

## 深度报告

## 金融工程

## 量化投资

## 数据挖掘专题报告

2010年05月20日

本报告的独到之处

■支持向量机改变了经验风险最小化原则，优于神经网络、决策树等传统数据挖掘方法。

■首次使用支持向量机对A股市场的个股进行价格预测方面的应用说明。

## 专题报告

## 支持向量机在股票价格预测方面的应用

支持向量机(support vector machine, SVM)是数据挖掘中的一项新技术,是借助于最优化方法解决机器学习问题的新工具。它成为克服“维数灾难”和“过学习”等传统困难的有效办法,虽然他还处在飞速发展的阶段,但它的理论基础和实现途径的基本框架已经形成。支持向量机目前主要用来解决分类问题(模式识别,判别分析)和回归问题。而股市行为预测通常为预测股市数据的走势和预测股市数据的未来数值。而当我们把走势看作两种状态(涨、跌),问题便转化为分类问题,而预测股市未来的价格是指为典型的回归问题。我们有理由相信支持向量机可以对股市进行预测。

本报告是支持向量机对股票价格预测应用报告的综述,旨在介绍预测股票价格走势的SVM简单预测模型。该模型可以用来预测未来若干天股票价格的大体走势,这对于股票投资可以起到很好的指导性作用。

## 实证结果表明:

1、SVM简单预测模型在当所预测的未来若干天的股票价格趋于平稳上升或者下降状态时,该模型预测的效果最佳,用作预测股票价格的大体走势可信度较高。

2、然而在股票价格突然快速上涨或者下跌时,模型的预测往往不能立即跟上真实行情的变化速度,导致模拟走势与实际情况存在差异。

3、实证例子还表明在牛市和熊市中,预测效果比震荡市中的效果更为准确。原因是震荡市中所要预测的实际股票价格时涨时跌,预测的曲线存在很多的急转弯,影响了其正确性。

但一般来讲,从长远来看(50交易日),该模型还是能够描述出实际股票价格的大体走势,只是如果当预测走势的趋势不确定时,说明实际股票数据可能变化较小或震荡可能性较大,因此可以考虑不对该股票进行投资或者抛出该股票。

在应用SVM进行股票预测时,如何有效地选取输入向量的分量是决定预测模型准确性的关键。样本向量的各个分量应该选取能充分反应股票市场交易特征的定量指标,不加选择则会增加期望误差的上界。反之,选取指标过少难以刻画股票市场的特点。

分析师: 黄志文

电话: 0755-82130833-6210

E-mail: huangzw@guosen.com.cn

SAC 执业证书编号: S0980206110185

分析师: 葛新元

电话: 0755-82130833-1870

E-mail: gexy@guosen.com.cn

SAC 执业证书编号: S0980200010107

## 独立性声明:

作者保证报告所采用的数据均来自合规渠道,分析逻辑基于本人的职业理解,通过合理判断并得出结论,力求客观、公正,结论不受任何第三方的授意、影响,特此声明。

## 内容目录

股市预测的发展概况 .....	4
数据挖掘模型的优点 .....	4
数据挖掘的客观性和实用性 .....	4
数据挖掘模型 .....	5
支持向量机介绍 .....	5
概述 .....	5
支持向量机的分类算法 .....	6
支持向量机在股票市场的应用 .....	7
基于支持向量机的简单预测模型 .....	8
研究对象 .....	8
原始数据的标准化 .....	8
核函数的选择 .....	9
输入向量的选取方法 .....	9
参数的选择 .....	9
股票价格趋势预测方法 .....	9
SVM 模型的实际运用 .....	10
实证例子分析 .....	11
模型的改进方案和后续拓展 .....	15
SVM 简单预测模型存在的问题 .....	15
改进方案 .....	16
后续拓展方向 .....	16

## 图表目录

图 1: 线性可分问题 .....	6
图 2: 近似线性可分问题 .....	6
图 3: 线性不可分问题 .....	6
图 4: 浦发银行的预测走势 1, 07-8-6 至 07-8-31 (20 交易日) .....	12
图 5: 浦发银行的预测走势 2, 07-6-27 至 07-9-4 (50 交易日) .....	12
图 6: 浦发银行的预测走势 3, 08-7-10 至 08-9-18 (50 交易日) .....	13
图 7: 浦发银行的预测走势 4, 10-2-1 至 10-4-27 (45 交易日) .....	13
图 8: 常山股份的预测走势, 10-1-14 至 10-3-31 (50 交易日) .....	14
图 9: 中信证券的预测走势, 10-2-24 至 10-5-6 (50 交易日) .....	14
表 1: 实证结果统计 .....	15

## 股市预测的发展概况

时至今日，在金融经济学的发展上人们对金融预测作了大量的探索，取得了丰硕的成果。典型的金融预测是时间序列预测。时间序列的典型特征是相邻观测之间的依赖性。为了研究这种依赖性，人们提出了许多时间序列模型，并对这模型的性质及分析方法进行了深入的研究。传统的金融时间序列大致上有两种研究方法，一种方法是从基本的经济原理出发建立金融时间序列服从的数学模型，像 Markovitz 的投资组合理论，资本资产定价模型（CAPM）、套利定价理论（APT）、期权定价模型等。实际上，这部分成果就是确定金融时间序列的趋势项。这些理论，理论上很成功（有三项获得诺贝尔经济学奖），但它们都是建立在很理想的假设上，而这些假设与市场的实际情况有很大差距，所以这些理论在实际效果中并不理想。另一种方法是从统计角度对金融时间序列进行研究。这种方法直接从实际数据出发，应用概率统计推断出市场未来的变化规律。虽然这种方法从经济学角度来讲缺乏理论性，但是在实际应用中效果较好。而且，统计方法还可以对经济模型的好坏进行检验和评价。二十世纪 80 年代以前，人们对时间序列的研究主要集中在一种线性模型，即自回归移动平均模型（AutoRegressive Moving Average Models, ARMA），这种模型结构简单，有着完善的统计推断技术，应用非常广泛。但是 ARMA 模型毕竟是一种线性模型，有些实际现象在模型中得不到反映。在这种情况下人们开始提出并研究非线性时间序列，最重要的就是 R.F. Engle 在八十年代初提出的自回归条件异方差模型（AutoRegressive Conditionally Heteroscedastic models, ARCH），由于 ARCH 模型将方差看作随时间变动的量，而不是一个常量，从某种程度上克服了线性模型的局限性。与实际情况更相符，从而得到了广泛的应用。

