Journal of Computer Applications
计算机应用,2013,33(11):3052-3056
ISSN 1001-9081
CODEN JYIIDU
2013-11-01
http://www.joca.cn

文章编号:1001-9081(2013)11-3052-05
doi:10.11772/j.issn.1001-9081.2013.11.3052

基于区间相似度的模糊时间序列预测算法

刘　芬[1,2],郭躬德[1,2*]
(1.福建师范大学 数学与计算机科学学院,福州 350007;　2.福建师范大学 网络安全与密码技术福建省高校重点实验室,福州 350007)
(*通信作者电子邮箱 ggd@fjnu.edu.cn)

摘　要:针对现有模糊时间序列预测算法无法适应预测中新关系出现的问题,提出了一种基于区间相似度的模糊时间序列预测(ISFTS)算法。首先,在模糊理论的基础上,采用基于均值的方法二次划分论域的区间,在论域区间上定义相应模糊集将历史数据模糊化;然后建立三阶模糊逻辑关系并引入逻辑关系相似度的计算公式,计算未来数据变化趋势值得到预测的模糊值;最后对预测模糊值去模糊化得到预测的确定值。由于ISFTS算法是预测数据变化趋势,克服了目前预测算法的逻辑关系的缺陷。仿真实验结果表明,与同类的预测算法相比,ISFTS算法预测误差更小,在误差相对比(MAPE)、绝对误差均值(MAE)和均方根误差(RMSE)三项指标上均优于同类的对比算法,因此ISFTS算法在时间序列预测中尤其是大数据量情况下的预测具有更强的适应性。

关键词:模糊时间序列;模糊集;相似度;逻辑关系;预测
中图分类号:TP399　　**文献标志码**:A

Interval-similarity based fuzzy time series forecasting algorithm

LIU Fen[1,2], GUO Gongde[1,2*]
(1. *School of Mathematics and Computer Science, Fujian Normal University, Fuzhou Fujian* 350007, *China*;
2. *Key Laboratory of Network Security and Cryptography, Fujian Normal University, Fuzhou Fujian* 350007, *China*)

Abstract: There are limitations in establishing fuzzy logical relationship of the existing fuzzy time series forecasting methods, which makes it hard to adapt to the appearance of new relationship. In order to overcome the defects, an interval-similarity based fuzzy time series forecasting (ISFTS) algorithm was proposed. Firstly, based on fuzzy theory, an average-based method was used to redivide the intervals of the universe of discourse. Secondly, the fuzzy sets were defined and the historical data were fuzzified, then the third-order fuzzy logical relationships were established and a formula was used to measure the similarity between logical relationships. By computing the changing trend of future data, the fuzzy values were obtained. Finally, the fuzzy values were defuzzified and the forecasting values were obtained. The proposed algorithm makes up for the shortcomings in logical relationship of the existing forecasting algorithms because it forecasts the changing trend of future data. The experimental results show that the proposed algorithm ISFTS is superior to other forecasting algorithms on forecasting error, including Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Therefore, the algorithm ISFTS is more adaptive in time series forecasting, especially in the case of large data.

Key words: fuzzy time series; fuzzy set; similarity; logical relationship; forecast

0　引言

预测在人们日常生活中扮演着十分重要的角色,人们通常对与其生活息息相关的事物进行预测,例如股票指数、农作物产量、天气预报、人口数量、医疗观察等。而由于时间序列数据是普遍存在的,因此时间序列的预测问题显得尤为重要。传统的时间序列预测往往从经典集合论的角度进行分析,主要针对完整的历史数据进行相关的预测,但是现实生活中,存在大量的历史数据不完整或不确定的预测问题[1-2],而传统的时间序列预测方法或模型则表现出了局限性,因此便开创了以时间序列和模糊集合的有机结合为基础的模糊时间序列预测方法。

模糊集合的概念是由Zadeh[3]于1965年首次提出的,模糊理论本身具有语义变量蕴含特性,可以减少在处理具体问题时可能出现不确定性对问题的困扰[4]。Song等[5]于1993年提出了模糊时间序列的概念,并且提出了模糊时间序列的一阶非时变模型和时变模型,他们将其用于Alabama大学的登记预测问题。接着于1994年他们[6]又提出了一个新的模糊时间序列模型,预测精度得到很大提高。Chen[7]使用数学运算代替了Song等[5]模型中的最大化-最小化(max-min)操作,从而提高了模型的适用性,具有较高的鲁棒性。Huarng[8]指出不同的区间长度能够得到不同的模糊关系,进而得到不同的预测结果,并且各种实证分析显示有效的区间长度能够改善预测结果。Sun等[9]考虑了Huarng[8]所提出的观点之后,以绝对误差的均值为基础确定区间间隔长度,提出了基于均值的模糊时间序列模型并通过实例进行论证。Tsaur等[10]在建立模糊逻辑关系的过程中统计历史数据出现的频率,并以此作为预测步骤的权重的赋值依据;同时,他还将指数平滑

