(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108009674 A
(43)申请公布日 2018.05.08

(21)申请号 201711203233.1

(22)申请日 2017.11.27

(71)申请人 上海师范大学
地址 200234 上海市徐汇区桂林路100号

(72)发明人 张波 雍睿涵 李美子 赵勤 秦东明

(74)专利代理机构 上海科盛知识产权代理有限公司 31225
代理人 赵志远

(51)Int.Cl.
G06Q 10/04(2012.01)
G06Q 50/26(2012.01)
G06N 3/04(2006.01)
G01N 15/06(2006.01)

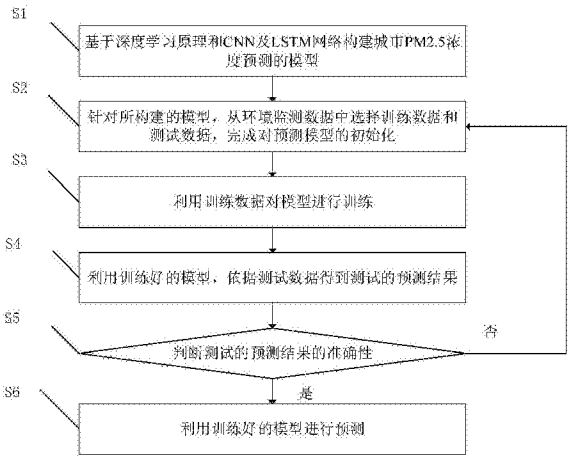
权利要求书2页 说明书6页 附图3页

(54)发明名称

基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法

(57)摘要

本发明涉及一种基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法，包括：步骤S1：基于深度学习原理和CNN及LSTM构建城市PM2.5浓度预测的模型；步骤S2：针对所构建的模型，从环境监测数据中选择训练数据和测试数据，完成对预测模型的初始化；步骤S3：利用训练数据对模型进行训练；步骤S4：利用训练好的模型，依据测试数据得到测试的预测结果；步骤S5：判断测试的预测结果的准确性，若准确性超过阈值，则执行步骤S6，若为否，则返回步骤S2；步骤S6：利用训练好的模型进行预测。与现有技术相比，本发明预测的准确度较传统的预测方法高，在同样的工作时长和工作条件下，可以产生更好的结果。



CN 108009674 A

1.一种基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法，其特征在于，包括：

步骤S1：基于深度学习原理和CNN及LSTM构建城市PM2.5浓度预测的模型；

步骤S2：针对所构建的模型，从环境监测数据中选择训练数据和测试数据，完成对预测模型的初始化；

步骤S3：利用训练数据对模型进行训练；

步骤S4：利用训练好的模型，依据测试数据得到测试的预测结果；

步骤S5：判断测试的预测结果的准确性，若准确性超过阈值，则执行步骤S6，若为否，则返回步骤S2；

步骤S6：利用训练好的模型进行预测。

2.根据权利要求1所述的一种基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法，其特征在于，所述模型包括：

CNN，用于接收输入数据，压缩和提取输入数据重要特征；

LSTM，用于接收CNN层的输出，提取时间序列特征，产生最终预测结果。

3.根据权利要求2所述的一种基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法，其特征在于，所述CNN，其卷积层和池化层均取单层，卷积核的数量为500；所述LSTM为单层，神经元数量为128。

4.根据权利要求3所述的一种基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法，其特征在于，对于CNN，其训练阶段的损失函数如下：

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(O_i-P_i)^2}{N}}$$

其中：RMSE为均方根误差，O_i为目标污染物的真实值，P_i为目标污染物的预测值，i为时间序号，N为预测的总时长。

对于模型，其训练阶段的损失函数如下：

$$E(\varphi)=\sqrt{\frac{\sum_{i=1}^{N}(O_i-P_i)^2}{N}}+\lambda[\zeta\varphi^T\varphi+(1-\zeta)|\varphi|]$$

其中：$E(\varphi)$为训练阶段的损失函数，λ为非负超参数，φ为网络中连接权值的集合，ζ为比例参数。

5.根据权利要求3所述的一种基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法，其特征在于，所述步骤S2中对模型的初始化过程具体包括：