股市预测，是金融经济预测的一个重要分支。它对股票交易所反映的各种资讯进行收集、整理、综合等工作，从股市的历史、现状和规律性出发，运用科学的方法，对股市未来发展前景进行测定。

股市预测一般基于以下三点假设：

(1) 有效市场假设：指股票市场会对每一条有可能影响股价的信息都会作出反映，而各种价格的变动正是这种反映的结果。

(2) 供求决定假设：指一切信息都会对股票市场的供求双方力量对比产生影响，供求决定交易量和交易价格。

(3) 历史相似原则：指由历史资料所概括出来的规律已经包含了未来股票市场的一切变动趋势。

股市预测按不同的标准可以有不同的分类。按涉及的范围不同可分为：指数预测和个股预测；按预测时间长短不同可分为：长期预测、中期预测和短期预测；按预测方法的不同可分为：定性预测和定量预测等等。

## 数据挖掘模型的优点

数据挖掘是通过自动或半自动化的工具对大量的数据进行探索和分析的过程，其目的是发现其中有意义的模式和规律。采用数据挖掘建立的模型具有一系列优点。

## 数据挖掘的客观性和实用性

由于数据挖掘是在经济金融理论指导下基于现实数据的实证方法，相对于完全基于经济金融理论的方法，它的客观性和实用性更强。一般基于估值理论的选股模型，其未来现金流的确定是一大难点，对其的估计涉及到很多主观判断及财务报表之外的信息，这就使得不同的研究员在同一时期对于同一家公司的判断有可能大相径庭。同时，由于估值方法需要研究员对于各种信息深入的研究，所以每个研究员所能研究的股票是非常有限的，因此为了取得更好的研究效果，在这种深入研究之前对于海量的信息进行客观实证的挖掘就非常有必要。数据挖掘的选股模型一定程度上弥补了估值研究的不足，能与之起到互补互助的作用。

## 数据挖掘模型

数据挖掘的核心是建立预测模型，换一句说法就是找出数据中的规律，通过这些规律来对数据集中一个或多个变量值进行估测的过程。不同的模型适用于不同的研究领域，模型的种类包括了：Logistic 回归、决策树（Decision Trees）模型、人工神经网络（Artificial Neural Networks）、小波分析和支持向量机等。在数据挖掘系列报告的这一篇里，我们重点介绍的是基于支持向量机在股票价格预测方面的应用。

## 支持向量机介绍

### 概述

较常用的时间序列分析方法主要是基于线性模型的自回归移动平均模型，但现实中多数数据都是非线性的且含有复杂的噪声数据，因而这一模型就会表现出较大的不适应性，产生较大的误差。目前金融时间序列的预测方法如神经网络，在非线性逼近函数的能力很强，但是其容易陷入局部最小，网络的层数和每层节点的个数依据经验来设定，其优化目标是经验风险最小化，不能达到泛化的能力，这种方法基于经验结构风险最小化，无法实现期望风险最小，而且经验最小也不一定意味着实际风险也最小，很难保证学习的精度。支持向量机最大的特点就是改变了传统神经网络中经验风险最小化原则，针对于结构风险最小化原则，具有良好的泛化性，改变了神经网络中局部最小的缺点，通过最优化理论得到全局最优解，这对股市进行准确分析预测是非常必要的。

支持向量机（Support Vector Machine, SVM）最初于 20 世纪 90 年代由 Vapnik 提出，近些年来在其理论研究和算法实现方面都取得了突破性进展，并开始成为克服“维数灾难”和“过学习”等传统困难的有效办法，虽然他还处在飞速发展的阶段，但它的理论基础和实现途径的基本框架已经形成。**支持向量机的最大特点**是改变了传统的经验风险最小化原则，而是针对结构风险最小化原则提出的，因此具有很好的泛化能力。另外，支持向量机在处理非线性问题时，首先将非线性问题转化为高维空间中的线性问题，然后用一个核函数来代替高维空间中的内积运算，从而巧妙地解决了复杂计算问题，并且有效地克服了维数灾难以及局部极小问题。

尽管支持向量机有以上的优点，但是在金融时间序列预测这方面研究还很少。Mukherjee et al. 证明支持向量机可以应用于金融时间序列分析。Tay 和 Cao 在 2002 年时证明了 5 种金融时间序列数据是可以支持向量机进行预测，并指出，支持向量机在标准均方误差，均方绝对误差，趋势正确率，加权趋势正确率标准下优于人工神经网络。Kyoung-jae Kim 用支持向量机对股市指数的运动趋势进行预测。

支持向量机目前主要用来解决分类问题（模式识别，判别分析）和回归问题。而股市行为预测通常为预测股市数据的走势和预测股市数据的未来数值。而当我们走势看作两种状态（涨、跌），问题便转化为分类问题，而预测股市未来的价格是指为典型的回归问题。我们有理由相信支持向量机可以对股市进行预测。

## 支持向量机的分类算法

一般来说，预测股市的行为通常为预测股市数据的走势和预测股市数据的未来数值。而当我们讲走势看作两种状态（涨、跌），问题便转化为分类的问题，一类是涨一类是跌。而预测股市未来的价格趋势是典型的回归问题，由于在本报告的实证中并没有运用到回归算法，在此就不给予详细说明。

### 1、什么是分类问题

分类问题一般可以表示为，考虑  $n$  维空间上的分类问题，它包含  $n$  个指标（即  $x \in R^n$ ）和  $l$  个样本点。记这  $l$  个样本点的集合为

$$T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l, \quad (1-1)$$

其中  $x_i \in X = R^n$  是输入指标向量，或称输入，其分量称为特征，或输入指标； $y_i \in Y = \{1, -1\}$  是输出指标，或称输出， $i=1, \dots, l$ 。这  $l$  个样本点组成的集合称为训练集。这时我们的问题是，对任意给定的一个新的模式  $x$ ，根据训练集，推断它所对应的输出  $y$  是 1 还是 -1。

