

(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
2019 年 2 月 28 日 (28.02.2019)



(10) 国际公布号
WO 2019/037260 A1

- (51) 国际专利分类号:
G06Q 10/04 (2012.01)
- (21) 国际申请号: PCT/CN2017/108801
- (22) 国际申请日: 2017 年 10 月 31 日 (31.10.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201710715445.1 2017 年 8 月 20 日 (20.08.2017) 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八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。
- (72) 发明人: 徐亮(XU, Liang); 中国广东省深圳市福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。 李弦(LI, Xian); 中国广东省深圳市福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。 商瑾(SHANG, Jin); 中国广东省深圳市福田区八卦岭八卦三路平安大厦

厦吴东勤, Guangdong 518000 (CN)。 肖京(XIAO, Jing); 中国广东省深圳市福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理事务所 (普通合伙) (SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦四路 10 号中浩大厦 1528-1530 室于志光, Guangdong 518000 (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, 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,

(54) Title: PREDICTIVE MODEL ESTABLISHMENT METHOD AND DEVICE, AND COMPUTER-READABLE STORAGE MEDIUM

(54) 发明名称: 预测模型建立装置、方法及计算机可读存储介质

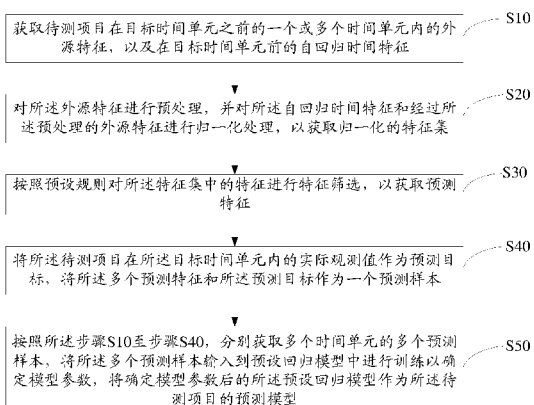


图 3

S10 ACQUIRE AN EXOGENOUS FEATURE OF AN ITEM UNDER PREDICTION IN ONE OR MORE TIME UNITS BEFORE A TARGET TIME UNIT, AND AN AUTO-REGRESSION TIME FEATURE BEFORE THE TARGET TIME UNIT
S20 PERFORM PRE-PROCESS ON THE EXOGENOUS FEATURE AND PERFORM A NORMALIZATION PROCESS ON THE AUTO-REGRESSION TIME FEATURE AND THE PRE-PROCESSED EXOGENOUS FEATURE TO ACQUIRE A NORMALIZED SET OF FEATURES
S30 PERFORM FEATURE FILTERING ACCORDING TO A PRESET RULE ON FEATURES IN THE SET OF FEATURES TO ACQUIRE A PREDICTIVE FEATURE
S40 USE AN ACTUAL OBSERVED VALUE OF THE ITEM UNDER PREDICTION IN THE TARGET TIME UNIT TO BE A PREDICTIVE TARGET AND USE THE PLURALITY OF PREDICTIVE FEATURES AND THE PREDICTIVE TARGET TO BE ONE PREDICTIVE SAMPLE
S50 ACCORDING TO STEPS S10 TO S40, ACQUIRE A PLURALITY OF PREDICTIVE SAMPLES OF A PLURALITY OF TIME UNITS RESPECTIVELY, INPUT THE PLURALITY OF PREDICTIVE SAMPLES TO A PRESET REGRESSION MODEL FOR TRAINING TO DETERMINE A MODEL PARAMETER, AND USE THE PRESET REGRESSION MODEL HAVING THE DETERMINED MODEL PARAMETER TO BE A PREDICTIVE MODEL OF THE ITEM UNDER PREDICTION

(57) Abstract: A predictive model establishment method and device, and a computer-readable storage medium. The device comprises : a memory, a processor and a predictive model establishment procedure stored in the memory and operable by the processor, the procedure, when executed by the processor, performing the following steps : acquiring an exogenous feature of an item under prediction in one or more time units before a target time unit, and an auto-regression time feature before the target time unit; performing pre-process and a normalization process on the features to acquire a normalized set of features; performing feature filtering on the features according to a preset rule to acquire a predictive feature; using an actual observed value of the item under prediction to be a predictive target and using the predictive feature and the predictive target to be a predictive sample; and acquiring a plurality of predictive samples according to the above steps and inputting the predictive samples to a preset regression model for training to generate a predictive model. The present invention increases the prediction accuracy of a predictive model.

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: 一种预测模型建立装置、方法及计算机可读介质, 该装置包括: 存储器、处理器及存储在存储器上并可在处理器上运行的预测模型建立程序, 该程序被处理器执行时实现如下步骤: 获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征, 以及在目标时间单元前的自回归时间特征; 对上述特征进行预处理和归一化处理, 以获取归一化的特征集; 按照预设规则对特征进行特征筛选获取预测特征; 将待测项目的实际观测值作为预测目标, 将预测特征和预测目标作为预测样本; 按照以上步骤获取多个预测样本, 将预测样本输入预设回归模型进行训练生成预测模型。提高了预测模型的预测准确率。

预测模型建立装置、方法及计算机可读存储介质

优先权申明

5 本申请基于巴黎公约申明享有2017年08月20日递交的申请号为201710715445.1、名称为“预测模型建立装置、方法及计算机可读存储介质”的中国专利申请的优先权，该中国专利申请的整体内容以参考的方式结合在本申请中。

技术领域

10 本申请涉及终端技术领域，尤其涉及一种预测模型建立装置、方法及计算机可读存储介质。

背景技术

15 目前，通过机器学习对数据进行预测的技术在越来越多的领域得到应用，例如广告点击率的预测、某种流行性疾病的发病率的预测等等，目前普遍采用的方式时，采集这些待预测项目的历史数据构成时间序列，基于这个时间序列本身的特征建立自相关回归时间序列模型（ARIMA）进行预测，但该模型只用到待预测项目自身的趋势特征进行预测，无法结合预测外部特征使用，导致预测准确性不高。

发明内容

20 本申请提供一种预测模型建立装置、方法及计算机可读存储介质，其主要目的在于结合待测项目的外源特征与自回归时间特征建立预测模型，提高预测模型的预测精准度。

25 为实现上述目的，本申请提供一种预测模型建立装置，该装置包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的预测模型建立程序，所述预测模型建立程序被所述处理器执行时实现如下步骤：

