(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 108564110 A
(43)申请公布日 2018.09.21

(21)申请号 201810254517.1

(22)申请日 2018.03.26

(71)申请人 上海电力学院
地址 200090 上海市杨浦区平凉路2103号

(72)发明人 张挺

(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 叶敏华

(51)Int.Cl.
G06K 9/62(2006.01)
G01N 33/00(2006.01)

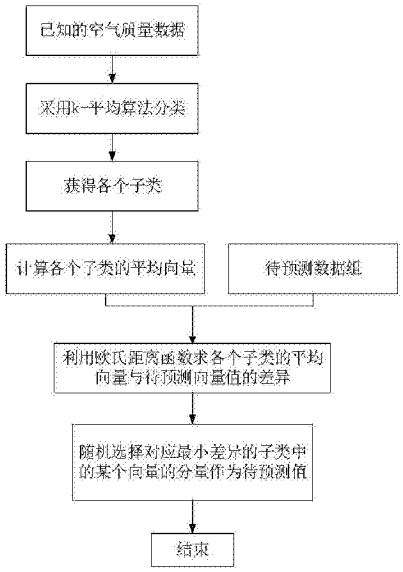
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种基于聚类算法的空气质量预测方法

(57)摘要

本发明涉及一种基于聚类算法的空气质量预测方法，包括以下步骤：S1：确定多个时间序列，对不同时间序列的等间隔时间点采集各个空气污染物的浓度；S2：将不同空气污染物的浓度组成污染物浓度向量，每个时间序列的所有污染物浓度向量作为一个数据组；S3：采用聚类算法对每个数据组进行分类，获取多个子类数据，并对每个子类数据求取平均向量；S4：分别求取各平均向量与待预测数据组中各个浓度向量之间的差异值，选取出差异值最小的平均向量，并选出其所对应的子类数据；S5：根据待预测数据组中缺失的污染物浓度种类，在选出的子类数据中找出该污染物种类的浓度值作为待预测数据组的预测值。与现有技术相比，本发明具有提高预测质量等优点。



CN 108564110 A

1.一种基于聚类算法的空气质量预测方法，其特征在于，包括以下步骤：

S1：确定多个时间序列，对不同时间序列的等间隔时间点采集各个空气污染物的浓度；

S2：将各等间隔时间点采集的不同空气污染物的浓度组成污染物浓度向量，每个时间序列的所有污染物浓度向量作为一个数据组；

S3：采用聚类算法对每个数据组进行分类，获取多个子类数据，并对每个子类数据求取平均向量；

S4：分别求取各平均向量与待预测数据组中各个浓度向量之间的差异值，选取出差异值最小的平均向量，并选出其所对应的子类数据；

S5：根据待预测数据组中缺失的污染物浓度种类，在选出的子类数据中找出该污染物种类的浓度值作为待预测数据组的预测值。

2.根据权利要求1所述的一种基于聚类算法的空气质量预测方法，其特征在于，所述的聚类算法采用K-平均算法。

3.根据权利要求2所述的一种基于聚类算法的空气质量预测方法，其特征在于，采用K-平均算法进行分类的处理过程为：

31)将每个数据组的数据分为k个簇，随机选择k个对象作为初始的k个簇的质心；

32)计算各个对象与各个簇的质心的距离，将各个对象划分到与其自身距离最近的簇；

33)重新计算每个新簇的均值，即质心；

34)若簇的质心不再变化，则返回划分结果，否则执行步骤32)。

4.根据权利要求1所述的一种基于深度神经网络的空气质量预测方法，其特征在于，所述的步骤S2中，污染物浓度向量Vec_{ab}的表达式为：

$$Vec_i=\{C_i(A),C_i(B),C_i(C),C_i(D),C_i(E)\cdots,T\}$$

式中，i为时间点编号，$i=1,2,\cdots,N$，N为时间点的总数；$C_i(A)$，$C_i(B)$，$C_i(C)$，$C_i(D)$，$C_i(E)$分别为不同污染物种类的浓度值；T为采集污染物浓度的时间点。