步骤S21：对选取的监测数据进行归一化的预处理，并将数据集按照60％，20％，20％的比例划分训练集、验证集和测试集。

步骤S22：设置模型的误差阈值，将输入的训练集的污染物数据和气象数据转化为二维矩阵，矩阵的每一行为一个站点的各种污染物信息和气象信息，每一列为某一种特定的污染物信息或者特定的气象信息。

6.根据权利要求5所述的一种基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法，其特征在于，所述步骤S3具体包括：

步骤S31：将被转化成二维矩阵的输入特征输入到CNN中；

步骤S32：CNN的卷积层对输入特征进行处理得到多个特征图，作为池化层的的输入，池

化层将输出同样个数的缩小后的特征图；

步骤S33：CNN将所有特征图展开成同样个数的一维向量，再经过全连接层的解码，得到输出的污染物浓度值；

步骤S34：将得到的时序性的一维向量作为LSTM的输入，将已训练好的CNN和未训练的LSTM融合，对整个模型进行训练。

7.根据权利要求1所述的一种基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法，其特征在于，所述步骤S5具体为：

将测试的预测结果与观测值作对比，若计算所得的真实值与观测值的误差在预先设定的阈值内，且较预设的其他预测方法能够产生精确度更高的结果，则认为准确性超过阈值，反之，认为准确性不足阈值。

基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法

技术领域

[0001] 本发明涉及一种空气质量浓度预测方法，尤其是涉及一种基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法。

背景技术

[0002] 空气污染在日常生活中受到广泛的关注。特别是如PM2.5这样粒径小，面积大，活性强，易附带重金属、微生物等有害物质，且在大气中难被消除、传播距离远的污染物，更需要被重点关注。当前空气污染问题日益突出，空气污染分析和预测具有复杂性和动态性，涉及多部门、多地区和多领域，对空气污染进行准确的预测，需要处理大量与之相关的环境数据和环境信息。因此，当前形势下，面对种类繁多的空气污染源、污染物及日益增加的环境监测数据，需要实现对这些大数据的充分利用，从而做出更精确的城市空气污染物浓度预测。

[0003] 目前，国内外许多学者都对城市空气污染物浓度预测进行了研究，但大多使用的仍为非深度学习的传统预测方法。如唐晓等通过经验统计方法，建立了月均目标污染物浓度与其他污染物浓度之间的关系；M Dong等人运用隐半马尔科夫模型，加入了时间结构，将过去时间的气象测量值和过去PM2.5的历史观测浓度水平纳入训练数据集，为每个浓度水平都训练出对应的HSMM，可以使预报精度达到24小时以上；Balachandran等利用贝叶斯算法对不同来源的污染物对其浓度的影响；以及王俭等人以沈阳市监测中心的数据集为原始数据，选择1999年秋季气象数据及NOX小时浓度数据中的120组数据为训练集，以2000年的气象数据为测试集，建立污染物浓度的预测模型，用于NOX的预测等等。这些传统的预测方法曾在该类预测工作中有着突出的表现，但缺乏对数据更深程度的分析，从而无法提取数据深层次的联系；另一方面，污染物浓度在时间维度上存在关联，而基于传统预测方法的工作通常是在历史数据和经验的长期积累上进行的，不能很好地符合多变的空气污染情况。

发明内容

[0004] 本发明的目的就是为了克服上述现有技术存在的缺陷而提供一种基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法。

[0005] 本发明的目的可以通过以下技术方案来实现：

[0006] 一种基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法，包括：

[0007] 步骤S1：基于深度学习原理和CNN及LSTM构建城市PM2.5浓度预测的模型；

[0008] 步骤S2：针对所构建的模型，从环境监测数据中选择训练数据和测试数据，完成对预测模型的初始化；

[0009] 步骤S3：利用训练数据对模型进行训练；

[0010] 步骤S4：利用训练好的模型，依据测试数据得到测试的预测结果；

[0011] 步骤S5：判断测试的预测结果的准确性，若准确性超过阈值，则执行步骤S6，若为否，则返回步骤S2；

[0012] 步骤S6:利用训练好的模型进行预测。

[0013] 所述模型包括:

[0014] CNN,用于接收输入数据,压缩和提取输入数据重要特征;

[0015] LSTM,用于接收CNN层的输出,提取时间序列特征,产生最终预测结果。

[0016] 所述CNN,其卷积层和池化层均取单层,卷积核的数量为500;所述LSTM为单层,神经元数量为128。

[0017] 对于CNN,其训练阶段的损失函数如下:

[0018] $$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(O_i-P_i)^2}{N}}$$

[0019] 其中:RMSE为均方根误差,O_i为目标污染物的真实值,P_i为目标污染物的预测值,i为时间序号,N为预测的总时长。

[0020] 对于模型,其训练阶段的损失函数如下:

[0021] $$E(\varphi)=\sqrt{\frac{\sum_{i=1}^{N}(O_i-P_i)^2}{N}}+\lambda[\zeta\varphi^T\varphi+(1-\zeta)|\varphi|]$$

[0022] 其中:$E(\varphi)$为训练阶段的损失函数,λ为非负超参数,φ为网络中连接权值的集合,ζ为比例参数。

[0023] 所述步骤S2中对模型的初始化过程具体包括:

[0024] 步骤S21:对选取的监测数据进行归一化的预处理,并将数据集按照60%,20%,20%的比例划分训练集、验证集和测试集。

[0025] 步骤S22:设置模型的误差阈值,将输入的训练集的污染物数据和气象数据转化为二维矩阵,矩阵的每一行为一个站点的各种污染物信息和气象信息,每一列为某一种特定的污染物信息或者特定的气象信息。

[0026] 所述步骤S3具体包括:

[0027] 步骤S31:将被转化成二维矩阵的输入特征输入到CNN中;

[0028] 步骤S32:CNN的卷积层对输入特征进行处理得到多个特征图,作为池化层的的输入,池化层将输出同样个数的缩小后的特征图;

[0029] 步骤S33:CNN将所有特征图展开成同样个数的一维向量,再经过全连接层的解码,得到输出的污染物浓度值;

[0030] 步骤S34:将得到的时序性的一维向量作为LSTM的输入,将已训练好的CNN和未训练的LSTM融合,对整个模型进行训练。

[0031] 所述步骤S5具体为:

[0032] 将测试的预测结果与观测值作对比,若计算所得的真实值与观测值的误差在预先设定的阈值内,且较预设的其他预测方法能够产生精确度更高的结果,则认为准确性超过阈值,反之,认为准确性不足阈值。

[0033] 与现有技术相比,本发明具有以下有益效果:

[0034] 1)并未将污染物的预测工作仅仅依赖于大量历史数据总结出的经验和历史经验归纳污染物的变化规律,从而能够充分考虑大气环境复杂多变这一问题。

[0035] 2)对污染数据的特征能够进行深层次的分析,从而可以提取数据间深层次的联

系，有效地利用环境大数据，实现环境管理水平的提升。

[0036] 3)预测的准确度较传统的预测方法高，在同样的工作时长和工作条件下，可以产生更好的结果。

附图说明

[0037] 图1为本发明方法的主要步骤示意图；

[0038] 图2为本发明实施例的流程示意图；

[0039] 图3为本发明所构建的预测模型的结构示意图。

具体实施方式

[0040] 下面结合附图和具体实施例对本发明进行详细说明。本实施例以本发明技术方案为前提进行实施，给出了详细的实施方式和具体的操作过程，但本发明的保护范围不限于下述的实施例。

[0041] 本发明首先对空气污染物浓度预测进行定义：

[0042] 定义1空气污染物浓度预测：主要是通过历史污染物和气象信息，对PM2.5、PM10等一系列空气污染在未来一定时间内的浓度进行预测，是环境科学、气象科学、计算机科学等都在重点研究的课题之一，因而具有一定的学科交叉性。

[0043] 定义2传统预测法：非深度学习的空气污染物浓度预测方法统称为传统的预测方法，如基于历史数据和统计学方法的经验模型的预测；基于统计学和数学方法或模型建立概率模型的预测；利用综合方法的预测；以及基于传统机器学习建立的预测模型等，均属于传统预测法。

