文章编号:1003－207(2013)03－0035－05

基于支持向量机的中国股指期货回归预测研究

赛 英[1],张凤廷[1],张 涛[2]

(1. 山东财经大学管理科学与工程学院,山东 济南 250014;2. 山东财经大学会计学院,山东 济南 250014)

摘 要:本文针对股指期货预测的特点,选择对股指期货指数有重要影响的相关指标,首次提出用支持向量机(SVM)方法对其进行回归预测,并用遗传算法(GA)和粒子群算法(PSO)分别优化四种不同核函数的支持向量机,构建了八种不同的中国股指期货回归预测方案,用实证研究的方法对这八种方案的准确性和时效性进行了比较。实验结果表明粒子群算法优化的线性核函数支持向量机作为中国股指期货回归预测的模型,具有更好的预测效果。

关键词:中国股指期货;支持向量机;遗传算法;粒子群算法;回归预测

中图分类号:F830.91 **文献标识码**:A

1 引言

在我国,沪深 300 股指期货合约已经于 2010 年 4 月 16 日上市交易,两年以来的日交易数据为对其进行预测研究提供了必要条件。但是,股指期货市场是一个极其复杂的动力学系统,高噪声、非线性和投资者的主观性等因素决定了对其进行预测的困难。神经网络以其良好的非线性逼近能力和自学习、自适应等特性,成为目前比较流行的预测手段。然而由于其基于经验风险最小化准则,在训练中最小化样本点误差,因而不可避免地出现过学习现象,模型的泛化能力受到了限制,同时易出现陷入局部最小点的问题,这都极大地影响了预测效果。

针对以往方法的这些不足,我们提出了一种基于支持向量机的股指期货预测方法。支持向量机(SVM)是一种机器学习方法,它的基础是 Vapnik 创建的统计学习理论,采用了结构风险最小化准则,而且由于它是一个凸二次优化问题,能保证找到的极值解就是全局最优解。借助 SVM 的这些特点,利用基于支持向量机的预测模型克服了神经网络过学习和易陷入局部最小的两大弱点,并在实证研究中取得了较大的精度。

目前,虽然已有一些学者将支持向量机应用于金融方面的预测研究,例如,Trafalis 和 Ince[1－2]用支持向量回归机对股票价格进行了预测研究;Cao 和 Tay[3－5]对支持向量机和 BP 神经网络在金融预测方面进行了比较研究;Kim[6]用支持向量机对股指趋势进行了预测研究,但是,还未有学者用支持向量机对股指期货指数进行回归预测。另外,上述学者的研究只是证明了支持向量机相比于其他方法具有更强的容错能力和泛化能力,具有更好的预测效果,却没有就支持向量机参数的选择进行深入研究,而 SVM 的惩罚参数 c 和核函数参数 γ 对预测准确率有着重要影响。本文提出用遗传算法和粒子群算法这两种智能优化算法对四种核函数的支持向量机进行优化,选择出合适的参数 c 和 γ,找出最适合对中国股指期货进行回归预测的模型,从而提高了预测的准确性和时效性。

2 相关技术的基本原理

2.1 支持向量机的基本原理

支持向量机是由线性可分情况下的最优分类平面发展而来,其基本思想可用图 1 所示的两类线性可分情况说明。

图 1 中,实心点和空心点分别代表两类样本,H 为最优分类超平面,H_1 和 H_2 分别为过两类离分类超平面最近的样本且平行于最优分类超平面的平面,它们之间的距离叫做分类间隔。所谓最优分类超平面就是要求分类面不但能将两类训练样本正确分开,而且使分类间隔最大。H_1 和 H_2 上的距离最优分类超平面最近的样本向量称为支持向量。使分

收稿日期:2012－07－30;**修订日期**:2013－01－20

基金项目:国家自然科学基金资助项目(70840018)

作者简介:赛英(1970－),女(汉族),山东济南人,山东财经大学管理科学与工程学院,教授,博士,研究方向:数据挖掘与商务智能.