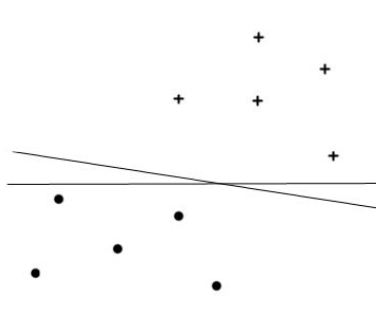
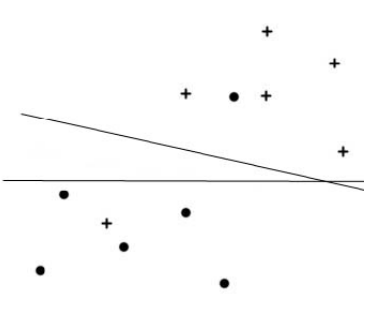
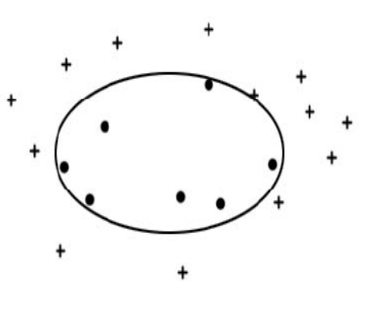
我们可以用数学语言把分类问题描述如下：

根据给定的训练集  $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$ ，其中  $x_i \in X = R^n$ ， $y_i \in Y = \{1, -1\}$   $i=1, \dots, l$ ，寻找  $X = R^n$  上的一个实际值函数  $g(x)$ ，以使用决策函数

$$f(x) = \text{sign}(g(x)) \quad (1-2)$$

推断任一模式  $x$  对应的  $y$  值。由此可见，求解分类问题，实质上就是找一个把  $R^n$  上的点分成两部分的规则。确切的说，上述分类问题是分成两类的问题。与此类似的还有分成多类的分类问题，它们的不同之处仅仅是前者输出只有两个值，而后者可能有多个值。后面我们研究股票价格变化均指分成两类的分类问题。

按照机器学习领域的术语，我们把解决上述分类问题的方法称为分类学习机。分类问题大体有三种类型，对于不同类型的问题，可能需要采用不同的分类学习机。下面三幅图分别诠释了线性可分问题，近似线性可分问题和线性不可分问题。本报告所关注的就是最后一种类型，即线性不可分问题。

图 1：线性可分问题	图 2：近似线性可分问题	图 3：线性不可分问题
		
资料来源：国信证券经济研究所	资料来源：国信证券经济研究所	资料来源：国信证券经济研究所



## 2、支持向量机解决非线性分类问题

支持向量机分类的目标是开发计算有效途径，从而能在高维特征空间中学习“好”的分类超平面。支持向量机的研究最初是针对模式识别中的二类线性可分问题提出来的。由于股市的数据是非线性的，超平面的分类能力毕竟有限，因而 SVM 对数据进行非线性映射，通过映射  $\phi: x \rightarrow f$ ，将数据映射到一个更高维的特征空间  $F$  中，从而使数据线性可分，然后再  $f$  中构造最优超平面。由于优化函数和分类函数都涉及样本空间的内积运算  $\langle x_i \cdot x_j \rangle$ ，因此在变换后的高维特征空间  $E$  中需进行内积运算  $\langle \phi(x_i) \cdot \phi(x_j) \rangle$ ，根据满足 Mercer 定理，对应线性变换空间中的内积， $\langle \phi(x_i) \cdot \phi(x_j) \rangle = k(x_i, x_j)$ 。采用适当的核函数  $k(x_i, x_j)$ ，就能代替向高维空间中的非线性映射，实现非线性变换后的线性分类。据最优化理论，得到的优化问题

$$\begin{aligned} \text{maximise } w(\alpha) &= \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j \text{Ker}(x_i, x_j) \\ \text{subject to } \sum_{i=1}^l y_i \alpha_i &= 0 \quad c \geq \alpha_i \geq 0 \quad i=1, \dots, l \end{aligned} \quad (2-1)$$

式中： $x_i \in R^n$ ， $y_i \in R$ ； $x_i$  是输入， $y_i$  是  $x_i$  对应的输出值； $l$  是样本个数， $\alpha$  是拉格朗日系数； $c$  是正则化参数，实现最大间隔和分类错误的折中。这里令  $f(x) = \alpha_i^* k(x_i, x_j) + b^*$ ，选择  $b^*$  使得  $y_i f(x_i) = 1$  成立，其中对任意  $i$  有  $c > \alpha_i^* > 0$  决策规则由  $\text{sgn}(f(x))$  给出，它等价于解决优化问题的核  $k(x, z)$  隐式定义的特征空间中的超平面，这里松弛变量的定义与几何间隔相关

$$\gamma = \left( \sum_{i,j \in \text{sv}} y_i y_j \alpha_i^* \alpha_j^* \text{Ker}(x_i, x_j) \right)^{-\frac{1}{2}} \quad (2-2)$$

相应的决策函数为

$$g(x) = \text{sign} \sum_{i=1}^l y_i \alpha_i \text{Ker}(x_i, x_j) + b \quad (2-3)$$

由优化函数 (2-1) 可以看出，SVM 的方法训练复杂度与维数无关。

## 支持向量机在股票市场的应用

自 2000 年以来，国内外对支持向量机的研究不断地增加，随着它为越来越多的人所熟知，许多相关领域的工作者都试图采用该方法解决一些实际问题，计算机领域首当其冲。在证券领域，也有很多人对于支持向量机在该领域的应用做出了很多贡献。

通过对前人工作的参考和总结，我们发现历史上在进行这方面研究时主要需进行如下几个步骤，即：

- 股票指标的选取和原始数据的处理；
- 数据的标准化；
- 核函数及其参数的选取；

#### ➤ 计算结果的分析与比较。

从股票指标的选择来看，大多采用三种指标选择法，一种是采用历史上在股票市场上经常被使用的技术指标，另一种是直接利用股票的基本指标或其简单组合，第三种就是将前两者混合起来使用。没有什么科学的理由说明哪种指标的选择是最好的，因为即便是股票的技术指标也是从基本指标推导得出的，所以在 SVM 的应用中没有能够说明前者比后者更能提高预测精度的理论根据。