5.根据权利要求4所述的一种基于深度神经网络的空气质量预测方法，其特征在于，所述的步骤S2中，数据组的表达式为：

$$X_a=\{Vec_1,Vec_2,Vec_3,\cdots,Vec_N\}$$

式中，a为时间序列编号。

6.根据权利要求1所述的一种基于聚类算法的空气质量预测方法，其特征在于，所述的步骤S4中，采用欧式距离求取各平均向量与待预测数据组中各个浓度向量之间的差异值。

7.根据权利要求4所述的一种基于聚类算法的空气质量预测方法，其特征在于，各个污染物浓度向量中采集污染物浓度的时间点T为统一换算标准。

一种基于聚类算法的空气质量预测方法

技术领域

[0001] 本发明涉及空气质量预测技术领域，尤其是涉及一种基于聚类算法的空气质量预测方法。

背景技术

[0002] 现有的空气质量预测方式主要分为两类：数值预报与统计预报。数值预报的基本原理是通过计算物质守恒方程来对大气中的污染物进行数值计算，数值预报意在模拟一种真实的大气环境，用数学与化学公式尽可能的接近于真实大气的运作机理，充分考虑大气的污染源清单、气象因子、粒子化学、光化学反应过程、二次污染物、污染物传输、清除等因素，来模拟真实的大气环境，并根据大气环境的变化呈现出不同的精确结果。该方式的缺点在于，因数值预报技术依赖于气象因子与污染物排放等因素，气象因子的不确定性会导致数值预报系统在不同区域可能呈现出预报准确程度的差异。

[0003] 统计预报的发展得益于人类开始建立比较完善合理的大气监测研究网络，收集大气系统中的污染物因子、气象因子、污染源因子等变量。根据现有的积累的大量的历史监测数据，运用统计学的相关方法模型如：人工神经网络、灰色系统理论、聚类与多元回归等，分析出大气预测中潜在的规律规则，来对未来空气质量进行预测。

[0004] 空气质量的时空分布受到气象场、排放源、复杂下垫面、理化过程的耦合等多种因素的影响，具有较强的非线性特性。现有的统计预报方式虽然建立简单，业务运行方便、易普及，但缺乏坚实的物理基础，需要大量的监测资料；数值预测虽然物理基础坚实，预测结果全面，但模式所需要的边界、初始条件不易给出，使得预测结果精度不高。

发明内容

[0005] 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聚类算法的空气质量预测方法。

[0006] 本发明的目的可以通过以下技术方案来实现：

[0007] 本发明采用基于聚类算法进行空气质量预测，聚类算法的原理为：

[0008] 聚类(clustering)是数据挖掘的一个重要方法，即把一个数据对象集合分成多个簇(或划分成组)，使得每簇中的点彼此相似，但与其他簇中的点尽可能不同。通常，聚类被认为是一种无监督的学习过程，因为它不需要事先定义类别信息或是给出训练样例来指明数据应该具有怎样的关系。然而，在实际应用中，对于数据集的最终划分需要一定的评价体系来衡量其优劣。这也就意味着聚类更多的被当作是一种增强学习模型来使用。

[0009] 聚类技术可以粗分为两大类：划分方法(如k-平均、k-中心、CLARANS等)和层次方法(如BIRTH、CURE、CHAMELEON等)。层次方法可以找到一个可以继续分割成子聚类的结构，然后递归的进行下去；而划分方法都基于某个目标函数的最优化问题。基于划分的聚类算法的基本思路是：试图找到一个最优划分以把数据分成指定数量聚类，其本质是组合优化的一种形式。聚类过程就是采用一定的启发式方法来搜索全部空间中的一个很小部分，找

到局部最优解。比较流行的启发式方法是k-平均算法和k-中心点算法。本专利提出一种利用k-平均算法预测空气质量的方法。

[0010] 一种基于聚类算法的空气质量预测方法，包括以下步骤：

[0011] S1：确定多个时间序列，对不同时间序列的等间隔时间点采集各个空气污染物的浓度；

[0012] S2：将各等间隔时间点采集的不同空气污染物的浓度组成污染物浓度向量，每个时间序列的所有污染物浓度向量作为一个数据组；

[0013] S3：采用k-平均算法对每个数据组进行分类，获取多个子类数据，并对每个子类数据求取平均向量；

[0014] S4：分别求取各平均向量与待预测数据组中各个浓度向量之间的差异值，选取出差异值最小的平均向量，并选出其所对应的子类数据；

[0015] S5：根据待预测数据组中缺失的污染物浓度种类，在选出的子类数据中找出该污染物种类的浓度值作为待预测数据组的预测值。

[0016] 优选地，采用K-平均算法进行分类的处理过程为：

[0017] 31)将每个数据组的数据分为k个簇，随机选择k个对象作为初始的k个簇的质心；

[0018] 32)计算各个对象与各个簇的质心的距离，将各个对象划分到与其自身距离最近的簇；

[0019] 33)重新计算每个新簇的均值，即质心；

[0020] 34)若簇的质心不再变化，则返回划分结果，否则执行步骤32)。

[0021] 优选地，所述的步骤S2中，污染物浓度向量Vec_{ab}的表达式为：

[0022] $Vec_i=\{C_i(A),C_i(B),C_i(C),C_i(D),C_i(E)\cdots,T\}$

[0023] 式中，i为时间点编号，$i=1,2,\cdots,N$，N为时间点的总数；$C_i(A)$，$C_i(B)$，$C_i(C)$，$C_i(D)$，$C_i(E)$分别为不同污染物种类的浓度值；T为采集污染物浓度的时间点。

