基于 n 阶原子模板的时间序列相似匹配算法*)

汤 胤 彭 宏 郑启伦
(暨南大学管理学院电子商务教研室 广州 510632)

摘 要 本文以时间序列一、二阶原子模式的定义为基础,推导基于 n 阶原子模式的构造,研究了原子模式之间的偏序相似关系,使得序列能进行细腻的比较,并由此提出了基于模板匹配的算法。实验表明,基于模板匹配的算法与同类方法和传统方法比较在精度上和性能上都有较大优势。

关键词 时间序列,模式识别,原子模式

n-th Order Template-Based Matching Algorithm for Time Series

TANG Yin PENG Hong ZHENG Qi-Lun
(Lab of E-Commerce, College of Management, Jinan Univ., Guangzhou 510632)

Abstract Based on definitions of 1st and 2nd order of irreducible pattern of time series, this paper deduces n-th order of irreducible pattern, where partial relationship of similarity within these patterns is examined. These patterns enable more refined comparison between sequences, based on which we propose Template-Based Matching Algorithm. The experimental result has verified distinct advantages over some similar and classical approaches both in accuracy and performance.

Keywords Time series, Pattern recognition, Atomic pattern

1 引言

经典的时间序列分析方法主要包括:移动平均法(Moving Average)[1,2]、指数平滑方法(Exponential Smoothing)[20]、Box-Jenkins 的 ARMA/ARIMA 方法[21],以及 Ge and Smith 的 Deformable Markov 模板[7]、贝叶斯模型[18,19]以及一些模糊集方法[8~10]等等。随着计算技术的快速发展,时间序列建模方法从早期的数学建模,逐渐转向降维与外形建模。例如离散傅里叶变换(DFT)[13~17]将序列转换为一系列多维空间点;动态时间弯曲(Dynamic Time Warping)[11,12]将时间拉长或缩短,使得不同时间长度的序列可以比较;路标(Landmark Similarity[4])采集序列特征点,形成最小特征集合,大大减少了存储和计算量。Branko Pecar 的 APRE 方法[6]也从模式识别角度对时序预测做了初步的尝试。欧式距离(Euclidean Distance)将每个序列视为 n 维点并计算其欧式距离,其复杂度为 O(n);Goldin 与 Kanellakis[23]的距离标准化方法(Normalization method)则将均值和方差标准化,以方便比较。

经典的时间序列建模方法,往往假定时间序列服从概率分布,或服从一定的数学模型。本文提出的方法则使用一些基本原子模式组合去进行匹配,称为 TBM(Template-Based Matching),类似于模式识别。本文基于"由粗到精"进行序列匹配的思想,给出了考量序列外形"偏序"相似的标准,并给出一个时间序列挖掘和推理的算法。实验结果证明,该算法提高了时序预测的精度和速度。

2 核心思想

时间序列是由一系列相关的点组成的序列,每个点值与前一个发生的点不外乎"升"、"降"、"平"三种关系。Branko Pecar 的 APRE 方法[6]将数值序列做差分后,按照"升"、"降"、"平"用英文字母"P","N","Z"代替,然后进行序列搜索和匹配,匹配的原则还包括两个序列的均值、标准差等等一些指标[2]。这种方法相对有些苛刻。如果两个序列十分相似,仅仅是在某个点上的一些升降的区别,就容易影响它们相似的计算。

本文的工作从"P"、"N"、"Z"这些基本的模式(定义为一阶原子模式)入手,定义二元运算,从而得到 13 个二阶原子模式,再往上进行复合,又可以得到 169 个三阶原子模式。考察它们之间的偏序关系,在序列挖掘时用这些原子模式组合成的模板去匹配序列,在匹配时根据这些偏序关系做比较宽容但合理的判断。这种方法的好处在于可以将不该漏掉的序列容纳在考虑范围。随着阶数的升高,精度提高,可匹配的序列不会减少。