## 基于支持向量机的简单预测模型

结构模型和时间序列模型是股市预测模型的两大分类。采用结构模式主要有基本面数据方法和金融资产法，前者是以影响股市的宏观经济指标或企业财务指标为影响参数。我们知道，股票的价格关键是由公司经营状况和财务状况决定，并受到宏观经济状况，制度政策等因素影响。这些因素非常适合对股票中长期预测。同时，由于我国股市发展时间较短，数据量不是很丰富，加之宏观数据，和财务数据发布的时间间隔大，最短的也需要每月公布一次，而可用月度数据非常有限，直接通过基本面数据进行预测数据不够充分，意义不是很大。后者，根据股票和各种相关的金融资产的协同运动规律来进行预测。例如，应用这种模型，通常要有完善的金融市场，需要有良好的金融资产价格形成机制，使各种资产的风险价格化。在我国，资本市场和货币市场还很不成熟，还不能应用这种模型。还有一种是通过技术指标进行预测，由于大多数技术指标是通过日 K 线数据计算形成的，而且可用的日 K 线数据的数据量很充分，足以进行预测和检验。

如果用支持向量机的方法来分析股票，指标的选取将会更加灵活方便。本文将提出几种用支持向量机分析股票的方法，在参数的选择方面将遵循以下原则：

- 参数的选择要包含尽量多的特征，但所选特征必须充分的代表问题。
- 选择的参数要尽可能多的用到已知的信息量。
- 对于预测而言，选择的参数要尽量用到最新的数据。

## 研究对象

本文的研究对象是股市中的个别股票（个股），本报告将选取国内一些有代表性，有一定知名度的上市股票的数据。一般来讲，相对于大盘指数而言，个股的波动“随机性”较强，因此本文不打算选取一个通用的解决方案来分析所有的个股股票，也没有打算为每一个选择的股票提供一个固定的解决方案，而只是提出一套可行的方法，在实际的操作中，至于参数的选取，数据指标的选取等都将根据实际情况灵活进行。

对于个股的研究如果能够取得一定的成果，将会对投资起到非常大的指导性作用，但是个股自身的特点也给研究带来了很大的困难，通过本文的研究，希望能够得到一套解决方案，使该方案能够适用于一些典型的个股股票，也为将来的进一步研究打下一定的基础。

## 原始数据的标准化

一个  $m$  维的样本数据里，各个分量来自不同的领域，数值的取值范围各不相同，因此不可避免的情况是，不同的分量的数据处在不同的数量级上，数量级大的分量对模型的影响就较大，而数量级小的分量对模型的影响就小，小数“淹没”在



了大数里而失去了意义，这相当于在构造模型之前便人为的给不同指标加上了权重，甚至是丢失了一些信息，这是十分不科学的；另外，由于计算机所能表示的数据精度有限，如果个别数据过大或者过小，也容易使计算后的数据越界而丢失信息。基于以上考虑，所选择的原始数据需要进行一定的处理，将数据标准化是在本文将要采用的策略。

本文采用的是统计标准化，即：

$$x'_i = \frac{x_i - \bar{x}}{\sqrt{\text{var}(x)}}, i = 1, 2, \dots, l,$$

其中  $\bar{x}$  和  $\sqrt{\text{var}(x)}$  分别为变量  $x$  的平均值和标准差。

## 核函数的选择

我们参考了国内外关于支持向量机预测金融市场的报告，在以往对股票数据研究中一般采用的是 Gauss 核函数。在本报告的初始系列里，我们暂且不为核函数的选择进行详细的讨论，而是按早前人的经验采用 Gauss 核函数。

## 输入向量的选取方法

在应用 SVM 进行股票预测时，如何有效地选取输入向量的分量是决定预测模型准确性的关键。样本向量的各个分量应该选取能充分反应股票市场交易特征的定量指标，不加选择则会增加期望误差的上界。反之，选取指标过少难以刻画股票市场的特点。

模型里所用到的输入向量是由 7 个股票基础指标计算得出，这 7 个指标分别是：前收盘价、开盘价、最高价、最低价、均价、成交量、成交额、涨跌、涨跌幅、换手率、收盘价。选取上述的指标去构造出 8 个我们认为对股价具有较强相关性的输入向量：

- VolumeIndex :- 今日总手与过去M天平均总手的比值；
- CloseIndex :- 今日收盘价与过去M天平均收盘价的比值；
- TradeAmountIndex :- 今日交易额与过去M天平均交易额的比值；
- TurnoverIndex :- 今日换手率与过去M天平均换手率的比值；
- MeanIndex :- 过去M天的平均股价；
- RateChangeIndex :- 过去M天的涨跌幅；
- HighIndex :- 过去M天当中的最高股价；
- LowIndex :- 过去M天当中的最低股价。

M 为预测天数。

## 参数的选择

在我们的 SVM 简单预测模型里，核函数采用的是市场公认的效果较好的 Gauss 径向基核函数，因此模型有两个参数需要调试，分别是 Gauss 核函数的参数  $\sigma$  和模型最优化求解时所用的惩罚参数 C。

在我们下面实证例子中的浦发银行实验里，这两个参数经过枚举法调试后确认了最优组合。这样的参数组合在我们测试的不同市场环境下均得到稳定的效果。

## 股票价格趋势预测方法

预测未来一些天的股票价格是增长还是下降的时候,除非具有非常高的正确率,否则对于实际投资并没有太大的指导意义。如果能够描绘出未来一些天股票价格的价格走势或者说是股票价格曲线的大致形状,即便预测出的股票价格与实际股票价格有一定的出入,对于股票投资来说也会有很大的指导意义。因此,本文的目标就是寻找一种方法来得到上述的预测效果。

先来简单说明一下 SVM 预测模型的运作流程:

选取  $l$  个数据作为训练样本,经过推导得到如下公式,我们称它为决策函数。

$$f(x) = \text{sign} \sum_{i=1}^l \alpha_i y_i \text{Ker}(x_i, x) + b \quad (*)$$

若  $x_0$  为其中一个输入参数,将  $x_0$  代入上述公式,如果  $f(x_0)=1$  则认为  $x_0$  属于正类,否则如果  $f(x_0)=-1$  则认为  $x_0$  属于负类。

