文章编号：1006-4354（2006）06-0008-04

气象时间序列规则发现及其应用

左爱文[1,2]，郭宏武[3]，王保保[1]

（1. 西安电子科技大学，西安 710000；
2. 陕西省人工影响天气办公室，西安 710014；3. 西安市气象局，西安 710016）

中图分类号：P409 文献标识码：A

1 时间序列及挖掘

时间序列是指将某一指标在不同时间上的不同数值，按照时间的先后顺序排列而成的数列[1]。随着信息技术的广泛使用以及人们获取数据手段的多样化，人类所拥有的时间序列信息急剧增加。按照研究对象和问题的不同，可以得到各种时间序列。例如产品销售记录、股票价格数据、气象数据、医疗信息等。目前计算机存储的数据中，时间序列数据占据了相当大（约 80%）的比例，面对如此海量的时间序列数据，人们想找到有效的方法或技术来揭示这些数据内部所隐藏的知识或信息。例如股票经纪人想从某一种股票每日收盘价格的历史记录中发现其变化规律，以预测该股票未来行情走势；气象工作者想从降水量的历史变化中发现其变化规律，以预测未来降水量的变化趋势等等。

时间序列挖掘通过对过去历史客观记录分析，揭示其内在规律（如波动周期、趋势等），进而完成预测未来行为的决策性工作。人们希望通过对时间序列的分析，从大量的数据中发现和揭示某一现象的发展规律以及与其他现象之间的内在关系，并分析未来变化趋势，这就是对时间序列数据的挖掘。数据挖掘是从大量的、不完全的、模糊的、甚至是随机的数据中，提取隐含在其中潜在的有用信息和知识的过程。而时间序列数据的挖掘是挖掘时序数据中潜在的有用的知识或信息，时序数据挖掘已经成为信息领域的研究热点之一。

2 气象时间序列的研究

气象部门积累了大量的基于时间序列的数据，对气象资料时间序列的研究由来已久。传统的气象时间序列研究主要是建立时间序列的预测或分析模型，以数学表达式形式表示的时间模型对时间序列进行趋势分析或预测[2]。如运用自回归模型（AR）、动平均模型（MA），自回归动平均模型（ARMA）、累计式自回归——动平均模型（ARIMA）等对气象历史数据进行分析，判断何种模型适合气象数据变化过程，对模型和参数进行估计，对模型进行合理性检验，建立预报模型，在预报过程中对预报模型进行校正等等。当模型的准确率可以为预报员接受时，模型就以数学公式的形式确定下来，对未来的天气、降水量或气候变化等进行预测。对未来预测只是一种机械的固定的映射关系，灵活性较差。然而现实中气象时间序列数据非常复杂，观测值通常是以离散的跳跃式出现，并不像连续函数所描述的那么简单，因而不可能用连续函数来精确的描述其变化趋势。因此，提出利用时间序列的关联挖掘技术及其挖掘算法对气象年降水量时间序列进行分析来发现其上升、下降的子序列的特点，从中发现关联规则，以指导年降水量的预测或预报。

3 相关概念及规则发现方法

设 $U=\{x_t|t=1, 2, \cdots, n\}$ 为一时间序列，按照 x_t 发生时间的先后顺序，把 U 改写成向量形式 $U=(x_1, x_2, \cdots, x_n)$，即有 x_i 发生在 x_j 之前（$i<j$），称 x_i 与 x_{i+1} 在序列 U 中是相邻的（$i=$

收稿日期：2006-07-18

作者简介：左爱文（1969-），女，陕西泾阳人，高级工程师，学士，从事计算机应用开发。

1，2，…，$n-1$)[3]。

定义1　称$<x_i, x_{i+1}, \cdots x_{i+k-1}>$为时间序列$U$的一个子时间序列，$k$称为该子时间序列的长度。其中$x_{i+j}$和$x_{i+j+1}$是相邻的且其先后顺序不变($j=1, 2, \cdots, k-1$)。

由定义1可知，U的长度为k的全部子序列为$\{<x_i, x_{i+1}, \cdots, x_{i+k-1}> \mid i=1, 2, \cdots, n-k+1\}$。$U$的长度为$k$的子时间序列集记为$S(U, k)$，此时，$|S(U, k)|=n-k-1$。

