K 线能量计算的股市生命期态势预测方法*

姚宏亮，周光辉，李俊照
（合肥工业大学 计算机与信息学院，合肥 230009）

摘　要：股市中 K 线特征是股价涨跌的因果信息，基于支持向量机（SVM）的股价预测模型没有考虑 K 线特征知识，对于股价态势难以有效预测。提出基于 K 线能量计算的股市生命期支持向量机态势预测算法（LPF-SVM），首先，提取典型 K 线特征，通过引入特征的孕育成熟度和爆发力定义，给出 K 线特征支持向量机算法（KLF-SVM）；进而，在 KLF-SVM 算法基础上定义特征的能量计算模型给出一种 K 线能量计算的 SVM 股价预测算法。为了有效地预测态势，引入股价波动的生命期概念，通过 K 线组合特征判定股价所处的生命期的阶段，进而结合生命期阶段之间的时序影响关系给出一种基于生命期的股价态势预测算法。在上证和深证数据集上的实验结果表明，LPF-SVM 算法对于股价上升波段和下跌波段的股价预测取得了很好的效果。

关键词：K 线特征；孕育成熟度；爆发力；能量；股市生命期
中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1001-3695(2016)06-1637-05
doi:10.3969/j.issn.1001-3695.2016.06.009

K line energy calculation method for stock market lifetime prediction

Yao Hongliang, Zhou Guanghui, Li Junzhao
(*School of Computer & Information*, *Hefei University of Technology*, *Hefei* 230009, *China*)

Abstract: In the stock market, K line feature is the causal information for the rise and fall of stock price. Stock price prediction model of support vector machine (SVM), which does not consider K line features, can not predict the stock trend effectively. Based on K line energy calculation, this paper put forward a lifetime support vector machine (LPF-SVM) algorithm, forecasting stock market situation. First, it extracted the typical K line, with the introduction of maturity and explosive force definition, obtained the K line feature of support vector machine algorithm (KLF-SVM). Then on the basis of KLF-SVM, it defined typical energy calculation model, gave a kind of SVM prediction algorithm for K line energy calculation. In order to predict the situation effectively, it introduced the lifetime concept of stock price volatility. The stage of lifetime of the stock price could be judged through the K line combination features, and then combing with the timing effect relationship of life stage, it gave a stock price trend prediction algorithm based on lifetime. The experimental results on the Shanghai and Shenzhen data set show that, the LPF-SVM algorithm can predict the rising and falling band of stock price effectively.

Key words: K line feature; maturity; explosive force; energy; lifetime of the stock

0　引言

随着金融市场的发展，股票投资在人们生活中的地位越来越重要，使用高效的挑选方法可以帮助人们在股票投资中取得丰厚的回报。但是中国股市因为受法制建设不健全、市场机制不完善以及投资者心理不成熟等因素影响容易产生大的波动[1]，使得股票价格难以有效地预测。

对于股票价格的预测，早期常使用 AR 模型、ARMA 模型以及 ARIMA 模型，它们在正态、平稳时间序列数据中展现出了优越性[2]；后来 Engle 和 Bollerslev[3,4] 分别提出 ARCH 模型和 GARCH 模型，但是它们更适合处理线性数据，对非线性和时变数据效果不够理想[5]。Hung[6] 提出将 GARCH 模型和模糊系统相结合能够处理非线性和时变的股市数据，预测准确率得以提高，但是系统参数的选取过于复杂。

股价是非线性的，容易受国家宏观经济政策等因素影响而出现非正态和非线性。随着非线性科学的发展有人提出用神经网络模型处理股价，神经网络模型在处理非线性数据时优势明显[7,8]，但是初始网络的节点规模大小、节点之间权值的大小很难给出一个合适的确定方法，此外神经网络容易陷入局部最优也导致数据处理结果不够理想[9]。Vapnik 等人[10] 在统计学理论基础上提出了支持向量机，它是建立在结构风险最小化理论基础之上，通过巧妙地应用核函数将低维数据映射到高维[11]，使得在低维上线性不可分的数据在高维上变得线性可分，对解决小样本、非线性、高维数据有着很好的作用[12]。此外，支持向量机在原理上避免了神经网络易陷入局部最优的问题，大量实验结果证明支持向量机的预测精度高于神经网络。

股票价格数据仅仅是从价格层面反映了股票的状况，为了深入了解股价变化的原因，需要分析股市表现出的一些特征，这些特征往往能从更深层次揭示股价波动的机制。Zhang 等人[13] 提出用支持向量机和粗糙集相结合提取技术指标的特征来预测股票价格取得了良好的效果；Mansukhbhai 等人[14] 提出用软计算技术提取特征来预测股价，相比普通的技术具有更快的运算速度和更高的准确率。但是他们的研究都没有分析股市中 K 线特征对于股价走势的影响。

收稿日期：2015-01-18；**修回日期**：2015-03-09　**基金项目**：国家自然科学基金资助项目(61175051,61070131,61175033)

作者简介：姚宏亮(1972-)，男，副教授，博士，主要研究方向为人工智能、知识工程(dmicyhl@163.com)；周光辉(1989-)，硕士，主要研究方向为人工智能和知识工程；李俊照(1975-)，男，讲师，博士研究生，主要研究方向为机器学习、人工智能.