收稿日期:2013-05-22;**修回日期**:2013-07-23。
基金项目:国家自然科学基金资助项目(61070062,61175123);福建高校产学合作科技重大项目(2010H6007)。
作者简介:刘芬(1989-),女,福建宁德人,硕士研究生,主要研究方向:时间序列数据挖掘;　郭躬德(1965-),男,福建龙岩人,教授,博士生导师,主要研究方向:数据挖掘、机器学习。

方法运用于修改预测值,在自适应模型的基础上提出了一种自适应的模糊时间序列模型。Egrioglu[11]认为多数文献中的方法由于采用模糊集的阶数建立模糊关系表,从而忽略了模糊集的隶属函数,因此提出了一个充分考虑隶属度的基于一阶模糊时间序列预测模型的新方法。

预测的准确率几乎不可能达到 100%,研究人员所能做的是尽最大努力减少预测误差[12]。现有的诸多模糊时间序列预测模型各有所长,但它们均以提高预测的精确度为目标。目前,多数预测算法(例如 Sun 等[9]和 Tsaur 等[10]的模型),需要完全统计历史数据的模糊规则,在预测未来数据时对历史数据规则有过强依赖性,新规则的出现导致模糊逻辑规则零命中而无法预测,从而大大降低预测精度。本文提出一种基于区间相似度的模糊时间序列预测算法,将模糊逻辑关系确定为三阶,并且在寻找匹配的模糊逻辑组的过程中,引入相似度计算来预测时间序列点的变化趋势,而非点对点的规则,突破现有算法中对历史规则的过度依赖,大大提高了预测的适用范围。在预测阶段,利用相似度得到合理的预测偏移值,提高了预测值的精确度。

1　相关概念

令 U 为给定论域,将论域划分为 n 个子区间,则 $U=\{u_1,u_2,\cdots,u_n\}$,采用 Zadeh 表示法建立隶属函数,令 A_i 为定义在 U 中的一个模糊集,定义如式(1):

$$A_i=\frac{\mu_{A_i}(u_1)}{u_1}+\frac{\mu_{A_i}(u_2)}{u_2}+\cdots+\frac{\mu_{A_i}(u_n)}{u_n} \tag{1}$$

其中:μ_{A_i} 是模糊集 A_i 的隶属函数,则 $\mu_{A_i}(u_j)$ 表示 u_j 对于模糊集合 A_i 的隶属程度,$\mu_{A_i}(u_j)\in[0,1]$。

模糊时间序列以及相关概念[5,11]如下所示:

定义 1　令 R 中一子集 $Y(t)(t=1,2,\cdots,n)$ 为给定论域,$f_i(t)(i=1,2,\cdots,m)$ 为定义在此论域上的模糊集合,令 $F(t)$ 为所有模糊集合 $f_i(t)(i=1,2,\cdots,m)$ 的集合,则 $F(t)(t=1,2,\cdots,n)$ 称为定义在 $Y(t)$ 上的模糊时间序列。

定义 2　若 $F(t)$ 仅仅由 $F(t-1)$ 引起,即 $F(t-1)\rightarrow F(t)$,这个关系可表示为 $F(t)=F(t-1)\times R(t,t-1)$,其中 $R(t,t-1)$ 表示一个模糊关系。则这两个连续的观测值之间的关系被称作一阶模糊逻辑关系,其中:$F(t-1)$ 称为前件,$F(t)$ 称为后件。若 $F(t)$ 同时由 $F(t-1),F(t-2),\cdots,F(t-n)$ 引起,则 n 阶模糊逻辑关系表示为 $F(t-1),F(t-2),\cdots,F(t-n)\rightarrow F(t)$,其中 $F(t-1),F(t-2),\cdots,F(t-n)$ 称为 n 阶模糊逻辑关系的当前状态。

2　基于区间相似度的模糊时间序列预测算法

本文在对现有的模糊时间序列模型进行研究的基础上,主要从论域的区间划分、模糊逻辑关系建立、预测三个方面进行改进,提出一种基于区间相似度的模糊时间序列预测(Interval-Similarity based Fuzzy Time Series forecasting, ISFTS)算法。