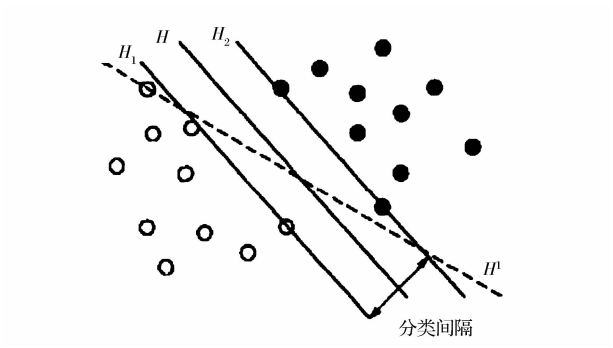

图 1 线性可分 SVM

类间隔最大从而提高模型的推广能力，这是 SVM 的核心思想。

设样本 x_i 为 d 维向量，$i=1,2,\cdots,n$，n 为训练样本数。根据每个 x_i 属于 $y_i=1$ 或者 $y_i=-1$，组成样本集 (x_i,y_i)，$x_i\in R^d$，$y_i\in\{-1,1\}$。设分类超平面方程为 $w^T\varphi(x_i)+b=0$，则归一化的样本集满足：

$$y_i(w^T\varphi(x_i)+b)\geq 1\quad i=1,2,\cdots,n \tag{1}$$

根据最优分类面的定义，可得分界面的分类间隔为 $d(x_i,w,b)=\dfrac{|w^T\varphi(x_i)+b|}{\|w\|^2}$，此时使分类间隔最大等价于使 $\|w\|$ 最小，从而我们可知寻找最优超平面问题就转化为在条件(1)约束下的二次规划问题：

$$\min\varphi(w)=\frac{1}{2}\|w\|^2 \tag{2}$$

解决这个问题需求出拉格朗日函数 $L_{p1}=\dfrac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i[y_i(w^T\varphi(x_i)+b)-1]$ 的鞍点，其中 α_i 是非负的拉格朗日乘子。

如果训练样本是线性不可分的，可以在条件(1)中加一个松弛因子 ξ_i，变为：

$$y_i(w^T\varphi(x_i)+b)\geq 1-\xi_i(\xi_i\geq 0,i=1,2,\cdots,n) \tag{3}$$

折中考虑最少错分样本和最大分类间隔，将目标函数变为：

$$\min\varphi(w,\xi)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i \tag{4}$$

其中，C 是一个大于零的常数，它控制对错分样本的惩罚程度，称为惩罚参数。

对于非线性问题，可以通过非线性变换转化为对偶问题，变换成某个高维空间中的线性问题，再在变换空间求最优分类面，这种非线性变换可以通过核函数来实现。主要的核函数有：线性核函数 $k(x,x_i)=\gamma x^Tx_i$，多项式核函数 $k(x,x_i)=(\gamma x^Tx_i+r)^d$，径向基 RBF 核函数 $k(x,x_i)=\exp(-\gamma\|x-x_i\|)$，Sigmoid 核函数 $k(x,x_i)=\tanh(\gamma x^Tx_i+r)$。其中的 γ 为本文要优化的核函数参数。此时分类函数变为：$f(x)=Sign(\sum_{i=1}^{n}\alpha_iy_iK(x_i,x)+b)$ 其中 $b=\dfrac{1}{N_s}\sum_{0<\alpha_i<C}(y_j-\sum_{i=1}^{n}\alpha_iy_iK(x_i,x_j))$。公式中 $0<\alpha_i<C$，C 为惩罚参数，$K(x_i,x)$ 为核函数[7-9]。

Vapnik 于 1997 年 Vanik1997 年将用于求解分类问题的支持向量机应用于回归问题，提出了 ε-支持向量回归机。支持向量回归机的解同样依赖于核函数和惩罚参数。

2.2 遗传算法的基本原理

遗传算法模仿生物界“优胜劣汰”的生物进化原理，通过选择、交叉、变异操作，及其适应度函数(本文为 SVM 训练函数)对个体进行筛选，适应度值(本文为训练得到的均方误差)好的被保留下来，差的被淘汰，新群体继承了上一代的信息，又优于上一代。这样反复循环，直到满足条件为止。遗传算法的基本操作有选择、交叉和变异操作。选择操作是指按照概率(与适应度值成正比)从旧种群中选择个体到新种群中。交叉操作是指从种群中选择两个个体，通过染色体的交换组合，形成新的优秀个体。变异操作是指选择一个个体，通过自身的染色体变异产生出更优秀的个体[10]。