K线特征是股市多空双方博弈的结果,能够反映股市上涨能量和下跌能量的强弱,不同的K线特征对后续股价涨跌有着不同的影响。对于针对现有的股市预测算法没有深入分析K线特征对于股市走势造成影响的这一问题,提出基于股市生命期中不同阶段的K线特征的股市预测算法(stock life different period K line feature-SVM,LPF-SVM)。LPF-SVM算法分为两个步骤:a)根据K线特征出现之前的环境对于K线特征孕育情况建立特征的孕育成熟度模型,然后根据K线特征出现时的各种指标表现情况建立特征的爆发力模型,根据K线特征的孕育成熟度和爆发力预测特征之后的短期的股价走势(KLF-SVM);b)为了能够从整体的角度预测价格走势,将股市一个上涨波段或一个下跌波段记为一个生命周期,根据一个生命周期中股价走势的不同将一个生命期分为四个不同的阶段。根据K线特征的孕育成熟度和爆发力凝练出K线特征的能量模型,再根据K线特征的能量大小和特征的组合方式判断股价当前所处的阶段,将当前阶段的价格走势作为先验知识加入SVM预测未来股价。最后在上证指数和深证成指的数据集上分别验证算法的有效性。

1 支持向量机

支持向量机方法是建立在统计学习理论的VC维理论和结构风险最小原理基础上的,很大程度地解决了传统机器学习中的维数灾难、过学习、非线性以及易陷入局部最优问题。支持向量机通过使用核函数将输入的变量从低维空间映射到高维空间中,在新的高维特征空间寻找最优分类面进行线性回归。假定给定训练样本集$(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)$,其中$x_i\in R^N$为$N$维特征向量,$y_i\in\{-1,1\}$或$y_i\in\{1,2,\cdots,k\}$;当$y_i\in\{-1,1\}$时为二分类问题,当$y_i\in\{1,2,\cdots,k\}$为$K$分类问题。

若数据是线性可分的,则寻找一个超平面方程为$w\cdot x+b=0$,对数据集进行归一化,使得对于线性可分的样本集满足$y_i(w\cdot x_i+b)-1\geqslant 0, i=1,2\cdots m$;分类间隔为$2/\|w\|$,要求出最优超平面就要是使得分类间隔最大化,$\varepsilon_i$是松弛变量表示样本被错分的程度,$C$是惩罚因子表示对样本加载错分点上的惩罚。原始的求解问题就转换为凸二次规划的求解问题,约束条件如下:

$$\min_{w,b,\varepsilon_i}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\varepsilon_i$$
$$\text{s.t.}\quad y_i((w\cdot x_i)+b)\geqslant 1-\varepsilon_i\quad i=1,2,\cdots,m \tag{1}$$

由于凸二次规划问题存在全局最优解,用Lagrange乘子将其转换为对偶形式求解。得到最优超平面决策函数如下:

$$f(x)=\text{sgn}\left(\sum_{i=1}^{m}\alpha_i^* y_i(x_i\cdot x)+b^*\right) \tag{2}$$

其中:α_i^*、b^*为确定最优超平面划分的参数。若面对处理非线性数据的情况,SVM通过核函数将数据样本从低维空间映射到高维空间进行线性划分,为了降低高维特征空间中的计算复杂度,支持向量机采用核函数$K(x_i\cdot x)$来代替高维空间中的内积运算。最优超平面的分类决策函数转变为

$$f(x)=\text{sgn}\left(\sum_{i=1}^{m}\alpha_i^* y_i K(x_i\cdot x)+b^*\right) \tag{3}$$

2 基于K线特征的股价预测

K线特征是股市中的重要信息,不同的K线特征对于股价涨跌有着不同的影响。首先根据K线特征出现之前的环境对于特征的孕育情况得出特征的孕育成熟度模型,然后根据K线特征出现时的各种指标表现情况得出特征的爆发力模型,最后根据K线特征的孕育成熟度和爆发力判断特征之后短期的股价波动范围,并将其作为先验知识加入SVM模型,在股票价格数据集上预测未来股价。

2.1 K线特征的描述

K线又称做阴阳线、棒线或蜡烛线,起源于日本德川幕府时代的米市交易,以交易时间为横坐标,价格为纵坐标。每日的K线图包括开盘价、收盘价、最高价、最低价四项数据。

定义1 K线特征。在股市的K线图中会出现各种类型的特征,不同的K线特征对股价涨跌有着不同的影响。对于常见的重要K线特征描述如下:

a)长阳,实体长度大于2%的阳线,用特征f_1表示,如图1所示。

b)长阴,实体长度大于2%的阴线,用特征f_2表示,如图2所示。

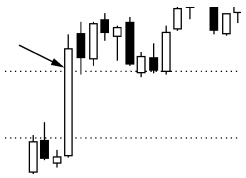

图1 K线特征:长阳

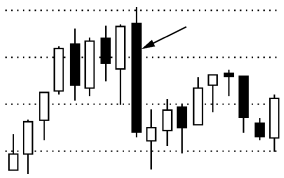

图2 K线特征:长阴

c)n连阳($n\geqslant 2$),连续N天的阳线,每天的收盘价比前一天高,用特征f_3表示,如图3所示为3连阳。

d)n连阴($n\geqslant 2$),连续N天的阴线,每天的收盘价比前一天低,用特征f_4表示,如图4所示为3连阴。

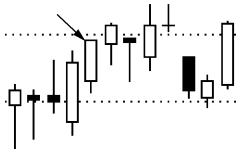

图3 K线特征:3连阳

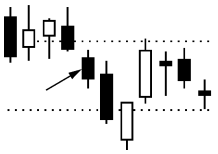

图4 K线特征:3连阴

e)阳包阴,第二天阳实体完全包容第一天的阴实体,用特征f_5表示,如图5所示。

f)阴包阳,第二天阴实体完全包容第一天的阳实体,用特征f_6表示,如图6所示。

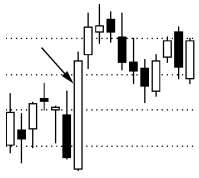

图5 K线特征:阳包阴

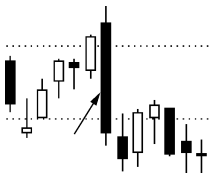

图6 K线特征:阴包阳

在股市中长阳、n连阳($n\geqslant 2$)、阳包阴等特征是看涨的K线特征,简称上涨特征;长阴、n连阴($n\geqslant 2$)、阴包阳等特征是看跌的K线特征,简称下跌特征。

