第 30 卷第 4 期　　四川理工学院学报(自然科学版)　　Vol. 30　No. 4
2017 年 8 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Aug. 2017
文章编号:1673-1549(2017)04-0096-05　　DOI:10.11863/j.suse.2017.04.17

# 熵在空气质量指数(AQI)预测中的应用

文　琴,罗　飞

(成都信息工程大学软件工程学院,成都　610225)

**摘　要**:为了更准确地找出影响空气质量指数的气象因子与提高其预测精度,提出了基于熵、BP 神经网络和时间序列模型的组合预测模型。该方法利用增加了特征变量的转移熵方法,得到影响 AQI 的气象因子及其影响度,将得到的气象因子与 AQI 实测值作为 BP 神经网络的输入因子和时间序列分析模型的特征因子,影响度作为 BP 神经网络输入因子的初始权重,构建 BP 神经网络预测模型和时间序列分析预测模型,最后用熵值法组合各个预测模型的预测结果。实验表明利用该方法对空气质量指数进行预测可提高其预测精度。

**关键词**:空气质量指数预测;转移熵;熵值法

**中图分类号**:TP311　　**文献标志码**:A

## 引　言

近年来城市空气污染问题越来越严重,对自然环境和人民的生活带来巨大冲击,因此,建立科学的空气质量指数预测模型尤为重要。开展空气质量指数预测可以让人们对影响空气质量指数的因素以及未来城市空气质量指数的变化有所了解和把握,为其出行提供健康指引,同时为政府相关部门制定空气污染处理方案提供辅助材料。

张学文[1-2]给出了计算气象要素场熵值的方法,同时认为开展熵气象学研究可以为气象学找出新的出路。在空气质量指数预测的文献中,用于分析影响空气质量指数的气象要素的方法主要有主成分分析[3]、统计对比分析[4]、统计和个例分析等。但是主成分分析主要是用于分析变量之间的线性关系,而大气环境质量的预测和评价是一个多变量和非线性问题;统计对比分析会由于采集的数据集不同而导致最终得出不同的结论。因此,本文采用转移熵方法找出影响 AQI 的气象因子,在该方法中增加一个特征变量风场,因为在不同的风力风速情况下,气象条件对空气质量指数值的影响是不同的,即在考虑了影响 AQI 的主要因素风场的条件下再判断其他气象要素对 AQI 值的信息转移。转移熵[5]是能够分析系统之间信息相互作用的一种有效工具,同时能够解决非线性系统问题以及描述两因素之间的相关度,这是因为在转移熵的模型中考虑了系统之间的不对称性以及动态特性。近年来转移熵被广泛应用于神经电信号与市场股票的时间序列分析研究中[6-7],并都取得了不错的成果。

目前用于空气质量指数的预测模型主要是时间序列分析模型和神经网络模型。于萍[8]提出利用时间序列分析 ARMA(1,1)模型对大连市未来 10 天的空气质量进行预测,该模型短期预测较为准确,但一旦测试天

**收稿日期**:2017-04-25

**基金项目**:国家公益性行业(气象)科研专项(GYHY201506025)

**作者简介**:文　琴(1991-),女,四川成都人,硕士生,主要从事气象信息化方面的研究,(E-mail)15928662936@163.com;

罗　飞(1977-),女,四川成都人,副教授,主要从事气象信息化技术、数据集成与可视化方面的研究,(E-mail)luofei@cuit.edu.cn

数增加,预测结果可能会不稳定。南亚翔[9]等人利用自回归移动平均模型(ARMA 算法)为卡尔曼滤波建立模型,提出将 RBF 神经网络融合于卡尔曼滤波的方法,实现对空气质量指数的混合预测,但卡尔曼滤波一般用于线性系统。王珍[10]提出采用因子分析法先将多指标进行降维,然后再用 BP 神经网络模型进行综合评价,而因子分析法是主成分分析法的推广,这两种分析方法均适用于线性关系的分析,但大气环境质量的预测是一个非线性问题,用因子分析法可能会对预测结果造成一定的影响。祝翠玲[11]和郭庆春[12]等人将人工神经网络应用到空气质量预测及大气污染预测中,由于每日空气质量指数及污染物浓度呈非线性变化且受多种因素的影响,并且神经网络具有较强的非线性处理能力和自学习能力,实验表明将神经网络模型用于空气质量指数的预测,能够提高预测结果的精度和正确率。为了准确地提取影响空气质量指数的气象因子,提出在考虑风场对污染物扩散影响的条件下再提取影响空气质量指数的气象因子。为了进一步提高空气质量指数的预测精度,将时间序列分析模型与神经网络预测模型的预测结果进行组合。