3 原子模式的定义和模板构造

3.1 一阶原子模式

定义 1 一阶原子模式 $X'\{x'_i\}$ 为时间序列 $X\{x_i\}$ 进行一阶差分后的序列,序列 $M\{m_i\}$ 定义为:

$$\text{当时 } x'_i=\begin{cases}>0\\<0\\=0\end{cases}\text{时,分别有 } m_i=\begin{cases}\text{"}P\text{"}\\\text{"}N\text{"}\\\text{"}Z\text{"}\end{cases} \quad (1)$$

这里的"P"、"N"、"Z"称为一阶原子模式,序列 $M\{m_i\}$ 称为一阶模式序列。一阶原子模式十分简单,对于一个时间序列来说,其一阶模式序列丢失了很多原序列内在的信息。因此光是研究一阶模式是远远不够的,还需要研究其曲线的形状,比如峰谷、凹凸等等,这些形状往往包含着丰富的信息。对一个时间序列点,其影响作用最大的是其最临近序列点。因此不仅需要描述增长关系,还要描述增长率的关系,即研究序列点前后的点组成的夹角。因此,原子模式设计遵循下列

*)本文受暨南大学博士启动基金资助(51104653)。汤 胤 博士,讲师,研究方向:人工智能、电子商务、知识管理。

原则:(1)要区别增长和下降;(2)要区别加速增长和减速增长;(3)要能体现与 X 轴和与 Y 轴夹角的明显区别。

3.2 二阶原子模式

$X''\{x''_i\}$ 为时间序列 $X'\{x'_i\}$ 进行一阶差分后的序列,即 $X\{x_i\}$ 进行二阶差分后的序列。序列 $M'\{m'_i\}$ 定义见表 1。

这里的 m'_i 称为二阶原子模式,序列 $M'\{m'_i\}$ 则称为二阶模式序列。

注:这里的 PP+, PP-, NN+, NN-等是为行文方便所取的代号,可以十分直观地联想到相应的形状,在计算机程序里会用一个单独的字母代替。另外,Z 和 ZZ 在一般的序列里都是较难遇到的,在实际的算法里,$x_i - x_{i-1}$ 只要小于某个极小的数值就算"Z",这样更具有实际意义。

表1 二阶原子模式的定义

m'_i	符号	条件	形状解释
PP+	E	$x'_i>0$, $x'_{i+1}>0$, $x''_i>0$	上升接加速上升
PP-	C	$x'_i>0$, $x'_{i+1}>0$, $x''_i<0$	上升接减速上升
PN	A	$x'_i>0$, $x'_{i+1}<0$	上升接下降
NN-	K	$x'_i<0$, $x'_{i+1}<0$, $x''_i>0$	下降接减速下降
NN+	I	$x'_i<0$, $x'_{i+1}<0$, $x''_i<0$	下降接加速下降
NP	V	$x'_i<0$, $x'_{i+1}>0$	下降接上升
PP0	D	$x'_i>0$, $x'_{i+1}>0$, $x''_i=0$	上升接减速上升
NN0	J	$x'_i<0$, $x'_{i+1}<0$, $x''_i=0$	下降接减速下降
ZZ	O	$x'_i=0$, $x'_{i+1}=0$	平行+平行
ZP	F	$x'_i=0$, $x'_{i+1}>0$	平行+上升
ZN	H	$x'_i=0$, $x'_{i+1}<0$	平行+下降
PZ	B	$x'_i>0$, $x'_{i+1}=0$	上升+平行
NZ	L	$x'_i<0$, $x'_{i+1}=0$	下降+平行

3.3 原子模式的偏序关系

一般模式识别中,可以简单地将这些算子直接运用到时序的匹配当中,也可以得到比较好的效果,但还不能完全满足需要。如图 1,PZ 与 PN,PN 与 ZN 往往只是一些细微的差别,而 NN-与 ZN 的差别似乎也并不大,或许可以将它们归为一类,但 PZ 与 NN-似乎差别越来越大了,如果在匹配过程中忽略这些相近和相远的情况,显然容易造成错误。不难发现,上述 13 种二阶模式之间,是有着相似距离的远近关系的,而这些关系不是像"1<2,2<3 则 1<3"这样完全线性的。因此有必要研究这些模式之间的偏序相似关系。