ISFTS 的算法步骤如下。

1)一次划分论域。

统计历史数据 $x(t)(t=1,2,\cdots,n)$ 的最小值 $D_{\min}$ 和最大值 $D_{\max}$。为了防止预测值溢出,需要适当扩大论域范围,确定历史数据论域范围 $U=[D_{\min}-D_1,D_{\max}+D_2]$,$D_1,D_2$ 为适当的正数。本文中的论域划分采用 Sun 等[9]模型中确定区间间隔长度的方法将论域 U 划分成若干个长度相同的不相交子区间。区间长度确定方法如下:

① 按式(2)计算 n 个历史数据中相邻两个数 $x(i)$ 与 $x(i+1)(i=1,2,\cdots,n-1)$ 的差的绝对值,将这 $n-1$ 个绝对值累加求平均值的一半作为长度:

$$Range=\frac{\sum_{i=1}^{n-1}|x(i)-x(i+1)|}{2(n-1)} \tag{2}$$

② 根据表 1[9] 对 *Range* 进行四舍五入取整运算。

表 1　区间间隔基数 *Base* 取值

Range	*Base*	*Range*	*Base*
[0.1,1)	0.1	[10,100)	10
[1,10)	1	[100,1 000]	100

例如,当计算出 *Range* = 33.45,则 *Range* 落在 10 ~ 100 区间范围内,其四舍五入的基数为 10,将 *Range* 以 10 为基数四舍五入,最终取 *Range* = 30。

③ 第一次划分的结果:将论域 U 划分得到 m 个长度相等的子区间 $U_1,U_2,\cdots,U_m$,m 由式(3)计算得到,其中 $Max=D_{\max}+D_2$,$Min=D_{\min}-D_1$,则第 i 个子区间的范围表示为 $U_i=[Min+(i-1)\times Range, Min+i\times Range]$,这里 $i=1,2,\cdots,m$。

$$m=\left\lceil\frac{Max-Min}{Range}\right\rceil \tag{3}$$

2)二次划分论域。

根据上一步中的论域子区间划分结果,时间序列中的每个点都会落在某一个子区间 U_i 中。再次统计时间序列历史数据落在每个子区间中的点的数量,对每个子区间进行二次划分。划分规则如下。

求 m 个子区间平均节点个数 *avgNum*,统计落在区间 U_i 的点的个数 $numU_i$,则把 U_i 区间划分成 k_i 个子区间,k_i 按式(4)计算得到:

$$k_i=\lceil numU_i/avgNum\rceil \tag{4}$$

经过二次划分,把 U_i 划分成 k_i 个长度相等的小的子区间 $U_{i\_1},U_{i\_2},\cdots,U_{i\_k_i}$。

经过以上两个步骤的划分,将整个论域划分成若干个区间长度不同的子区间,能更合理地反映数据在论域区间上的分布规律。

3)定义模糊集。

根据第一次论域划分,假设将论域划分成 n 个区间,定义模糊集 $A=\{A_1,A_2,\cdots,A_n\}$,其中 A_i 表示模糊集 A 的语义变量。模糊集表示如式(5)所示:

$$\begin{cases}A_1=1/u_1+0.5/u_2+0/u_3+\cdots+0/u_n\\A_2=0.5/u_1+1/u_2+0.5/u_3+\cdots+0/u_n\\\vdots\\A_{n-1}=0/u_1+\cdots+0.5/u_{n-2}+1/u_{n-1}+0.5/u_n\\A_n=0/u_1+\cdots+0/u_{n-2}+0.5/u_{n-1}+1/u_n\end{cases} \tag{5}$$

4)数据模糊化。

根据二次划分,对所有的历史数据进行模糊化处理。例如当历史数据 $x(t)\in U_{i\_m}$ 时,将 $x(t)$ 模糊化成 A_{i-1+b},其中 $b=m/k_i$,k_i 为区间 U_i 的子区间个数。

5)建立模糊逻辑。

经过以上模糊化,每个数据都被模糊化成一个模糊值。在时间序列中,历史数据对数据预测有着重要的影响,越靠近预

测点的数据和预测点的值的关联性越强，对预测值的影响也越大。预测模型中，若所选历史数据太多，则模型的计算量太大，若所选历史数据过少又会丢失太多数据，预测精度不高，因此在时间序列预测中无需对所有的历史数据建立预测模型，本文考虑三阶模糊逻辑关系[1]，即一个数据由前面三个数据引起。