A、获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征，以及在目标时间单元前的自回归时间特征；

30 B、对所述外源特征进行预处理，并对所述自回归时间特征和经过所述预处理的外源特征进行归一化处理，以获取归一化的特征集；

C、按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征；

D、将所述待测项目在所述目标时间单元内的实际观测值作为预测目标，将所述多个预测特征和所述预测目标作为一个预测样本；

35 E、按照所述A至D的步骤分别获取多个时间单元的多个预测样本，将所述多个预测样本输入到预设回归模型中进行训练以确定模型参数，将确定模型参数后的所述预设回归模型作为所述待测项目的预测模型。

可选地，若所述外源特征包括所述待测项目对应的搜索指数集，所述预处理为极值优化处理，则所述对所述外源特征进行预处理的步骤包括：

获取所述搜索指数集的第一四分位数、第三四分位数和四分差；

根据所述第一四分位数、所述第三四分位数和所述四分差确定搜索指数的第一阈值和第二阈值，所述第一阈值小于所述第二阈值；

5 将所述搜索指数集中大于所述第二阈值的搜索指数转换为所述第二阈值，将所述搜索指数集中小于所述第一阈值的搜索指数转换为所述第一阈值

可选地，所述对所述外源特征进行预处理的步骤还包括：

根据邻近 kNN 算法，对经过极值优化处理的搜索指数进行缺失值补充。

可选地，所述按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征的步骤包括：

10 计算所述特征集中的各个特征值的 pearson 相关系数；

将所述特征集中选择 pearson 相关系数小于或者等于预设相关系数的特征，作为预测特征。

可选地，所述待测项目为流感预测项目，所述外源特征包括搜索指数、天气特征和环境特征，所述预设回归模型为 LASSO 回归模型。

15 此外，为实现上述目的，本申请还提供一种预测模型建立方法，该方法包括：

A、获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征，以及在目标时间单元前的自回归时间特征；

20 B、对所述外源特征进行预处理，并对所述自回归时间特征和经过所述预处理的外源特征进行归一化处理，以获取归一化的特征集；

C、按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征；

D、将所述待测项目在所述目标时间单元内的实际观测值作为预测目标，将所述多个预测特征和所述预测目标作为一个预测样本；

25 E、按照所述 A 至 D 的步骤分别获取多个时间单元的多个预测样本，将所述多个预测样本输入到预设回归模型中进行训练以确定模型参数，将确定模型参数后的所述预设回归模型作为所述待测项目的预测模型。

可选地，若所述外源特征包括所述待测项目对应的搜索指数集，所述预处理为极值优化处理，则所述对所述外源特征进行预处理的步骤包括：

30 获取所述搜索指数集的第一四分位数、第三四分位数和四分差；

根据所述第一四分位数、所述第三四分位数和所述四分差确定搜索指数的第一阈值和第二阈值，所述第一阈值小于所述第二阈值；

将所述搜索指数集中大于所述第二阈值的搜索指数转换为所述第二阈值，将所述搜索指数集中小于所述第一阈值的搜索指数转换为所述第一阈值。

35 可选地，所述对所述外源特征进行预处理的步骤还包括：

根据邻近 kNN 算法，对经过极值优化处理的搜索指数进行缺失值补充。

可选地，所述按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征的步骤包括：

计算所述特征集中的各个特征值的 pearson 相关系数;

将所述特征集中选择 pearson 相关系数小于或者等于预设相关系数的特征, 作为预测特征。

5 此外, 为实现上述目的, 本申请还提供一种计算机可读存储介质, 所述计算机可读存储介质上存储有预测模型建立程序, 所述预测模型建立程序可被一个或多个处理器执行, 以实现如下步骤:

A、获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征, 以及在目标时间单元前的自回归时间特征;

10 B、对所述外源特征进行预处理, 并对所述自回归时间特征和经过所述预处理的外源特征进行归一化处理, 以获取归一化的特征集;

C、按照预设规则对所述特征集中的特征进行特征筛选, 以获取预测特征;

D、将所述待测项目在所述目标时间单元内的实际观测值作为预测目标, 将所述多个预测特征和所述预测目标作为一个预测样本;

15 E、按照所述 A 至 D 的步骤分别获取多个时间单元的多个预测样本, 将所述多个预测样本输入到预设回归模型中进行训练以确定模型参数, 将确定模型参数后的所述预设回归模型作为所述待测项目的预测模型。

本申请提出的预测模型建立装置、方法及计算机可读存储介质, 获取待
20 测项目在目标时间单元之前的一个或多个时间单元内的外源特征, 以及在目标时间单元前的自回归时间特征, 对外源特征和自回归时间特征进行预处理和归一化处理, 以获取归一化的特征集, 按照预设规则对特征集中的特征进行筛选得到预测特征, 该预测特征与目标时间单元对应的实际观测值构成一个预测样本, 按照上述过程获取多个时间单元的多个预测样本, 将上述多个
25 预测样本输入到预设回归模型中进行训练以确定模型参数, 将确定模型参数后的预设回归模型作为预测模型, 该发明的方案将待测项目的外源特征与自回归时间特征进行结合并进行预处理构成一个特征集, 从特征集筛选出符合条件的特征作为预测特征对回归模型进行训练生成预测模型, 避免了样本特征的单一性, 提高了预测模型的预测精准度。

30

附图说明

图1为本申请预测模型建立装置较佳实施例的示意图;

图2为本申请预测模型建立装置一实施例中预测模型建立程序的程序模块示意图;

35 图3为本申请预测模型建立方法第一实施例的流程图。

本申请目的的实现、功能特点及优点将结合实施例, 参照附图做进一步说明。

具体实施方式

应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

本申请提供一种预测模型建立装置。参照图 1 所示，为本申请预测模型建立装置较佳实施例的示意图。

在本实施例中，预测模型建立装置可以是 PC (Personal Computer, 个人电脑)，也可以是智能手机、平板电脑、电子书阅读器、MP3 (Moving Picture Experts Group Audio Layer III, 动态影像专家压缩标准音频层面 3) 播放器、MP4 (Moving Picture Experts Group Audio Layer IV, 动态影像专家压缩标准音频层面 4) 播放器、便携计算机等具有显示功能的可移动式终端设备。

该预测模型建立装置包括存储器 11、处理器 12，通信总线 13，以及网络接口 14。

其中，存储器 11 至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、磁性存储器、磁盘、光盘等。存储器 11 在一些实施例中可以是预测模型建立装置的内部存储单元，例如该预测模型建立装置的硬盘。存储器 11 在另一些实施例中也可以是预测模型建立装置的外部存储设备，例如预测模型建立装置上配备的插接式硬盘，智能存储卡 (Smart Media Card, SMC)，安全数字 (Secure Digital, SD) 卡，闪存卡 (Flash Card) 等。进一步地，存储器 11 还可以既包括预测模型建立装置的内部存储单元也包括外部存储设备。存储器 11 不仅可以用于存储安装于预测模型建立装置的应用软件及各类数据，例如预测模型建立程序的代码等，还可以用于暂时地存储已经输出或者将要输出的数据。

处理器 12 在一些实施例中可以是一中央处理器 (Central Processing Unit, CPU)、控制器、微控制器、微处理器或其他数据处理芯片，用于运行存储器 11 中存储的程序代码或处理数据，例如执行预测模型建立程序等。

通信总线 13 用于实现这些组件之间的连接通信。

网络接口 14 可选的可以包括标准的有线接口、无线接口 (如 WI-FI 接口)，通常用于在该装置与其他电子设备之间建立通信连接。

图 1 仅示出了具有组件 11-14 以及预测模型建立程序的预测模型建立装置，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。