通过观察我们注意到,假如把上述公式中的正负判断函数  $\text{sign}$  去掉,令

$$g(x) = \sum_{i=1}^l \alpha_i y_i \text{Ker}(x_i, x) + b \quad (**)$$

显然  $g(x)$  比  $f(x)$  含有更多的信息量,在一般的分类问题中,这些信息都被忽略掉了。从模型的理论角度来考虑,  $g(x)$  实际上体现的是预测点与训练所到的超平面的距离。所以当我们所要研究的样本是股票数据,而且预测的目标是股票价格的涨跌幅时,这些信息便可以很好的利用起来。

我们假设  $\bar{x}$  为一个预测用样本(输入参数),用决策函数  $(*)$  可以预测出对应的未来某一天(假设为第  $m$  天)的股票价格是涨还是跌。若使用公式  $(**)$  则可以得到股价变化的幅度。

$$z = g(\bar{x}) = \sum_{i=1}^l \alpha_i y_i \text{Ker}(x_i, \bar{x}) + b$$

假如第  $m$  天的股票价格实际变化为  $y$ ,而  $\hat{y}$  是预测得到的变化值,那么假设存在映射  $\varphi$ ,使得  $z$  与  $\hat{y}$  存在一一对应关系,即:

$$\hat{y} = \varphi(z).$$

这样便可以通过  $\bar{x}$  预测第  $m$  天的股票价格相当于当天的涨跌变化幅度。

## SVM 模型的实际运用

- 1) 利用股票原始数据选取并构造预测指标,标准化指标向量,从而形成向量  $x_i (i = 1, 2, \dots)$ 。
- 2) 定义  $N$  为训练数据的长度,  $M$  为预测数据的长度。
- 3) 令  $y_i = \text{第}(i + M \text{ 天的收盘价}) - \text{第} i \text{ 天的收盘价}$ 。
- 4) 选取  $N(x_1, x_2, \dots, x_N)$  个数据作为训练数据,选取适当的核函数对这  $N$  个数据进行训练,得到  $\alpha_1, \alpha_2, \dots, \alpha_N$  和  $b$ ;
- 5) 计算  $z_i = g(x_i) = \sum_{j=1}^N \alpha_j \bar{y}_j \text{Ker}(x_j, x_i) + b, i = 1, 2, \dots, N$ 。
- 6) 选取这  $N$  个数据随后的  $M(x_{N+1}, x_{N+2}, \dots, x_{N+M})$  个数据作为预测数据,用该  $M$  个数据,计算  $z_i = g(x_i) = \sum_{j=1}^N \alpha_j \bar{y}_j \text{Ker}(x_j, x_i) + b$ , (其中  $\bar{y}_i = \text{sign}(y_i), i = N+1, N+2, \dots, N+M$ )。

7) 令

$$\delta_i =$$

$$\begin{cases} (\text{第}i\text{天的收盘价} + y_i) - \text{第}i + (M - 1)\text{天的收盘价}, i = N + 1 \\ (\text{第}i\text{天的收盘价} + y_i) - (\text{第}i - 1\text{天的收盘价} + y_{i-1}), i = N + 2, \dots, N + M \end{cases}$$

则 $\delta_i(i=N+1, N+2, \dots, N+M)$ 就代表了第 $i+M$ 天相对于第 $i+M-1$ 天的实际涨跌幅度。

8) 令

$$\hat{\delta}_i =$$

$$\begin{cases} (\text{第}i\text{天的收盘价} + \hat{y}_i) - \text{第}i + (M - 1)\text{天的收盘价}, i = N + 1 \\ (\text{第}i\text{天的收盘价} + \hat{y}_i) - (\text{第}i - 1\text{天的收盘价} + \hat{y}_{i-1}), i = N + 2, \dots, N + M \end{cases}$$

则 $\hat{\delta}_i(i=N+1, N+2, \dots, N+M)$ 就代表了第 $i+M$ 天相对于第 $i+M-1$ 天的实际涨跌幅度。

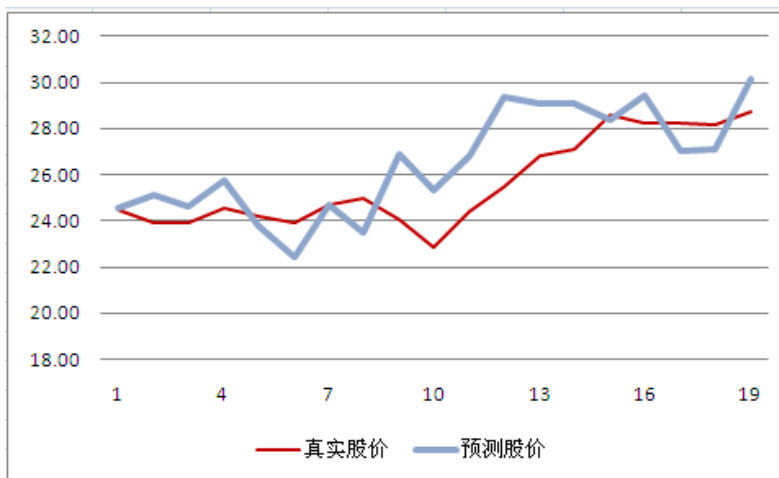
我们将利用上述计算得出的结果来绘制股票价格曲线,通过对比真实与预测行情的走势,直观的来检验模型预测结果是否具有一定的说服力。

## 实证例子分析

我们选择了预测 20 个交易日和 50 个交易日,这样可以检验模型随着时间的推移是否具有好的延续性。并且划分了牛市、熊市和震荡市,分别检验模型在不同市场环境下的预测效果。