对于任意的三阶模糊关系，可以表示为 $A_i,A_j,A_k \to A_l$，因此本文的模糊关系如式(6)所示：

$$\begin{cases} A_{t_1},A_{t_2},A_{t_3} \to A_{t_4} \\ A_{t_2},A_{t_3},A_{t_4} \to A_{t_5} \\ \vdots \\ A_{t_i},A_{t_{i+1}},A_{t_{i+2}} \to A_{t_{i+3}} \\ \vdots \\ A_{t_{n-3}},A_{t_{n-2}},A_{t_{n-1}} \to A_{t_n} \end{cases} \tag{6}$$

根据式(6)中的三阶逻辑关系，确定逻辑关系的分布如表2[1]所示。

表2 逻辑关系分布

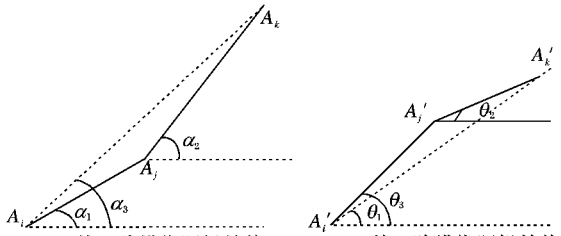
类型	关系	模糊关系	命名
1	$i<j<k$	$A_i,A_j,A_k \to A_l$	“上-上”趋势模糊逻辑组
2	$i<j=k$	$A_i,A_j,A_k \to A_l$	“上-等”趋势模糊逻辑组
3	$i=j<k$	$A_i,A_j,A_k \to A_l$	“等-上”趋势模糊逻辑组
4	$i=j=k$	$A_i,A_j,A_k \to A_l$	“等-等”趋势模糊逻辑组
5	$i>j>k$	$A_i,A_j,A_k \to A_l$	“下-下”趋势模糊逻辑组
6	$i>j=k$	$A_i,A_j,A_k \to A_l$	“下-等”趋势模糊逻辑组
7	$i=j>k$	$A_i,A_j,A_k \to A_l$	“等-下”趋势模糊逻辑组
8	$i<j>k$	$A_i,A_j,A_k \to A_l$	“上-下”趋势模糊逻辑组
9	$i>j<k$	$A_i,A_j,A_k \to A_l$	“下-上”趋势模糊逻辑组

6)统计相似度。

① 根据表2中的模糊逻辑分组，将三阶模糊关系分成9种类型，将历史数据归类到相应的逻辑分组中：根据9种模糊类型设置9个相应类型集合，统计历史数据的三阶逻辑，加入到匹配的分组集合中。例如：历史数据中的一个三阶模糊逻辑关系为“上-上”型，则加入到“上-上”型集合中。

② 确定预测点 $x(n+1)$ 前三阶 $x(n-2)$、$x(n-1)$、$x(n)$ 对应模糊集 $S(A_i,A_j,A_k)$ 的模糊逻辑组类型，计算相应类型集合元素的相似度。假设 $S(A_i,A_j,A_k)$ 为“上-上”型模糊逻辑，若“上-上”型集合 $TypeS$ 中有 num 个元素，分别为：$\{S_1,S_2,\cdots,S_{num}\}$，将这 num 个元素与 S 计算相似值。计算如下所示：

设 $S_a(1 \leqslant a \leqslant num)$ 为 $x(m-2),x(m-1),x(m)$ 三个连续时间序列点的三阶模糊逻辑，记为 (A_i',A_j',A_k')，S 与 S_a 的趋势如图1中的(a)、(b)所示。



(a) S的三阶模糊逻辑趋势 (b) S_a的三阶模糊逻辑趋势

图1 两个三阶模糊逻辑的趋势

则二者的相似度 $similar(S,S_a)$ 按式(7)计算：

$$similar(S,S_a) = 1 - distance(S,S_a) \tag{7}$$

$similar(S,S_a)$ 反映 S 与 S_a 趋势增减的相似度，其值越接近1，表明 S 与 S_a 的趋势增减越相似，其预测偏差也越接近。其中 $distance(S,S_a)$ 见定义3。

定义3 任意两个三阶模糊逻辑 a 和 b 的距离 $distance(S,S_a)$ 表示如式(8)：

$$distance(S,S_a) = (|\sin\alpha_1 - \sin\theta_1| + |\sin\alpha_2 - \sin\theta_2| + |\sin\alpha_3 - \sin\theta_3|)/3 \tag{8}$$