可选地，该装置还可以包括用户接口，用户接口可以包括显示器 (Display)、输入单元比如键盘 (Keyboard)，可选的用户接口还可以包括标准的有线接口、无线接口。可选地，在一些实施例中，显示器可以是 LED 显示器、液晶显示器、触控式液晶显示器以及 OLED (Organic Light-Emitting Diode, 有机发光二极管) 触摸器等。其中，显示器也可以适当的称为显示屏或显示单元，用于显示在预测模型建立装置中处理的信息以及用于显示可视化的用户界面。

可选地，该装置还可以包括触摸传感器。所述触摸传感器所提供的供用

户进行触摸操作的区域称为触控区域。此外，这里所述的触摸传感器可以为电阻式触摸传感器、电容式触摸传感器等。而且，所述触摸传感器不仅包括接触式的触摸传感器，也可包括接近式的触摸传感器等。此外，所述触摸传感器可以为单个传感器，也可以为例如阵列布置的多个传感器。该装置的显示器的面积可以与所述触摸传感器的面积相同，也可以不同。可选地，将显示器与所述触摸传感器层叠设置，以形成触摸显示屏。该装置基于触摸显示屏侦测用户触发的触控操作。

可选地，该装置还可以包括摄像头、RF (Radio Frequency, 射频) 电路，传感器、音频电路、WiFi 模块等。其中，传感器比如光传感器、运动传感器以及其他传感器。具体地，光传感器可包括环境光传感器及接近传感器，其中，若该装置为移动终端，环境光传感器可根据环境光线的明暗来调节显示屏的亮度，接近传感器可在移动终端移动到耳边时，关闭显示屏和/或背光。当然，移动终端还可配置陀螺仪、气压计、湿度计、温度计、红外线传感器等其他传感器，在此不再赘述。

在图 1 所示的装置实施例中，存储器 11 中存储有预测模型建立程序；处理器 12 执行存储器 11 中存储的预测模型建立程序时实现如下步骤：

A、获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征，以及在目标时间单元前的自回归时间特征；

B、对所述外源特征进行预处理，并对所述自回归时间特征和经过所述预处理的外源特征进行归一化处理，以获取归一化的特征集；

C、按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征；

D、将所述待测项目在所述目标时间单元内的实际观测值作为预测目标，将所述多个预测特征和所述预测目标作为一个预测样本；

E、按照所述 A 至 D 的步骤分别获取多个时间单元的多个预测样本，将所述多个预测样本输入到预设回归模型中进行训练以确定模型参数，将确定模型参数后的所述预设回归模型作为所述待测项目的预测模型。

在该实施例中，以流感预测项目作为待测项目对本实施例的方案进行说明。由于在对模型进行训练时采用的都是历史数据，即已经过去的时间段内的数据，因此，在从历史数据中采集特征值作为测试样本时，每一个测试样本中的时间单元也是过去的一个时间段。在本实施例中，以周作为一个时间单元。假设将过去的 100 个周的外源特征以及流感样病例百分比作为历史数据，其中，流感样病例百分比为某地区各哨点医院流感样病例总数/某地区各哨点医院门诊总就诊人次。

从上述历史数据中获取一个预测样本的过程如下：从上述历史数据中确定一个周作为目标时间单元，例如，以过去的 100 个周中的第 100 周，即与当前时间点最邻近的一个周，作为目标时间单元，则将第 100 周的流感样病例百分比的实际观测值作为预测目标。从第 100 周之前的多个周的数据中采

集特征作为预测特征。

具体地，获取与目标时间单元之前的一个或者多个时间单元对应的外源特征，优选地，外源特征包括搜索指数、天气特征和环境特征等对流感样病例的百分比造成一定程度的影响的外部特征。

5 搜索指数为：在目标时间单元之前的一个时间单元内，优选地为目标时间单元的上一个时间单元，搜索引擎用户在该搜索引擎上搜索相关关键词的频度，搜索引擎一般会对该频度进行记录，例如，百度的百度指数。其中，相关关键词为用户预先设置的与该流感的预测具有一定因果关系的关键词，例如“鼻塞”、“打喷嚏”、“抗病毒口服液”、“嗓子痛”等与流感相关的检索
10 词，在此不再一一列举，用户可以根据需要设置多个关键词，并从一些搜索网站上按照关键词提取对应的搜索指数构成搜索指数集。

天气特征包括但不限于以下几个特征：每周晴天数、每周多云天数、每周阴天数、每周雨量、每周风量、每周平均中间气温、每周中间气温最大差以及周日中间气温与周一中间气温的差值。其中，每周平均中间气温为一周
15 七天的中间气温的平均值，中间气温为一天内的温度区间的中间值。在获取天气特征时，可以取第 100 周的前一周，第 99 周的天气特征，也可以取第 100 周之前的多个周的天气特征，即历史时间单元的个数可以是多个。

环境特征为 SO₂、NO₂、PM₁₀、O₃、CO 和 PM_{2.5} 的浓度分别在历史时间单元内的均值。

20 对于目标时间单元，即第 100 周的预测目标来说，其自回归时间特征可以包括但不限于以下特征：上一周的流感样病例百分比、上两周的流感样病例百分比、上三周的流感样病例百分比、临近三周的流感样病例百分比均值、临近三周的流感样病例百分比方差、去年同期的流感样病例百分比。

在获取到上述外源特征和自回归时间特征后，对这些特征进行预处理以及
25 归一化处理，以去掉上述特征中的异常特征，避免这些异常特征值对预测模型的训练带来的一些异常影响，进而提高预测模型的预测精准度。

可选地，预处理可以包括极值优化处理和/或缺失值补充处理。

由于搜索指数的数量级比较大，可能出现的数据范围也比较大，因此优选地对搜索指数进行极值优化处理，具体地，获取所述搜索指数集的第一四分位数 Q₁、第三四分位数 Q₃ 和四分差 IQR (interquartile range, 又称为四分
30 位距)；根据所述第一四分位数、所述第三四分位数和所述四分差确定搜索指数的第一阈值和第二阈值，所述第一阈值小于所述第二阈值；将所述搜索指数集中大于所述第二阈值的搜索指数转换为所述第二阈值，将所述搜索指数集中小于所述第一阈值的搜索指数转换为所述第一阈值。上述第一阈值可以
35 取 $Q_1 - k * IQR$ ，上述第二阈值可以取 $Q_3 + k * IQR$ ，其中， $k \geq 0$ 。优选地，在一实施例中， $k = 1.5$ 。

关于缺失值补充处理，可以根据 kNN(k-NearestNeighbor, k 最近邻)算法，对经过极值优化处理的搜索指数进行缺失值补充。具体地，在获取到搜索指

数后,检测搜索指数中是否有关键字没有对应的指数值,若有,则获取与该目标时间单元相邻的多个时间单元中搜索指数没有缺失值时间单元,用这些时间单元的搜索指数中对应关键字的搜索指数的平均值或者中位数或者众数代替缺失的数据。例如,第 99 周的搜索指数中关键字“抗病毒口服液”的搜索指数缺失,而第 98 周、第 97 周、第 95 周的搜索指数中关键字“抗病毒口服液”的搜索指数没有缺失,则可以用这三个周的关键字“抗病毒口服液”的搜索指数的平均值或者中位数或者众数代替第 99 周的搜索指数中缺失的“抗病毒口服液”的搜索指数。在其他实施例中,也可以采用其他方式对缺失值进行补充,例如均值法,取除缺失值之外的其他关键字的搜索指数的平均值,用获取的平均值替代目标单元中的缺失值。