[0024] 优选地，所述的步骤S2中，数据组的表达式为：

[0025] $X_a=\{Vec_1,Vec_2,Vec_3,\cdots,Vec_N\}$

[0026] 式中，a为时间序列编号。

[0027] 优选地，所述的步骤S4中，采用欧式距离求取各平均向量与待预测数据组中各个浓度向量之间的差异值。

[0028] 优选地，各个污染物浓度向量中采集污染物浓度的时间点T为统一换算标准。

[0029] 与现有技术相比，本发明并对条件数据采用k-平均算法进行分类，将采集的污染物浓度向量划分为多个条件数据，利用k-平均算法的自身优势，尝试找出使平方误差函数值最小的k个划分，当结果簇是密集的，而簇与簇之间区别明显时，它的效果较好；面对大规模的污染物浓度数据集，k-平均算法相对可扩展，并且具有较高的效率；因划分获取的条件数据越多，预测的效果则越好，进一步提高了空气质量预测的预测质量。

附图说明

[0030] 图1为本发明方法的流程图。

具体实施方式

[0031] 下面结合附图和具体实施例对本发明进行详细说明。

[0032] 实施例

[0033] 如图1所示，本发明涉及一种基于聚类算法的空气质量预测方法，包括以下步骤：

[0034] 步骤一、确定多个时间序列，对不同时间序列的等间隔时间点采集各个空气污染物的浓度；

[0035] 步骤二、将各等间隔时间点采集的不同空气污染物的浓度组成污染物浓度向量，每个时间序列的所有污染物浓度向量作为一个数据组；

[0036] 步骤三、采用聚类算法对每个数据组进行分类，获取多个子类数据，并对每个子类数据求取平均向量；

[0037] 步骤四、分别求取各平均向量与待预测数据组中各个浓度向量之间的差异值，选取出差异值最小的平均向量，并选出其所对应的子类数据；

[0038] 步骤五、根据待预测数据组中缺失的污染物浓度种类，在选出的子类数据中找出该污染物种类的浓度值作为待预测数据组的预测值。

[0039] 利用污染物浓度作为衡量空气质量的主要评价标准。主要的空气污染物包括二氧化氮、二氧化硫、臭氧、一氧化碳、悬浮微粒(PM2.5、PM10)等。本实施例的污染物浓度以二氧化氮、二氧化硫、臭氧、一氧化碳、悬浮微粒中的PM2.5、PM10为例对本发明方法进行说明。利用已知的污染物浓度数据和测试污染物浓度的时间点作为训练数据，每组数据由若干污染物浓度向量组成，每个污染物浓度向量设计格式如下：

[0040] $Vec_i=\{C_i(NO_2),C_i(SO_2),C_i(O_3),C_i(CO),C_i(PM2.5),C_i(PM10),T\}$

[0041] 式中，$C_i(NO_2),C_i(SO_2),C_i(O_3),C_i(CO),C_i(PM2.5),C_i(PM10)$分别为不同时间点所采集的二氧化氮浓度、二氧化硫浓度、臭氧浓度、一氧化碳浓度以及PM2.5、PM10浓度。i为时间点编号，$i=1,2,\cdots,N$，N为时间点的总数；T为采集污染物浓度的时间点。测试污染物浓度的时间点换算成统一格式，例如23点45分30秒可以换算成23.75833点(45/60＝0.75，30/60/60＝0.00833)，即将所有时间换算成用小时表示。假设已知一组空气污染物浓度数据组Data1，将该数据组作为训练数据进行聚类划分，即对Data1进行分类，本发明采用的聚类方法是经典的k-平均(k-means)方法，该方法的主要思想是：

[0042] k-平均算法以k为参数，把n个对象分成k个簇，使簇内具有较高的相似度，而簇间的相似度较低。相似度的计算根据一个簇中对象的平均值，即一个簇的质心来进行的。k-平均算法的处理过程如下：首先，随机选取k个对象作为初始的k个簇的质心；然后，将其余对象根据其与各个簇质心的距离分配到最近的簇；再后重新计算每个簇的质心。这个过程不断重复，直到目标函数最小化为止。目标函数使生成的簇尽可能地紧凑和独立，它使用的距离度量可以是欧几里得距离，也可以采用其他距离度量。本发明采用K-平均算法进行分类的处理过程为：