则式(8)满足距离的三大性质：1)非负性；2)对称性；3)满足三角不等式。

$distance(S,S_a)$ 的非负性和对称性显然满足，其满足三角不等式的证明如下：

证明 假设有任意三个三阶模糊逻辑组，参考图1，其对应三个线段的三个角分别为 $\alpha_1,\alpha_2,\alpha_3$；$\beta_1,\beta_2,\beta_3$；$\gamma_1,\gamma_2,\gamma_3$，则 p,q,r 两两距离分别表示为式(9)、(10)、(11)：

$$distance(q,p) = (|\sin\beta_1 - \sin\alpha_1| + |\sin\beta_2 - \sin\alpha_2| + |\sin\beta_3 - \sin\alpha_3|)/3 \tag{9}$$

$$distance(q,r) = (|\sin\beta_1 - \sin\gamma_1| + |\sin\beta_2 - \sin\gamma_2| + |\sin\beta_3 - \sin\gamma_3|)/3 \tag{10}$$

$$distance(p,r) = (|\sin\alpha_1 - \sin\gamma_1| + |\sin\alpha_2 - \sin\gamma_2| + |\sin\alpha_3 - \sin\gamma_3|)/3 \tag{11}$$

三角不等式中，不失一般性，只需证明式(12)成立：

$$distance(p,q) + distance(q,r) \geqslant distance(p,r) \tag{12}$$

式(12)成立即可表示 $distance(a,b)$ 满足三角不等式。

在三组角中，只需证明不等式(13)成立，则不等式(12)便可成立。

$$|\sin\alpha_i - \sin\beta_i| + |\sin\beta_i - \sin\gamma_i| \geqslant |\sin\alpha_i - \sin\gamma_i| \tag{13}$$

根据本文线段趋势的划分，α_i，β_i，γ_i 为满足条件的任意角，根据划分得结果 α_i，β_i，$\gamma_i \in (-90°,90°)$。

在不等式(13)中，根据角度关系共分成6种情况。

a) 当 $\alpha_i \geqslant \beta_i \geqslant \gamma_i$ 时，$\sin\alpha_i \geqslant \sin\beta_i$，$\sin\beta_i \geqslant \sin\gamma_i$，$\sin\alpha_i \geqslant \sin\gamma_i$，则不等式(13)等价于：

$\sin\alpha_i - \sin\beta_i + \sin\beta_i - \sin\gamma_i \geqslant \sin\alpha_i - \sin\gamma_i \Leftrightarrow$

$\sin\alpha_i - \sin\gamma_i \geqslant \sin\alpha_i - \sin\gamma_i$

由于 $\sin\alpha_i - \sin\gamma_i \geqslant \sin\alpha_i - \sin\gamma_i$ 成立，故不等式(12)成立。

b) 当 $\alpha_i \geqslant \gamma_i \geqslant \beta_i$ 时，$\sin\alpha_i \geqslant \sin\beta_i$，$\sin\beta_i \leqslant \sin\gamma_i$，$\sin\alpha_i \geqslant \sin\gamma_i$，则不等式(13)等价于：

$\sin\alpha_i - \sin\beta_i + \sin\gamma_i - \sin\beta_i \geqslant \sin\alpha_i - \sin\gamma_i \Leftrightarrow$

$2\sin\gamma_i \geqslant 2\sin\beta_i \Leftrightarrow \sin\gamma_i \geqslant \sin\beta_i$

因为 $\beta_i \leqslant \gamma_i$，并且 α_i，β_i，$\gamma_i \in (-90°,90°)$，

所以 $\sin\gamma_i \geqslant \sin\beta_i$ 成立，故不等式(12)成立。

c) 当 $\beta_i \geqslant \alpha_i \geqslant \gamma_i$ 时，$\sin\alpha_i \leqslant \sin\beta_i$，$\sin\beta_i \geqslant \sin\gamma_i$，$\sin\alpha_i \geqslant \sin\gamma_i$，则不等式(13)等价于：

$\sin\beta_i - \sin\alpha_i + \sin\beta_i - \sin\gamma_i \geqslant \sin\alpha_i - \sin\gamma_i \Leftrightarrow$

$2\sin\beta_i \geqslant 2\sin\alpha_i \Leftrightarrow \sin\beta_i \geqslant \sin\alpha_i$

因为 $\beta_i \geqslant \alpha_i$，并且 α_i，β_i，$\gamma_i \in (-90°,90°)$，

所以 $\sin\beta_i \geqslant \sin\alpha_i$ 成立，故不等式(12)成立。

同理可证：d) 当 $\beta_i \geqslant \gamma_i \geqslant \alpha_i$ 时；e) 当 $\gamma_i \geqslant \alpha_i \geqslant \beta_i$ 时；f) $\gamma_i \geqslant \beta_i \geqslant \alpha_i$ 时不等式(12)均成立。

