支持向量机选择及其在股票走势预测中的应用*

郭 辉
(重庆师范大学 数学与计算机科学学院, 重庆 400047)

摘 要:支持向量机(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。在研究了股票数据的特点以及对股票预测的研究结果后,本文根据传统的 SVM 算法原理,提出一种在线选择训练样本的在线增量训练的方式完成模型更新的动态预测模型(DMDI),使得仅增加较小工作量为代价而获得更高的预测精度成为可能。应用 DMDI 对股市的大盘和个股的走势分别进行中短期预测,并跟神经网络的预测结果进行了比较。大量数值实验表明, DMDI 模型比不进行选择的静态模型和神经网络模型对股票走势的预测更为有效,具有明显的优越性。

关键词:支持向量机(SVM);股票预测;核函数;动态选择
中图分类号:TP311.5 文献标识码:A 文章编号:1672-6693(2007)04-0045-05

支持向量机(Support Vector Machine,SVM)理论源于 Vapnick 在 1963 年提出的解决模式识别问题的支持向量法[1]。1995 年 Vapnick 提出统计学习理论(Statistical Learning Theory,SLT),较好地解决了线性不可分问题[2]。近年来,在理论研究和算法实现方面都取得了突破性进展,开始成为克服"维数灾难"和"过学习"等传统困难的有力手段。关于支持向量机在经济预测中特别是股票市场的应用却是刚刚起步,很多问题有待研究和探索。

针对股票走势预测这种有监督的分类问题,SVM 比神经网络等传统的学习方法具有以下几个特点和优点:1)它综合考虑了分类器的经验风险和置信风险,在一定概率意义下是推广能力最好的分类器,这种结构风险最小化的设计思路可以避免陷入"欠学习"和"过学习"等情况;2)它有全局最优解,不会陷入局部最优;3)它利用核函数的方法解决了非线性的分类问题,其算法复杂程度主要取决于训练样本的个数,而与特征维数基本无关。

股票市场是一个复杂的非线性动态系统,利用传统的时间序列预测技术很难揭示其内在的规律。当前关于股市预测的研究方法仍然以神经网络和时间序列为主,而这些方法基本上采用的是通过批量学习建立静态模型的方法。对于类似股票指数的非线性时间序列,随着新样本的不断获得,如果能根据新样本实现模型动态更新,则能够适应问题的变化,提高预测精度。如果能实现增量学习,则能避免大量样本的重复训练,缩短训练时间,提高训练效率。

本文基于以上考虑,结合传统的 SVM 算法原理和时间序列在线增量学习的思想,提出了一种在线选择训练样本,在线增量训练的方式完成模型的更新,建立动态预测模型(DMDI),将使仅增加较小工作量为代价而获得更高的预测精度成为可能,对股市的大盘和个股的走势(涨或跌)进行中短期预测。

1 基于 SVM 的多种预测模型的构造

对于预测连续 K 天的时序问题,有 3 种预测模型构造方法[3]。

1)静态模型 - 静态输入(SMSI)。该方法根据时序连续的 L 个训练样本来建立 K 个 SVM,使其中第 K 个 SVM 来预测第 k 天的输出。这样,仅使用一个测试输入即可以获取连续的 K 天的测试输出。相应地,其训练样本中的 Y 也应修正为 k 步后的输出。基于该方法得到的模型特点是:静态的模型,静态的输入。

2)静态模型 - 动态输入(SMDI)。该方法使用传统的 SVM 构造方法,使用 L 个训练样本通过一次离线批量学习,仅构造一个 SVM,对给定的测试输入仅进行后一步的预测输出,要得到 K 天连续输出则需要给 K 个连续的输入。基于该方法得到的模型特点是:静态的模型,动态的输入。

3)动态模型 - 动态输入(DMDI)。该方法最初只

* 收稿日期:2007-04-01
作者简介:郭辉(1980-),女,河南郑州人,助教,硕士,研究方向为支持向量机在数据挖掘方面的应用、最优化算法及应用。

使用 L 个训练样本通过一次离线学习构造一个SVM。每当根据一个测试输入进行一次输出后,就将该测试数据作为训练样本进行在线增量学习,并用在线学习后的新模型进行下一次预测并再次进行在线学习,如此重复。在进行完第 k 步预测输出的同时,完成第 $k+1$ 个动态模型的生成。基于该方法得到的模型特点是:动态的模型,动态的输入。