本文首先用转移熵找出影响空气质量指数的气象要素,再用找出的气象因子与实测空气质量指数数据构建 BP 神经网络预测模型与时间序列分析模型。然后用熵值法对 BP 神经网络和时间序列分析模型的预测结果进行评价,确定各个预测模型的权重,将各个预测模型的预测结果进行组合。最终得到较单一预测模型更加准确的预测结果。

## 1　转移熵及熵值法在预测中的应用

### 1.1　转移熵在预测中的应用

在污染源不变的条件下,空气质量的变化主要与气象条件有关。为了提高空气质量指数预测的准确率,本文提出利用转移熵找出影响空气质量指数的主要气象要素。在信息理论中,转移熵是测量两个随机过程之间的有向信息传递量的统计量。转移熵不仅能反映两个随机变量之间的相互作用,同时也能描述两因素之间的相关度。根据信息理论,假设有两个随机时间序列 $x_n^k=\{x_n,x_{n-1},x_{n-2}\cdots x_{n-k+1}\}$, $y_n^l=\{y_n,y_{n-1},y_{n-2}\cdots y_{n-l+1}\}$,定义转移熵[13]:

$$T_{Y->X}=\sum_{x_{n+1}}p(x_{n+1},x_n^k,y_n^l)\log\left(\frac{p(x_{n+1}\mid x_n^k,y_n^l)}{p(x_{n+1}\mid x_n^k)}\right)$$

$$T_{X->Y}=\sum_{y_{n+1}}p(y_{n+1},x_n^k,y_n^l)\log\left(\frac{p(y_{n+1}\mid x_n^k,y_n^l)}{p(y_{n+1}\mid y_n^l)}\right)\quad(1)$$

其中,$T_{Y\to X}$表示在考虑了时间序列 $Y_n$ 值的情况下,状态 $X_n$ 到状态 $X_{n+1}$ 的变化,即 $X_n$ 到 $X_{n+1}$ 的变化是否与因素 $Y_n$ 的值有关。在本文中 $X_n$ 表示空气质量指数的时间序列值,$Y_n$ 表示要考察的某一个气象因子的时间序列值。时间序列 $Y_n$ 到 $X_n$ 的转移熵,实际上是 $Y_n$ 传递给 $X_n$ 的信息量。转移熵的值越大,说明 $Y_n$ 传递给 $X_n$ 的信息量越大,若转移熵的值为 0,说明 $Y_n$ 对 $X_n$ 的变化没有任何影响。本文将转移熵用于气象要素与空气质量指数的因果关系分析,相关性分析。

式(1)中 $p(x_{n+1},x_n^k,y_n^l)$ 是时间序列从状态 $(x_n^k,y_n^l)$ 到状态 $x_{n+1}$ 的转移概率,$p(x_{n+1}\mid x_n^k,y_n^l)$ 与 $p(x_{n+1}\mid x_n^k)$ 为条件概率,可以写:

$$p(x_{n+1}\mid x_n^k,y_n^l)=p(x_{n+1},x_n^k,y_n^l)/p(x_n^k,y_n^l)$$

$$p(x_{n+1}\mid x_n^k)=p(x_{n+1},x_n^k)/p(x_n^k)\quad(2)$$

在式(2)中 $x_n^k=\{x_n,x_{n-1},x_{n-2}\cdots x_{n-k+1}\}$ 与 $y_n^l=\{y_n,y_{n-1},y_{n-2}\cdots y_{n-l+1}\}$ 之间相互作用的时间延迟为 1,但 $x_n^k=\{x_n,x_{n-1},x_{n-2}\cdots x_{n-k+1}\}$ 与 $y_n^l=\{y_n,y_{n-1},y_{n-2}\cdots y_{n-l+1}\}$ 之间的相互作用可能有较长的时间延迟,因此,可以将式(1)写成:

$$T_{Y->X}=\sum_{x_{n+1}}p(x_{n+u},x_n^{(k)},y_n^{(l)})\log\left(\frac{p(x_{n+u},x_n^{(k)},y_n^{(l)})p(x_n^{(k)})}{p(x_n^{(k)},y_n^{(l)})p(x_{n+u},x_n^{(k)})}\right)$$

$$T_{X->Y}=\sum_{y_{n+1}}p(y_{n+u},x_n^{(k)},y_n^{(l)})\log\left(\frac{p(y_{n+u},x_n^{(k)},y_n^{(l)})p(y_n^{(l)})}{p(x_n^{(k)},y_n^{(l)})p(y_{n+u},y_n^{(l)})}\right)\quad(3)$$

其中,u 表示的是 $x_n^k=\{x_n,x_{n-1},x_{n-2}\cdots x_{n-k+1}\}$ 与 $y_n^l=\{y_n,y_{n-1},y_{n-2}\cdots y_{n-l+1}\}$ 之前相互作用的延迟时间。

为了更加准确地找出影响空气质量指数的气象因子,本文在使用转移熵的同时,提出使用增加特征变量的转移熵找影响空气质量指数的主要气象要素,公式中增加的特征变量为风场。

由于现有的提取影响空气质量指数的气象因子的方法如:统计分析方法与因子分析方法等没有考虑风速和风向即风场因素对污染物扩散的影响,而空气质量指数又是依据空气中污染物浓度的高低判断的,故会对预测的结果造成一定的影响。因为在不同的风速与风向情况下,各个气象因子对污染物的扩散有变化。因此,

在转移熵公式中增加一个特征变量即影响污染物浓度的主要气象因子风场的基础上再考虑其他的气象因子对空气质量指数 AQI 的影响,这有利于提高预测结果的准确度。

增加了特征变量风场以后,转移熵[14]的公式可以写成:

$$T_{Y->X|Z} = \sum_{x_{n+1}} p(x_{n+1},x_n^k,y_n^l,z_n) \log\left(\frac{p(x_{n+1}|x_n^k,y_n^l,z_n)}{p(x_{n+1}|x_n^k)}\right)$$

$$T_{X->Y|Z} = \sum_{y_{n+1}} p(y_{n+1},x_n^k,y_n^l,z_n) \log\left(\frac{p(y_{n+1}|x_n^k,y_n^l,z_n)}{p(y_{n+1}|y_n^l)}\right) \quad (4)$$

式(4)中的时间序列 $Z_n$ 在本文中代表风场,式(4) 在该预测中表示的含义是在考虑风场 $Z_n$ 的条件下,气象要素 $Y_n$ 对空气质量指数 $X_n$ 的信息转移熵。

### 1.2 熵值法在预测中的应用

为了克服单一预测模型的不准确,本文提出利用组合预测模型提高空气质量指数的预测。组合预测模型中权重由熵值法确定。

在信息论中,熵是不确定性和无序性的度量,熵值的大小就代表不确定性大小,如果熵值小,则不确定性就小,那么所包含的信息量就越多;如果熵值大,则不确定性就大,那么所包含的信息量就越少,根据此特性,本文将熵值作为各个预测模型的预测精度的度量,用熵值法确定各个预测模型在组合预测中的权重,具体的方法如下[14]:

(1)日 AQI 数据序列为 $\{x_t,t = 1,2,\ldots,n\}$,定义第 i 种预测模型第 t 时刻的相对误差为 $e_{it}(i = 1,2,\ldots,m,t = 1,2,\ldots,n)$ 且 $e_{it} \in [0,1]$,$\{e_{it},t = 1,2,\ldots,n\}$ 为第 i 种预测模型第 t 时刻预测相对误差序列。

$$e_{it} = \begin{cases} 1, |x_t\text{-}x_{it}/x_t| \geqslant 1 \\ |x_t\text{-}x_{it}/x_t|, 0 \leqslant |x_t\text{-}x_{it}/x_t| < 1 \end{cases}$$

