第46卷第3期 中国科学技术大学学报 Vol. 46, No. 3
2016年3月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Mar. 2016

文章编号:0253-2778(2016)3-0238-09

基于SVM修正的模糊时间序列模型在沪指预测中的应用

李小琳,孙 玥,刘 洋

(南京大学管理学院,江苏南京 210093)

摘要:传统股票指数研究方法多停留在经验判断或简单的数据分析阶段,主要方法有基本面分析法、交易指标分析法等,这类分析方法或是对以往数据包含的信息使用效率比较低,或是对使用者的经验积累要求很高.近年来,数据挖掘方法在股市中已有很多成功的应用.在上述工作的基础上,从以下三方面提出一种改进的糊时间序列(fuzzy time series , FTS)模型并将其应用于股市预测中:一是提出了新的区间划分方法;二是提出新的模糊集权重公式;三是运用SVM分类算法进行模型修正,提出组合FTS模型.样本是选取1996~2003年上证指数数据,利用提出模型进行指数预测.实验结果表明,与多种重要FTS模型进行比较,本文提出的改进模型效果更优.

关键词:模糊时间序列; SVM算法; 股指预测

中图分类号:TP311 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2016.03.009

引用格式:LI Xiaolin, SUN Yue, LIU Yang. Forecasting Shanghai stock index using FTS model based on SVM-modify[J]. Journal of University of Science and Technology of China, 2016,46(3):238-246.
李小琳,孙 玥,刘 洋. 基于SVM修正的模糊时间序列模型在沪指预测中的应用[J]. 中国科学技术大学学报,2016,46(3):238-246.

Forecasting Shanghai stock index using FTS model based on SVM-modify

LI Xiaolin, SUN Yue, LIU Yang

(*School of Management, Nanjing University, Nanjing* 210093, *China*)

Abstract: Traditional methods for stock index research are still at the stage of judging by experience or relying on simple data analysis, among which fundamental analysis and trading indicator analysis frequently used. These methods have noticeable disadvantages: Inefficient utilization of existing information or requirement for highly experienced users. A modified fuzzy time series (FTS) model was proposed based on the following three aspects. Firstly, a new method of interval division was developed. Secondly, a new weight formula for fuzzy set was devised. Thirdly, a modified FTS model was built with the application of SVM classification model. Predictions for stock index were made by using the proposed model. Experiment results from Shanghai index data ranging from 1996 to 2003 indicate that compared with other important FTS models; the proposed model provides better performance.

Key words: fuzzy time series; SVM algorithm; stock index prediction

收稿日期:2015-09-12;**修回日期**:2015-12-29

基金项目:国家自然科学基金(60803055),教育部人文社会科学一般项目(08JC630041)资助.

作者简介:李小琳(通讯作者),女,1978年生,博士/副教授. 研究方向:商务智能,数据挖掘,企业信息化. E-mail: lixl@nju.edu.cn

0　引言

股票市场是国家经济的重要监控指标，从金融市场发展的角度，反映了经济发展的健康和繁荣程度. 同时，股票市场也是人们日常理财、投资的重要渠道之一. 股市的高收益与高风险并存的特点，促使投资者个人投入较高的精力研究股价趋势变化，规避非系统风险，获取更高的投资利润. 因此对股票内在性质及走势预测的研究具有重大理论意义和应用前景.

股市分析方法主要有两类，一类是传统分析方法，主要有 4 种：基于股市基本面的分析方法；基于股市走势；各种交易指标的技术分析方法；指标量化分析法；另一类是量化、智能分析方法，前者包括以统计学为基础的计量研究方法，后者主要是数据挖掘算法在股市分析中的应用. 主流的数据挖掘算法包括挖掘频繁模式算法、SVM 算法、遗传算法和模糊集算法.

从股市进入中国到今天，投资者经历了大大小小近十次牛市与熊市，中国股市的周期性已经被很多学者证明[1-3]. 还有学者关注证券市场波动周期与经济指标波动周期的相关关系，试图找出两者之间的内在联系[4-5]. 本文结合股票市场周期性的特征选择以模糊时间序列模型为研究方法.

1　模糊时间序列模型研究现状

1.1　模糊时间序列模型的提出

模糊时间序列模型早期主要用于对天气、固定的实验数据集的预测，而后发展到网络用户数量[6]、医学领域[7]、黄金价格[8]等各行各业. 最早引入模糊时间序列模型的是学者 Song[9]. Lee[10]、Chen[11] 等在 Song 的基础上对该模型进行了简单的改进，使得模型预测更加便捷、准确. 这些学者共同创建了早期的模型框架，他们构建的模型的共同点在于模糊区间的划分方法一致，即都是等距划分，隶属度函数设置简单. 早期模型都会产生精度上的损失，模糊化的方法过于随意，缺乏理论基础，模糊区间的上限与下限的设置存在缺陷，无法确定上下限的准确位置. 往往在明确所有数据集（含需要预测的数据集）的前提下设置上下限，存在数据溢出的风险.