手工寻找二阶模式之间的关系并不困难,但为了给三阶以及更高阶的模式复合提供严密的基础,必须从一阶开始定义。

令 H 代表二阶原子模式的集合,$\prec$ 表示这个集合上的一个关系,对于 $\forall x, y \in H$,如果把 $x \prec y$ 定义为原子模式 x 向下的夹角和小于等于原子模式 y 向下的夹角和。思想大概是这样:希望能证明 $\prec$ 为 H 上的偏序关系,并且定义出集合 H 上的操作,使得寻找上界的时候,为各取 x, y 两枝与竖直方向夹角之大者 $x \vee y$;寻找下界的时候为各取 x, y 两枝与竖直方向夹角之小者 $x \vee y$,最终得到完全格,完全格的乘积是完全格,这样就可以往上一阶一阶地复合原子模式,得到越来越复杂而精细的结果。

为证明方便,将这一系列模式替换成{1,-1,0,.9,-.9}的组合(图 1)。不难看出,证明完全等价的:

(1)定义 2 个简单的完全格

设 $S=\{-1,-.9,0,.9,1\}$,S 对于数值关系"大于或等于"($\geqslant$)也做成完全格,记为 $L_1(S, \prec_2, \cap_2, \cup_2)$,其中:① $\prec_2$ 即为"大于或等于"($\geqslant$);②对任意的 $x, y \in S$, $x \cup_2 y = \min(x,y)$, $x \cap_2 y = \max(x,y)$。

对于数值性质的"小于或等于"($\leqslant$),S 做成一个完全格,记为格 $L_2(S, \prec_1, \cup_1, \cap_1)$,其中:① $\prec_1$ 为"小于或等于";②对任意的 $x, y \in S$, $x \cup_1 y = \max(x,y)$, $x \cap_1 y = \min(x,y)$。

(2)格积 $L_1 \times L_2$

令 $L_1 \times L_2 = (S \times S, \prec_\times, \cap_\times, \cup_\times)$,其中 $(x_1, x_2) \prec_\times (y_1, y_2) \Leftrightarrow x_1 \prec_1 y_1$ 且 $x_2 \prec_2 y_2 \Leftrightarrow x_1 \geqslant y_1$ 且 $x_2 \leqslant y_2$。

定义 $\cup_\times$ 和 $\cap_\times$:对任意的 (x_1, x_2), $(y_1, y_2) \in S \times S$, $(x_1, x_2) \cup_\times (y_1, y_2) = (\min(x_1, y_1), \max(x_2, y_2))$, $(x_1, x_2) \cap_\times (y_1, y_2) = (\max(x_1, y_1), \min(x_2, y_2))$。

由于完全格的积也是完全格,故 $L_1 \times L_2$ 为完全格。

(3)二阶模式

令 T 为 S×S 的一个子集{(-1,1), (0,1), (0.9,1), (1,1), (1,0.9), (1,0), (1,-1), (0,0), (-1,0), (-1,-0.9), (-1,-1), (-0.9,-1), (0,1)},为方便甄别,$L_1 \times L_2$ 中的关系 $\prec_\times$ 在 T 上改为 $\prec_T$,则 T 也是一个完全格,记为 $(T, \prec_T, \cap, \cup_T)$。

证明:显然对于 $L_1 \times L_2$ 中的关系 $\prec_\times$ 在 T 上,记为 $\prec_T$,是 T 上的一个关系。

自反性:对任意的 $x \in T$,显然有 $x \prec_T x$ 成立;

反对称性:若 $x \prec_T y$,即 $x \prec_\times y$,且 $x \neq y$,则定不存在 $y \prec_T x$。

传递性:若 $x \prec_T y$, $y \prec_T z$,则 $x \prec_T z$。因为 $x \prec_T y$, $y \prec_T z$,则 $x \prec_\times y$, $y \prec_\times z$,故 $x \prec_\times z$,则 $x \prec_T z$。