2.3 粒子群算法的基本原理

每个粒子代表解空间中的一个潜在最优解，有速度、位置和适应度值三项指标。速度决定粒子移动的方向和距离，并可根据自身和其他粒子的移动经验(个体极值和群体极值，也就是个体和群体所经历的适应度值最优的位置)动态调整，从而实现个体在解空间中的寻优。根据目标函数(本文为 SVM 训练函数)，可计算出位置所对应的适应度值(本文为训练得到的均方误差)，适应度值的好坏表示了粒子的优劣。

假设在 D 维解空间中，由 n 个粒子组成的种群为 $X=(X_1,X_2,\cdots\cdots,X_n)$，第 i 个粒子的位置为 $X_i=(x_{i1},x_{i2},\cdots\cdots,x_{iD})^T$，根据目标函数和每个粒子的位置可以计算出每个粒子的适应度值。第 i 个粒子的速度为 $V_i=(V_{i1},V_{i2},\cdots\cdots,V_{iD})^T$，其个体极值为 $P_i=(P_{i1},P_{i2},\cdots\cdots,P_{iD})^T$，群体极值为 $P_g=(P_{g1},P_{g2},\cdots\cdots,P_{gD})^T$。在每次迭代过程中，粒子通

过个体极值和群体极值更新自身的速度和位置,更新公式如下:

$$V_{id}^{k+1} = \omega V_{id}^{k} + c_1 r_1 (P_{id}^{k} - X_{id}^{k}) + c_2 r_2 (P_{gd}^{k} - X_{id}^{k}); X_{id}^{k+1} = X_{id}^{k} + V_{id}^{k+1}$$

其中 ω 表示惯性权重;$d = 1,2,\cdots\cdots,D$;k 为当前迭代次数;V_{id} 为粒子的速度分量;c_1、c_2 为加速度因子,是非负常数;r_1、r_2 是[0,1]之间的随机数[11]。

3　模型构建

中国股指期货回归预测模型的构建过程如图 2 所示。

第一步,选取用于中国股指期货回归预测的指标和数据集。

第二步,数据预处理。

第三步,用遗传算法和粒子群算法分别优化支持向量机,选取最佳惩罚参数 c 和核函数参数 g。

第四步,训练不同核函数的支持向量机。

第五步,用训练好的支持向量机模型进行回归预测。

第六步,根据回归预测的结果选择最佳方案,建立中国股指期货回归预测的模型。

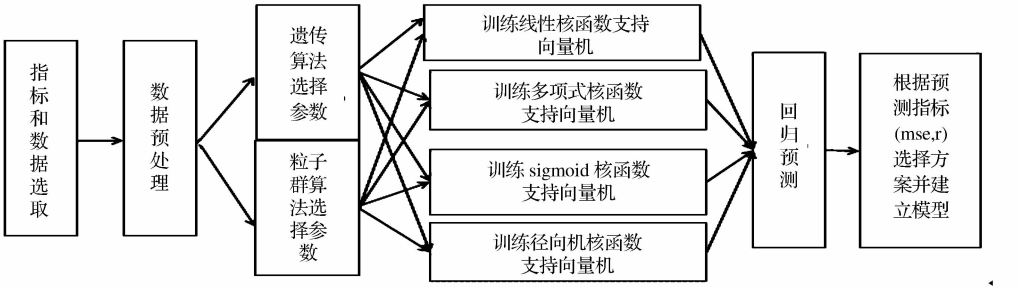

图 2　中国股指期货回归预测

4　实证研究

在实证研究中,我们利用 MATLAB 和台湾大学林智仁教授等人开发的 LIBSVM 软件包。

第一步,模型假设股指期货指数每日的开盘价与前一日的开盘价、最高价、最低价、收盘价、涨幅、振幅、总手、成交金额有关,本文选取了反映当日交易情况的这八个指标,用来回归预测第二天的开盘价。并从同花顺股指期货交易信息系统上下载了从 2010 年 4 月 19 日至 2012 年 6 月 15 日的所有日交易数据作为研究数据。