综上所述，对满足 α_i，β_i，$\gamma_i \in (-90°,90°)$ 条件下的任意角，不等式(12)均成立。

则$\left(\sum_{i=1}^{3}|\sin\alpha_i-\sin\beta_i|\right)/3+\left(\sum_{i=1}^{3}|\sin\beta_i-\sin\gamma_i|\right)/3\geqslant\left(\sum_{i=1}^{3}|\sin\alpha_i-\sin\gamma_i|\right)/3$成立。即模糊逻辑组的距离公式$distance(S,S_a)$满足三角不等式性质。

③ 按式(14)计算S与$TypeS$中所有元素的相似概率:

$$similarP(S,S_a)=\frac{similar(S,S_a)}{\sum_{a=1}^{num}similar(S,S_a)} \tag{14}$$

7) 预测模型。

时间序列具有相似关联性,相同类型的模糊逻辑出现的概率高。两个三阶模糊逻辑关系其相似度越高,则预测偏差也越接近。

定义 4　在任意一个三阶模糊逻辑$S(A_i,A_j,A_k\rightarrow A_m)$中,称$D=(m-k)$为$S$的模糊关系偏移值。

$TypeS$中num个元素$S_1,S_2,\cdots,S_{num}$对应的模糊预测偏移分别为:$D_1,D_2,\cdots,D_{num}$。

预测点$x(n+1)$的模糊逻辑为$S(A_i,A_j,A_k\rightarrow A_{(n+1)})$,$S$与$TypeS$类型匹配。则其模糊预测偏移$D(n+1)$是$TypeS$中模糊偏移值的数学期望。按式(15)计算:

$$D(n+1)=\sum_{i=1}^{num}D_i\times similarP(S,S_i) \tag{15}$$

则预测点的模糊值$A(n+1)$为:

$$A(n+1)=A_k+D(n+1) \tag{16}$$

8) 去模糊化。

得到预测点的模糊值$A(n+1)$之后,计算$A(n+1)$所在的区间U_i,以及U_i对应的模糊集A_i,则模糊值在区间内的偏移值$fuzzyDif(A(n+1),A_i)$按式(17)计算得到:

$$fuzzyDif(A(n+1),A_i)=k+D(n+1)-i \tag{17}$$

计算U_i的中间值,则预测值$F(n+1)$为:

$$F(n+1)=M[U_i]+fuzzyDif(A(n+1),A_i)\times length(U_i) \tag{18}$$

其中:$length(U_i)$表示区间U_i的长度,$M[U_i]$表示区间U_i的中间值。

3　不同预测算法分析及实验对比

本文的实验数据分别采用上证指数从 2006-12-01 至 2007-07-17 的收盘价共 150 个数据以及澳大利亚从 1956 年起每个月的硫酸产量的 150 个数据,均以前 120 个数据作为历史数据训练,划分论域,确定模糊逻辑关系,预测后面 30 个数据。

3.1　实验 1

上证指数历史数据(从2006-12-01至2007-06-04的收盘价,1 个工作日 1 个数据,共 120 个数据)如图 2 所示。

在本节实验分析中,本文的 ISFTS 算法、Tsaur 等[10]、Sun 等[9]的策略中第一次论域划分均采用 Sun 等[9]中的论域划分方法,即第 2 章中介绍的一次论域划分方法。

在预测时:

1)以历史数据中从 2006-12-01 至 2007-06-05 的前 120 个数据作为历史训练数据,计算后可得区间的划分间隔长度$Range=27.9784$。根据表 1 可得$Base=10$。将$Range$以 10 为基数四舍五入,确定$Range$的值为 30。以 30 为区间长度划分论域区间。

2) 由于时间序列的数据具有时间递进性,越远离预测点的历史数据对预测点的影响越小,过于久远的数据对于预测点的数据预测并无太大意义,本文将预测数据窗口大小M设置成 120,即预测第$i(i>120)$个数据时,以从$x(i-120)$到$x(i-1)$个数据作为规则学习。

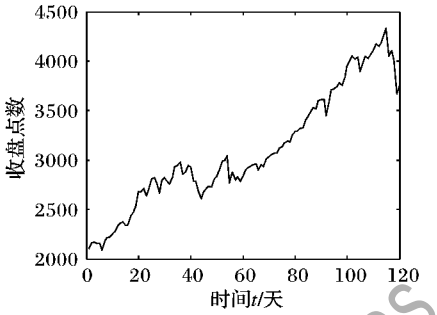


图 2　上证指数 120 个历史训练数据

3) 预测完$x(i+1)(i>120)$时,将最新的$x(i)$真实值按照第 1) 步中的论域划分进行模糊化,加入到模糊逻辑规则中,并剔除最陈旧的一个模糊逻辑关系:在本文算法中剔除$x(i-117)$的模糊逻辑规则;在 Tsaur 等[10]、Sun 等[9]的算法中剔除$x(i-119)$的模糊逻辑;使预测的数据窗口始终保持在 120 的大小。预测时由于 Tsaur 等[10]、Sun 等[9]算法可能出现模糊规则零命中情景,当出现零命中时,本文将论域区间的中间值作为其预测值。