2.2 K线特征的孕育成熟度

定义2 K线特征的孕育成熟度。K线特征出现之前的环境对于K线特征的孕育情况决定了该特征的出现是否合理。若特征出现合理,则后续的股价会延续特征的走势,若不够合理则不久股价就会背离该特征的走势。

对于长阳、n连阳($n\geqslant 2$)、阳包阴等上涨特征,它们在出现

之前的环境对它们的孕育成熟度越高,则上涨越真实,后续的股价会一直往上走。在上涨特征出现之前 5 日的收盘价距离 BOLL 线的上轨道平均越远,则股价后续向上轨道靠近的可能性越大,因此上涨特征孕育越充分。此外,上涨特征出现前 5 日的平均成交量明显缩小,则为上涨特征蓄积能量越充分,因此上涨特征孕育越充分。上涨 K 线特征孕育成熟度定义为

$$b=\frac{\mathrm{Vol}_{10}}{\mathrm{Vol}_5}\cdot\frac{1}{5}\sum_{i=1}^{5}\frac{\mathrm{up}_{t-i}-\mathrm{close}_{t-i}}{\mathrm{up}_{t-i}-\mathrm{down}_{t-i}} \tag{4}$$

其中:b 表示 K 线特征的孕育成熟度;Vol_5 和 Vol_{10} 分别代表特征出现之前 5 日的平均成交量和 10 日平均成交量;up_{t-i} 表示特征出现前 i 日 BOLL 线的上轨指数;close_{t-i} 表示特征出现前 i 日的收盘价;down_{t-i} 表示特征出现前 i 日 BOLL 线的下轨指数。

同理,对于长阴、n 连阴($n\geqslant 2$)、阴包阳等下跌特征,在它们出现前 5 日的收盘价距离 BOLL 线的下轨道平均距离越远,则股价后续向下轨道靠近的可能性越大,下跌特征孕育越充分。在下跌特征出现前 5 日成交量放大,则后续上证的能量消耗过多,下跌的可能性越大,下跌特征孕育越充分。下跌 K 线特征孕育成熟度定义为

$$b=\frac{\mathrm{Vol}_5}{\mathrm{Vol}_{10}}\cdot\frac{1}{5}\sum_{i=1}^{5}\frac{\mathrm{close}_{t-i}-\mathrm{down}_{t-i}}{\mathrm{up}_{t-i}-\mathrm{down}_{t-i}} \tag{5}$$

2.3　K 线特征爆发力

定义 3　K 线特征的爆发力。股市中 K 线特征的涨跌幅、成交量、资金流入流出可以从三个不同方面反映该特征的强弱,若这三个指标表现得强势,则该特征的爆发力强,对于后续价格走势有着较强的影响,反之,对于后续价格走势影响较弱。综合上述三个因素对长阳、三连阳、阳包阴的爆发力定义如下,长阴、三连阴、阴包阳的定义同理。

a)长阳的爆发力定义为

$$y=w_1\cdot a+w_2\cdot v+w_3\cdot z \tag{6}$$

其中:w_1、w_2、w_3 代表权值;a、v、z 分别代表涨跌幅、成交量、资金流入流出。为了方便结果处理,对 a、v、z 均归一化为[0,1],归一化公式如下:

$$x'=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{7}$$

其中:x'表示归一化后的值,$x_{\max}$表示 x 的数据集中最大值,$x_{\min}$ 表示 x 的数据集中最小值。

b)n 连阳($n\geqslant 2$)的爆发力定义为

$$y=w_1\cdot\frac{1}{n}\sum_{j=1}^{n}a_j+w_2\cdot\frac{1}{n}\sum_{j=1}^{n}v_j+w_3\cdot\frac{1}{n}\sum_{j=1}^{n}z_j \tag{8}$$

其中:w_1、w_2、w_3 代表权值;n 为连阳持续的天数;a_j、v_j、z_j 分别表示其持续时间内第 $j(j=1,2,\cdots,n)$ 日的涨跌幅、成交量、资金流入流出。为了方便结果处理,对于 a_j、v_j、z_j 使用式(7)归一化为[0,1]。

c)阳包阴的爆发力定义为

$$y=w_1\cdot(a_t-a_{t-1})+w_2\cdot(v_t-v_{t-1})+w_3\cdot(z_t-z_{t-1}) \tag{9}$$

对于 K 线特征阳包阴,由于第二天决定未来的走势,爆发力的强弱主要决定于第二天的涨跌幅、成交量、资金流入流出相对于第一天的强弱。其中 w_1、w_2、w_3 代表权值;a_t-a_{t-1}、v_t-v_{t-1}、z_t-z_{t-1} 分别表示第二天与第一天的涨跌幅度之差、成交量之差、资金流入流出之差。为了方便计算,对 a_t-a_{t-1}、v_t-v_{t-1}、z_t-z_{t-1} 分别使用式(7)归一化为[0,1]。