然后,对于自回归时间特征以及经过上述预处理的外源特征进行归一化处理,将上述具有不同量纲的特征值转换为无量纲的符合正态分布的特征值,避免因为特征严重歪曲而造成预测结果的精准度低。

获取到归一化处理后的特征集之后,对特征集中的特征进行筛选。可选地,在一些实施例中,采用 pearson 相关系数 (Pearson correlation coefficient, 皮尔森相关系数)对特征进行过滤。计算所述特征集中的各个特征值的 pearson 相关系数;将所述特征集中选择 pearson 相关系数小于或者等于预设相关系数的特征,作为预测特征。在一些实施例中,预设相关系数优选地取 0.2。可以理解的是,在其他实施例中,也可以采用其他的特征筛选方法对特征集进行筛选。将经过特征筛选后剩余的特征作为预测特征,将目标时间单元的流感样病例百分比的实际观测值作为预测目标,上述预测特征和预测目标构成一个预测样本。

按照上述方式,从历史数据中获取多个预测时间单元对应的预测样本,例如,分别将第 99 周、第 98 周、第 97 周……等作为目标时间单元,分别获取他们的预测样本,将获取的这些预测样本输入到预设回归模型中进行模型训练,得到该模型的模型参数,其中,在一些实施例中,可以采用 k-折交叉验证的方式对模型进行训练,将上述多个预测样本中的一个样本作为测试集,剩余的样本作为训练集,对模型进行训练。或者在其他实施例中,预设的回归模型可以是 LASSO 回归模型、岭回归模型等。在一实施例中,优选地采用 LASSO 回归模型。将确定了模型参数的回归模型作为该待测项目的预测模型。

可以理解的是,上述实施例中,以流感样病例百分比的预测为例对本申请的方案进行了说明。但是本申请提出的方案并不仅限于此,还可以应用于其他的项目的预测,例如,天气的预测、广告点击率的预测等等,对应不同的预测项目,需要采集能够反映该项目的变化的数据作为外源特征,并按照上述方案对外源特征进行处理,然后结合项目本身在时间维度上的自回归时间特征对回归模型进行训练以生成预测模型。

本实施例提出的预测模型建立装置,获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征,以及在目标时间单元前的自回归时间

特征，对外源特征和自回归时间特征进行预处理和归一化处理，以获取归一化的特征集，按照预设规则对特征集中的特征进行筛选得到预测特征，该预测特征与目标时间单元对应的实际观测值构成一个预测样本，按照上述过程获取多个时间单元的多个预测样本，将上述多个预测样本输入到预设回归模型中进行训练以确定模型参数，将确定模型参数后的预设回归模型作为预测模型，该发明的方案将待测项目的外源特征与自回归时间特征进行结合并进行预处理构成一个特征集，从特征集筛选出符合条件的特征作为预测特征对回归模型进行训练生成预测模型，避免了样本特征的单一性，提高了预测模型的预测精准度。

可选地，在其他的实施例，预测模型建立程序还可以被分割为一个或者多个模块，一个或者多个模块被存储于存储器 11 中，并由一个或多个处理器（本实施例为处理器 12）所执行以完成本申请，本申请所称的模块是指能够完成特定功能的一系列计算机程序指令段。

参照图 2 所示，为本申请预测模型建立装置一实施例中的预测模型建立程序的程序模块示意图，该实施例中，预测模型建立程序可以被分割为获取模块 10、处理模块 20、筛选模块 30、形成模块 40 以及训练模块 50，其中：

获取模块 10 用于获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征，以及在目标时间单元前的自回归时间特征；

处理模块 20 用于对所述外源特征进行预处理，并对所述自回归时间特征和经过所述预处理的外源特征进行归一化处理，以获取归一化的特征集；

筛选模块 30 用于按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征；

形成模块 40 用于将所述待测项目在所述目标时间单元内的实际观测值作为预测目标，将所述多个预测特征和所述预测目标作为一个预测样本；

训练模块 50 用于分别获取多个时间单元的多个预测样本，将所述多个预测样本输入到预设回归模型中进行训练以确定模型参数，将确定模型参数后的所述预设回归模型作为所述待测项目的预测模型。

此外，本申请还提供一种预测模型建立方法。参照图 3 所示，为本申请预测模型建立方法第一实施例的流程图。该方法可以由一个装置执行，该装置可以由软件和/或硬件实现。

在本实施例中，预测模型建立方法包括：

步骤 S10，获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征，以及在目标时间单元前的自回归时间特征；

步骤 S20，对所述外源特征进行预处理，并对所述自回归时间特征和经过所述预处理的外源特征进行归一化处理，以获取归一化的特征集；

步骤 S30，按照预设规则对所述特征集中的特征进行特征筛选，以获取预

测特征;

步骤 S40, 将所述待测项目在所述目标时间单元内的实际观测值作为预测目标, 将所述多个预测特征和所述预测目标作为一个预测样本;

5 步骤 S50, 按照所述步骤 S10 至步骤 S40, 分别获取多个时间单元的多个预测样本, 将所述多个预测样本输入到预设回归模型中进行训练以确定模型参数, 将确定模型参数后的所述预设回归模型作为所述待测项目的预测模型。

在该实施例中, 以流感预测项目作为待测项目对本实施例的方案进行说明。由于在对模型进行训练时采用的都是历史数据, 即已经过去的时间段内的数据, 因此, 在从历史数据中采集特征值作为测试样本时, 每一个测试样
10 本中的时间单元也是过去的一个时间段。在本实施例中, 以周作为一个时间单元。假设将过去的 100 个周的外源特征以及流感样病例百分比作为历史数据, 其中, 流感样病例百分比为某地区各哨点医院流感样病例总数/某地区各哨点医院门诊总就诊人次。

从上述历史数据中获取一个预测样本的过程如下: 从上述历史数据中确
15 定一个周作为目标时间单元, 例如, 以过去的 100 个周中的第 100 周, 即与当前时间点最邻近的一个周, 作为目标时间单元, 则将第 100 周的流感样病例百分比的实际观测值作为预测目标。从第 100 周之前的多个周的数据中采集特征作为预测特征。

具体地, 获取与目标时间单元之前的一个或者多个时间单元对应的外源
20 特征, 优选地, 外源特征包括搜索指数、天气特征和环境特征等对流感样病例的百分比造成一定程度的影响的外部特征。