4)在对比实验中,本文增加了 Sun 等[9]策略的自适应机制调整预测结果,从实验结果可以得出,增加自适应机制后其精确度有很大提高。

最终预测结果(从从 2007-06-06 至 2007-07-17 的收盘价,1 个工作日 1 个数据,共 30 个数据)如图 3 所示,对应的实验结果见表 3。

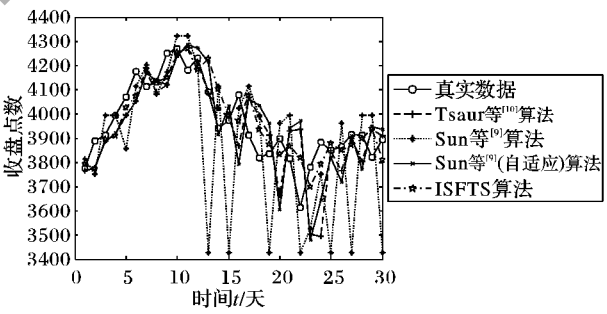


图 3　上证指数实验预测结果

表 3　上证指数实验结果对比

算法	MAPE	MAE	RMSE
Tsaur 等[10]算法	0.027 125 8	105.900 0	143.088 0
Sun 等[9]算法	0.048 599 5	190.977 0	255.280 0
Sun 等[9](自适应)算法	0.030 524 6	119.332 0	152.268 0
ISFTS 算法	0.019 008 0	74.831 8	90.254 2

本文采用误差相对比(Mean Absolute Percentage Error, MAPE)、绝对误差均值(Mean Absolute Error, MAE)和均方根误差(Root Mean Squared Error, RMSE)三个指标对不同预测方法进行比较[10,13],其中:

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{AuctalValue(t)-ForecaseValue(t)}{AuctalValue(t)}\right| \tag{19}$$

$$MAE=\frac{\sum_{i=1}^{n}|AuctalValue(t)-ForecaseValue(t)|}{n} \tag{20}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}[AuctalValue(t) - ForecaseValue(t)]^2}{n}} \tag{21}$$

其中:$AuctalValue(t)$ 和 $ForecaseValue(t)$ 分别为第 $t(t = 1,2,\cdots,n)$ 个真实数据和预测结果。

3.2 实验 2

硫酸产量历史数据(从 1956-10 至 1966-09 的硫酸产量,1 个月 1 个数据,共 120 个数据)见图 4 表示。

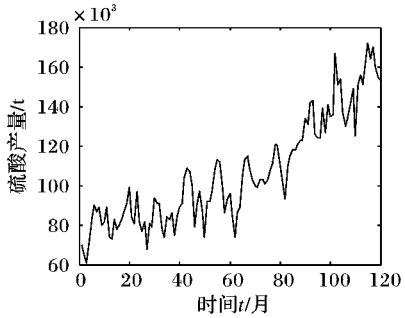


图 4 硫酸产量 120 个历史训练数据

最终预测(从 1966-10 至 1969-03 的硫酸产量,1 个月 1 个数据,共 30 个数据)结果如图 5 所示,对应的实验结果见表 4。

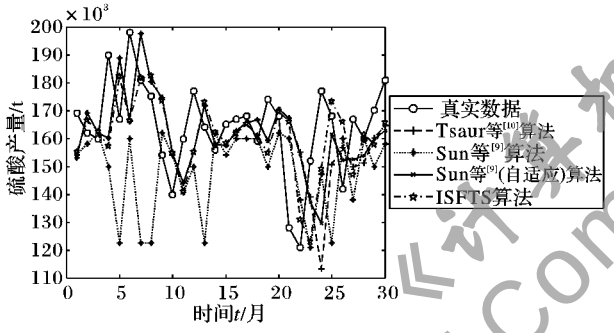


图 5 硫酸产量实验预测结果

表 4 硫酸产量实验结果对比

算法	MAPE	MAE	RMSE
Tsaur 等[10]算法	0.100725 0	16.214 0	20.790 6
Sun 等[9]算法	0.134 871 0	22.400 0	27.498 9
Sun 等[9](自适应)算法	0.094 005 7	15.101 8	18.911 4
ISFTS 算法	0.086 761 7	14.011 7	17.458 6