定义2　设$s=<x_i, x_{i+1}, \cdots, x_{i+k-1}> \in S(U, k)$，如$x_i \leqslant x_{i+1} \leqslant \cdots \leqslant x_{i+k-1}$，则称$s$为$U$的一个长度为$k$的上升子时间序列；反之，如$x_i \geqslant x_{i+1} \geqslant \cdots \geqslant x_{i+k-1}$，则称$s$为$U$的一个长度为$k$的下降子时间序列。

定理　设$s=<x_i, x_{i+1}, \cdots, x_{i+k-1}> \in S(U, k)$，如果$s$为$U$的长度为$k$的上升(下降)子时间序列，$<x_i, x_{i+1}, \cdots, x_{i+k-1}, x_{i+k}>$(当$i+k \leqslant n$)和$<x_{i-1}, x_i, \cdots x_{i+k-1}>$(当$i>2$时)都不是$U$的上升(下降)的子时间序列，此时称$s$为$U$的一个长度为$k$的极大上升(下降)的子时间序列。

例1　设$U=$(2，3，4，5，5，6，5，3，2，1)，则$S(x, 4)=\{<2, 3, 4, 5>, <3, 4, 5, 5>, <4, 5, 5, 6>, <5, 5, 6, 5>, <5, 6, 5, 3>, <6, 5, 3, 2>, <5, 3, 2, 1>\}$，其中$<2, 3, 4, 5>$，$<3, 4, 5, 5>$，$<4, 5, 5, 6>$是$U$的长度为4的上升子时间序列，而$<2, 3, 4, 5, 5, 6>$是$U$的长度为6的极大上升子时间序列；$<6, 5, 3, 2>$，$<5, 3, 2, 1>$是$U$的长度为4的下降子时间序列，而$<6, 5, 3, 2, 1>$是$U$的长度为5的极大下降子时间序列。

定理结论的正确是显然的。定理为判断U的长度为k的子时间序列是否为极大上升(下降)子时间序列提供了依据。如果已知时间序列U的极大上升子时间序列为k，而现在U的第$n-k+1$到第n的观测值分别满足$x_{n-k+1} \leqslant x_{n-k+2} \leqslant \cdots \leqslant x_n$，，且$x_{n-k+1}<x_n$，则可以推测未来的$x_{n+1}$应该满足$x_n>x_{n+1}$，即得到这样的规则：

“if$x_{n-k+1} \leqslant x_{n-k+2} \leqslant \cdots \leqslant x_n$ and $x_{n-k+1}<x_n$ then $x_n>x_{n+1}$”，从而确定了时间序列的趋势走向，达到了对时间序列进行趋势分析或预测的目的。

研究更一般的情形。设$s=<x_i, x_{i+1}, \cdots, x_{i+k-1}>$为$U$的一个长度为$k$的上升子时间序列，现在研究规则“如果$s$是上升的子时间序列，则$x_{i+k-1}>x_{i+k}$”的支持度和置信度问题。支持度和置信度是衡量规则的兴趣度，分别反映发现的规则的有用性和可信的程度。很显然，规则“如果s是上升子时间序列，则$x_{i+k-1}>x_{i+k}$”为真，当且仅当$<x_i, x_{i+1}, \cdots, x_{i+k-1}, x_{i+k}>$不是$U$的一个长度为$k+1$的上升的子时间序列，而规则“如果$s$是上升的子时间序列，则$x_{i+k-1}>x_{i+k}$”为假，当且仅当$<x_i, x_{i+1}, \cdots, x_{i+k-1}, x_{i+k}>$是$U$的一个长度为$k+1$的上升子时间序列。因而对于长度为$k$的上升子时间序列集$|s_u(k)| \geqslant 1$，规则“如果$x_i \leqslant x_{i+1} \leqslant \cdots \leqslant x_{i+k-1}$，则$x_{i+k-1}>x_{i+k}$”的置信度(Confidence)为$\frac{|s_u(k)|-|s_u(k+1)|}{|s_u(k)|}$，支持度(Support)为$\frac{|s_u(k)|-|s_u(k+1)|}{n-k+1}$。同样，对于$U$的长度为$k$的下降子时间序列的情形，如果$|s_d(k)| \geqslant 1$，规则“如果$x_i \geqslant x_{i+1} \geqslant \cdots \geqslant x_{i+k-1}$，则$x_{i+k-1}<x_{i+k}$”的置信度为$\frac{|s_d(k)|-|s_d(k+1)|}{|s_d(k)|}$，支持度为$\frac{|s_d(k)|-|s_d(k+1)|}{n-k+1}$。