为了确定式(6)(8)(9)中的权值,先设 w_1、w_2、w_3 为 1/3,再计算这三个指标与爆发力 y 间的欧氏距离,欧氏距离公式为

$$d_{x_i,y}=\sqrt{\sum_k(x_{i,k}-y_k)^2}\quad i=1,2,3;k=1,2,3,\cdots \tag{10}$$

其中:y 代表爆发力的值,对于特征长阳和长阴,$x_i(i=1,2,3)$ 分别代表 a、v 和 z;对于 n 连阳和 n 连阴,$x_i(i=1,2,3)$ 分别代表 $\frac{1}{n}\sum_{j=1}^{n}a_j$、$\frac{1}{n}\sum_{j=1}^{n}v_j$、$\frac{1}{n}\sum_{j=1}^{n}z_j$;对于特征阳包阴和阴包阳,$x_i(i=1,2,3)$ 分别代表 a_t-a_{t-1}、v_t-v_{t-1}、z_t-z_{t-1}。根据欧氏距离计算相应的近似度为

$$d_i=\mathrm{e}^{-d_{x_i,y}} \tag{11}$$

根据近似度得出对应的权值为

$$w_i=d_i/\sum_i d_i \tag{12}$$

2.4　K 线形态特征作为先验知识加入 SVM

在当前的数据集中,使用时序滑动窗口捕捉 K 线形态特征,当捕捉到一个特征 f_i 后,计算该特征的孕育成熟度 b_i 和特征的爆发力 y_i,根据特征的类型 ID、特征的孕育成熟度 b_i 和特征的爆发力 y_i 来判断未来 n 日的波动范围[u,v]。

将[u,v]作为先验知识加入 SVM 模型中,在支持向量机公式中加入约束条件使预测结果在[u,v]间。

$$\begin{aligned}&\min\quad\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)\\&\text{s.t.}\quad y_i-w\cdot\varphi(x_i)-b\leqslant\varepsilon+\xi_i\\&w\cdot\varphi(x_i)+b-y_i\leqslant\varepsilon+\xi_i^*,\ \varepsilon,\xi_i,\xi_i^*\geqslant 0;i=1,2,\cdots,n\end{aligned} \tag{13}$$

最终股价预测结果约束在[u,v],即 $u\leqslant w\cdot\varphi(x_i)+b\leqslant v$,记当前价格为 p_t,后续向下和向上波动大小分别为 a_1 和 a_2,u 和 v 分别等于 p_t-a_1 和 p_t+a_2,C 是惩罚因子,w 为分类间隔。为解决约束最优化问题,引入拉格朗日函数

$$\begin{aligned}L(w,b,\xi_i,\xi_i^*)=&\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}(\xi_i+\xi_i^*)-\\&\sum_{i=1}^{m}\alpha_i[\varepsilon+\xi_i-y_i+(w\cdot\varphi(x_i))+b]-\\&\sum_{i=1}^{m}\alpha_i^*[\varepsilon+\xi_i^*+y_i-(w\cdot\varphi(x_i))-b]-\\&\sum_{i=1}^{m}(\beta_i\xi_i+\beta_i^*\xi_i^*)-\sum_{i=1}^{m}\gamma_i[c+(w\cdot\varphi(x_i))+b]-\\&\sum_{i=1}^{m}\theta_i[d-(w\cdot\varphi(x_i))-b]\end{aligned} \tag{14}$$

其中:α_i、α_i^*、β_i、β_i^*、γ_i、θ_i 为拉格朗日系数。令 $\lambda_i=\alpha_i-\alpha_i^*-\gamma_i+\theta_i$,最优化问题转换为

$$\begin{aligned}&\max\ -\frac{1}{2}\sum_{i,j=1}^{m}\lambda_i\lambda_jK(x_i,x_j)+\\&\sum_{i=1}^{m}y_i(\alpha_i-\alpha_i^*)-\varepsilon\sum_{i=1}^{m}(\alpha_i-\alpha_i^*)-\sum_{i=1}^{m}\gamma_ic-\sum_{i=1}^{m}\theta_id\\&\text{s.t.}\quad\sum_{i=1}^{m}\lambda_i=0\end{aligned} \tag{15}$$

融合先验知识的支持向量机表示为

$$w=\sum_i\lambda_i\varphi(x_i) \tag{16}$$

平面决策函数为

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{m}\lambda_iK(x_i\cdot x)+b) \tag{17}$$

将先验知识作为预测结果的约束条件加入 SVM 后,再利用股票的开盘价、收盘价、最高价、最低价、MACD、KDJ 指标等因素预测股票的收盘价。

2.5　KLF-SVM 算法描述

输入:股票的价格数据集 dataset。

输出:股票的收盘价 close 的预测结果。

a)以滑动窗口的形式捕捉上证指数和深证成指数据集上

的 K 线特征 f_i，记录每个特征的类型 ID、持续的时间 T，并计算特征的孕育成熟度 b_i 和特征的爆发力 y_i 以及未来 n 日股价向下和向上波动大小 a_1 和 a_2，将它们构成一个元组 f_i(ID、T、b_i、y_i、a_1、a_2)。

b)将在数据集中收集的 K 线特征元组 f_i(ID、T、b_i、y_i、a_1、a_2)组成数据集，利用 SVM 训练该数据集得到训练模型，使得根据一个特征元组的 ID、T、b_i、y_i 就能预测该特征之后 n 日的向下和向上波动大小 a_1 和 a_2。

c)将股票的每日开盘价 open、收盘价 close、最高价 max、最低价 min、MACD、KDJ 等技术指标作为输入变量，通过 SVM 建立基本的股票收盘价预测模型。

d)在数据集上以滑动窗口形式捕捉 K 线特征，若未捕捉到 K 线特征，则用传统 SVM 导入 dataset 数据集预测收盘价 close；若捕捉到 K 线特征之后，根据特征的 ID、T、b_i、y_i 预测其之后 n 日股价向下和向上波动大小 a_1 和 a_2，再根据 a_1 和 a_2 计算其波动范围$[u,v]$并作为先验知识加入 SVM，代入步骤 c)得出收盘价 close 的预测值。