1.2　国外研究现状

Huarng 等[12]第一个将模糊时间序列模型引入到股票指数预测中，他们在 Song 等提出的一阶模糊集的基础上提出了 2 阶模糊模型，即将模糊集的隶属度进行再一次模糊化，使之能更好地适应预测结果不确定性这一目的，与未来的不确定性更好契合. Cheng 等[13]将基于权重的模糊时间序列模型应用于对台湾股市的预测，他们将股市的波动分为三个类别，即上升、不变、下降，然后依次对三种类别计数，计算出每种变化类别的权重，作为最后预测模型中使用的权重向量. Lee[10] 指出，以往的一些模型无法充分利用指数波动中蕴含的信息，在预测方面存在缺陷，所以在传统模型的基础上，他将模糊时间序列模型与 Japanese CandleStick Theory（烛台理论）相结合，提出了模式识别模糊时间序列模型. Chu 等[14] 指出，以往研究很少考虑股市指数以外的其他因素，而仅仅依靠股市以前的信息，这些信息无法包含主要影响股市的因素，在不完全市场的假说下，无法充分预测未来的股价走势，误差较为显著，因此需要引入其他属性. 与此同时 Ting[15] 通过计量分析，发现股价与交易量之间存在双向因果关系，得出交易量是股价波动的一个显著影响指标，可能会对模型的精确度有提高作用，所以 Chu 在该文中引入了交易量属性，提出了双因素模糊时间序列模型，对台湾股市进行预测，结果好于 Chen[11] 与 Huarng 等[12] 的模型. 这些模型的应用大多采用单变量，是对基础模型的简单改进，在实际应用中效果不尽理想，存在模糊区间长度划分随意，没有考虑重复 FLRs 隐含信息，三角隶属度函数构建缺乏针对性等问题. 其中，隶属度函数表示当前数据与历史数据相近程度，一般来说，隶属度越大，则说明该历史数据包含的信息对基于当前数据的预测行为作用越大.

1.3　国内研究现状

国内在模糊时间序列 FTS 模型领域的研究起步较晚，但是发展很迅速. 近年，模糊时间序列模型在股指预测的应用研究中，重点研究方向包括：

（Ⅰ）加入其他联动股指指标，包括股票交易量指标、开盘价、同一国家内的其他股指变动等. Karpoff[16] 证明了股市中同时存在着两种关系，即“交易量与股价变动正相关”、“交易量与股价波动幅度正相关”. 国内学者金春雨[17]、王重[18] 等使用 VAR 模型或者格兰杰因果关系检验等发现指数与交易量之间存在因果关系.

（Ⅱ）模糊区间的进一步优化. Aladag[19]、Egrioglu[20] 等使用最优化算法对数据集进行训练，

从而找出最优区间的划分点. Chen 等[21]通过平均误差构造聚类算法,将历史数据聚成不同的区间类别. 这些新的区间划分方法使得数据信息能够被充分应用,预测效果显著强于传统区间划分方法.

(Ⅲ)提出新的隶属度公式.

(Ⅳ) 对模型预测结果进行修正. Chen 等在基于权重 FTS 模型预测结果提出可以加入调整参数 α,通过调节公式 Adapted_forecast(t)= Actual($t-1$)+α(Forecast(t)-Actual($t-1$))进行结果修正.

(Ⅴ)多种模型方法同时使用. Qiu 等[22]不仅使用了自己先前文献中提出的隶属度函数改进方法,还进一步提出可以优化区间划分方法,提出 C-fuzzy decision trees 方法,实验结果与 Hurang 等提出的模型进行比较,模型在精度上得到大幅提高.

2 对模糊时间序列模型的改进

2.1 模型的基本定义

定义 2.1 数据集 U,由 n 个数据区间组成,$U=\{u_1,u_2,\cdots,u_n\}$,U 的模糊集 A 被定义为

$$A=\frac{f_A(u_1)}{u_1}+\frac{f_A(u_2)}{u_2}+\cdots+\frac{f_A(u_n)}{u_n} \quad (1)$$

式中,f_A 是模糊集 A 的隶属度函数,f_A 在[0,1]之间取值,$f_A(u_i)$ 表示区间 $u_i(1\leqslant i\leqslant n)$ 相对于模糊集 A 的隶属度,符号"+"表示联合,而非求和.

定义 2.2 $Y(t)(t=0,1,2,\cdots)$ 表示定义 $f_i(t)(t=1,2,\cdots)$ 的值域,是一组实数的集合,$F(t)$ 是 t 时刻 $f_i(t)(t=1,2,\cdots)$ 的集合. 那么,$F(t)$ 是 $Y(t)(t=0,1,2,\cdots)$ 上的一个模糊时间序列.

定义 2.3 假设在 $F(t)$ 与 $F(t-1)$ 之间存在一个模糊关系 $R(t,t-1)$,也就是 $F(t)=F(t-1)\odot R(t,t-1)$,$\odot$是结合符号. 此时,$F(t)$ 可以被称作 $F(t-1)$ 引起的,可以用一个逻辑关系表示:$F(t-1)\to F(t)$.