所以 $\langle T, \prec_T \rangle$ 为偏序集。

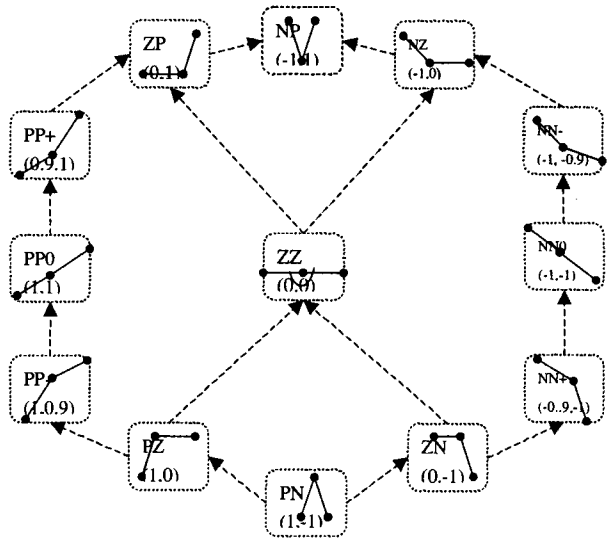


图 1 二阶原子模式偏序关系图

(4)证明 $\langle T, \leqslant_T \rangle$ 为完全格

设 w 为 T 的任一非空子集,令 $w=\{(x_1,y_1), (x_2,y_2), \cdots, (x_n,y_n)\}$,显然(-1,1)为 w 的一个上界,令:

$p=\max(y_1, y_2, \cdots, y_n) \in \{-1,-0.9,0,0.9,1\}$

$q=\min(x_1, x_2, \cdots, x_n) \in \{-1,-0.9,0,0.9,1\}$

若 $(q,p) \in T$,则 (p,q) 为 w 的上确界;若 (q,p) 不属于 T,不妨设 $q=.9$, $p=-0.9$,依次检查(.9,0), (.9,.9), (.9,1); (-1,.9), (-1,0), (-1,.9), (-1,1)是否属于 T,第一个属于 T 者,它必是 w 的上确界。同理可得 w 也有下确界,因而 $\langle T, \leqslant_T \rangle$ 是完全格。

表2 序列 X，Y 及其二阶模式

x_i	x'_i	x''_i	m'_i	y_i	y'_i	y''_i	n'_i	σ_i
1130				1137.5				
1112	-18	5	NN-	1114.5	-23	11	NN-	0
1099	-13	29	NP	1102.5	-12	14	NP	0
1115	16	-16	PZ	1104.5	2	8.5	PP+	3
1115	0	-20	ZN	1115	10.5	-13	PN	1
1095	-20	7	NN-	1112.5	-2.5	-17.5	NN+	2
1082	-13	26.5	NN-	1092.5	-20	41.5	NP	2
1095.5	13.5	-26	PN	1114	21.5	-58	PN	0
1083	-12.5	10.5	NN-	1077.5	-36.5	38	NP	0
1081	-2	2.5	NP	1079	1.5	21.5	PP+	1
1081.5	0.5	-25.5	PN	1102	23	-56	PN	0
1056.5	-25	13	NN-	1069	-33	18	NN-	0
1044.5	-12	11.5	NN-	1054	-15	3.5	NN-	0
1044	-0.5			1042.5	-11.5			

(5)定义运算符

对任意的$(x_1, y_1), (x_2, y_2) \in T$,

$\bigcup_T$ 运算符:令 $q=\min(x_1, x_2)$, $p=\max(y_1, y_2)$。若 $(q,p) \square T$,则$(x_1, y_1) \square T (x_2, y_2) = (q,p)$;否则,把$(q,p)$中出现$-0.9$的地方替换成$-1$,出现0.9的地方替换成1,则$(x_1, y_1) \bigcup_T (x_2, y_2) = (q,p)$

$\bigcap_T$ 运算符:令 $q=\max(x_1, x_2)$, $p=\min(y_1, y_2)$。若$(q,p) \in T$,则$(x_1, y_1) \bigcap_T (x_2, y_2) = (q,p)$;否则,把$(q,p)$中出现$-0.9$的地方替换成$-1$,出现0.9的地方替换成1,则$(x_1, y_1) \bigcap_T (x_2, y_2) = (q,p)$