3 基于股市生命期中不同阶段的股市价格预测

KLF-SVM 算法根据 K 线特征的孕育成熟度和爆发力预测后续的股价可以很大程度上提高特征之后股价走势预测的准确率。但是单独根据一个特征是从局部角度把握股价的趋势，为了从整体的角度更好地把握股市的趋势，本章提出一种基于 K 线能量计算的股市生命期不同阶段态势预测算法(LPF-SVM)。

LPF-SVM 算法将股市中一个上涨波段或一个下跌波段记为一个生命周期。将一个生命期分为四种不同阶段，即态势确立阶段、态势发展阶段、态势震荡阶段、衰退结束阶段。在 KLF-SVM 算法基础上建立 K 线特征能量模型，根据当前的特征能量强弱以及特征的组合方式来判断当前股价所处的阶段；根据所处的阶段判断未来股价的走势及其波动范围，把未来股价波动范围作为先验知识加入 SVM，预测未来股价。

3.1 股市生命期中不同阶段的描述

股市的上涨波段和下跌波段都可以看成股市的一个生命期。股价在上涨波段刚开始时上涨能量充足、上涨速度快，随着时间推移上涨能量减弱，股价在某一个价格区间上下震荡，到了末期则上涨能量非常弱，下跌能量主导股价走势，股价进入下跌波段，开始另一个生命期。

定义 4 特征的能量。K 线特征的能量大小影响着后续股价的涨跌，能量越大则后续上涨或下跌的空间越大。K 线特征出现之前的背景对特征孕育是否充分决定了特征出现的真实性，进而影响未来股价走势。特征的爆发力则反映了特征出现时各种指标的表现情况，它的强弱也会影响未来股价的走势。根据 K 线特征的成熟度和爆发力，综合得出 K 线特征的能量计算公式如下：

$$E_i = w_1 \cdot b_i + w_2 \cdot y_i \tag{18}$$

其中：E_i 表示特征 f_i 的能量；b_i 和 y_i 分别表示特征 f_i 的孕育成熟度和爆发力，为了方便计算，将 b_i 和 y_i 使用式(7)归一化为 $[0,1]$；w_1、w_2 代表相应权值，其确定方法使用式(8)～(10)。

以上涨波段为例，将股市生命期中态势确立阶段、态势发展阶段、态势震荡阶段、衰退结束阶段分别设为 S_1、S_2、S_3、S_4。在不同阶段，K 线特征的能量强弱有所不同。例如，统计 2010 年来上证指数中的 78 个长阳案例，根据特征的能量计算公式得出在 S_1 阶段长阳的能量均值为 0.69，在 S_2 阶段的长阳能量均值为 0.77，在 S_3 阶段的长阳能量均值为 0.46，在 S_4 阶段的长阳能量均值为 0.31。因此可以根据特征的能量值接近哪一阶段该特征的能量均值以及捕捉到的 K 线特征的组合方式判断当前处于哪一阶段。每个阶段的判断方法如下：

a)以滑动窗口形式捕捉 K 线形态特征，若先捕捉到下跌特征的组合，且每个特征 f_j 能量值 E_j 和该特征在 S_1 阶段的平均能量值 $\overline{E}_{1j}$ 满足 $|E_j - \overline{E}_{1j}| < \alpha_1$，则股价可能进入 S_1 阶段。继续用滑动窗口捕捉 K 线形态特征，若股价呈 V 型反转，捕捉到上涨特征的组合，上涨特征 f_k 的能量值 E_k 和该特征在 S_1 阶段的平均能量值 $\overline{E}_{1k}$ 满足 $|E_k - \overline{E}_{1k}| < \alpha_1$，则确定当前阶段为 S_1 阶段。

b)继续捕捉 K 线特征，若连续捕捉到上涨特征的组合，且上涨特征 f_k 的能量值 E_k 和该特征在 S_2 阶段的平均能量值 $\overline{E}_{2k}$ 满足 $|E_k - \overline{E}_{2k}| < \alpha_2$，则确定当前阶段为 S_2 阶段。

c)继续捕捉 K 线特征，若捕捉到的 K 线特征的收盘价 close_i 和近日某一价格 p 满足 $|\text{close}_i - p|/p < \beta$ 且捕捉到上涨特征和下跌特征交替出现的组合方式，每个特征 f_k 的能量值 E_k 和 S_3 阶段该特征的平均能量值 $\overline{E}_{3k}$ 满足 $|E_k - \overline{E}_{3k}| < \alpha_3$，则确定当前阶段为 S_3 阶段。

d)继续捕捉 K 线特征，若先连续捕捉到上涨特征的组合，且每个上涨特征 f_k 的能量值 E_k 和 S_4 阶段该特征的平均能量值 $\overline{E}_{4k}$ 满足 $|E_k - \overline{E}_{4k}| < \alpha_4$，接着连续捕捉到下跌特征的组合，且每个下跌特征 f_j 的能量值 E_j 和 S_4 阶段该特征的平均能量值 $\overline{E}_{4j}$ 满足 $|E_j - \overline{E}_{4j}| < \alpha_4$，则判断当前阶段为 S_4 阶段。