在以上三个定义的基础上,两个连续观测对象之间的模糊关系可以定义如下.

定义 2.4 假设 $F(t-1)=A_i$,$F(t)=A_j$,那么 $F(t)$ 与 $F(t-1)$ 之间的关系称为模糊逻辑关系(FLR),可以用 $A_i\to A_j$ 表示. A_i 称作左模糊集(LHS),A_j 称作右模糊集(RHS).

定义 2.5 训练数据集中所有的模糊逻辑关系都可以依据相同的左边模糊集(LHS)进一步聚成多个模糊逻辑关系组,每个组的 LHS 是相同的,而 RHS 可以是不同的.

例如,有 4 个模糊逻辑关系,$A_i\to A_{j1}$,$A_i\to A_{j2}$,$A_m\to A_{j3}$,$A_m\to A_{j4}$,那么,$A_i\to A_{j1}$,$A_i\to A_{j2}$ 可以聚成一个组,$A_m\to A_{j3}$,$A_m\to A_{j4}$ 可以聚成一个组.

在上述相关定义的前提下,基本的模糊时间序列算法一般描述如下:

Step 1 定义数据集和模糊区间,数据集表示为 $U=$[starting, ending],根据事先定义的跨度将数据集划分为 n 个模糊区间 $U=\{u_1,u_2,\cdots,u_n\}$,m_i 是 u_i 的中间值,u_i 的模糊集是 A_i.

Step 2 根据数据集定义模糊集,将历史数据模糊化,模糊集 A_i 能够表示为 $A_i=\frac{a_{i1}}{u_1}+\frac{a_{i2}}{u_2}+\cdots+\frac{a_{in}}{u_n}$,或者 $A_i=(a_{i1},a_{i2},\cdots,a_{in})$,$a_{ij}\in[0,1]$,$a_{ij}$ 表示模糊区间 u_j 与模糊集 A_i 的隶属度,所有的历史数据都依据上述定义进行模糊化. 例如,一个属于 u_i 区间的值 k 被模糊化为 A_j,如果 k 与 A_j 的隶属度最大,换句话,$a_{ij}=\max\{a_{i1},a_{i2},\cdots,a_{in}\}$,$1\leqslant j\leqslant n$. 在 Song 的模型中,使用的是公式(1)所示的三角隶属度函数.

Step 3 创建模糊逻辑关系.

Step 4 预测模型:$A_{t+1}=A_t*R$.

A_t 与 A_{t+1} 是时间 t 和 $t+1$ 时的模糊集,R 是两者的模糊关系,'$*$'是运算方式,不同的算法有不同的运算方式.

2.2 区间划分

时间序列数据在模糊化过程中,模糊区间的选择与划分至关重要. 以往模型为了简化运算,基本采用等距划分,但是等距划分存在比较大的随意性. 本文引入信息熵的概念. "信息熵"由香浓于 1948 年提出,是从物理学中的热力学领域借用过来,用于解决信息的量化度量问题. "信息熵"被用来描述信息的不确定性,这是"熵"的概念在信息领域的拓展. 就具体数据集来说,假定我们能充分利用这些数据包含的信息来做预测,预测时能够使用到所有已知的信息,对于未知信息我们不做任何假定. 在这种情况下,概率分布最均匀则预测的风险也就最小. 那么我们用什么来度量信息划分之后,每个区间信息分布是否最均匀? 香浓指出,我们可以用"信息增益"这样一个概念来测算通过每个分界点划分之后的区间包含的信息是否最均匀. 信息增益表示为:

$$\left.\begin{aligned}&\mathrm{Gain}(A)=\mathrm{Info}(S)-\mathrm{Info}_A(S)\\&\mathrm{Info}(S)=-P_i*\log_2(P_i)\\&\mathrm{Info}_A=P_j*\mathrm{Info}(S_j)\end{aligned}\right\}\tag{2}$$

式中，Gain(A)就是信息增益，值越大越好. 相对应的 Info$_A$(S)越小越好，如果我们预先设定将某区间划分成 N 个区间，那么我们需要选择(N-1)个分界点，若这些分界点的信息增益最大，则各区间的信息分布越平均.

本文的区间划分方法可通过 Matlab 编程实现，程序算法如下：

(Ⅰ)从－10 到 10，每 0.5 划分一个区间，这是初始区间；

(Ⅱ)将训练集数据按从小到大排序，同时将所有数据分别归入步骤(Ⅰ)各区间范围中；

(Ⅲ)将第一个区间与第二个区间合并，依次计算该区间每两个数中间作为分界点的信息熵增益数值；

(Ⅳ)如果该合并区间没有数据，则跳过该合并区间，将下一个区间作为第一个区间，直到合并区间存在数据；

(Ⅴ)返回信息熵增益最大的数据的位置，将该位置作为合并区间的新的分界点；

(Ⅵ)合并新产生的第二个区间和第三个区间，重复步骤(Ⅲ)至(Ⅴ)). 接下来依次合并新产生的区间与相邻的下一个区间，直到相邻区间均重新划分新的分界点；

(Ⅶ)重复步骤(Ⅲ)至步骤(Ⅵ).