第二步,使用 MATLAB 中的 [X,Xps] = mapminmax(X,Ymin,Ymax) 函数对研究数据进行归一化,此函数的数学形式为:$x_k = (x_k - \text{Xmin}) * (\text{Ymax} - \text{Ymin}) / (\text{Xmax} - \text{Xmin}) + \text{Ymin}$。其中 X_k 表示需要归一化的向量 X 中的第 k 个值,Xmax 和 Xmin 分别表示向量 X 中的最大值和最小值。本文中我们需要将数据归一化到[1,2]区间内,所以 Ymin 取值为 1,Ymax 取值为 2。

第三步,先用遗传算法和粒子群算法粗略选择参数 c 和 g,然后根据粗略选择结果调整终止代数,种群数量,c 和 g 的取值范围等参数,不断实验得到精细选择结果。通过遗传算法精细选择出的参数 $c = 88.3455$,$g = 0.0095368$;通过粒子群算法精细选择出的参数 $c = 0.1$,$g = 0.59308$。图 3 为粒子群优化算法精细选择参数 c 和 g 的效果图。

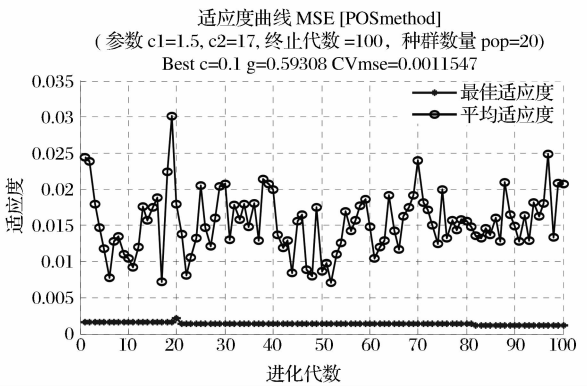

图 3　粒子群算法精细选择参数 c 和 g 的效果图

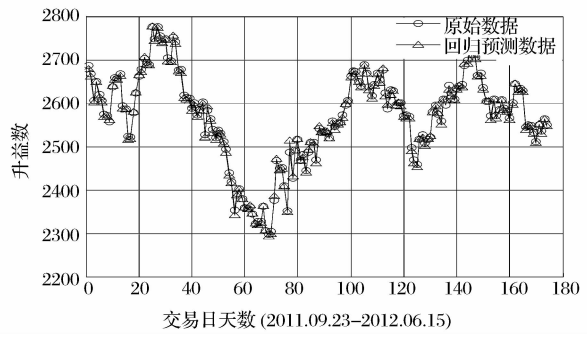

图 4　粒子群算法优化线性核函数支持向量机模型回归预测结果

表 1 方案比较

核函数		线性	多项式	Sigmoid	RBF
优化策略	效果对比指标	核函数	核函数	核函数	核函数
不使用优化策略	MSE	0.000973958	0.00836502	0.821863	0.00101805
	R	96.4637%	R = 90.3465%	R = 12.6517%	96.0003%
	T	25 秒	23 秒	19 秒	27 秒
遗传优化算法	MSE	4.58658e−005	0.00317093	0.000379234	1.95594e−005
	R	99.887%	91.9873%	99.4337%	99.8024%
	T	9 分 6 秒	8 分 42 秒	7 分 44 秒	9 分 54 秒
粒子群优化算法	MSE	1.81484e−005	0.00140003	2.25122	0.000311692
	R	99.8021%	93.2007%	11.3322%	99.2394%
	T	4 分 23 秒	4 分 29 秒	4 分 34 秒	4 分 36 秒

第四步，用约三分之二的研究数据（前 350 个交易日的数据）对不同核函数的支持向量机进行训练，训练函数为 model＝svmtrain（TSI，TSXI，*cmd*），其中 model 为训练得到的模型，TS1 为期望输出，TSX1 为训练输入，*cmd* 为参数设置。由于篇幅所限，训练所得到的各模型的决策函数的系数和支持向量不再给出。

第五步，用训练好的模型对剩余的约三分之一的研究数据（后 174 个交易日的数据）进行回归预测，检测模型的预测效果。图 4 给出了粒子群算法优化线性核函数支持向量机模型的回归预测结果，其中蓝色圆圈表示原始数据，红色三角表示预测数据。

第六步，根据预测结果从相关性（R），均方误差（MSE），时效性（从数据加载到打印预测结果用时 T）上对是否需要优化策略，及选择两种优化策略时所形成的八种方案进行比较，如表 1 所示。

