274	A	17 August 2018	None			
CN	108647249	A	12 October 2018	WO	2019200786	A1	24 October 2019
CN	109241987	A	18 January 2019	None			
CN	109656918	A	19 April 2019	None			
WO	2017186048	A1	02 November 2017	CN	105930934	A	07 September 2016
				US	2019147350	A1	16 May 2019
				EP	3451250	A1	06 March 2019
				CN	108960514	A	07 December 2018

国际检索报告

国际申请号
PCT/CN2019/116942

A. 主题的分类

G16H 50/80(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; G06N; G06Q; G06K

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））

CNPAT,EPODOC,WPI,CNKI: 疾控，疾病，疫病，疫情，传染病，流行病，舆情，预测，样本，填充，清洗，滞后，时间，随机森林，深度森林，粒度，粒子，池化，卷积，下采样，disease, epidemic, illness, flu, public sentiment, forecast, predict, sample, data cleaning, lagging, time, forest, gcForest, deep, pooling, convolution, down sampling, subsampling, decimation

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 108417274 A (东南大学 等) 2018年 8月 17日 (2018-08-17) 说明书第[0039]-[0079]段、第[0099]-[0102]段	1-20
A	CN 108647249 A (平安科技深圳有限公司) 2018年 10月 12日 (2018-10-12) 全文	1-20
A	CN 109241987 A (南京邮电大学) 2019年 1月 18日 (2019-01-18) 全文	1-20
A	CN 109656918 A (平安科技深圳有限公司) 2019年 4月 19日 (2019-04-19) 全文	1-20
A	WO 2017186048 A1 (第四范式北京技术有限公司) 2017年 11月 2日 (2017-11-02) 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

“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	国际检索报告邮寄日期
2020年 4月 30日	2020年 5月 18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	受权官员 孙娟 电话号码 (86-10)53961383

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/116942

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	108417274	A	2018年 8月 17日		无		
CN	108647249	A	2018年 10月 12日	WO	2019200786	A1	2019年 10月 24日
CN	109241987	A	2019年 1月 18日		无		
CN	109656918	A	2019年 4月 19日		无		
WO	2017186048	A1	2017年 11月 2日	CN	105930934	A	2016年 9月 7日
				US	2019147350	A1	2019年 5月 16日
				EP	3451250	A1	2019年 3月 6日
				CN	108960514	A	2018年 12月 7日