由此,二阶以及更高阶的原子模式之间的关系都可以搭建起来。

3.4 偏序关系下的相似度量

往更高阶的原子模式完全格搭建的时候,需要低阶的完全格的性质,而在这一阶原子模式的实际应用中,只需用到原子模式之间的偏序关系。建立了算子间的相似性度量后,就可以对基本模式所组成序列建立相似性度量方法了。

现在可以定义原子模式之间的距离为其哈斯图两个位置之间的路代价的总和。比如,PN与PP+的距离为4,NN+与PP−无法比较,距离约定为6(哈斯图中最大距离为6)。根据序列中原子模式之间的距离,可以按照给定公式计算出两个模式序列的相似程度:

$$\sigma=\frac{\sum_{i=2}^{n-1}\sigma_i}{(n-2)*6} \tag{2}$$

其中,σ_i 为每一对对应模式之间距离,$0 \leqslant \sigma_i \leqslant 6$,$n$ 为序列的长度。显然,σ 是介于0和1之间的一个实数。$\sigma=0$ 时序列外形完全相同,$\sigma=1$ 时序列外形完全不同。

3.5 TBM的核心算法

在程序实现中,二阶原子模式一共才13个,没必要用算法动态计算其偏序关系,可直接将这些模式哈斯图的两两之间的代价作为一个数组或列表保存在内存中。

Input: sequence X, sequence Y, simiarity table S //S为原子模式两两的距离

Output: 序列 X,Y 的外形距离 σ

(1) 对 X 进行1,2阶差分,得到二阶模式序列 m'

(2) 对 Y 进行1,2阶差分,得到二阶模式序列 n'

(3) $q=0$

(4) for $i=2$ to 序列长度 $l-1$

(5) $p=$ 查 table S 得到 m'_i, n'_i 之间的距离

(6) $q=q+p$

(7) next i

(8) return $\sigma=q/((l-2)*6)$

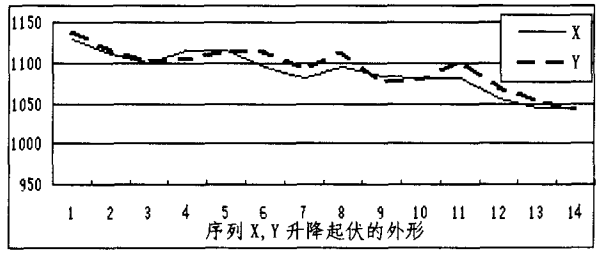


图2 序列 X,Y 起伏升降的外形

简例 序列 X,Y 及其二阶模式参看表2,最后一列为 x_i, y_i 所在位置的二阶模式的距离 σ_i。由表2,$\sigma=0.125$,可以下结论:光从升降起伏的外形上看,两个序列还是比较相似的(图2)。当然,最终的相似还是需要进一步考察其欧式距离,或利用文[22]基于拟合代价的相似度来衡量。

还可继续将二阶原子模式进行复合,得到三阶原子模式共169种,模式之间的关系处理更加细腻,以取得较高的精度。

3.6 三阶原子模式

设3.3中所述的两个完全格 $L_1(T, \prec_T, \bigcup_T, \bigcap_T)$和 $L_2(T, \prec_T, \bigcap_T, \bigcup_T)$,定义 $L(T \times T, \prec, \bigcup, \bigcap)$,其中对任意的$(x_1, y_1), (x_2, y_2) \in T \times T$,有$(x_1, x_2) \prec (y_1, y_2) \Leftrightarrow x_1 \prec_T y_1$ 且 $x_2 \prec_T y_2$。将 L 称为三阶原子模式,易证 L 是个完全格,仍然可以利用其中的偏序关系进行相似度量。

由于时间问题,本文只用到了二阶原子模式,取得了较好效果。如果能应用三阶模式,精度会更高。

4 相似序列搜索与预测算法

4.1 时序范例挖掘

(1)序列预处理:利用差分或移动平均相减等方法将序列分解,得到平稳时间序列;