4　挖掘算法研究

基于以上分析，给出求U的上升子时间序列集$s_u(k)$、极大上升子时间序列、下降子时间序列集$s_d(k)$、极大下降子时间序列、以及规则的支持度和置信度的算法。

算法1：求U的长度为k的子时间序列集$S(U, k)$

input U，k

$S(U, k) \leftarrow \Phi$

For ($t=1$；$t \leqslant n-k+1$；$t++$) {

$x(t)=<x_t, x_{t+1}, \cdots, x_{t+k-1}>$；

$s(U, k) \leftarrow S(U, k) \cup \{x(t)\}$；}

output　$S(U, k)$

算法 2：求 X 的上升的子时间序列集 s_u (k)

Rising function (U, k) //求上升子时间序列集 s_u (k) 的函数

{input U, k

s_u (k) ←Φ //s_u (k) 初始值为空

for (t=1; t≤n−k+1; t++) {

x (t) =<x_t, x_{t+1}, …, x_{t+k-1}>;

For (i=t; i≤t+k−1; i++)

If (x_i>x_{i+1}) break;

If (i>t+k−1) s_u (k) ←s_u (k) ∪ {x (t)};}

Return s_u (k)}

算法 3：求 x 的极大上升子时间序列

step1 input U, m //m∈ (1, n)

step2 rising function (U, m) //求极大上升子时间序列的 k 值

while (s_u (k) ≠Φ)

{ m=m+1;

rising function (U, m);

if (s_u (k) =Φ) k=m−1 goto step3;}

while (s_u (k) =Φ)

{ m=m−1;

rising function (U, m)

if (s_u (k) ≠Φ) k=m+1 goto step3;}

step3 rising function (U, k) // 求极大上升的子时间序列 s_u (k)

算法 4：求 x 的下降的子时间序列集 s_d (k)

descending function (x, k) //求下降子时间序列集 s_d (k) 的函数

{ input x, k

s_d (k) ←Φ //s_d (k) 初始值为空

for (t=1; t≤n−k+1; t++) {

x (t) =<x_t, x_{t+1}, …, x_{t+k-1}>;

for (i=t; i≤t+k−1; i++)

if (x_i<x_{i+1}) break;

if (i>t+k−1) s_d (k) ←s_d (k) ∪ {x (t)};}

return s_d (k)}

算法 5：求 x 的极大下降子时间序列

step1 input x, m //m∈ (1, n)

step2 descending function (U, m) //求极大下降子时间序列的 k 值

while (s_d (k) ≠Φ)

{ m=m+1;

descending function (U, m);

if (s_d (k) =Φ) k=m−1 goto step3;}

while (s_d (k) = Φ)

{ m=m−1;

descending function (U, m)

if (s_d (k) ≠Φ) k=m+1 goto step3;}

step3 descending function (U, k) // 求极大下降子时间序列 s_d (k)

算法 6：求上升规则的支持度和置信度

input U, k

rising function (U, k) //求 s_u (k)

rising function (U, k+1) //求 s_u (k+1)

support= (|s_u (k) |−|s_u (k+1) |) / (n−k+1) //求支持度

confidence= (|s_u (k) |−|s_u (k+1) |) /|s_u (k) | //求置信度

output support, confidence

算法 7：求下降规则的支持度和置信度

input U, k

descending function (U, k) //求 s_d (k)

descending function (U, k+1) //求 s_d (k+1)