3.2 当前阶段 K 线特征的能量确定未来股价涨跌

在不同阶段，股价的上涨速度有较大的不同，根据特征的能量和组合方式确定当前股价所处的阶段 S_i 之后，为了确定该阶段未来的波动范围，需要结合每个阶段的平均涨跌速度 $\bar{v}_i$ 和该阶段特征能量波动大小来判断该阶段未来的涨跌速度，根据 v'_i 计算未来 n 日的涨跌范围$[u,v]$，并将其作为约束条件加入 SVM，预测未来的股价。

通过计算某阶段 K 线特征的能量方差来判断该阶段能量波动的平稳性，并根据能量的波动平稳性确定未来的涨跌幅度。每个阶段特征能量平稳性检测公式如下：

$$\sigma_E = \frac{\sum_{j=1}^{n}(E_{ij} - \overline{E}_{ij})^2}{n} \tag{19}$$

其中：σ_E 表示一个阶段 K 线特征能量的方差，E_{ij} 表示 S_i 阶段 K 线特征 f_j 的能量，$\overline{E}_{ij}$ 表示 S_i 阶段 K 线特征 f_j 的能量均值。

若 $\sigma_E \leqslant \theta_i$($\theta_i$ 为 S_i 阶段的平稳性阈值)，则说明当前阶段 K 线特征的能量值波动平稳，后续股价的涨跌速度按照该阶段平均涨跌速度 $\bar{v}_i$ 计算，即未来涨跌速度 $v'_i = \bar{v}_i$。

若 $\sigma_E > \theta_i$，则说明当前阶段特征的能量值波动较为剧烈，后期的涨跌速度会与该阶段平均的涨跌速度有所差别，若当前特征的能量值 $E_i > \overline{E}_i$，则未来涨跌速度 $v'_i = \bar{v}_i + \partial_i \cdot \sigma_E$；若当前特征的能量值 $E_i < \overline{E}_i$，则未来涨跌速度 $v'_i = \bar{v}_i - \partial_i \sigma_E$。

通过计算 2000 年 1 月 1 日以来的 36 个案例，得出不同阶段参数 ∂_i 和 θ_i 的实验误差 RMSE 的关系分别如图 7 和 8 所示。每个阶段的平均涨跌速度 $\bar{v}_i$ 值和参数 ∂_i、θ_i 最优值如表 1 所示。

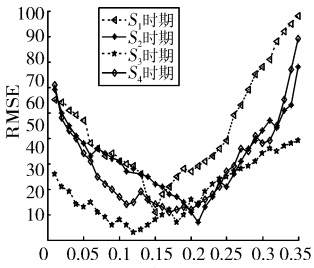

图 7　参数∂_i 与实验误差 RMSE 的关系

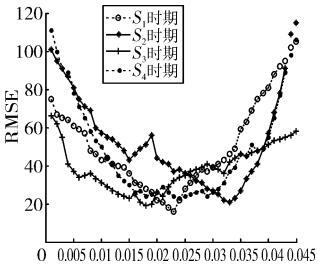

图 8　参数 θ_i 与实验误差 RMSE 的关系

表 1　不同阶段的参数最优值

参数	S_1	S_2	S_3	S_4
$\bar{v}_i$	0.98%	1.21%	0	-1.12%
∂_i	0.15	0.21	0.12	0.17
θ_i	0.023	0.033	0.018	0.025

使用收盘价、开盘价、最高价、最低价、MACD、KDJ 等指标通过 SVM 建立基础的对数收益率预测模型,再根据当前阶段的特征的能量和特征的组合方式判断当前所处的阶段,进而判断未来 n 日的波动范围$[u,v]$,并作为先验知识加入 SVM 模型,得出 LPF-SVM 算法对收盘价的预测结果。

3.3　LPF-SVM 算法描述

输入:股票的价格数据集 dataset。

输出:股票的收盘价 close 的预测结果。

a)将 dataset 中的每日开盘价 open、收盘价 close、最高价 max、最低价 min、MACD、KDJ 等技术指标作为输入变量,通过 SVM 建立基本的股票收盘价 close 预测模型。

b)用滑动窗口捕捉上证指数和深证成指数据集上的 K 线特征 f_i,计算 K 线特征的孕育成熟度 b_i 和爆发力 γ_i,进而得出每个 K 线特征的能量 E_i;若未捕捉到 K 线特征则用传统 SVM 根据 dataset 数据集预测收盘价 close。

c)根据当前 K 线特征的能量 E_i 和特征的组合方式判断当前的股价所处的阶段,再根据当前阶段特征的能量波动情况判断未来股价的涨跌速度 v'_i。

d)通过未来股价的涨跌速度 v'_i 判断未来股价的波动范围$[u,v]$,将波动范围$[u,v]$作为先验知识加入 SVM,重复步骤 a)预测收盘价 close。

4　实验数据和实验结果

实验在 MATLAB 2010 环境下,使用 LIBSVM-2.91 工具包。在上证指数和深证成指上各选取两个数据集作为实验数据,一共 2 132 组数据,其中上证指数选取 1 061 组数据,深证成指选取 1 071 组数据,对每个数据集分别设定训练数据集 train-data 和测试数据集 test-data。