上述步骤划分的区间可能存在两个问题，一是即当某数据的个数过多时，可能某个区间会只包含一种数据，所以根据实际划分情况，本文将对区间划分做二次处理. 即合并那些只包含一种数据的相邻区间，直至该区间、上一区间、下一区间均包含至少两种数. 二是某个区间跨度比较大，我们需要对该区间再次重新划分，重复上述步骤(Ⅰ)至(Ⅶ)，唯一的不同在于步骤(Ⅰ)中的初始区间为当前需重新划分的区间依据 0.5 的步伐划分的新区间，最后再观察区间是否存在第一个问题，如果存在，则依据第一个问题的处理方法对新的区间进行处理.

2.3　隶属度函数

隶属度函数表示当前数据与历史数据相近程度，一般来说隶属度越大则说明该历史数据包含的信息对基于当前数据的预测行为作用越大. 传统模型多采用 Song 等提出的三角隶属度函数，在构建预测模型时不同的模糊集给予了相同的隶属度，这看上去并不合理. Qiu 等[28]指出三角隶属度函数存在着比较明显的不足，具体来看，如果有多个数据经模糊化之后属于同一模糊集 A_i，而输出数据模糊化之后会有 $n(n\geqslant 2)$ 个模糊集 A_{j1}、A_{j2}、…、A_{jn}，每个输出模糊集的权重全部一样. 这样的处理方式导致原本包含不同重要程度信息的输入变量输出了包含相同重要信息的输入变量，所以需要提出新的隶属度函数.

如果将数据区间平均划分为 n 个等距区间，间距为 l_{in}. $\mu_{A_i}(t)$ 代表隶属度，则在 t 时刻模糊集 A_i 的隶属度函数表达式为

$$\mu_{A_i}(t)=\begin{cases}1,\ \text{if } i=1 \text{ and } x_t\leqslant m_1\\1,\ \text{if } i=n \text{ and } x_t\geqslant m_n\\\max\left\{0,1-\dfrac{|x_t-m_i|}{2l_{in}}\right\},\text{otherwise}\end{cases}\tag{3}$$

如果基于 t 时刻值 ，我们需要人为确定用于预测对象 $t+1$ 时刻值的历史数据，假设需要利用 k 个历史数据. 那么我们需要利用新的隶属度函数计算所有历史数据的隶属度，然后按从小到大顺序进行排序，选取 k 个隶属度最大的数据及相应模糊集. K ($K\leqslant n$，n 为历史数据总数)个隶属度选取方法的数学表达式为

$$\mu^k_{A_i}(t)=\begin{cases}\mu_{A_i},\text{if } \mu_{A_i}\geqslant\mu^k(t)\\0,\ \text{otherwise}\end{cases}\tag{4}$$

式中，$\mu^k(t)$是 中第 k 大的数. 经新隶属度公式计算后，我们需要构建隶属度向量，构建公式如下：

$$(\mu^k_1,\mu^k_2,\cdots,\mu^k_n)=\frac{((\mu^k_{A_1})^\alpha,(\mu^k_{A_2})^\alpha,\cdots,(\mu^k_{A_n})^\alpha)}{\sum_{i=1}^{n}(\mu^k_{A_i}(t))^\alpha}\tag{5}$$

式中，α 为调节参数，$\alpha\in(0,+\infty)$.

2.4　预测

2.4.1　加入权重

以往模型对属于一个区间的所有数据均给予了相同的权重，这样做容易损失以往信息包含的关键内容. 第一个引入权重公式的是学者 Cheng[20]，根据不同数据的关键性给予不同的权重就可能增加模型精度，提升预测能力. 学者一般可以依据预测对象的性质、类别选择权重公式. 本文主要对数据重要性进行加权.

当历史数据与当前数据处于同一个模糊区间，

并且用于对下一刻数据进行预测时，它们在重要性上的差异可以用赋予它们的权重进行衡量. 已有文献在考虑权重时很少考虑单个数据的权重，Cheng等[13]也仅仅是对趋势赋予权重. 本文提出新的权重公式如下：

$$w_k(t)=\frac{|(b_{l+1}-b_l)-(x_k-x_i)|}{b_{l+1}-b_l} \tag{6}$$

$$w(t)=\{w_1(t),w_2(t),\cdots,w_n(t)\} \tag{7}$$

式中，$x_1,x_2,\cdots,x_n$ 是与 x_i 处于同一个模糊区间 $[b_{l+1},b_l]$ 的历史数据，$w_k(t)$ 是第 k 个数据相对于 x_i 的权重，公式赋予了与当前数值越接近的数值所属模糊集的下一个模糊集越大的权重.