搜索指数为: 在目标时间单元之前的一个时间单元内, 优选地为目标时
间单元的上一个时间单元, 搜索引擎用户在该搜索引擎上搜索相关关键词的
25 频度, 搜索引擎一般会对该频度进行记录, 例如, 百度的百度指数。其中, 相关关键词为用户预先设置的与该流感的预测具有一定因果关系的关键词, 例如“鼻塞”、“打喷嚏”、“抗病毒口服液”、“嗓子痛”等与流感相关的检索词, 在此不再一一列举, 用户可以根据需要设置多个关键词, 并从一些搜索网站上按照关键词提取对应的搜索指数构成搜索指数集。

天气特征包括但不限于以下几个特征: 每周晴天数、每周多云天数、每
30 周阴天数、每周雨量、每周风量、每周平均中间气温、每周中间气温最大差以及周日中间气温与周一中间气温的差值。其中, 每周平均中间气温为一周七天的中间气温的平均值, 中间气温为一天内的温度区间的中间值。在获取天气特征时, 可以取第 100 周的前一周, 第 99 周的天气特征, 也可以取第 100 周之前的多个周的天气特征, 即历史时间单元的个数可以是多个。

35 环境特征为 SO₂、NO₂、PM₁₀、O₃、CO 和 PM_{2.5} 的浓度分别在历史时间单元内的均值。

对于目标时间单元, 即第 100 周的预测目标来说, 其自回归时间特征可以包括但不限于以下特征: 上一周的流感样病例百分比、上两周的流感样病

例百分比、上三周的流感样病例百分比、临近三周的流感样病例百分比均值、临近三周的流感样病例百分比方差、去年同期的流感样病例百分比。

在获取到上述外源特征和自回归时间特征后，对这些特征进行预处理以及归一化处理，以去掉上述特征中的异常特征，避免这些异常特征值对预测模型的训练带来的一些异常影响，进而提高预测模型的预测精准度。

可选地，预处理可以包括极值优化处理和/或缺失值补充处理。

由于搜索指数的数量级比较大，可能出现的数据范围也比较大，因此优选地对搜索指数进行极值优化处理，具体地，获取所述搜索指数集的第一四分位数 $Q1$ 、第三四分位数 $Q3$ 和四分差 IQR (interquartile range, 又称为四分位距); 根据所述第一四分位数、所述第三四分位数和所述四分差确定搜索指数的第一阈值和第二阈值，所述第一阈值小于所述第二阈值; 将所述搜索指数集中大于所述第二阈值的搜索指数转换为所述第二阈值，将所述搜索指数集中小于所述第一阈值的搜索指数转换为所述第一阈值。上述第一阈值可以取 $Q1-k*IQR$ ，上述第二阈值可以取 $Q3+k*IQR$ ，其中， $k \geq 0$ 。优选地，在一实施例中， $k=1.5$ 。

关于缺失值补充处理，可以根据邻近 kNN 算法，对经过极值优化处理的搜索指数进行缺失值补充。具体地，在获取到搜索指数后，检测搜索指数中是否有关键字没有对应的指数值，若有，则获取与该目标时间单元相邻的多个时间单元中搜索指数没有缺失值时间单元，用这些时间单元的搜索指数中对应关键字的搜索指数的平均值或者中位数或者众数代替缺失的数据。例如，第 99 周的搜索指数中关键字“抗病毒口服液”的搜索指数缺失，而第 98 周、第 97 周、第 95 周的搜索指数中关键字“抗病毒口服液”的搜索指数没有缺失，则可以用这三个周的关键字“抗病毒口服液”的搜索指数的平均值或者中位数或者众数代替第 99 周的搜索指数中缺失的“抗病毒口服液”的搜索指数。在其他实施例中，也可以采用其他方式对缺失值进行补充，例如均值法，取除缺失值之外的其他关键字的搜索指数的平均值，用获取的平均值替代目标单元中的缺失值。

然后，对于自回归时间特征以及经过上述预处理的外源特征进行归一化处理，将上述具有不同量纲的特征值转换为无量纲的符合正态分布的特征值，避免因为特征严重歪曲而造成预测结果的精准度低。

获取到归一化处理后的特征集之后，对特征集中的特征进行筛选。可选地，在一些实施例中，采用 pearson 相关系数对特征进行过滤。计算所述特征集中的各个特征值的 pearson 相关系数; 将所述特征集中选择 pearson 相关系数小于或者等于预设相关系数的特征，作为预测特征。在一些实施例中，预设相关系数优选地取 0.2。可以理解的是，在其他实施例中，也可以采用其他的特征筛选方法对特征集进行筛选。将经过特征筛选后剩余的特征作为预测特征，将目标时间单元的流感样病例百分比的实际观测值作为预测目标，上述预测特征和预测目标构成一个预测样本。

按照上述方式，从历史数据中获取多个预测时间单元对应的预测样本，例如，分别将第 99 周、第 98 周、第 97 周……等作为目标时间单元，分别获取他们的预测样本，将获取的这些预测样本输入到预设回归模型中进行模型训练，得到该模型的模型参数，其中，在一些实施例，可以采用 k-折交叉验证的方式对模型进行训练，将上述多个预测样本中的一个样本作为测试集，剩余的样本作为训练集，对模型进行训练。或者在其他实施例，预设的回归模型可以是 LASSO 回归模型、岭回归模型等。在一实施例中，优选地采用 LASSO 回归模型。将确定了模型参数的回归模型作为该待测项目的预测模型。

可以理解的是，上述实施例中，以流感样病例百分比的预测为例对本申请的方案进行了说明。但是本申请提出的方案并不仅限于此，还可以应用于其他的项目的预测，例如，天气的预测、广告点击率的预测等等，对应不同的预测项目，需要采集能够反映该项目的变化的数据作为外源特征，并按照上述方案对外源特征进行处理，然后结合项目本身在时间维度上的自回归时间特征对回归模型进行训练以生成预测模型。该预测模型可以用于对未来一段时间内待测项目的预测。

本实施例提出的预测模型建立方法，获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征，以及在目标时间单元前的自回归时间特征，对外源特征和自回归时间特征进行预处理和归一化处理，以获取归一化的特征集，按照预设规则对特征集中的特征进行筛选得到预测特征，该预测特征与目标时间单元对应的实际观测值构成一个预测样本，按照上述过程获取多个时间单元的多个预测样本，将上述多个预测样本输入到预设回归模型中进行训练以确定模型参数，将确定模型参数后的预设回归模型作为预测模型，该发明的方案将待测项目的外源特征与自回归时间特征进行结合并进行预处理构成一个特征集，从特征集筛选出符合条件的特征作为预测特征对回归模型进行训练生成预测模型，避免了样本特征的单一性，提高了预测模型的预测精度。