下面是利用上述算法绘制的浦发银行预测图形。图 1 为浦发银行 20 个交易日的预测走势与真实走势的对比。

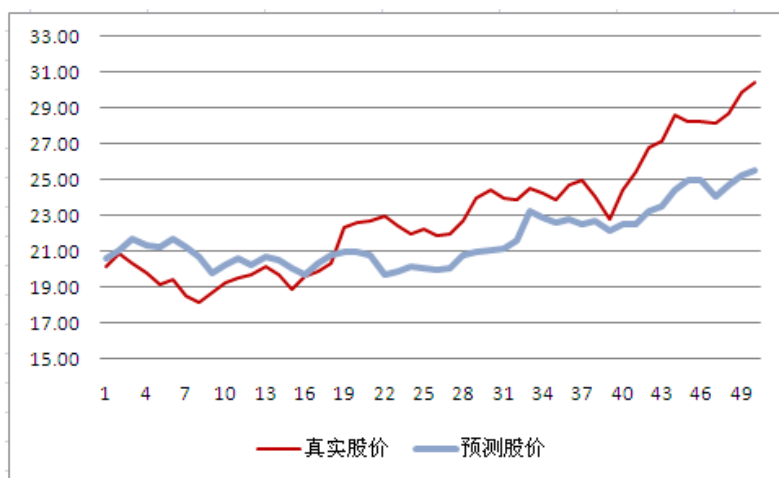
图 4：浦发银行的预测走势 1，07-8-6 至 07-8-31（20 交易日）



资料来源：国信证券经济研究所

图 2 为浦发银行牛市中 50 个交易日的预测走势。

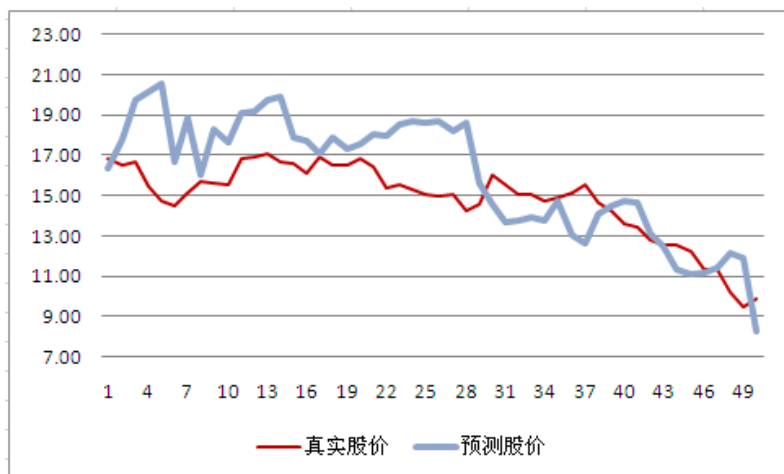
图 5：浦发银行的预测走势 2，07-6-27 至 07-9-4（50 交易日）



资料来源：国信证券经济研究所

图 3 为浦发银行熊市中 50 个交易日的预测走势。

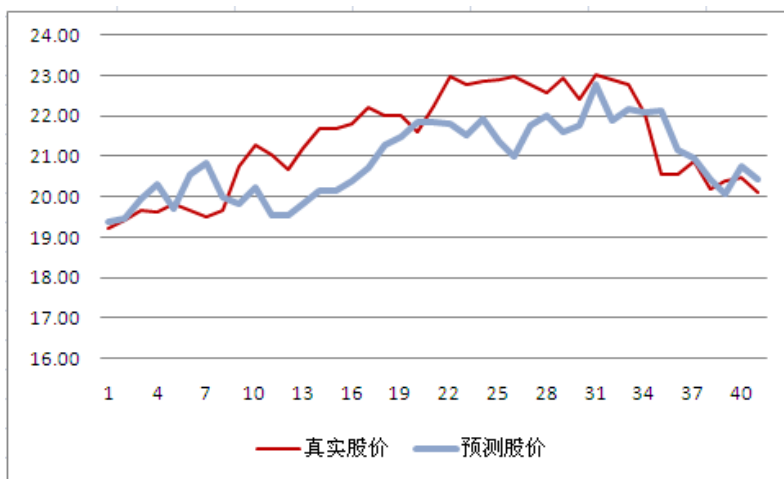
图 6：浦发银行的预测走势 3，08-7-10 至 08-9-18（50 交易日）