具体来看，如果当前股票指数为1000点，属于[900,1100]区间范围内，选取2个历史数据，假定分别为990点与910点，那么上述公式会赋予990点历史数据更大的权重，权重值为95%，910点的权重为55%. 构建该权重公式的依据在于“历史会重演”，这也是对象能够进行预测的重要前提之一.

2.4.2 使用平均值

已有的研究，包括Song、Chen、Lee等，基本都是用模糊区间的中间值来代替该模糊区间，并用于预测. 这种做法无法体现该模糊区间所包含的的真实信息，特别是当该模糊区间包含的数据具有明显偏向的时候，例如，某模糊区间[100,200]，其中间值为150，其中共包含50个数据，其中48个在[100,130]之间，只有2个在[130,200]之间，如果用150代替该区间，那么将会扭曲区间的真实信息；所以本文提出第一个改进就是用模糊区间包含的所有数据的平均值来表示该区间. 即 u_i 区间用 $m_i=\frac{y_{i1}+y_{i2}+\cdots+y_{ik}}{k}$ 表示.

2.5 结果修正

模糊时间序列模型用于股市指数的预测已经有近十年的历史. 在这段期间，研究领域大多在对模型技术层面的改进，很少涉及对预测指标的改进；然而，股指波动不仅受到自身内在价值的影响，而且交易量对价格同样有着巨大的影响.

在对股指影响因素的探究中，交易量的影响在实证研究中应用比较多. “量在价之前”是股市中的一句谚语，大量的相关研究表明，股票交易量与股价之间存在在显著的相关性. Osborn[23]是最早研究交易量与股价波动之间关系的学者. Clark[24]与Tauchen等[25]发现股价变动的绝对值与交易量之间存在正相关的结论. Gervais等[26]的研究表明，当天或本周交易量大时，未来一个月的股价将会上升，反之未来会下降. 国内学者田利辉等[27]指出，在国内股票市场上，交易金额与股票收益率呈负相关. 这些文献充分说明交易量与股票指数之间必然存在着某种联系.

“交易量”主要体现在两个方面：一是股票交易数量，二是交易金额. 本文通过对交易量、交易金额、开盘价、收盘价等指标构造新的指标，运用SVM模型进行涨跌预测，类别号为涨或跌，分别用1与−1表示，模型分类结果用于对预测数值符号的修正.

SVM是近年来发展起来的一种新的用于模式分类的人工智能方法，它适合于小样本数据分类，对样本数据的分布没有严格要求，而且预测准确率高，因此选用该算法修正模型初始预测结果.

3 实证研究

本文研究对象为上证指数的涨跌幅，数据均来自于新浪财经. 数据范围从1996年1月1日至2003年12月31日，共8年数据，比较单位为每一年数据预测误差，数据类型包括开盘价、收盘价、成交量、成交金额等. 选择这段时间数据主要基于几方面考虑：一是这段时间数据包含了中国股市几个阶段的牛市与熊市，数据样本具有代表性；二是这些数据在文献中的应用频率较高，基本所有涉及上证指数预测的FTS模型领域文献均采用这些数据，便于比较；三是这段数据包含了中国股市从产生到发展再到成熟的整个过程，这对检验模型在股市发展各阶段的预测效果具有很好的参考价值. 2003年至2014年中，在2005年到2007年短暂的牛市后，经历了2007年至2014年的长熊市，数据样本不具有代表性，且很少有上证指数预测的FTS模型使用此阶段数据，难以作比较，故本文舍弃此阶段数据.

选择涨跌幅作为研究对象，而不是指数本身，原因在于指数变动范围不确定. 再者，上证指数的上限与下限并没有确定的边界，如果将指数本身作为研究对象，则用于划分的区间选择有比较大的随机性. 几乎所有构建股指预测模型的学者在了解所有数据集的前提下去设定一个区间，使所有数据均不会超出数据集的范围，但是我们构建模型的目的是能够在实际应用中使用，而实际中的未来股指我们是无法确定的，股指在未来是否会创新高是一个未知数，所以本文选取股指变动率作为研究对象，主要原因在于中国股市的变动范围为[−10%,10%]，这是一个确定的范围，所以将股指变动率作为研究对象更具合理性.

3.1　数据预处理与度量标准

数据预处理是挖掘数据规律的前提与基础. 本文对数据的预处理主要包括以下内容：

(Ⅰ)数值处理

本文在划分数据区间时，需要使用一定时间范围内的所有数据，并对所有数据保留一位小数，划分区间以外的其他数据均保留两位小数.

(Ⅱ)构建 SVM 输入变量

主要包括 4 个构造的输入变量，构造公式与数值范围如表 1(t 日为当前日，t-1 日为上一日，t+1 日为下一日)所示.