(2)TBM二阶模式转换:对需要分析的时间序列进行二阶模式转换;

(3)模式挖掘:确定滑动窗口长度 w 和间距,遍历序列,在每个滑动窗口用TBM算法进行匹配,提取相似序列并进行索引,存入知识库。

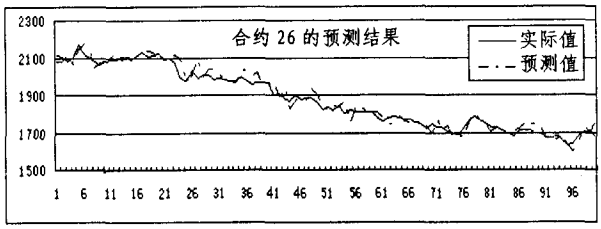


图 3 合约 26 的预测结果

4.2 时序搜索

给定一个需要匹配的序列 T,

(1)将 T 进行二阶模式转换为 T';

(2)在知识库中,使用 TBM 算法对 T' 进行搜索,找到与之二阶模式相匹配的序列集,并利用文[22]中的 FFCBS 方法确定最优序列 T'_1;

(3)计算 FFCBS(T,T')时候的 α,β,ε。这些 α,β,ε 将成为类比转换的算子。

4.3 时序类比转换与趋势外推

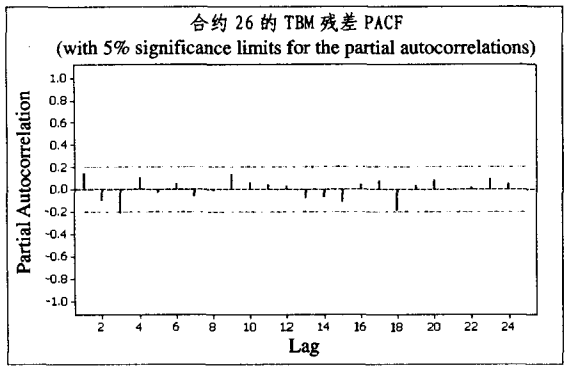
得到搜索结果的时序范例之后,进行序列趋势外推,得到序列 T 的外推结果 S。

(1)在实际序列中以 T'_1 的结束日期为起始日期,取得该序列后段序列 S',所取的序列长度由用户指定;

(2)将得到的类比转换算子 α,β,ε 逆应用于 S',获得 S,获得外推的结果。

5 实验结果

精度实验数据来自于 1997～2000 年铜期货收盘价走势,以 ARMA、DES 和 APRE 方法做对比。从 TBM 精度实验的误差比较中发现,本文的 TBM 方法比同类方法 APRE 提高了精度将近 1 倍,比经典的 ARMA 模型也有了很大提高。误差的方差较小,说明 TBM 也比较稳定。由于篇幅限制,这里只列出其中一个序列(合约 26)的一些结果。



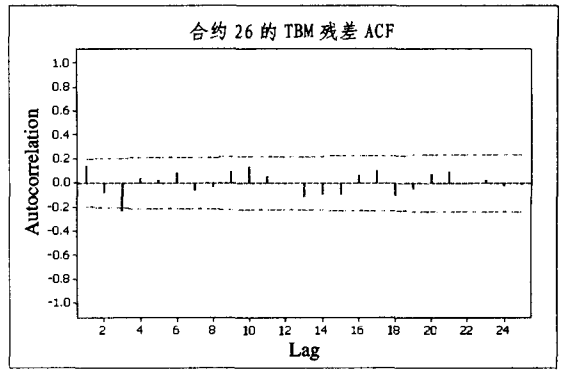


图 4 合约 26 的 TBM 预测的残差 ACF 和 PACF

从合约 26 的 TBM 预测的残差 ACF 和 PACF 图来看,残差基本上是随机序列,充分说明了 TBM 方法的有效性。从 TBM 实验数据发现,模型在预测精度上比几种经典方法高,稳定性好。随着序列长度增加,精度呈稳定上升趋势,具有较大应用价值。