2 改进的算法原理

支持向量机分类问题可以表述如下[4]。

设给定的训练集:$T = \{(x_1,y_1),\cdots,(x_l,y_l)\} \in (X \times Y)^l$,其中,$x_i \in X = \mathbf{R}^n$,$y_i \in Y = \{1,-1\}$,$i = 1,\cdots,l$ 选择适当的惩罚参数 $C > 0$,构造并求解最优化问题

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i,x_j) - \sum_{j=1}^{l}\alpha_j$$

$$\text{s.t.} \quad \sum_{i=1}^{l} y_i \alpha_i = 0$$

$$0 \leqslant \alpha_i \leqslant C, i = 1,\cdots,l$$

得到最优解 $\alpha^* = (\alpha_1^*,\cdots,\alpha_l^*)^{\mathrm{T}}$ 选择 α^* 的一个小于 C 的正分量 α_j^*,并据此计算 $b^* = y_i - \sum_{i=1}^{l} y_i \alpha_i^* K(x_i,x_j)$;求得决策函数 $f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i^* K(x_i,x) + b^*\right)$。其中 $K(x,x')$ 为核函数,α_i 为对应的拉格朗日乘子。

根据VC维理论[4],设 h 为假设集 S 的VC维,l 是训练集所含的样本个数,当 $l > h$,$h\left(\ln\frac{2l}{h}+1\right) + \ln\frac{4}{\delta} \geqslant \frac{1}{4}$ 成立时,则对任意的概率分布 $P(x,y)$ 和任意的 $\delta \in (0,1]$,假设集 S 中的任意假设 f 都可使得下列不等式

$$R(f) \leqslant R_{emp}(f) + \sqrt{\frac{8}{l}\left(h\left(\ln\frac{2l}{h}+1\right) + \ln\frac{4}{\delta}\right)} \quad (1)$$

至少以 $1-\delta$ 的概率成立,特别地,当 $R_{emp}(f) = 0$ 时,有 $R(f) \leqslant \sqrt{\frac{8}{l}\left(h\left(\ln\frac{2l}{h}+1\right) + \ln\frac{4}{\delta}\right)}$

(1)式给出了期望风险 $R(f)$ 的一个定量估计,它的右端的第二项 $\sqrt{\frac{8}{l}\left(h\left(\ln\frac{2l}{h}+1\right) + \ln\frac{4}{\delta}\right)}$ 称为置信区间,而右端的两项之和称为结构风险,它是期望风险 $R(f)$ 的一个上界。

结构风险最小化归纳原理的基本思想是:不仅要使经验风险最小,还要使VC维尽量小,对未来样本才会有较好的推广能力。即,如果要求风险最小,就需要不等式(1)式右边中的两项相互平衡,共同趋于极小;另外,在获得的学习模型经验风险最小的同时,希望学习模型的推广能力尽可能大,这样就需要 h 值尽可能小,即置信风险尽可能小。而文献[1]定理说明,对于 n 维空间中的线性指示函数集合来说,它的VC维是 $h = n+1$,而对于固定的输入空间维数 n,则 h 是确定的。

根据(1)式,期望风险 $R(f)$ 是训练样本数目 l 的增函数,在时序股票数据中,笔者是使用历史的数据来预测未来的股票指数,所以,要选择的向量只能是要预测的连续 k 天前的股票数据。而且,离所需预测时间第 k 天的股市指数的距离不同,则对其的涨跌影响就不同。有人就固定选择离所要预测时间最近的固定 l 个,而抛弃离其最远的训练样本。也曾有人研究加权稳健支持向量回归方法和加权支持向量机在证券指数预测中的应用,但是如何确定合适的权重很难解决。随着时间的延续,不加选择地增加训练样本,则期望风险会逐渐增大。随着新样本的不断获得,如果能根据新样本实现模型的动态更新,则能够适应问题的变化,提高预测精度;如果能实现增量学习,则能避免大量样本的重复训练,缩短训练时间,提高训练效率。

本文根据以上分析,设计以下这个动态输入-动态输出模型,对股票大盘和个股的走势进行连续 k 天的预测,固定训练样本数目,随着时间延续,对训练样本进行选择。

基于DMDI的选择准则如下。

1)首先选定 $(X_{s_1},Y_{s_1}),(X_{s_2},Y_{s_2}),\cdots,(X_{s_k},Y_{s_k})$ 为初始选定的 k 个训练向量(按时间顺序排列),且有 $1 \leqslant S_1 \leqslant S_2 \leqslant \cdots \leqslant S_k \leqslant l$ 这 k 个训练向量是要预测的第 $k+1$ 天的前 k 天的股票数据;