表 1　SVM 输入-输出指标

Tab. 1　SVM input-output index

指标类别	指标说明	类别	变量类型
开-收盘价指标	t 日开盘价高于 t−1 日开盘价	1	输入变量
	t 日开盘价高于 t−1 日开盘价	−1	
构造变量 A	t 日成交量与成交金额均小于(t−1)日，且(t+1)日开盘价高于 t 日	1	输入变量
	t 日成交量与金额均大于(t−1)日，且 t+1 日开盘价小于 t 日	−1	
	其他情况	0	
构造变量 B	t 日成交量与金额均大于(t−1)日，且 t+1 日开盘价小于 t 日收盘价	−1	输入变量
	t 日成交量与成交金额均小于(t−1)日，且(t+1)日开盘价高于 t 日收盘价	1	
	t 日成交量与成交额均小于 t−1 日，且 t+1 人开盘价也小于 t 日收盘价	2	
	t 日成交量与成交额均大于 t−1 日，且 t+1 日开盘价也大于 t 日收盘价	3	
	其他情况	0	
构造变量 C	t 日成交量与成交额均大于 t−1 日	4	输入变量
	t 日成交量与成交额均小于 t−1 日	3	
	t 日成交量大于 t−1 日，成交额小于 t−1 日	2	
	其他情况	1	
收盘价−类别	t+1 日收盘价高于 t 收盘价	1	输出变量
	其他情况	−1	

上述 4 个构造变量为经过属性选择之后剩余变量；原有变量还包括上一日收盘价指标、交易量、交易金额、上一日涨跌幅. 最终数据格式如表 2 所示.

表 2　SVM 预测结果

Tab. 2　SVM prediction results

Time	开−收盘价	构造变量 A	构造变量 B	构造变量 C	class
1996/1/2	−1	0	0	3	1
1996/1/3	−1	0	3	4	1
1996/1/4	1	0	0	4	−1
1996/1/5	−1	−1	−1	4	1
1996/1/8	1	1	1	3	1
1996/1/9	−1	0	3	4	−1
1996/1/10	−1	0	0	3	−1
1996/1/11	−1	0	0	3	1
1996/1/12	−1	0	0	3	1
1996/1/15	1	0	0	1	−1

续表 2

Time	开−收盘价	构造变量 A	构造变量 B	构造变量 C	class
1996/1/16	−1	0	0	3	−1
1996/1/17	−1	−1	−1	4	−1
1996/1/18	−1	−1	−1	4	1
1996/1/19	−1	1	0	3	−1
1996/1/22	−1	0	0	3	1
1996/1/23	1	1	1	3	1
1996/1/24	1	0	0	4	−1
1996/1/25	−1	−1	−1	4	1
1996/1/26	1	1	1	3	1
1996/1/29	1	0	0	4	1
1996/1/30	−1	0	3	4	1
1996/1/31	−1	0	3	4	−1
1996/2/1	−1	0	0	3	−1
1996/2/2	1	0	0	4	−1
⋮					

预测结果的比较标准有两个:一是RMSE(均方根误差),RMSE值越小说明预测值与真实值误差越小,模型效果越好;二是预测涨跌符号正确率,涨跌符号预测正确率计算公式为:FR=预测符号正确个数/总预测数据个数.

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{t=1}^{n}(\mathrm{Forcast}(t)-\mathrm{Actual}(t))^2}{n}} \quad (8)$$

3.2 实验结果与分析

3.2.1 初始改进FTS模型预测结果

本文提出的模糊时间序列模型与传统模型的不同之处在于:一是使用了区间划分结果;二是使用了前文提出的新的隶属度公式;三是用区间所有数的平均值代替该区间,而不是用区间的平均值.经上述三点改进之后的模型预测结果即为本文初始预测结果,部分数据如表3(其中历史数据为2000年度数据,预测数据为2001年上半年数据)所示.

表3 本文模型初始预测结果

Tab.3 Prediction results of the model proposed in this paper

训练模型数据		初始预测结果	
Time	涨跌幅	Time	涨跌幅
2000/1/4	2.91	2001/1/2	0.42
2000/1/5	0.24	2001/1/3	−1.06
2000/1/6	3.85	2001/1/4	0.42
2000/1/7	3.60	2001/1/5	−0.09
2000/1/10	1.88	2001/1/8	0.10
2000/1/11	−4.23	2001/1/9	0.74
2000/1/12	−2.82	2001/1/10	0.17
2000/1/13	−0.94	2001/1/11	0.45
2000/1/14	−1.09	2001/1/12	−0.09
2000/1/17	1.74	……	
2000/1/18	−0.47	2001/6/29	−0.05
2000/1/19	0.99		
2000/1/20	1.81		
2000/1/21	−0.12		
2000/1/24	0.84		
⋮			
2000/12/29	0.96		

3.2.2 使用SVM进一步修正预测结果

对于3.2.1的预测结果,误差范围尽管比传统模型有所改善,但是通过观察,依旧发现很多情况下,预测结果与真实值的变动方向不一致,因此我们引入SVM算法对初始预测结果的符号进行修正.修正规则为若初始预测涨跌结果与SVM结果不一致,则改变初始预测结果涨跌符号.