[0043] 31)将每个数据组的数据分为k个簇，随机选择k个对象作为初始的k个簇的质心；

[0044] 32)计算各个对象与各个簇的质心的距离，将各个对象划分到与其自身距离最近的簇；

[0045] 33)重新计算每个新簇的均值，即质心；

[0046] 34)若簇的质心不再变化，则返回划分结果，否则执行步骤32)。

[0047] k-平均算法尝试找出使平方误差函数值最小的k个划分。当结果簇是密集的，而簇

与簇之间区别明显时，它的效果较好。面对大规模数据集，该算法是相对可扩展的，并且具有较高的效率。

[0048] 在利用k-平均算法对已知污染物浓度数据Data1分类后，获得M个“子类”，每个子类命名为$Class_i$(i＝1,2,…,M)，对于每个$Class_i$，可以定义一个“平均向量”，它是属于该$Class_i$的所有向量的均值，即平均向量Ave_i。

[0049] 在预测污染物浓度时，若一组数据Data2中的各个污染物浓度向量Vec_j(j＝1,2,…,K)并未获得全部分量的值，例如臭氧和一氧化碳的浓度数据缺失，那么可以利用欧氏距离比较各个Vec_j与各个平均向量的差异；选择与某个Vec_j差异最小的平均向量Ave_i所对应的子类$Class_i$中某个向量的臭氧和一氧化碳的浓度数据作为Vec_j缺失的臭氧和一氧化碳的浓度数据，即完成了污染物的浓度预测。

[0050] 例如，假设求取第一时间点污染物浓度向量Vec_1(缺少臭氧和一氧化碳的浓度数据)与平均向量Ave_i的差异，其中，Vec_1与Ave_4的差异值是所有差异中最小的，那么就以子类$Class_4$中的某个向量中的臭氧和一氧化碳的浓度数据作为Vec_1缺失的臭氧和一氧化碳的浓度数据。然后依次对所有的Vec_j执行上述操作。

[0051] 以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的工作人员在本发明揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以权利要求的保护范围为准。

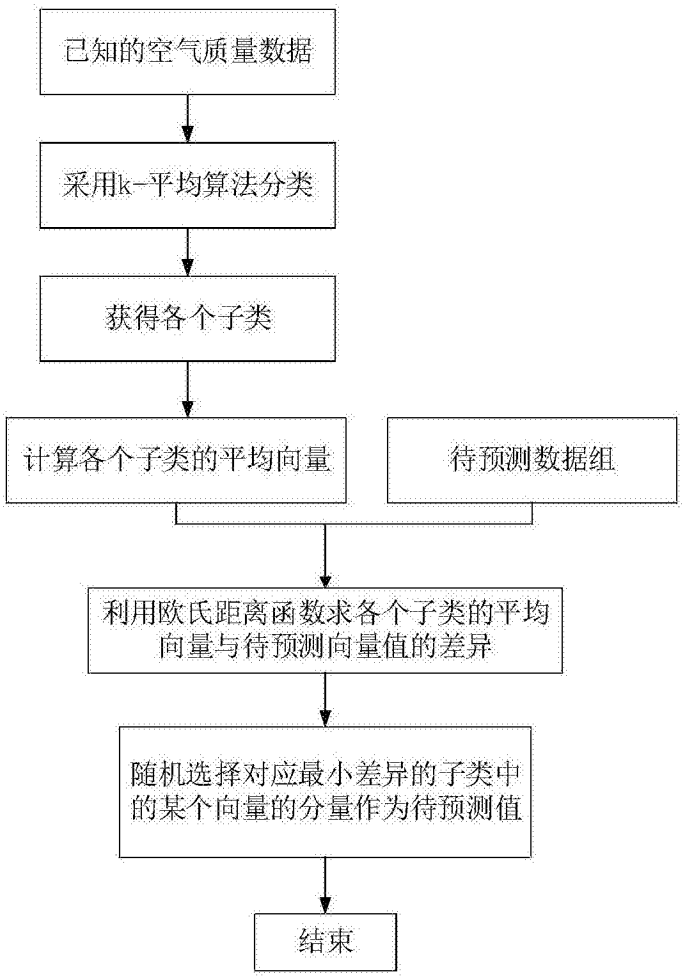

已知的空气质量数据
采用k-平均算法分类
获得各个子类
计算各个子类的平均向量
待预测数据组
利用欧氏距离函数求各个子类的平均向量与待预测向量值的差异
随机选择对应最小差异的子类中的某个向量的分量作为待预测值
结束

图1