(2)将各个预测模型的预测相对误差序列单位化,即计算第 i 种预测模型在第 t 时刻时的预测相对误差的比重 $p_{it}$。

$$p_{it} = \frac{e_{it}}{\sum_{t=1}^{n} e_{it}}, t = 1,2,\ldots,n$$

其中,$\sum_{t=1}^{n} p_{it} = 1, i = 1,2,\ldots,m$

(3)计算各个预测模型的预测相对误差的熵值,$h_i$ 表示第 i 种预测模型的预测相对误差的熵值。

$$h_i = -k\sum_{t=1}^{n} p_{it}\ln p_{it}, i = 1,2,\ldots,m$$

其中 k 为常数且 k >0,熵值。$h_i \geqslant 0, i = 1,2,\ldots,m$。

对第 i 种预测模型而言,如果 $p_{it}$ 全部相等,即 $p_{it} = 1/n, t = 1,2,\ldots,n$,那么 $h_i$ 取极大值,将 $p_{it} = 1/n$ 带入熵值公式得 $h_i = k\ln(n)$,取 $k = 1/\ln(n)$ 则 $h_i$ 的取值范围为:$0 \leqslant h_i \leqslant 1$。

(4)计算第 i 种预测模型的预测相对误差序列的变异程度系数 $d_i$,根据系统某项指标的熵值的大小与其变异程度相反的原则,定义第 i 种预测模型的预测相对误差序列的变异程度系数 $d_i$ 为:$d_i = 1 - h_i, i = 1,2,\ldots,m$。

(5)设各种预测模型的加权系数为 $w_1,w_2,\ldots,w_m$:

$$w_i = \frac{1}{m-1}\left(1 - \frac{d_i}{\sum_{i=1}^{m} d_i}\right)$$

并且权系数满足公式:$\sum_{i=1}^{m} w_i = 1$。

(6)计算组合预测值 $f_t$:

$$f_t = \sum_{i=1}^{m} w_i x_{it}, t = 1,2,\ldots,n$$

## 2 预测模型的建立

在大气污染源不变的条件下,空气质量指数的变化主要是由气象条件所引起,如何准确地找出影响空气质量指数变化的气象要素至关重要。本文提出使用增加特征变量的转移熵准确的找出影响空气质量指数的气象因子。再用找出的气象因子与实测 AQI 数据构建 BP 神经网络预测模型和时间序列分析模型。模型构建好了之后,用相应气象要素数的数值预报数据欧洲细网格数据(ecmwf_thin)及当前 AQI 数据对未来空气质量指数进行预报。最后用组合预测模型提高空气质量指数的预测,组合预测模型中各个预测模型的权重由熵值法确定。具体的预测模型的建立步骤如下:

(1)首先用增加特征变量的转移熵准确地找出影响空气质量指数的气象因子及其信息转移熵。

(2)用找出的气象因子及其信息转移熵与当前实测的 AQI 值构建 BP 神经网络预测模型和时间序列分析模型。

(3)预测模型构建好之后,用相应气象要素数的数值预报数据即欧洲细网格数据(ecmwf_thin)及当前 AQI

数据对未来空气质量指数进行预报。

(4)然后用熵值法将上述两种模型的预测结果进行组合,即用组合预测的方式来提高预测的精度。

(5)预测结果评价。

## 3　实例分析

### 3.1　数据来源

实验采用的数据包括空气质量指数数据和气象要素数据。空气质量指数数据是成都市 2016 年 10 月 ~ 2017 年 1 月公布的逐日实测 AQI 数据。用于训练的气象数据集采用成都市同期的实测气象要素数据。用于预测的气象要数数据采用欧洲细网格数据(ecmwf_thin)数值预报数据。