经过上述步骤,我们完成了对初始预测值的修正.最后的结果表明,经过修正之后的数据误差率要低于修正之前的误差率.

模型修正前后的数值绝对值相同,只是符号发生了变化,比较结果如表4所示.

表4 本文模型修正前后结果比较

Tab.4 Comparison of the results before and after the correction of the model

Time	未经修正模型结果		经修正模型结果	
	FR	RMSE	FR	RMSE
1997年	28.51	50.21%	26.30	58.85%
1998年	16.73	50.41%	16.12	59.76%
1999年	26.01	47.70%	25.32	61.09%
2000年	26.93	53.56%	25.31	60.67%
2001年	25.88	51.67%	25.05	59.83%
2002年	24.38	50.63%	24.22	56.12%
2003年	17.21	52.28%	16.78	55.66%
平均值	23.67	50.92%	22.73	58.85%

从表4可以看出,经过修正的模型预测结果显著优于未经修正的模型结果,RMSE平均值分别为22.73与23.67.具体来看,每一年的RMSE修正后的结果均优于未修正的结果,其中修正模型RMSE最高为1997年的26.30,最低为1998年的16.12;未修正模型RMSE最高为1997年的28.51,最低为1998年的16.73.FR也同样如此,平均值分别为58.85%与50.92%.具体来看,未经修正模型FR最高为2000年的53.56%,最低为1999年的47.47%.经修正模型FR最高为1999年的61.09%,最低为2003年的55.66%,且FR的结果提升效果要好于RMSE结果改善,由此可见模型修正的必要性.

3.2.3 修正模型与其他FTS模型结果比较

我们将采用RMSE指标评价各模型效果,结果如表5所示.

表 5　多种 FTS 模型 RMSE 指标结果比较

Tab. 5　Comparison results of RMSE index of various FTS models

Time	Song	Chen	Lee	Kuarng	证据理论 FTS	未修正模型	Proposed Model
1997	50.12	38.07	36.57	32.03	30.84	28.51	26.30
1998	44.19	34.51	32.02	24.80	25.60	16.73	16.12
1999	38.38	37.47	35.81	25.23	27.74	26.01	25.32
2000	46.16	42.15	40.33	30.26	30.07	26.93	25.31
2001	42.32	37.23	36.76	27.49	27.49	25.88	25.05
2002	38.80	47.19	35.39	39.90	27.65	24.38	24.22
2003	32.71	36.57	30.98	29.90	25.02	17.21	16.78
Mean	41.81	39.03	35.41	29.94	27.77	23.67	22.73

比较模型数据均来自于文献[28],从表 5 可以得出,本文提出的修正模型结果与未修正模型结果均要远好于其他比较模型结果,其中修正模型 RMSE 平均值降低百分比分别为 45.64%、41.76%、35.81%、24.08%、18.15%.其中 Song 模型误差率最大,达到 41.81,证据理论 FTS 模型在比较模型中平均误差率最低,为 27.77.本文模型无论是个体数据还是整体平均值基本均优于其他模型,平均 RMSE 为 22.73.

4　结论

本文是基于前人的研究成果,通过对已有模型的修正,主要从数据层面对股指进行预测.通过对模糊时间序列模型在预测方面的应用,特别是其基本模型以及学者改进模型在股市指数预测精度上的不断改善.本文通过三个方面对已有模型进行改进,一个是对模糊区间产生方法的改善,二是对隶属度计算公式的改善,三是模型修正方法.本文提出的改进模型在预测精度上要显著优于其他模型,因此本文模型在股指预测方面具有比较明显的优势.

尽管预测结果要优于比较模型,但在今后的研究中仍可以进一步改进.本文停留于数据层面,研究中未考虑广大股票市场投资者心理变化对股市的影响.Zhang 等[29]以 Tweeter 上不同情绪的词语为研究对象,统计每种情绪代表词语出现的次数对情绪进行量化,通过情绪量化指标与美国主要股市波动相关性研究对股指进行预测.学者 Bollen 等[30]通过 Tweeter 情绪词语构建了两个情绪跟踪器计算情绪跟踪器与道琼斯指数之间的关系,最后通过线性时间序列方程预测指数变动.本文用于 SVM 模型的构造变量基本是以股市自由参数为基础进行构造,即仅仅通过成交量、成交金额、开盘价、收盘价等构造变量.这些构造变量没有考虑影响股市波动的宏观经济指标、人口统计指标等因素,因此在未来的研究中可以考虑加入加入新的指标对模型进行修正.

参考文献(References)

[1] 陈迪红,杨湘豫,李华中. 中国证券市场指数波动的周期分析[J]. 湖南大学学报,2003,30(5):88-91.
CHEN Dihong, YANG Xiangyu, LI Huazhong. Analysis of cycle about fluctuation of stock index in Chinese securities market [J]. Journal of Hunan University, 2003, 30(5): 88-91.

[2] 周佰成,周建文,方炬. 中、美证券市场的波动周期比较[J]. 经济纵横,2006,(5):72-73.