4.1　实验数据集的选取

在上证指数和深证成指上分别选取两个数据集,上证数据集记为 DS-Ⅰ和 DS-Ⅱ,训练集和预测集安排如表 2 所示,深证成指数据集记为 DS-Ⅲ和 DS-Ⅳ,训练集和预测集安排如表 3 所示。在上证指数数据集上预测结果如图 9 和 10 所示,在深证指数数据集上预测结果如图 11 和 12 所示。

表 2　上证指数的训练集和预测集

datasets	train-data	test-data
DS-Ⅰ	2011/1/26 - 2012/11/22	2012/11/23 - 2013/2/27
DS-Ⅱ	2008/8/4 - 2010/6/25	2010/6/28 - 2010/11/17

表 3　深证成指的训练集和预测集

datasets	train-data	test-data
DS-Ⅲ	2011/1/26 - 2012/3/26	2012/3/27 - 2012/5/18
DS-Ⅳ	2010/11/17 - 2012/9/26	2012/9/27 - 2012/12/19

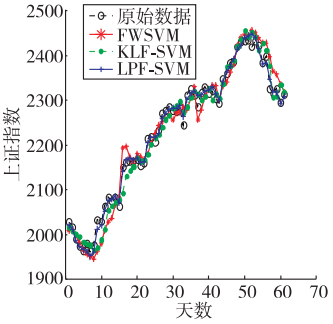

图 9　数据集 DS-Ⅰ上预测结果

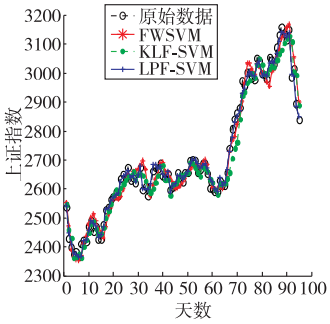

图 10　数据集 DS-Ⅱ上预测结果

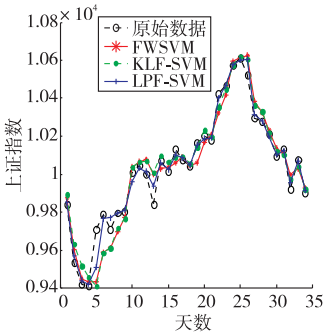

图 11　数据集 DS-Ⅲ上预测结果

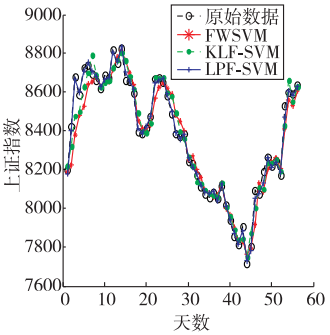

图 12　数据集 DS-Ⅳ上预测结果

4.2　实验结果

在数据集 DS-Ⅰ、DS-Ⅱ、DS-Ⅲ和 DS-Ⅳ上,将 LPF-SVM 算法与 KLF-SVM 算法以及 FWSVM 算法[15]、SVM 算法进行对比分析,使用 RMSE 评价算法性能,RMSE 定义如下:

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{t=1}^{n}(p_t-\hat{p}_t)^2} \tag{20}$$

其中:p_t 表示第 t 日的原始股价,$\hat{p}_t$ 表示第 t 日的股价预测结果,n 表示预测的天数。实验结果误差对比如表 4 所示。

表 4　上证和深证预测集上的实验误差

算法	上证指数(RMSE)		深证成指(RMSE)	
	DS-Ⅰ	DS-Ⅱ	DS-Ⅲ	DS-Ⅳ
SVM	41.352	47.781	56.231	45.316
FWSVM	28.464	36.415	41.559	38.461
KLF-SVM	21.068	28.263	29.397	32.162
LPF-SVM	18.119	22.317	23.913	25.378

4.3　实验结果对比分析

通过表 4 中的实验结果误差对比可以发现,LPF-SVM 算法的实验误差小于 KLF-SVM、FWSVM 以及 SVM。四种算法模型中,FWSVM 通过调节数据集中不同变量的权重使得预测效果比 SVM 算法更好;KLF-SVM 算法则通过加入 K 线特征的孕育成熟度和爆发力等先验知识作为预测结果的约束条件,使得预测结果更加精确,误差较 FWSVM 更小;LPF-SVM 算法则是在 KLF-SVM 算法的基础上,加入 K 线特征的能量以及股市生命期中的不同时期作为约束条件,通过调整不同时期的涨跌速度,使得在股市生命期较为规律的情况下,预测结果的约束条件更为严格,最终的预测结果误差比其余三种算法更小。

(下转第 1647 页)

[2] Nothman J, Ringland N, Radford W, *et al*. Learning multilingual named entity recognition from Wikipedia[J]. Artificial Intelligence, 2013, 194(1):151-175.

[3] Dong Xin, Halevy A Y, Madhavan J. Reference reconciliation in complex information spaces[C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2005: 85-96.

[4] Kim S, Toutanova K, Yu H. Multilingual named entity recognition using parallel data and metadata from Wikipedia[C]//Proc of the 50th Annual Meeting of the Association for Computational Linguistics. 2012:694-702.

[5] Naughton M, Stokes N, Carthy J. Sentence-level event classification in unstructured texts[J]. Information Retrieval, 2010, 13(2): 132-156.

[6] Fillmore C J. Frames and the semantics of understanding[J]. Quademi di Semantica, 1986, VI:222-253.

[7] 张传言, 洪晓光, 彭朝晖, 等. 基于 SVM 和扩展条件随机场的 Web 实体活动抽取[J]. 软件学报, 2012,23(10):2612-2627.