2)选择 C-支持向量分类机对训练向量进行训练,得到决策函数,将第 k 个向量的输入指标代入决策函数来预测第 $k+1$ 天的股市走势(涨或跌);

3)当预测第 $k+2$ 天时,计算 k 个训练向量中的第一个训练向量 (X_{s_1},Y_{s_1}) 在求解分类超平面中所对应乘子的值 α_1^*,如果 $\alpha_1^* \neq 0$,则表示其为支持向量,那么笔者采用 $(X_{s_2},Y_{s_2}),(X_{s_3},Y_{s_3}),\cdots,(X_{S_k},Y_{S_k})$ 作为训练向量进行训练,并求决策函数,将第 $k+1$ 天股市的各项指标,即输入向量代入决策函数来预测第 $k+2$ 天的股市走势;如果 $\alpha_1^* = 0$,则表示其

不是支持向量,则计算(X_{s_2},Y_{s_2}) 所对应的乘子 α_2^*,对其做类似于(X_{s_1},Y_{s_1}) 的处理。以此下去,直到找到 $\alpha_i^* \neq 0$ 为止,然后把其对应的训练向量(X_{s_i},Y_{s_i}) 抛弃,选择剩余的其他训练样本和(X_{S_k},Y_{S_k}) 作为新的训练向量,来预测第 $k+2$ 天的股市走势,依次下去。

3 实验结果及分析

3.1 样本的选取和预处理

股票交易市场是一个很不稳定的动态变化过程,不仅受国内外经济因素的影响,而且受投资人的行为(庄家的行为) 的影响。此外,政府的宏观调控也是影响未来走势的重要因素。因此,必须选取价格运行平稳的股市作为研究对象;否则,模型只能获取那些特殊的规律,失去了股市运动的主要规律,模型推广能力也不会很好。针对这些状况,本文选取沪市上证 180 指数,其规模大、流动性好、发展稳定、代表主体经济的行业龙头组成,能真正反映市场主体部分的价值水平,可以避免被庄家操纵等因素的影响,同时可以避免考虑除息除权的影响。对于个股,为了避免人为操纵和财务弄虚作假,选取那些业绩好诚信度高的企业股票作为预测对象。

在应用 SVM 预测股票走势时,如何有效地选取输入向量的每一个分量是决定预测模型准确性的关键之一。由于当前人们对股票波动等非线性问题的机理认识不足,经验在此起到重要的影响。本文所选的输入分量均是长期股票市场投资经验总结[5]。

表 1　确定的输入向量

项目	含义	项目	含义	项目	含义
x_1	今日最高指数	x_2	今日最低指数	x_3	今日开盘指数
x_4	今日收盘指数	x_5	今日成交额	x_6	昨日成交额
x_7	前日成交额	x_8	30d 平均成交额	x_9	今日成交量
x_{10}	30d 平均成交量	x_{11}	今日涨跌幅	x_{12}	昨日涨跌幅
x_{13}	前日涨跌幅	x_{14}	10d 平均涨跌幅	x_{15}	30d 平均涨跌幅

3.2 结果和讨论

3.2.1 对大盘走势预测进行数值实验和分析　沪市上证指数从 1998 年 8 月 31 日起,共 1693 个交易日的数据作为研究对象对股票市场的整体走势进行预测结果对比。在不同的核函数下,DMDI 预测结果比较(见表 2),从表中可以看出,径向基核函数 Rbf 的预测结果最好。

表 2　不同核函数 DMDI 股市整体走势预测结果比较

核函数	训练天数/d	预测天数/d	正确个数/个	回代正确率/%	预测正确率/%
线性	100	10	4	66.7	40
多项式($p=3$)	100	10	5	76.67	50
Erbf($p=100$)	100	10	6	71.5	60
Rbf($\sigma=25$)	100	10	9	98.7	90