此外，本申请实施例还提出一种计算机可读存储介质，所述计算机可读存储介质上存储有预测模型建立程序，所述预测模型建立程序可被一个或多个处理器执行，以实现如下步骤：

A、获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征，以及在目标时间单元前的自回归时间特征；

B、对所述外源特征进行预处理，并对所述自回归时间特征和经过所述预处理的外源特征进行归一化处理，以获取归一化的特征集；

C、按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征；

D、将所述待测项目在所述目标时间单元内的实际观测值作为预测目标，将所述多个预测特征和所述预测目标作为一个预测样本；

E、按照所述 A 至 D 的步骤分别获取多个时间单元的多个预测样本，将

所述多个预测样本输入到预设回归模型中进行训练以确定模型参数，将确定模型参数后的所述预设回归模型作为所述待测项目的预测模型。

进一步地，所述预测模型建立程序被处理器执行时还实现如下操作：

获取所述搜索指数集的第一四分位数、第三四分位数和四分差；

5 根据所述第一四分位数、所述第三四分位数和所述四分差确定搜索指数的第一阈值和第二阈值，所述第一阈值小于所述第二阈值；

将所述搜索指数集中大于所述第二阈值的搜索指数转换为所述第二阈值，将所述搜索指数集中小于所述第一阈值的搜索指数转换为所述第一阈值。

进一步地，所述预测模型建立程序被处理器执行时还实现如下操作：

10 根据邻近 kNN 算法，对经过极值优化处理的搜索指数进行缺失值补充。

进一步地，所述预测模型建立程序被处理器执行时还实现如下操作：

计算所述特征集中的各个特征值的 pearson 相关系数；

将所述特征集中选择 pearson 相关系数小于或者等于预设相关系数的特征，作为预测特征。

15 需要说明的是，上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。并且本文中的术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、装置、物品或者方法不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、装置、物品或者方法所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

20 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在如上所述的一个存储介质(如 ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，或者网络设备等等)执行本申请各个实施例所述的方法。

30 以上仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权利要求书

1、一种预测模型建立装置，其特征在于，所述装置包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的预测模型建立程序，所述预测模型建立程序被所述处理器执行时实现如下步骤：

A、获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征，以及在目标时间单元前的自回归时间特征；

B、对所述外源特征进行预处理，并对所述自回归时间特征和经过所述预处理的外源特征进行归一化处理，以获取归一化的特征集；

C、按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征；

D、将所述待测项目在所述目标时间单元内的实际观测值作为预测目标，将所述多个预测特征和所述预测目标作为一个预测样本；

E、按照所述 A 至 D 的步骤分别获取多个时间单元的多个预测样本，将所述多个预测样本输入到预设回归模型中进行训练以确定模型参数，将确定模型参数后的所述预设回归模型作为所述待测项目的预测模型。

2、根据权利要求 1 所述的预测模型建立装置，其特征在于，若所述外源特征包括所述待测项目对应的搜索指数集，所述预处理为极值优化处理，则所述对所述外源特征进行预处理的步骤包括：

获取所述搜索指数集的第一四分位数、第三四分位数和四分差；

根据所述第一四分位数、所述第三四分位数和所述四分差确定搜索指数的第一阈值和第二阈值，所述第一阈值小于所述第二阈值；

将所述搜索指数集中大于所述第二阈值的搜索指数转换为所述第二阈值，将所述搜索指数集中小于所述第一阈值的搜索指数转换为所述第一阈值。

3、根据权利要求 2 所述的预测模型建立装置，其特征在于，所述对所述外源特征进行预处理的步骤还包括：

根据邻近 kNN 算法，对经过极值优化处理的搜索指数进行缺失值补充。

4、根据权利要求 1 所述的预测模型建立装置，其特征在于，所述按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征的步骤包括：

计算所述特征集中的各个特征值的 pearson 相关系数；

将所述特征集中选择 pearson 相关系数小于或者等于预设相关系数的特征，作为预测特征。

5、根据权利要求 2 所述的预测模型建立装置，其特征在于，所述按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征的步骤包括：

计算所述特征集中的各个特征值的 pearson 相关系数；

将所述特征集中选择 pearson 相关系数小于或者等于预设相关系数的特征，

作为预测特征。

6、根据权利要求3所述的预测模型建立装置，其特征在于，所述按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征的步骤包括：

5 计算所述特征集中的各个特征值的pearson相关系数；

将所述特征集中选择pearson相关系数小于或者等于预设相关系数的特征，作为预测特征。

10 7、根据权利要求1所述的预测模型建立装置，其特征在于，所述待测项目为流感预测项目，所述外源特征包括搜索指数、天气特征和环境特征，所述预设回归模型为LASSO回归模型。

8、一种预测模型建立方法，其特征在于，所述方法包括：

15 A、获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征，以及在目标时间单元前的自回归时间特征；

B、对所述外源特征进行预处理，并对所述自回归时间特征和经过所述预处理的外源特征进行归一化处理，以获取归一化的特征集；

C、按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征；

20 D、将所述待测项目在所述目标时间单元内的实际观测值作为预测目标，将所述多个预测特征和所述预测目标作为一个预测样本；

E、按照所述A至D的步骤分别获取多个时间单元的多个预测样本，将所述多个预测样本输入到预设回归模型中进行训练以确定模型参数，将确定模型参数后的所述预设回归模型作为所述待测项目的预测模型。

25 9、根据权利要求8所述的预测模型建立方法，其特征在于，若所述外源特征包括所述待测项目对应的搜索指数集，所述预处理为极值优化处理，则所述对所述外源特征进行预处理的步骤包括：

获取所述搜索指数集的第一四分位数、第三四分位数和四分差；

30 根据所述第一四分位数、所述第三四分位数和所述四分差确定搜索指数的第一阈值和第二阈值，所述第一阈值小于所述第二阈值；

将所述搜索指数集中大于所述第二阈值的搜索指数转换为所述第二阈值，将所述搜索指数集中小于所述第一阈值的搜索指数转换为所述第一阈值。

35 10、根据权利要求9所述的预测模型建立方法，其特征在于，所述对所述外源特征进行预处理的步骤还包括：

根据邻近kNN算法，对经过极值优化处理的搜索指数进行缺失值补充。

11、根据权利要求8所述的预测模型建立方法，其特征在于，所述按照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征的步骤包括：

计算所述特征集中的各个特征值的 pearson 相关系数;

将所述特征集中选择 pearson 相关系数小于或者等于预设相关系数的特征, 作为预测特征。

- 5 12、根据权利要求 9 所述的预测模型建立方法, 其特征在于, 所述按照预设规则对所述特征集中的特征进行特征筛选, 以获取预测特征的步骤包括:
计算所述特征集中的各个特征值的 pearson 相关系数;
将所述特征集中选择 pearson 相关系数小于或者等于预设相关系数的特征, 作为预测特征。

- 10 13、根据权利要求 10 所述的预测模型建立方法, 其特征在于, 所述按照预设规则对所述特征集中的特征进行特征筛选, 以获取预测特征的步骤包括:
计算所述特征集中的各个特征值的 pearson 相关系数;
将所述特征集中选择 pearson 相关系数小于或者等于预设相关系数的特征,
15 作为预测特征。

- 20 14、根据权利要求 8 所述的预测模型建立方法, 其特征在于, 所述待测项目为流感预测项目, 所述外源特征包括搜索指数、天气特征和环境特征, 所述预设回归模型为 LASSO 回归模型。

- 25 15、一种计算机可读存储介质, 其特征在于, 所述计算机可读存储介质上存储有预测模型建立程序, 所述预测模型建立程序可被一个或多个处理器执行, 以实现如下步骤:

- 25 A、获取待测项目在目标时间单元之前的一个或多个时间单元内的外源特征, 以及在目标时间单元前的自回归时间特征;

B、对所述外源特征进行预处理, 并对所述自回归时间特征和经过所述预处理的外源特征进行归一化处理, 以获取归一化的特征集;

C、按照预设规则对所述特征集中的特征进行特征筛选, 以获取预测特征;

- 30 D、将所述待测项目在所述目标时间单元内的实际观测值作为预测目标, 将所述多个预测特征和所述预测目标作为一个预测样本;

E、按照所述 A 至 D 的步骤分别获取多个时间单元的多个预测样本, 将所述多个预测样本输入到预设回归模型中进行训练以确定模型参数, 将确定模型参数后的所述预设回归模型作为所述待测项目的预测模型。

- 35 16、根据权利要求 15 所述的计算机可读存储介质, 其特征在于, 若所述外源特征包括所述待测项目对应的搜索指数集, 所述预处理为极值优化处理, 则所述对所述外源特征进行预处理的步骤包括:

获取所述搜索指数集的第一四分位数、第三四分位数和四分差;

根据所述第一四分位数、所述第三四分位数和所述四分差确定搜索指数

的第一阈值和第二阈值，所述第一阈值小于所述第二阈值；

将所述搜索指数集中大于所述第二阈值的搜索指数转换为所述第二阈值，
将所述搜索指数集中小于所述第一阈值的搜索指数转换为所述第一阈值。

- 5 17、根据权利要求 16 所述的计算机可读存储介质，其特征在于，所述对
所述外源特征进行预处理的步骤还包括：

根据邻近 kNN 算法，对经过极值优化处理的搜索指数进行缺失值补充。

- 10 18、根据权利要求 15 所述的计算机可读存储介质，其特征在于，所述按
照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征的步骤包
括：

计算所述特征集中的各个特征值的 pearson 相关系数；

将所述特征集中选择 pearson 相关系数小于或者等于预设相关系数的特征，
作为预测特征。

- 15 19、根据权利要求 16 所述的计算机可读存储介质，其特征在于，所述按
照预设规则对所述特征集中的特征进行特征筛选，以获取预测特征的步骤包
括：

计算所述特征集中的各个特征值的 pearson 相关系数；

- 20 将所述特征集中选择 pearson 相关系数小于或者等于预设相关系数的特征，
作为预测特征。

- 25 20、根据权利要求 15 所述的计算机可读存储介质，其特征在于，所述待
测项目为流感预测项目，所述外源特征包括搜索指数、天气特征和环境特征，
所述预设回归模型为 LASSO 回归模型。