[0044] 一种基于CNN和LSTM融合神经网络的空气PM2.5浓度预测方法，如图1～图3所示，包括：

[0045] 步骤S1：基于深度学习原理和CNN及LSTM构建城市PM2.5浓度预测的模型，具体的：基于深度学习原理和CNN及LSTM的特点，根据环境监测各类污染物浓度和气象因子的数据，以PM2.5为预测的目标污染物，构建城市PM2.5浓度预测的模型，模型以CNN为底层，压缩和提取输入数据重要特征；其结果作为高层LSTM的输入，提取时间序列特征，产生最终预测结果。

[0046] 模型包括：CNN，用于接收输入数据，压缩和提取输入数据重要特征；LSTM，用于接收CNN层的输出，提取时间序列特征，产生最终预测结果。

[0047] 步骤S2：根据构建的融合神经网络预测模型的特点，从海量环境监测数据中选择合适的训练和测试数据，完成对预测模型的初始化，其中对模型的初始化过程具体包括：

[0048] 步骤S21：对选取的监测数据进行归一化的预处理以提高模型的训练速度和预测精度，并将数据集按照60％，20％，20％的比例划分训练集、验证集和测试集。本实施例中选取Z-score标准化方法进行归一化处理：

[0049] $$x' = \frac{x-\mu}{\sigma}$$

[0050] 其中：μ为原始数据均值，σ为原始数据标准差，x为监测数据，x'为归一化处理后的监测数据，经过处理的数据皆符合μ＝0，σ＝1的标准正态分布。

[0051] 步骤S22:设置模型的误差阈值,将输入的训练集的污染物数据和气象数据转化为二维矩阵,矩阵的每一行为一个站点的各种污染物信息和气象信息,每一列为某一种特定的污染物信息或者特定的气象信息。

[0052] 合理设置模型的误差阈值,取值范围在0.001-0.00001之间,学习速率在0.01-0.1之间取值,最大迭代次数为1000次,LSTM的自循环系数取0.001,λ取1e-4,ζ取0.9。对于CNN,其卷积层和池化层均取单层,卷积核的数量为500,LSTM也为单层,神经元数量为128。

[0053] 对于CNN,其训练阶段的损失函数如下:

[0054] $$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(O_i-P_i)^2}{N}} \quad (1)$$

[0055] 其中:RMSE为均方根误差,O_i为目标污染物的真实值,P_i为目标污染物的预测值,i为时间序号,N为预测的总时长,RMSE越小,证明预测的准确度越高。

[0056] 对于模型,其训练阶段的损失函数如下:

[0057] $$E(\varphi)=\sqrt{\frac{\sum_{i=1}^{N}(O_i-P_i)^2}{N}}+\lambda[\zeta\varphi^T\varphi+(1-\zeta)|\varphi|] \quad (2)$$

[0058] 其中:$E(\varphi)$为训练阶段的损失函数,λ为非负超参数,φ为网络中连接权值的集合,ζ为控制L1,L2penalty使用的比例参数,$\zeta\in(0,1)$。

[0059] 函数设置为均方根误差和正则项之和。公式(2)中,前半部分为均方根误差;后半部分引入Elastic Net算法进行正则化约束。

[0060] 步骤S3:利用训练数据对模型进行训练,训练预测模型,假设网络共有δ层,μ代表正在训练的层,x_i是一组输入数据,y_i是一组输出的未经全连接层解码的一维向量,i表示动态变化的时间,分别表示卷积层、池化层、LSTM层的输出,整个模型的训练可以用如下公式表示:

[0061] $$\begin{cases} C_i^{\mu}=g(u^Tx_i) & \mu=2 \\ P_i^{\mu}=g(v^TC_i) & 2<\mu<\delta-3 \\ L_i^{\mu}=g(w^TP_i+d^TL_i) & 3<\mu<\delta-2 \\ y_i^{\mu}=f(\xi^TL_i) & \mu=\delta-2 \end{cases} \quad (3)$$

[0062] 其中,u,v,w,d,ζ皆是网络的权值矩阵。u表示输入层到卷积层的权值矩阵,v表示卷积层到池化层的权值矩阵,w表示池化层到LSTM的权值矩阵,d表示LSTM内部神经元之间传递信息的权值矩阵,ξ表示LSTM到全连接层的权值矩阵。最终产生的y_i经全连接层的解码后翻译成污染物浓度值。