结论 本文在完整的理论框架内部对一阶、二阶以至更高阶的原子模式作出定义,提出了基于原子模式模板匹配的相似算法(TBM),并基于此给出了时间序列预测的预测方法。一系列实验表明,该算法的精度相对同类方法和经典方法都有很大提高。

参 考 文 献

1 Crato N, Ray B K. Model selection and forecasting for long-range dependent processes. Journal of Forecasting, 1996, 15: 107～125
2 Faloutsos C, Ranganathan M, Manolopoulos Y. Fast Subsequence Matching in Time-Series Database. In: Proc. 1994 ACM SIGMOD Conf., Minneapolis, 1994
3 Rafiei D, Mendelson A O. Querying Time Series Data Based on Similarity. IEEE Trans on Knowledge and Data Eng, 2000, 12(5)
4 Perng Chang-Shing, Wang Haixun, Zhang S R, et al. A New Model for Similarity-Based Pattern Querying in Time Series Databases. In: 16th International Conference on Data Engineering ICDE' 2000. 33～42
5 Xia B B. Similarity Search in Time Series Data Sets: [M. Sc. thesis]. Computing Science, Simon Fraser University, 1997
6 Pecar B. Case-based Algorighm for Pattern Recognition and Extrapolation. In: The 7th UK Case-Based Reasoning Workshop, 22nd SGAI Int'l Conf. on Knowledge Based Systems and Applied Artificial Intelligence, Cambridge, 2002. 10～12
7 Ge, Smyth. Deformable Markov model templates for time-series pattern matching. KDD, 2000. 81～90
8 von Constantin A. Fuzzy Logic and Neuro Fuzzy Applications Explained. Prentice Hall, Englewood Cliffs, 1995
9 Lefteri T H, Robert U E. Fuzzy and Neural Approaches in Engineering. New York, John Wiley, 1997
10 Lotfi Z A, Fu King-Sun, Kokichi T, et al. Fuzzy Sets and their Applications to Cognitive and Decision Processes. New York, Academic Press, 1975
11 Keogh E J, Pazzani M J. Scaling up Dynamic Time Warping for Datamining Applications. Knowledge Discovery and Data Mining, in ACM, 2000. 1～58
12 Berndt, Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. In: KDD Workshop, 1994. 359～370
13 Keogh E J. Pazzani M J A simple dimensionality reduction technique for fast similarity search in large time series databases. In: the Fourth Pacific- Asia Conference on Knowledge Discovery and Data Mining, 2000
14 Keogh E J, Pazzani M J. An Indexing Scheme for Fast Similarity Search in Large Time Series Databases. SSDBM 1999: 1998. 56～67
15 Keogh E J, Chakrabarti K, Pazzani M J, et al. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. Knowl Inf Syst, 2000, 3(3): 263～286
16 Agrawal R, Lin K I, IIS Sawhney S K. Fast Similarity Search in the Presence of Noise, Scaling, and translation in Time Series Database. In: Proc. 21st Int'l Conf. VLDB, 1995. 490～500
17 Agrawal R, Faloutsos C, Swami A. Efficient Similarity Search in Sequence Database. In: Proc. 4th Int'l Conf. Foundations of Data Organization and Algorithms, Oct. 1993. 69～84
18 West M, Harrison J. Basyesian Forecasting and Dynamic Models(2nd edn). NewYork: Springer, 1997
19 Finn J V. An Introduction to Bayesian Networks. London: UCL Press, 1996
20 Holt C C. Forecasting Seasonal and Trends by Exponentially Weighted Moving Averages. Carnegie Institute of Technology, Pittsburgh, Pennsylvania, 1957
21 George B E P, Jenkins G M. Time Series Analysis: Forecasting and Control, San Francisco: Holden Day, 1970
22 Ye Jia-Cheng, Tang Yin, Peng Hong, et al. Furthest Fitting Cost: Time Series Similarity from Another Angle. In: the IEEE Int'l Conf. on Information Reuse and Integration, Las Vegas, 2004
23 Kanellakis G. On Similarity Queries for Time Series Data: Constraint Specification and Implementation. CP 1995. 137～153