[3] 田俊刚,梁红漫. 中国股票市场周期性研究[J]. 武汉金融,2008,(7):14-16,36.

[4] 黄继平,黄良文. 中国股市波动的周期性研究[J]. 统计研究,2003,(11):9-14.

[5] 董直庆,夏小迪. 我国通货膨胀和股市周期波动共变性和非一致性再检验[J]. 经济学家,2010,(3):73-80.

[6] CHENG C H, CHEN Y S, WU Y L. Forecasting innovation diffusion of products using trend-weighted fuzzy time-series model [J]. Expert Systems with Applications, 2009, 36(2): 1826-1832.

[7] 张韬,冯子健,杨维中,等. 模糊时间序列分析在肾综合征出血热发病率预测的应用初探[J]. 中国卫生统计,2011,28(2):146-150.

[8] 钱冰冰. Type-2 模糊系统在黄金价格预测中的应用[J]. 佳木斯大学学报,2007,25(3):397-399.

[9] SONG Q, CHISSOM B S. Forecasting Enrollments With Fuzzy Time Series[J]. Fuzzy Sets and Systems, 1993, 54(93): 1-9.

[10] LEE M H, EFENDI R, ISMAIL Z. Modified weighted for enrollment forecasting based on fuzzy time series [J]. MATEMATIKA, 2009, 25(1): 67-78.

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

[12] HUARNG K, YU H K. A type 2 fuzzy time series model for stock index forecasting[J]. Physica A: Statistical Mechanics and its Applications, 2005, 353 (1-4): 445-462.

[13] CHENG C H, CHEN T L, CHIANG C H. Trend-weighted fuzzy time-series model for TAIEX forecasting[C]// Proceedings of the 13th International Conference on Neural Information Processing. Hong Kong, China: ACM Press, 2006, 4234: 469-477.

[14] CHU H H, CHEN T L, CHENG C H, et al. Fuzzy dual-factor time-series for stock index forecasting[J]. Expert Systems with Applications, 2009, 36(1): 165-171.

[15] TING J L. Causalities of the Taiwan stock market[J]. Physica A: Statistical Mechanics and its Applications, 2003, 324(1-2): 285-295.

[16] KARPOFF J M. The relation between price changes and trading volume: A survey[J]. The Journal of Financial and Quantitative Analysis, 1987, 22(1): 109-126.

[17] 金春雨，郭沛. 我国股票市场量价关系的实证研究--基于上证指数的 VAR 模型分析[J]. 价格理论与实践，2010，(9)：60-61.

[18] 王重，张文转. 股票指数与股票市场技术要素的实证分析[J]. 现代商贸工业，2008，20(2)：159-160.

[19] ALADAG C H, BASARAN M A, EGRIOGLU E, et al. Forecasting in high order fuzzy times series by using neural networks to define fuzzy relations[J]. Expert Systems with Applications, 2009, 36(3): 4228-4231.

[20] EGRIOGLU E, ALADAG C H, YOLCU U, et al. Finding an optimal interval length in high order fuzzy time series[J]. Expert Systems with Applications, 2010, 37(7): 5052-5055.

[21] CHEN S M, WANG N Y, PAN J S. Forecasting enrollments using automatic clustering techniques and fuzzy logical relationships[J]. Expert Systems with Applications, 2009, 36(8): 11070-11076.

[22] QIU W, LIU X, WANG L. Forecasting shanghai composite index based on fuzzy time series and improved C-fuzzy decision trees[J]. Expert Systems with Applications, 2012, 39(9): 7680-7689.

[23] OSBORNE M F M. Brownian motion in the stock market[J]. Operations Research, 1959, 7(2): 145-173.

[24] CLARK P K. A subordinated stochastic process model with finite variance for speculative prices[J]. General Information, 1973, 41(1): 135-155.

[25] TAUCHEN G E, Pitts M. The price variability-volume relationship on speculative markets[J]. Econometrica, 1983, 51(2): 485-505.

[26] GERVAIS S, Kaniel R, Mingelgrin D H. The high-volume return premium[J]. Journal Akuntansi Dan Keuangan, 2008, 56(3): 877-919.

[27] 田利辉，王冠英. 我国股票定价五因素模型:交易量如何影响股票收益率?[J]. 南开经济研究，2014，(2)：54-75.
TIAN L H, WANG G Y. Asset pricing model of the Chinese stock market: How trading volumes influence the returns[J]. Nankai Economic Studies, 2014, (2): 54-75.

[28] QIU W R, LIU X D, WANG L D. Forecasting Shanghai composite index based on fuzzy time series and improved C-fuzzy decision trees[J]. Expert Systems with Applications, 2012, 39(9): 7680-7689.

[29] ZHANG X, FUEHRES H, GLOOR P A. Predicting Stock market indicators through twitter "I hope it is not as bad as I fear"[J]. Procedia - Social and Behavioral Sciences, 2011, 26: 55-62.

[30] BOLLEN J, MAO H N, ZENG X J. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1): 1-8.