[0063] 具体,CNN的训练优先于整个模型的训练,损失函数采用公式(1),并通过反向传播算法进行误差传递和网络连接权值的更新;将训练好的CNN加入整个模型进行训练。输入的二维矩阵经过CNN的卷积层和池化层产生的是多个特征图,特征图经过展开得到对应的多个一维特征向量,输入到LSTM中,进行时间序列特征的提取,并输出最终的预测结果,全连接层用来为预测结果解码,得到PM2.5在t+1时刻的浓度值。损失函数采用公式(2),SGD表示随机梯度下降法,用来将误差反向传播到整个模型,更新各层和各节点的连接权值。

[0064] 整个预测模型的训练包括:

[0065] 步骤S31:将被转化成二维矩阵的输入特征输入到CNN中;

[0066] 将被转化成二维矩阵的输入特征输入到CNN中。令η为当前训练的层数,m代表特征图,卷积层的上一层输出的特征图由该卷积层的卷积核k进行学习,通过激活函数而得到输出特征图,i,j均为特征图下标,CNN从输入数据中自动学习特征,从而无需在训练前就对数据特征进行提取。

[0067] $$m_j^{\eta}=f\left(\sum_{i\in M} m_i^{\eta-1}*k_{ij}^{\eta}+b_j^{\eta}\right) \qquad (4)$$

[0068] 步骤S32:CNN的卷积层对输入特征进行处理得到多个特征图,作为池化层的的输入,池化层将输出同样个数的缩小后的特征图:

[0069] $$m_j^{\eta}=f\left(\beta_j^{\eta}down(m_j^{\eta-1})+b_j^{\eta}\right) \qquad (5)$$

[0070] 其中,β和b分别作为输出图像的乘性偏置和加性偏执,down表示下采样函数;

[0071] 步骤S33:CNN将所有特征图展开成同样个数的一维向量,再经过全连接层的解码,得到输出的污染物浓度值;

[0072] 以上是对CNN的训练,这一阶段输入的二维矩阵中,主要包括以下因子{PM2.5浓度,温度,风速,风向,湿度,降水量,其他污染物浓度,站点},将输入的二维矩阵进行压缩,得到真正的数据特征,网络可以准确地将输入值翻译成污染物浓度情况,建立一个输入到输出的映射。以公式(1)衡量预测的准确性,采用反向传播算法,将池化层作为考虑的因素并基于所有值更新卷积层的权重,优化网络预测性能,减少预测值和观测值之间的误差。当网络符合期望后,停止第一阶段网络的训练,进入第二阶段的训练。

[0073] 步骤S34:将得到的时序性的一维向量作为LSTM的输入,将已训练好的CNN和未训练的LSTM融合,对整个模型进行训练。

[0074] 具体的,二维输入矩阵经CNN压缩和特征提取后转化为的高度浓缩化的具有时序性的一维向量作为LSTM层的输入,模型具有时间序列预测的功能,t时刻之前D小时的值$O^{c}_{(t,\ldots,t-D)}$和$O^{f}_{(t,\ldots,t-D)}$作为整个模型的输入,预测的目标是t时刻之后设定小时的PM2.5的浓度值(t是设定好的时间窗口)。令x表示输入,表示动态的时间序列,W表示权值矩阵,h表示隐藏层信息,b表示偏置,用如下公式表示LSTM的训练过程:

[0075] S341.LSTM首先选择性遗忘过去的PM2.5及其他因子的某些数据信息,

[0076] $f_t=\sigma(W_f[h_{t-1},x_t]+b_f)$ (6)

[0077] B.决定在单元状态中储存新的信息,该信息来自两部分,“输入门限”的sigmoid层决定更新的信息,tanh层创建新的候选值向量,

[0078] $$\begin{aligned} i_t&=\sigma(W_i[h_{t-1},x_t]+b_i)\\ C_t'&=\tanh(W_C\cdot[h_{t-1},x_t]+b_t)\end{aligned} \qquad (7)$$

[0079] C.进行旧状态的更新,

[0080] $C_t=f_t*C_{t-1}+i_t*C'_t$ (8)