从表 3 ~ 4 的对比结果可以看出,本文的算法具有更高的预测精度,在 MAPE、MAE 和 RMSE 三个指标上均优于 Tsaur 等[10]的自适应策略、Sun 等[9]的策略以及 Sun 等[9]的改进自适应机制方案。由于 Tsaur 等[10]和 Sun 等[9]的策略中,受历史数据规则学习的限制,在预测未来数据时有可能遇到未曾学习到的规则,例如当历史数据中 A_i 没有至 A_j 的模糊逻辑,出现规则零命中情况,Tsaur 等[10]和 Sun 等[9]的策略无法预测。而本文的预测算法,其模糊逻辑规则是预测序列的趋势的匹配相似度度来预测数据,打破了 Tsaur 等[10]和 Sun 等[9]策略中规则学习的约束,尤其在大数据量的情况下,具有更好的适用性和更高的预测精度。

4 结语

本文根据时间序列模糊逻辑关系组之间的匹配相似度,提出了一种基于区间相似度的模糊时间序列新型预测算法,预测时间序列的趋势变化从而预测未来数据的值。该算法突破现有算法的完全统计模糊规则的限制,实验表明,该算法的 MAPE、MAE 和 RMSE 三个指标均优于同类预测算法,预测精度更高。

参考文献:

[1] 张慧. 自适应模糊时间序列预测模型的研究[D]. 大连: 大连海事大学, 2012.

[2] WONG H L, TU Y H, WANG C C. Application of fuzzy time series models for forecasting the amount of Taiwan export [J]. Expert Systems with Applications, 2010, 37(2): 1465 - 1470.

[3] ZADEH L A. Fuzzy sets [J]. Information and Control, 1965, 8(3): 338 - 353.

[4] 吴铭峰, 蒋勋. 基于模糊时间序列的预测模型——以上证指数为例[J]. 价值工程, 2008, 27(11): 165 - 168.

[5] SONG Q, CHISSOM B. Fuzzy time series and its models [J]. Fuzzy Sets and Systems, 1993, 54(3): 269 - 277.

[6] SONG Q, CHISSOM B. Forecasting enrollments with fuzzy time series—part 2 [J]. Fuzzy Sets and Systems, 1994, 62(1): 1 - 8.

[7] CHEN S M. Forecasting enrollments based on fuzzy time series [J]. Fuzzy Sets and Systems, 1996, 81(3): 311 - 319.

[8] HUARNG K. Effective lengths of intervals to improve forecasting in fuzzy time series [J]. Fuzzy Sets and Systems, 2001, 123(3): 387 - 394.

[9] SUN X, LI Y. Average-based fuzzy time series models for forecasting Shanghai compound index [J]. World Journal of Modelling and Simulation, 2008, 4(2): 104 - 111.

[10] TSAUR R C, KUO T C. The adaptive fuzzy time series model with an application to Taiwan's tourism demand[J]. Expert Systems with Applications, 2011, 38(8): 9164 - 9171.

[11] EGRIOGLU E. A new time-invariant fuzzy time series forecasting method based on genetic algorithm [J]. Advances in Fuzzy Systems, 2012, 2012: 1 - 6.

[12] CHEN S M, HWANG J R. Temperature prediction using fuzzy time series [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2000, 30(2): 263 - 275.

[13] HASSAN S, JAAFAR J, SAMIR B B, *et al*. A hybrid fuzzy time series model for forecasting [J]. Engineering Letters, 2012, 20(1): 88 - 93.

(上接第 3044 页)

[10] INOUE T, HEATH R W, Jr. Grassmannian predictive frequency domain compression for limited feedback beamforming [C]// Proceedings of the 2010 Information Theory and Applications Workshop. Washington, DC: IEEE Computer Society, 2010: 173 - 177.

[11] 夏欣, 武刚, 李少谦. MU-MIMO 系统中利用信道信息预测值的反馈机制[J]. 计算机工程与应用, 2010, 46(14): 5 - 7.

[12] 张贤达. 矩阵分析与应用[M]. 北京, 清华大学出版社, 2008.

[13] 程勇涛. 自适应调制系统中无线信道预测算法的研究[D]. 南京: 南京邮电大学, 2010.

[14] ZHOU S L, WANG Z D, GIANNAKIS G B, *et al*. Quantifying the power loss when transmit-beamforming relies on finite rate feedback [J]. IEEE Transactions on Wireless Communication, 2005, 4(4): 1948 - 1957.

[15] 王海洋. MIMO 系统中的有限反馈技术研究[D]. 杭州: 浙江大学, 2008.