### 3.2　实验分析

由于污染物的迁移传输需要一定的时间,故前日 AQI 对当日 AQI 有较大影响,因此前日 AQI 可以在一定程度上描述污染源的特征[15-17]。在污染源不变的条件下,污染物的扩散与沉降等能力和气象条件有着十分重要的关系。

首先利用增加特征变量的转移熵算法找出影响空气质量指数的气象要素,通过分析计算得出影响空气质量指数的主要气象要素及其信息转移熵见表 1。

**表 1　影响 AQI 的主要气象因子及其信息转移熵**

| 影响因子 | 风场 | 能见度 | 空气湿度 | 气温 | 气压 | 降雨 |
|---|---|---|---|---|---|---|
| 信息转移熵 | 0.061 | 0.040 | 0.035 | 0.033 | 0.024 | 0.015 |

用找出的影响空气质量指数的气象因子及其信息转移熵与当前实测的 AQI 值构建空气质量指数的 BP 神经网络预测模型(模型 1)和时间序列分析模型(模型 2)。预测成都市 2017 年 01 月 15 日 ~2017 年 01 月 27 日的 AQI 值。

用组合预测方法对空气质量指数进行预测,组合预测方法的权重由熵值法确定。首先求出各个预测模型的相对误差序列。再将相对误差序列单位化。然后计算各个预测模型预测相对误差的熵值 $h_i$ 和 $d_i$,模型 1 的熵值 $h_1$ =0.844 097 55 和 $d_1$ =0.155 902 45,模型 2 的熵值 $h_2$ =0.841 761 8 和 $d_2$ =0.158 238 17。于是得到各个预测模型的权重 $w_i$,其中模型 1 的权重 $w_1$ =0.503 717 634 478 470 2,模型 2 的权重 $w_2$ =0.496 282 365 521 529 7。分别设模型 1 和模型 2 的预测值为 PredictiveValue1 和 PredictiveValue2,最后得到组合预测模型(模型 3)的表达式为:

$$f_t = 0.5037176344784702 * \text{PredictiveValue1} + 0.4962823655215297 * \text{PredictiveValue2}$$

表 2 为成都市 2017 年 01 月 15 日 ~2017 年 01 月 27 日的实测值以及预测模型的预测值。

**表 2　实测值及其预测值**

| 日　期 | 实测值 | 预测值 | | |
|---|---|---|---|---|
| | | 模型 1 | 模型 2 | 模型 3 |
| 2017-01-15 | 186.20 | 188.09 | 186.18 | 187.15 |
| 2017-01-16 | 133.50 | 137.61 | 137.29 | 137.45 |
| 2017-01-17 | 142.58 | 145.98 | 151.29 | 148.61 |
| 2017-01-18 | 126.04 | 128.39 | 140.71 | 134.50 |
| 2017-01-19 | 124.08 | 131.79 | 127.78 | 129.80 |
| 2017-01-20 | 146.46 | 140.68 | 127.21 | 133.99 |
| 2017-01-21 | 227.83 | 151.94 | 155.19 | 153.55 |
| 2017-01-22 | 248.33 | 253.51 | 252.34 | 252.93 |
| 2017-01-23 | 226.62 | 207.76 | 249.42 | 228.43 |
| 2017-01-24 | 241.58 | 227.75 | 229.31 | 228.52 |
| 2017-01-25 | 271.38 | 241.43 | 252.28 | 246.81 |
| 2017-01-26 | 252.83 | 255.33 | 282.43 | 268.77 |
| 2017-01-27 | 245.25 | 238.81 | 252.11 | 245.41 |

本文采用平均绝对差(MAE)、均方差(MSE)、均方根差(RMSE)作为预测结果的误差评价指标。误差评价指标的取值范围是 0 到正无穷大,当误差指标值为 0 时,表示观测值与预报完全一致,误差指标值越小说明预报越精确。表 3 为三种模型预测结果的误差指标值。

**表 3　误差评价指标值**

| | MAE | MSE | RMSE |
|---|---|---|---|
| 模型 1 | 7.917 316 128 7 | 126.943 347 215 | 11.266 913 828 3 |
| 模型 2 | 11.455 166 829 5 | 206.253 418 232 | 14.361 525 623 4 |
| 模型 3 | 7.712 672 349 7 | 105.538 577 117 | 10.273 197 025 |