[0081] D.最后决定输出信息,即为预测的PM2.5浓度,

[0082] $$\begin{aligned} o_t&=\sigma(W_o[h_{t-1},x_t]+b_o)\\ h_t&=o_t*\tanh(C_t)\end{aligned} \qquad (9)$$

[0083] LSTM输出的预测值经过CNN的全连接层的解码,输出最终结果。在整个模型使用随

机梯度下降法的fine-tuning阶段，排除深层神经网络训练时易产生过拟合问题的影响。本发明使用结合了Ridge Regression和Lasso法优势的Elastic Net算法进行L1和L2正则化约束，通过误差反向传播的方式计算误差函数对网络全部权重和偏置值的梯度进行更新，避免过拟合问题，训练过程持续至模型的性能符合期望。模型训练结束后，各连接权值和参数也随之确定。

[0084] 步骤S4：利用训练好的模型，依据测试数据得到测试的预测结果；

[0085] 步骤S5：判断测试的预测结果的准确性，若准确性超过阈值，则执行步骤S6，若为否，则返回步骤S2，具体为：

[0086] 将测试的预测结果与观测值作对比，运用相关系数和平均绝对误差衡量模型误差，公式分别如(10)和(11)所示：

[0087] $$r(O,P)=\frac{Cov(O,P)}{\sqrt{Var[O]Var[P]}} \qquad (10)$$

[0088] 公式(10)中，r(O,P)为观测值和预测值的相关系数，Cov(O,P)为观测值和预测值的协方差，Var[O]，Var[P]分别是观测值和预测值的方差。

[0089] $$MAE=\frac{1}{N}\sum_{i=1}^{N}|P_i-O_i| \qquad (11)$$

[0090] 公式(11)中，MAE为平均绝对误差，i为时间序号，N为预测的总时长。P为预测值，O为观测值。

[0091] 若计算所得的真实值与观测值的误差在预先设定的阈值内，则说明模型满足预期，可以用于预测未来一定时间内的城市PM2.5的浓度，亦即：准确性超过阈值。

[0092] 对训练所得的预测模型进行性能评估。经实验分析和对比，相比于其他已有的方法，在相同情况下，本发明的模型能够产生精确度更高的结果，且能够充分利用海量的污染物和气象数据。

[0093] 步骤S6：利用训练好的模型进行预测。

[0094] 综上所述，本发明所构建的基于CNN和LSTM融合神经网络的预测模型是建立在已存在的两种深度神经网络的研究上的，利用两种网络的特点和优势，建立一种可以预测目标城市未来一定时间内的PM2.5浓度的模型。所使用的损失函数也为以往的研究中存在的，且被证明可以很好地衡量结果准确性。所以本发明针对以往的预测污染物浓度的方法的不足，充分利用了已存在的研究成果，提出了基于两种深度神经网络融合的预测模型。该模型以CNN为底层，对输入数据进行重要特征的提取，其输出结果作为高层LSTM的输入，提取污染物的时间序列特征，可以充分考虑到污染物的时间关联性，得到更为具有精确的预测结果，因而具有实际性的应用前景。