资料来源：国信证券经济研究所

下图中我们对 2010 年的其中一个时间段中浦发银行原先的 50 交易日预测做了 5 天的移动平均，绘制了 5 日预测股价的移动平均线和真实行情走势。

图 7：浦发银行的预测走势 4，10-2-1 至 10-4-27（45 交易日）  
5 日平均价

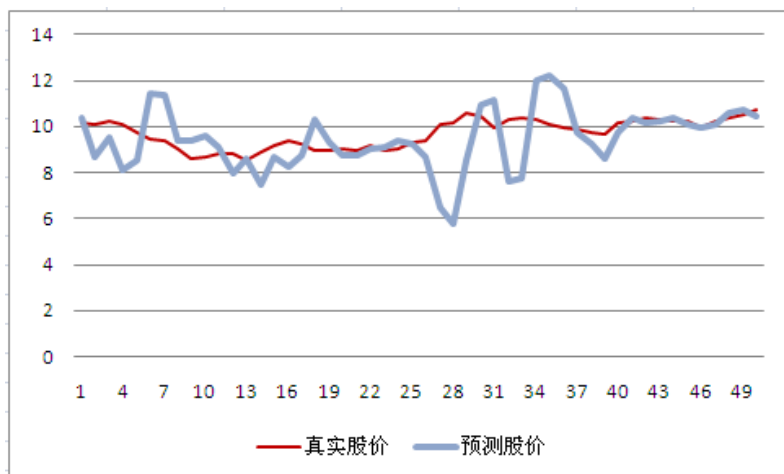


资料来源：国信证券经济研究所



图 5 为常山股份在震荡市中的预测走势。

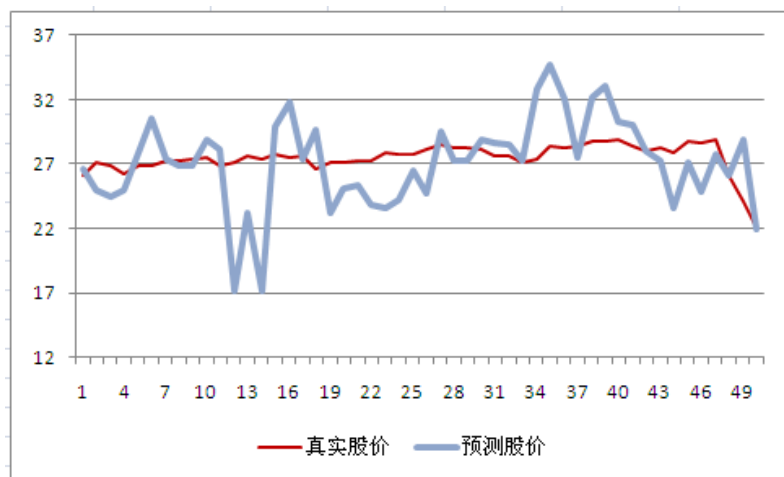
图 8：常山股份的预测走势，10-1-14 至 10-3-31（50 交易日）



资料来源：国信证券经济研究所

图 6 为中信证券在近期震荡市中的预测走势。

图 9：中信证券的预测走势，10-2-24 至 10-5-6（50 交易日）



资料来源：国信证券经济研究所

下表统计了上述案例的结果。相关性是指真实行情与预测行情两个时间序列在测量区间内的价格走势相关性；预测涨跌正确率是指我们预测回报的每一天相对于前一天是涨还是跌的正确比率，这里没有考虑涨跌的幅度，而关心的只是变化的方向。

**表 1：实证结果统计**

实证案例	市场环境	相关性	预测涨跌的正确率	实际回报	预测回报
浦发银行（20 日）	牛市	0.7023	52.6%	22.13%	14.33%
浦发银行（50 日）	牛市	0.8602	69.4%	51.27%	23.40%
浦发银行（50 日）	熊市	0.7643	46.9%	-41.43%	-41.89%
浦发银行（5 日均线）	震荡市	0.7517	50.0%	4.63%	5.34%
常山股份	震荡市	0.2453	49.0%	5.50%	1.08%
中信证券	震荡市	0.3084	63.3%	-15.12%	-17.58%

资料来源：国信证券经济研究所

以上的实证结果表明：

- 1) SVM 简单预测模型在当所预测的未来若干天的股票价格趋于平稳上升或者下降状态时，该模型预测的效果最佳，用作预测股票价格的大体走势可信度较高。
- 2) 然而在股票价格突然快速上涨或者下跌时，模型的预测往往不能立即跟上真实行情的变化速度，导致模拟走势与实际情况存在差异。
- 3) 实证例子还表明在牛市和熊市中，预测效果比震荡市中的效果更为准确。原因是震荡市中所要预测的实际股票价格时涨时跌，预测的曲线存在很多的急转弯，影响了其正确性。

但一般来讲，从长远来看（50 交易日），该模型还是能够描述出实际股票价格的大体走势，只是如果当预测走势的趋势不确定时，说明实际股票数据可能变化较小或震荡可能性较大，因此可以考虑不对该股票进行投资或者抛出该股票。

## 模型的改进方案和后续拓展

### SVM 简单预测模型存在的问题

支持向量机因其广泛的适应能力和学习能力在非线性系统的预测方面有之广泛的应用前景。股票市场预测就是一个非线性的建模问题，但是支持向量机在股市预测中还缺乏系统的研究，本报告在这方面做出了一些尝试，对支持向量机方法进行了进一步的挖掘，并将其应用于股票价格的预测，取得了基本令人满意地的结果。

尽管如此，支持向量机在股票市场预测中还有许多问题值得研究：

1. 对于一些情况预测模型泛化能力差，即拟合的非常好而预测的非常差。造成这一现象可能是训练集规模选择不当，进行短期预测，训练集过大会掩盖短期趋势，训练集过小很难把握运动趋势。

2. 奇异点问题。

由于我国股市发展还不完善，投机性和政策性特征明显，经常出现暴涨暴跌的情况，造成股价运行中出现了很多的奇异点。由于造成奇异点的因素很多而且难以

量化，仅靠支持向量机本身难以解决这个问题。

### 3. 模型的优化。

从应用经验上讲，几种常用的核函数已经足够，但对于股市，未必是最适合的核函数，因此，为实际问题构造适当的核函数也许是最好的选择。

### 4. 输入向量的选择。

股市的数据量非常庞大，各种指标层出不穷，这些数据和指标都有它的实际意义，都反映了一定的股市信息。但是要把这些都作为输入量是不现实的，选择哪些数据作为输入量可以获得最好的模型并没有可靠的结论。

### 5. 选取输入向量所用的时间段。

按照常理来思考，预测未来一天的股票走势，如果能利用越接近的相关指标准确性应该越高。在我们现在的模型里，假设要预测未来 20 天的股价走势，所要用到的数据是今天往前推的 20 天数据。即预测第 20 天的股票涨跌幅度能用到的最近数据是今天的，预测第 19 天只能用昨天以前的数据。

## 改进方案

解决问题 1，需要通过大量实证，从统计学的角度切入，寻找最恰当的训练集规模。

解决问题 2 可以考虑用傅立叶变换或小波变换换取出噪声和长周期因素。

解决问题 3 最好的方法可能还是通过累积的经验来选择最适合实际情况的核函数。或对熟悉常用的核函数用排除法处理。

解决问题 4，我们认为与输出结果有稳定关系的输入向量理论上来说能提供更准确的预测。比如对个股、行业或者大盘的资金净流量，和与股价密切相关的技术指标，如 MACD，KDJ，RSI，BIAS 等。

解决问题 5，我们可以采用另外一种建模方法。之前模型的算法是用一个训练集来预测未来 M 天的走势，我们可以简单的把 M 变小，趋近于 1，用循环的模式不断把训练集的时间段往后推移（这样往后的训练集会包含之前预测的数据），让训练集数据的时间可以尽可能的贴近预测数据的时间。

## 后续拓展方向

股指期货推出后，市场对其关注程度远超预期。除股指期货套利业务受到追捧外，具有较大风险偏好的投资者也希望通过期指的杠杆效应在投机业务上获利。此时对股指走势的判断显得比以往更为重要。我们将通过研究与股指存在密切相关的指标，将其作为模型的输入向量，运用 SVM 的简单预测模型对沪深 300 做出走势预测，希望给予股指期货投资者在投机交易业务上一些指导性建议。

本报告只是对支持向量机进行股价预测的一个综述，后面我们可以做的研究还有很多，例如输入向量的选择除交易性指标外，上市公司的财务指标也是很好的研究方向，这将给熟悉基本面分析的基金经理提供较好的参考信息。另外，结合傅立叶变换或小波变换过滤数据中的噪音和长周期因素，对 SVM 的预测准确性也会起到很大的帮助。