由表 3 可知,组合预测模型(模型 3)的误差评价指标值均低于其他两个模型的值。因此组合预测模型的预测精度相比于其他两个预测模型有所提高。

## 4　结束语

为了提高空气质量指数预测的精确度,本文将信息论中的熵引入空气质量指数预测的研究。首先采用增加了特征变量的转移熵方法找出影响空气质量指数的气象因子,再用找出的气象因子与实测 AQI 构建神经网络预测模型和时间序列分析模型。然后将熵值法确定组合预测模型中各个预测模型的权重,将传统单一预测模型转为组合预测模型。本文将该方法用于预测成都市空气质量指数的预测,实例结果表明该方法能提高预测精度。

**参 考 文 献:**

[1] 张学文.相对分布函数和气象熵[J].气象学报,1986(2):88-93.

[2] 张学文,马力.熵气象学简介[J].气象,1995,21(1):52-56.

[3] 刘萍.基于主成分分析和多元线性回归模型的空气质量评价方法研究[D].昆明:云南大学,2015.

[4] 普映娟,王琳邦.保山城区空气污染指数的时间序列分析[J].保山学院学报,2010,29(2):10-12.

[5] SCHREIBE T.Measuring information transfer[J].Phys Rev Lett,2000,85(2):461-464.

[6] 马超飞.基于转移熵的神经电信号分析研究[D].上海:华东理工大学,2013.

[7] 陈悦辰.基于转移熵方法的市场有效性评价及不同系统性风险股票与收益率之间的信息流分析[D].北京:北京交通大学,2014.

[8] 于萍.时间序列分析在空气质量指数(AQI)预测中的应用[D].大连:辽宁师范大学,2015.

[9] 南亚翔,李红利,修春波,等.基于卡尔曼滤波的空气质量指数预测方法[J].环境科学导刊,2016,35(3):80-84.

[10] 王珍.基于因子分析-BP 神经网络模型在空气质量综合评价中的应用[D].昆明:云南大学,2015.

[11] 祝翠玲,蒋志方,王强.基于 B-P 神经网络的环境空气质量预测模型[J].计算机工程与应用,2007,43(22):223-227.

[12] 郭庆春,何振芳,李力.人工神经网络在大气污染预测中的应用研究[J].工业仪表与自动化装置,2012,17(4):18-22.

[13] 叶中行.信息论基础[M].北京:高等教育出版社,2006.

[14] MONTALTO A,FAES L,MARINAZZO D.MuTE:A MATLAB Toolbox to Compare Established and Novel Estimators of the Multivariate Transfer Entropy[J].Plos One,2014,9(10):e109462.

[15] 陈华友.熵值法及其在确定组合预测权系数中的应用[J].安徽大学学报:自然科学版,2003,27(4):1-6.

[16] 周秀杰,苏小红,袁美英.基于 BP 网络的空气污染指数预报研究[J].哈尔滨工业大学学报,2004,36(5):582-585.

[17] 黎洁仪,梁之彦,杨国杰.广州市空气污染影响因子与预报建模[J].广东气象,2013,35(4):47-50.

## Application of Entropy in Air Quality Index (AQI) Prediction

*WEN Qin, LUO Fei*

(College of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

**Abstract**: In order to accurately extract the meteorological factors that affect the air quality index and improve the prediction accuracy, a prediction model based on entropy, BP neural network and time series model is proposed. This method uses the information transfer entropy with the characteristic variables to obtain the characteristic factor and the specific influence degree. The obtained characteristic factor and measured values of AQI are used as the input factor of the BP neural network and the characteristic factors of the time series analysis model, the influence degree is the initial weight of the BP neural network, construct BP neural network and time series analysis model, finally, the results of each prediction model are composed by the entropy method. The experiment shows that the This method can improve the stability and the predict accuracy of the forecast of air quality index.

**Key words**: air quality index forecasting; transfer entropy; entropy method