[8] Balasubramanyan R, Cohen W W. Block-LDA: jointly modeling entity-annotated text and entity-entity links[C]//Proc of the 11th SIAM International Conference on Data Mining. 2011:450-461.

[9] Monge A, Elkan C. The field matching problem: algorithms and applications[C]//Proc of the 2nd International Conference on Knowledge Discovery and Data Mining. 1996:267-270.

[10] Chandel A, Nagesh R, Sarawagi S. Efficient batch Top-*k* search for dictionary-based entity recognition[C]//Proc of the 22nd International Conference on Data Engineering. 2006: 28.

[11] 张永新, 李庆忠, 彭朝晖. 基于 Markov 逻辑网的两阶段数据冲突解决方法[J]. 计算机学报, 2012, 35(1):101-111.

[12] Zheng Lvexing, Lyu Xueqiang, Liu Kun, *et al*. Recognition of Chinese personal names based on CRFs and law of names[C]//Proc of International Workshop on Web Technologies and Applications. 2012: 163-170.

[13] Seker G A, Eryigit G. Initial explorations on using CRFs for Turkish named entity recognition[C]//Proc of the 24th International Conference on Computational Linguistics. 2012: 2459-2474.

[14] Sarawagi S, Cohen W W. Semi-Markov conditional random fields for information extraction[C]//Advances in Neural Information Processing Systems. 2004.

[15] Yang Bishan, Cardie C. Extracting opinion expressions with semi-Markov conditional random fields[C]//Proc of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012:1335-1345.

[16] Cohen W W, Sarawagi S. Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods[C]//Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2004: 89-98.

[17] 朱毅华,侯汉清,沙印亭.计算机识别汉语同义此的两种算法比较和测评[J].中国图书馆学报,2002,28(4):28-31.

[18] Crocetti G. Textual spatial cosine similarity[C]//Proc of Student-Faculty Research Day. 2014.

[19] Han Jiawei. Data mining: concepts and techniques[M]. 2nd ed. Beijing: China Machine Press, 2006:408-418.

(上接第 1641 页)

5　结束语

本文提出了一种基于股市生命期不同阶段的 K 线特征算法(LPF-SVM),根据不同 K 线特征对于股价走势的影响,将常见的 K 线特征分为上涨特征和下跌特征。利用 K 线特征产生前的环境对特征的孕育是否充分建立了 K 线特征的孕育成熟度模型,根据 K 线特征产生时的涨跌幅、成交量、资金流入流出情况建立 K 线特征的爆发力模型,融合孕育成熟度和爆发力组成 KLF-SVM 算法模型。为了从整体上预测股价走势,在 KLF-SVM 算法基础上提出股市生命期和生命期中不同阶段的概念,并综合 KLF-SVM 算法中 K 线特征的孕育成熟度和爆发力得出 K 线特征的能量,根据 K 线特征的能量和 K 线特征的组合方式判断当前所处的阶段,将该阶段的价格趋势作为先验知识加入 SVM 预测股价。在上证指数和深证成指数据集上与 SVM、FWSVM 和 KLF-SVM 算法实验效果作对比,验证了算法的有效性。如何优化 K 线特征组合方式以及分析不同阶段间的因果关系是下一步研究的课题。

参考文献:

[1] 王维国, 王霞. 基于贝叶斯推断的上证指数突变点研究[J]. 中国管理科学, 2009, 17(3): 8-16.

[2] Kazem A, Sharifi E, Hussain F K, *et al*. Support vector regression with chaos-based firefly algorithm for stock market price forecasting [J]. Applied Soft Computing, 2013, 13(2): 947-958.

[3] Engle R F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation[J]. Econometrica: Journal of the Econometric Society, 1982,50(4): 987-1007.

[4] Bollerslev T. Generalized autoregressive conditional heteroskedasticity [J]. Journal of Econometrics, 1986, 31(3): 307-327.

[5] Wang Jujie, Wang Jianzhou, Zhang Z G, *et al*. Stock index forecasting based on a hybrid model[J]. Omega, 2012, 40(6): 758-766.

[6] Hung J C. Applying a combined fuzzy systems and GARCH model to adaptively forecast stock market volatility[J]. Applied Soft Computing, 2011,11(5): 3938-3945.

[7] Hill T, O'conner M, Remus W. Neural network models for time series forecasts[J]. Management Science, 1996, 42(7): 1082-1092.

[8] Gheyas I A, Smith L S. A neural network approach to time series forecasting[C]//Proc of World Congress on Engineering. 2009:1-3.

[9] Kao Lingjing,Chiu Chihchou, Lu Chijie, *et al*. Integration of nonlinear independent component analysis and support vector regression for stock price forecasting[J]. Neurocomputing, 2013, 99(1):534-542.

[10] Vapnik V, Levin E, Le Cun Y. Measuring the VC-dimension of a learning machine[J]. Neural Computation, 1994, 6(5): 851-876.

[11] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[12] Zhang Ling, Zhang Bo. Relationship between support vector set and kernel functions in SVM[J]. Journal of Computer Science and Technology, 2002, 17(5): 549-555.

[13] Zhang Tao, Sai Ying, Yuan Zheng. Research of stock index futures prediction model based on rough set and support vector machine [C]//Proc of IEEE International Conference on Granular Computing. 2008: 797-800.

[14] Mansukhbhai P A, Patel J M. Role of soft computing techniques in predicting stock market direction [J]. Artificial Intelligence Research, 2012, 1(2): 198.

[15] 汪廷华, 田盛丰, 黄厚宽. 特征加权支持向量机[J]. 电子与信息学报, 2009, 31(3): 514-518.