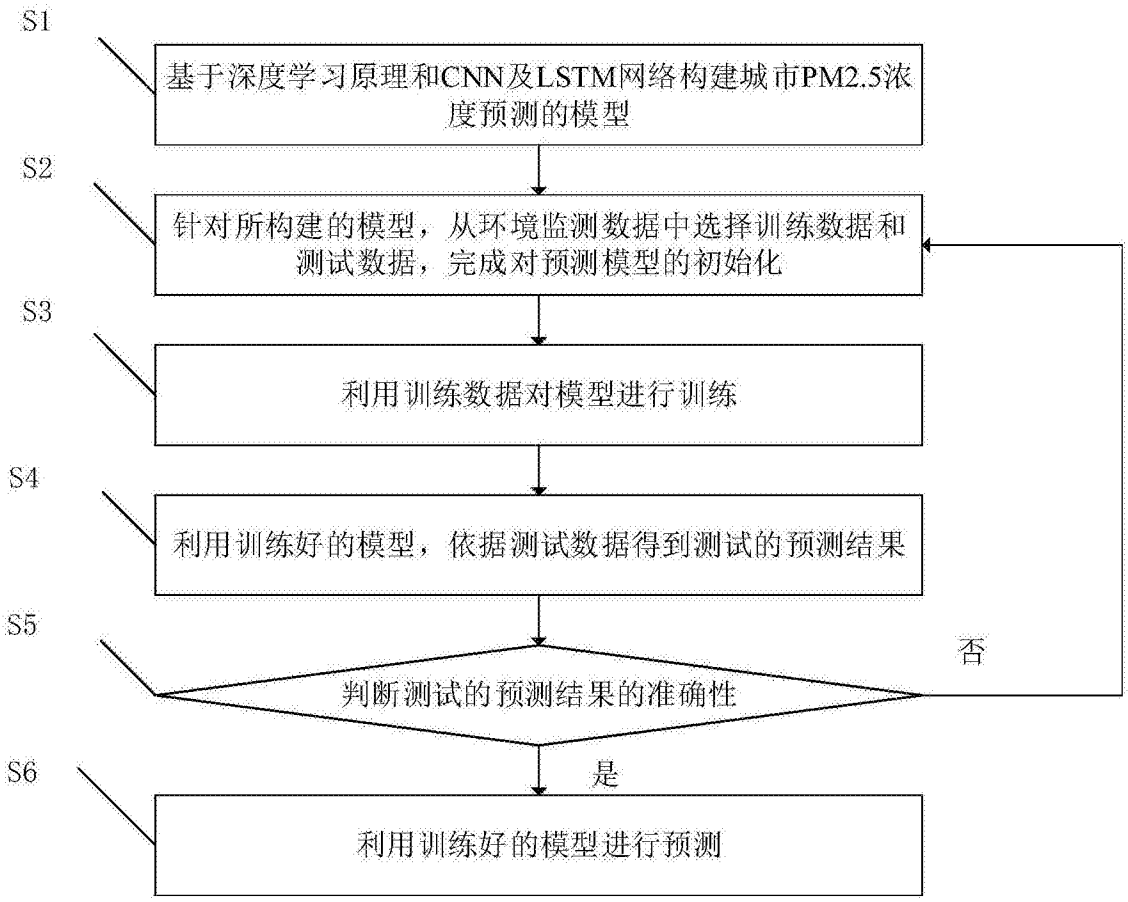

S1
基于深度学习原理和CNN及LSTM网络构建城市PM2.5浓度预测的模型
S2
针对所构建的模型，从环境监测数据中选择训练数据和测试数据，完成对预测模型的初始化
S3
利用训练数据对模型进行训练
S4
利用训练好的模型，依据测试数据得到测试的预测结果
S5
判断测试的预测结果的准确性
否
是
S6
利用训练好的模型进行预测