support= (|s_d (k) |−|s_d (k+1) |) / (n−k+1) //求支持度

confidence= (|s_d (k) |−|s_d (k+1) |) /|s_d (k) | //求置信度

output support, confidence

5 实例应用

利用西安气象观测站(57036)1951 年至 2000 年共 50 a 的年降水量进行时间序列分析研究，希望找出序列之间的联系与规则。U=(523.4，795，551.7，642.2，……，362，600.5，589.5，539)，利用上述算法得到|S (U, 3) |=48，s_u (3) ={<585.4，743.2，839>，<384.4，562.4，621.2>，……，<384.4，562.4，621.2>} 共计 5 个长度为 3 的上升子时间序列，|s_u|=5，s_d (3) ={ <642.2，591.1，585.4>，<621.2，556.5，494.1>，……，<600.5，589.5，539>} 共计 8 个长度为3的下降子时间序列，|s_d(3)|=8，

文章编号：1006-4354（2006）06-0011-04

小波分析在陕西省旱涝气候预测中的应用

高　炬[1]，王繁强[2]，黄祖英[3]

（1. 陕西省气象台，西安　710014；2. 陕西省气象科学研究所，西安　710014；
3. 陕西省气候中心，西安　710014）

摘　要：根据榆林、西安和汉中3个代表性测站534 a的旱涝资料，利用小波变换技术对陕西不同时间尺度旱涝周期比较和诊断，并对陕西未来的旱涝趋势作了初步分析。结果表明，小波分析方法在旱涝气候趋势预测中具有一定的应用前景。

关键词：小波分析；旱涝；气候预测

中图分类号：P456.3　　**文献标识码**：A

陕西省下垫面性质复杂，地形地貌特征多样化，区域内不同地域的降水存在明显差异。除陕南外其它地区降水量仍然不足，而且年际波动较大，时常出现旱涝灾害，尤其是干旱，使得对气候变化十分脆弱的生态环境更易于向恶化方向演变[1]，是该地区影响最大的气候灾害之一，往往会造成严重的经济损失。不少气象工作者对陕西省旱涝气候特征进行了研究[2,3]。为进一步了解陕西旱涝不同时间尺度上的频谱，并用以初步预测旱涝趋势，利用小波分析方法对陕西3个代表性测站（榆林、西安和汉中）的534 a旱涝资料进行分析，得到一些有意义的结果。

1　小波分析原理和方法

小波变换适用于非平稳信号的分析，与传统的Fourier变换相比，优点在于频域和时域上同时具有良好的局部化性质，而且由于其不同的频率成分在时域上的取样步长具有调节性，即高频小，低频大[4]，因此，对高频成分采用逐渐精细的时域或空域的取样步长，可以聚焦到任何细节，可称为数字显微镜[5,6]。

用Matlab小波工具箱提供的小波分析程序[7]，采用Mexican Hat（mexh）小波函数。用

收稿日期：2006-07-14

作者简介：高炬（1949-），男，陕西白水人，高工，从事天气预报工作。

$s_u(4)=\{<525.4, 547.7, 626.5, 671.6>\}$一个长度为4的极大上升子时间序列，$s_d(4)=\{<903.2, 665, 491.2, 402.8>\}$一个长度为4的极大下降子时间序列，$|s_d(4)|=1$。由于最大的上升子时间序列长度为4，因而得到规则："如果年降雨量连续4 a是上升的，则接下来的年降雨量是下降的"同理，最大的下降子时间序列长度为4，则得到规则"如果年降雨量连续4 a是下降的，则接下来的年降雨量是上升的"。但该规则的支持度和置信度较低，不利于实际工作中对年降雨量的预测。规则"如果年降雨量连续3 a是上升的，则接下来的年降雨量是下降的"的支持度为8.3%，置信度为80%；规则"如果年降雨量连续3 a是下降的，则接下来的年降雨量是上升的"的支持度为14.6%，置信度为87.5%。由此可知，连续3 a降雨量的上升或下降得出的规则的置信度高，可以在年雨量的预报或预测中作为参考。

参考文献：

[1]　毛国君，段立娟，王实，等．数据挖掘原理与算法[M]．北京：清华大学出版社，2006：183.

[2]　丁裕国，汪志红．气象数据时间序列信号处理[M]．北京：气象出版社，1998：55-69.

[3]　王勇，张新政，高向军．时序规则发现及其算法[J]．计算机应用研究，2005（6）．