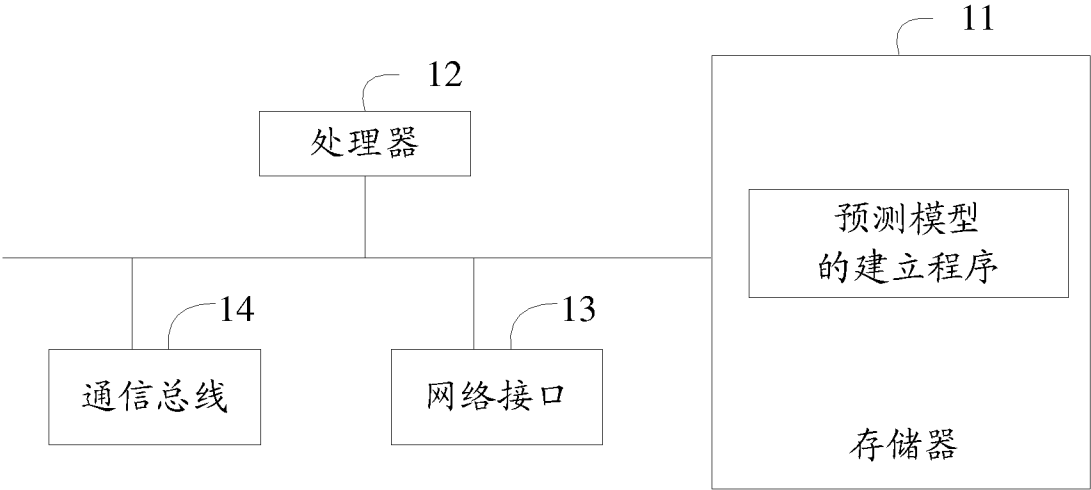


图 1

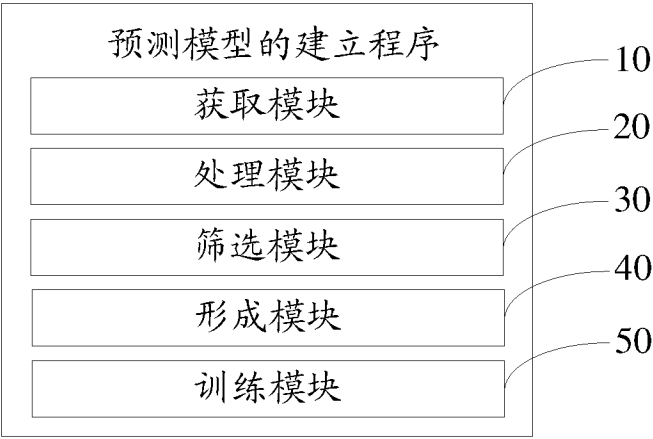


图 2

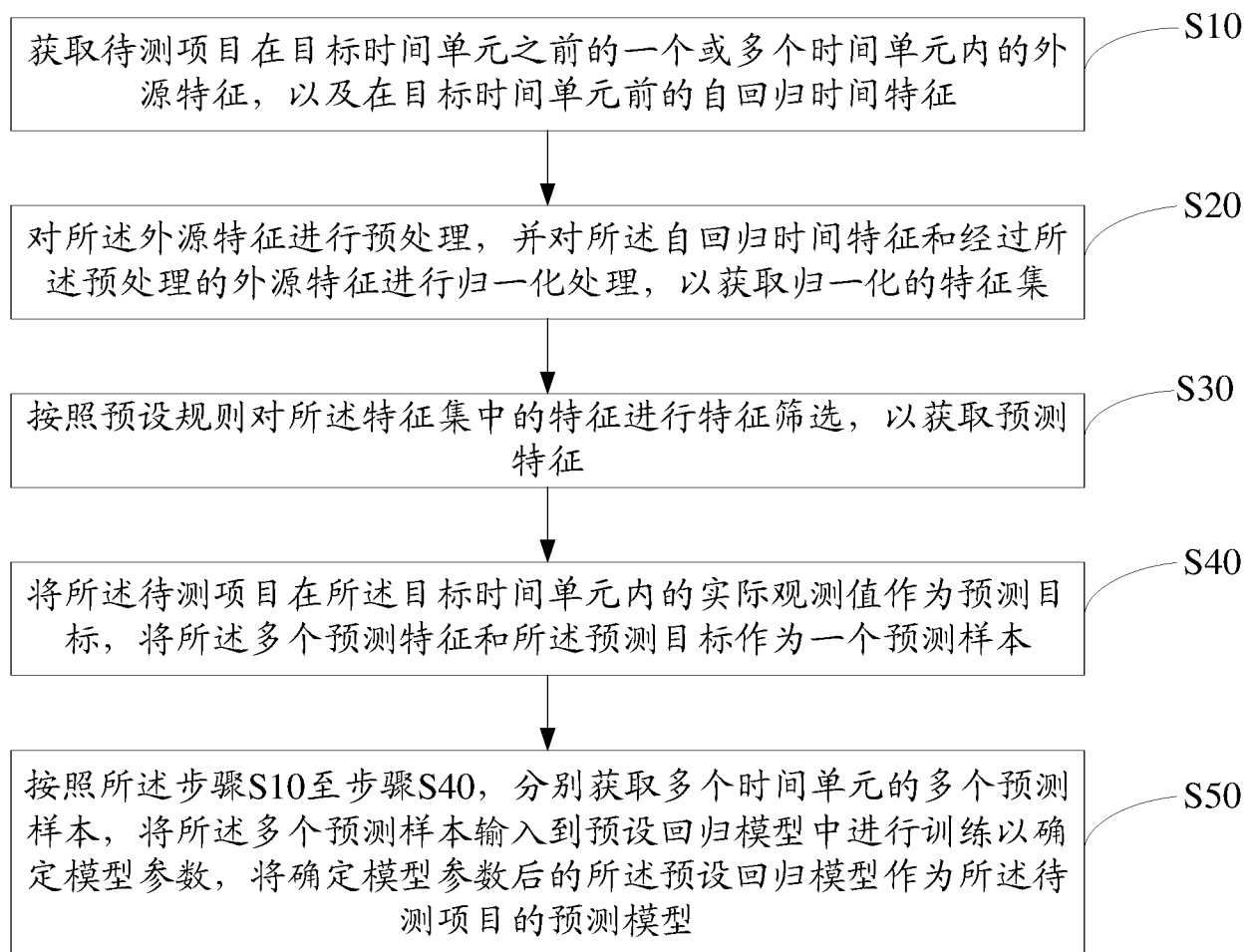


图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/108801

A. CLASSIFICATION OF SUBJECT MATTER

G06Q 10/04 (2012.01) 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)

G06Q

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, CNKI, WPI, EPODOC, IEEE: 预测, 模型, 训练, 回归, 特征, 筛选, 参数, 归一化, prediction, model, train, characteristic, select, parameter, normalization

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 106777891 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES) 31 May 2017 (31.05.2017), description, paragraphs [0040]-[0085], and figure 1	1, 4, 7, 8, 11, 14, 15, 18, 20
A	CN 106709588 A (NEC CORPORATION) 24 May 2017 (24.05.2017), entire document	1-20
A	CN 106777874 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES) 31 May 2017 (31.05.2017), entire document	1-20
A	CN 104809335 A (SHANGHAI HEALTH INFORMATION ENGINEERING TECHNOLOGY RESEARCH CENTER CO., LTD. et al.) 29 July 2015 (29.07.2015), entire document	1-20
A	CN 106980914 A (XIAMEN MEIYOU INFORMATION TECHNOLOGY CO., LTD.) 25 July 2017 (25.07.2017), entire document	1-20

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"&" document member of the same patent family
"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	

Date of the actual completion of the international search 02 May 2018	Date of mailing of the international search report 21 May 2018
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer ZHAO, Xiaomin Telephone No. (86-10) 53961332

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/108801

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 106777891 A	31 May 2017	None	
CN 106709588 A	24 May 2017	None	
CN 106777874 A	31 May 2017	None	
CN 104809335 A	29 July 2015	None	
CN 106980914 A	25 July 2017	None	

国际检索报告

国际申请号

PCT/CN2017/108801

A. 主题的分类 G06Q 10/04 (2012.01) 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) G06Q 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, CNKI, WPI, EPODOC, IEEE: 预测、模型、训练、回归、特征、筛选、参数、归一化、prediction、model、train、characteristic、select、parameter、normalization																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 106777891 A (中国科学院自动化研究所) 2017年 5月 31日 (2017 - 05 - 31) 说明书第[0040]-[0085]段, 图1</td> <td>1、4、7-8、11、14-15、18、20</td> </tr> <tr> <td>A</td> <td>CN 106709588 A (日本电气株式会社) 2017年 5月 24日 (2017 - 05 - 24) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 106777874 A (中国科学院自动化研究所) 2017年 5月 31日 (2017 - 05 - 31) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 104809335 A (上海卫生信息工程技术研究中心有限公司 等) 2015年 7月 29日 (2015 - 07 - 29) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 106980914 A (厦门美柚信息科技有限公司) 2017年 7月 25日 (2017 - 07 - 25) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 106777891 A (中国科学院自动化研究所) 2017年 5月 31日 (2017 - 05 - 31) 说明书第[0040]-[0085]段, 图1	1、4、7-8、11、14-15、18、20	A	CN 106709588 A (日本电气株式会社) 2017年 5月 24日 (2017 - 05 - 24) 全文	1-20	A	CN 106777874 A (中国科学院自动化研究所) 2017年 5月 31日 (2017 - 05 - 31) 全文	1-20	A	CN 104809335 A (上海卫生信息工程技术研究中心有限公司 等) 2015年 7月 29日 (2015 - 07 - 29) 全文	1-20	A	CN 106980914 A (厦门美柚信息科技有限公司) 2017年 7月 25日 (2017 - 07 - 25) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
X	CN 106777891 A (中国科学院自动化研究所) 2017年 5月 31日 (2017 - 05 - 31) 说明书第[0040]-[0085]段, 图1	1、4、7-8、11、14-15、18、20																		
A	CN 106709588 A (日本电气株式会社) 2017年 5月 24日 (2017 - 05 - 24) 全文	1-20																		
A	CN 106777874 A (中国科学院自动化研究所) 2017年 5月 31日 (2017 - 05 - 31) 全文	1-20																		
A	CN 104809335 A (上海卫生信息工程技术研究中心有限公司 等) 2015年 7月 29日 (2015 - 07 - 29) 全文	1-20																		
A	CN 106980914 A (厦门美柚信息科技有限公司) 2017年 7月 25日 (2017 - 07 - 25) 全文	1-20																		
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																				
<table border="0"> <tr> <td style="vertical-align: top;"> * 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 </td> <td style="vertical-align: top;"> “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 </td> </tr> </table>			* 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																
* 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																		
2018年 5月 2日		2018年 5月 21日																		
ISA/CN的名称和邮寄地址		受权官员																		
中华人民共和国国家知识产权局 (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		赵晓敏 电话号码 86-(10)-53961332																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/108801

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	106777891	A	2017年 5月 31日	无	
CN	106709588	A	2017年 5月 24日	无	
CN	106777874	A	2017年 5月 31日	无	
CN	104809335	A	2015年 7月 29日	无	
CN	106980914	A	2017年 7月 25日	无	