通过比较可以得出四点结论：第一，通过优化选择参数，能显著提高预测的准确性，虽然付出了时间的代价，但这样做是值得的。第二，在中国股指期货指数回归预测中，粒子群算法优化的支持向量机模型在时效性上要大大优于遗传算法优化的支持向量机模型，但准确性上却略差；第三，RBF 核函数和线性核函数的支持向量机在误差和相关性上优于另外两种核函数的支持向量机；第四，遗传算法优化的线性核函数支持向量机、遗传算法优化的 RBF 核函数支持向量机、粒子群算法优化的线性核函数支持向量机是效果最好的三个模型，三者在误差和相关性方面效果相差不大，但在时效性上粒子群算法优化的线性核函数支持向量机却大大优于前两者，所以本文选取粒子群算法优化的线性核函数支持向量机作为中国股指期货指数回归预测的模型，而且实验结果也证明了此模型用于中国股指期货回归预测能够取得很好的效果。

5 结语

本文采用基于支持向量机方法的回归预测模型对中国股指期货指数进行预测，并用遗传算法和粒子群算法两种优化策略分别优化了四种核函数的支持向量机，形成了八种方案。通过对这八种方案的实证研究，我们选取了用粒子群算法优化的线性核函数支持向量机作为中国股指期货回归预测的模型，实验结果也证明此模型用于中国股指期货回归预测能够取得很好的效果。中国股指期货刚推出，容易受到偶然因素的干扰，所以预测难度相对于其他成熟市场指数更大，但仍能取得较好的预测效果，因此该模型也适用于其他金融指数的预测。但是，目前模型只是从技术的角度对中国股指期货进行了短期回归预测，还有一定的局限性。下一步，从影响期货指数的各种指标入手，对中国股指期货进行长期预测将是我们的研究重点。

参考文献：

[1] Trafalis T，Ince H. Support vector machine for regression and applications to financial forecasting[C]. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks，Como，July 24 − 27，2000.

[2] Ince H，Trafalis T. Short term forecasting with support vector machines and application to stock price prediction [J]. International Journal of General Systems，2008，37 (6)：677−687.

[3] Tay F E H，Cao Lijuan. Improved financial time series forecasting by combining support vector machines with self-organizing feature map[J]. Intelligent Data Analysis，2001，5(4)：339−354.

[4] Cao Lijuan，Tay F E H. Financial forecasting using support vector machines[J]. Neural Computing & Applica-

tions,2001,10(2):184－192.
[5] Tay F E H, Cao Lijuan. Application of support vector machines in financial time series forecasting[J]. Omega, 2001,29:309－317.
[6] Kim K. Financial time series forecasting using support vector machines[J]. Neurocomputing,2003,55(1－2): 307－320.
[7] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2000,1(26):32－42.
[8] 邓乃扬,田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京:科学出版社,2004.
[9] Vapnik V. 统计学习理论的本质[M]. 张学工,译. 北京:清华大学出版社,2000.
[10] Holland J. Adaptation in natural and artificial systems [M]. Ann Arbor:University Michigan Press,1975.
[11] Kennedy J, Eberhart R C. Particle swarm optimization [C]. International Conference on Neural Networks, Perth, Nov. 27－Dec 01,1995.

Research of Chinese Stock Index Futures Regression Prediction Based on Support Vector Machines

SAI Ying[1], ZHANG Feng-ting[1], ZHANG Tao[2]

(1. School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan 250014, China;
2. School of Accounting, Shandong University of Finance and Economics, Jinan 250014, China)

Abstract: According to the characteristics of the stock index futures prediction, the indicators that have great influence on the development trend of stock index futures are selected and the support vector machines are firstly used to the regression prediction of stock index futures. Besides, genetic algorithm (GA) and particle swarm optimization algorithm (PSO) are employed to optimize the support vector machine (SVM) with four different kernel functions and eight different programs are attained. By comparing the accuracy and the time complexity of all the programs, the empirical study shows that the linear kernel function SVM optimized by PSO is the best model for regression prediction of Chinese stock index futures.

Key words: Chinese stock index futures; support vector machine (SVM); genetic algorithm (GA); particle swarm optimization algorithm (PSO); regression prediction