图1

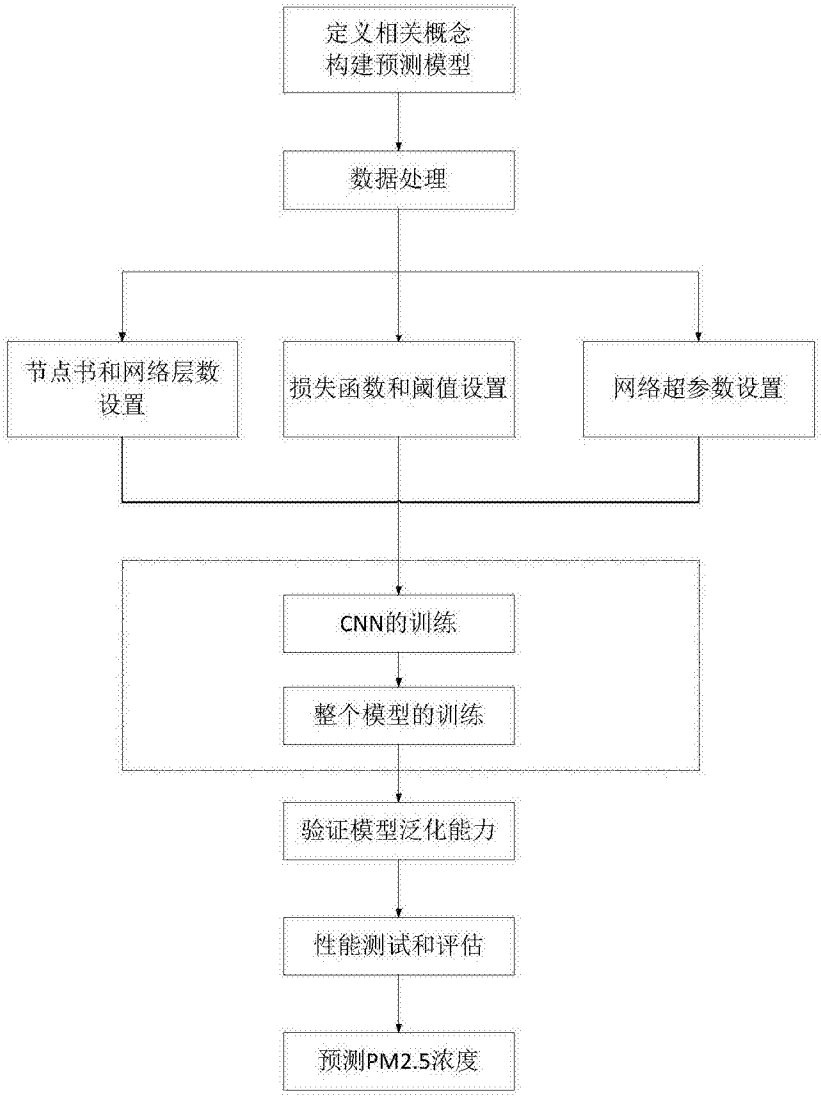

定义相关概念
构建预测模型
数据处理
节点书和网络层数设置
损失函数和阈值设置
网络超参数设置
模型初始化
CNN的训练
整个模型的训练
模型训练
验证模型泛化能力
性能测试和评估
预测PM2.5浓度

图2

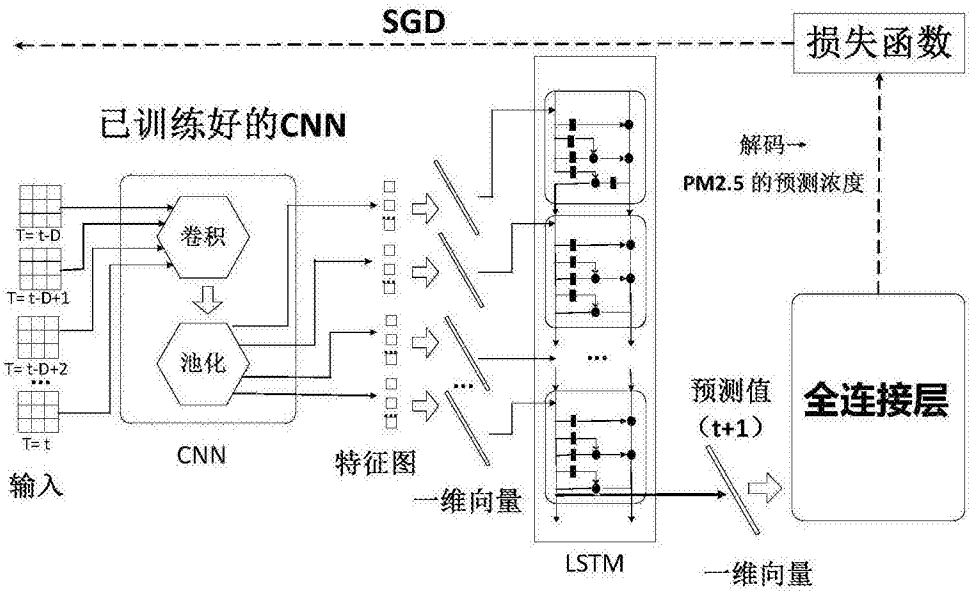

SGD
损失函数
已训练好的CNN
解码
PM2.5 的预测浓度
T= t-D
T= t-D+1
T= t-D+2
T= t
卷积
池化
CNN
输入
特征图
一维向量
LSTM
预测值
（t+1）
全连接层
一维向量

图3