### 国信证券投资评级

类别	级别	定义
股票 投资评级	推荐	预计 6 个月内，股价表现优于市场指数 20%以上
	谨慎推荐	预计 6 个月内，股价表现优于市场指数 10%-20%之间
	中性	预计 6 个月内，股价表现介于市场指数±10%之间
	回避	预计 6 个月内，股价表现弱于市场指数 10%以上
行业 投资评级	推荐	预计 6 个月内，行业指数表现优于市场指数 10%以上
	谨慎推荐	预计 6 个月内，行业指数表现优于市场指数 5%-10%之间
	中性	预计 6 个月内，行业指数表现介于市场指数±5%之间
	回避	预计 6 个月内，行业指数表现弱于市场指数 5%以上

### 免责声明

本报告信息均来源于公开资料，我公司对这些信息的准确性和完整性不作任何保证。报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价或询价。我公司及其雇员对使用本报告及其内容所引发的任何直接或间接损失概不负责。我公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。本报告版权归国信证券所有，未经书面许可任何机构和个人不得以任何形式翻版、复制、刊登。

**国信证券经济研究所研究团队(含联系人)**

<b>宏观</b>		<b>策略</b>		<b>交通运输</b>	
周炳林	0755-82133339	赵 谦	021-60933153	郑 武	0755- 82130422
林松立	010-82254212	崔 嵘	021-60933159	陈建生	0755- 82130422
		廖 喆	021-60933162	岳 鑫	0755- 82130422
		黄学军	021-60933142	高 健	0755-82130678
<b>银行</b>		<b>房地产</b>		<b>机械</b>	
邱志承	021-68864597	方 焱	0755-82130648	余爱斌	0755-82133400
黄 飙	0755-82133476	区瑞明	0755-82130678	黄海培	021-60933150
谈 焯	010- 82254212	黄道立	0755-82130833	陈 玲	0755-82133400
				杨 森	0755-82133343
				李筱筠	010-82254205
<b>汽车及零配件</b>		<b>钢铁</b>		<b>商业贸易</b>	
李 君	021-60933156	郑 东	010-82254160	孙菲菲	0755-82133400
左 涛	021-60933164	秦 波	010-66026317	吴美玉	010-82252911
				祝 彬	0755-82131528
<b>基础化工</b>		<b>医药</b>		<b>石油与石化</b>	
张栋梁	0755-82130532	贺平鸽	0755-82133396	李 晨	021-60875160
陈爱华	0755-82133397	丁 丹	0755-82130678	严蓓娜	021-60933165
邱 斌	0755-82130532	陈 栋	021-60933147		
<b>电力设备与新能源</b>		<b>传媒</b>		<b>有色金属</b>	
皮家银	021-60933160	陈财茂	021-60933163	彭 波	0755-82133909
				谢鸿鹤	0755-82130646
<b>电力与公用事业</b>		<b>非银行金融</b>		<b>通信</b>	
徐颖真	021-60875162	邵子钦	0755- 82130468	严 平	021-60875165
谢达成	021-60933161	田 良	0755-82130513	程 峰	021-60933167
		童成敦	0755-82130513		
<b>造纸</b>		<b>家电</b>		<b>计算机</b>	
李世新	0755-82130565	王念春	0755-82130407	段迎晟	0755- 82130761
邵 达	0755-82132098				
<b>电子元器件</b>		<b>纺织服装</b>		<b>农业</b>	
		方军平	021-60933158	张 如	021-60933151
<b>旅游</b>		<b>食品饮料</b>		<b>建材</b>	
廖绪发	021-60875168	黄 茂	0755-82133476	杨 昕	021-60933168
刘智景	021-60933148				
<b>煤炭</b>		<b>建筑</b>		<b>固定收益</b>	
李 然	010-66026322	邱 波	0755-82133390	李怀定	021-60933152
苏绍许	021-60933144	李遵庆	0755-82133343	高 宇	0755-82133528
陈 健	010-66215566			侯慧娣	021-60875161
				张 旭	010-82254210
				蔺晓熠	021-60933146
				刘子宁	021-60933145
<b>指数与产品设计</b>		<b>投资基金</b>		<b>量化投资</b>	
焦 健	0755-82131822	杨 涛	0755-82133339	葛新元	0755-82133332
王军清	0755-82133297	黄志文	0755-82133928	董艺婷	021-60933155
彭甘霖	0755-82133259	彭怡萍	0755-82133528	戴 军	0755-82133129
阳 瑾	0755-82131822	刘舒宇	0755-82131822	秦国文	0755-82133528
周 琦	0755-82131822	康 亢	010-66026337	林晓明	021-60933154
赵学昂	0755-82131822			赵斯尘	021-60875174
				程景佳	021-60933166
				徐左乾	0755-82133090



**国信证券机构销售团队**

华南区			华东区			华北区		
万成水	0755-82133147 13923401205 wancs@guosen.com.cn		盛建平	021-60875169 15821778133 shengjp@guosen.com.cn		王立法	010-82252236 13910524551 wanglf@guosen.com.cn	
邵燕芳	0755-82133148 13480668226 shaoyf@guosen.com.cn		马小丹	021-60875172 13801832154 maxd@guosen.com.cn		王晓建	010-82252615 13701099132 wangxj@guosen.com.cn	
林莉	0755-82133197 13824397011 Linli2@guosen.com.cn		郑毅	021-60875171 13795229060 zhengyi@guosen.com.cn		谭春元	010-82254209 13810118116 tancy@guosen.com.cn	
王昊文	0755-82130818 18925287888 wanghaow@guosen.com.cn		黄胜蓝	021-60875173 13761873797 huangsl@guosen.com.cn		焦骥	010-82254202 13601094018 jiaojian@guosen.com.cn	
甘墨	0755-82133456 15013851021 ganmo@guosen.com.cn		刘塑	021-60875177 13817906789 liusu@guosen.com.cn		李锐	010-82254212 13691229417 lirui2@guosen.com.cn	
			叶琳菲	021-60875178 13817758288 yelf@guosen.com.cn		徐文琪	010-82254210 13811271758 xuwq@guosen.com.cn	
			孔华强	021-60875170 13681669123				