在径向基核函数 Rbf 选定情况下,由于目前还没有明确的寻找参数 C 和 σ 值的最优值的方法,所以本文采用交叉检验和广义搜索的方法,从一个较大的取值范围内找到使交叉检验所得结果最好的一对参数(C,σ) = (100,25)。从 1998 年 8 月 26 日开始的 150 d 交易数据作为训练样本,紧接着的15 d、30 d 进行预测,回代正确率和预测正确率都很高。另外,数值实验中发现,不管是 SMDI($\sigma=40$) 还是 DMDI($\sigma=25$) 方法,不同的参数取值,预测正确率都很高。分析其原因,发现股市在此阶段比较稳定,没有太多的大涨大落。因此,可以推想,如果股市比较平稳,预测结果会更好。如表 3 所示。

表 3　SMDI 和 DMDI 股市整体走势预测结果比较

	训练天数/d	预测天数/d	正确个数/个	回代正确率/%	预测正确率/%
DMDI $\sigma=25$	150	15	15	100.00	96.09
	150	30	27	95.64	90.00
SMDI $\sigma=40$	150	15	9	70.00	60.00
	150	30	20	70.00	66.67

对比不同的训练样本下,最好的参数 SMDI($\sigma=40$) 和 DMDI($\sigma=25$) 下股市整体走势预测结果(表 4),发现在相对大样本量的情况下,DMDI 体现出了比 SMDI 要好的预测能力,在实际应用中,对于 SMDI 算法,对相同的训练样本,连续预测的天数越多,其正确率明显降低,而 DMDI 算法则体现其优越性。足够的样本训练才能得到更好的预测效果,这点得到了充分的说明。值得注意的是,根据结构风险最小化原则,由于期望风险的上界与样本个数成正比,则当训练样本个数过多时,期望风险就会增大,所以训练样本的个数不能过多。

另外在 DMDI 中,进行了模型的选择,几乎每次都要进行超平面的重新计算,所以计算时间比 SMDI

要略长,每次预测平均多需要50 s左右的时间。显然,所消耗的时间仍然可以满足股票预测的时效性,而正确率却大大提高了。

表4 SMDI和DMDI股市整体走势预测结果比较

	训练天数/d	预测天数/d	正确个数/个	回代正确率/%	预测正确率/%
SMDI	100	20	12	75.00	60.00
	100	30	17	75.00	56.67
	150	15	9	70.00	60.00
	150	30	20	70.00	66.67
	200	20	12	75.50	60.00
	200	30	16	75.50	53.33
	训练天数/d	预测天数/d	正确个数/个	平均回代正确率/%	平均预测正确率/%
DMDI	100	20	19	98.30	95.00
	100	30	28	97.97	93.33
	150	15	15	96.09	100.00
	150	30	27	95.64	90.00
	200	20	17	95.65	85.00
	200	30	25	95.40	83.33

3.2.2 对个股走势预测进行数值实验和分析 东方航空的数据采用的是2004年10月28日~2005年9月7日的211条数据作为实验对象,选定径向基核函数Rbf,$C=100$,$\sigma=10$采用统计标准化,把数据规范到$[-1,1]$这个范围内(预测结果见表5)。

表5 SMDI和DMDI东方航空涨跌预测结果比较

东方航空	训练天数/d	预测天数/d	正确个数/个	回代正确率/%	预测正确率/%
SMDI	100	10	6	100%	60.00%
	100	20	11	100%	55.00%
	100	30	16	100%	53.33%
DMDI	100	10	8	92.10%	80.00%
	100	20	15	93.30%	75.00%
	100	30	21	93.63%	70.00%

中视传媒的数据采用的是2003年12月19日~2005年10月12日的433条数据作为实验对象。向基核函数Rbf,$C=100$,$\sigma=25$采用统计标准化,把数据规范到$[-1,1]$这个范围内(预测结果见表6)。

从两个表中的预测结果,可以看到类似大盘的结论,DMDI算法比SMDI算法体现了明显的优越性,同时也说明了在线选择支持向量分类机应用于股票个股走势的可行性和有效性。

表6 SMDI和DMDI中视传媒涨跌预测结果比较

中视传媒	训练天数/d	预测天数/d	正确个数/个	回代正确率/%	预测正确率/%
SMDI	100	10	8	100	80.00
	100	20	14	100	70.00
	100	30	21	100	70.00
	150	15	13	100	86.67
	150	30	24	100	80.00
DMDI	100	10	8	100	80.00
	100	20	16	100	80.00
	100	30	21	93.63	70.00
	150	15	14	97.33	93.33
	150	30	26	96.93	86.67

4 与神经网络方法的比较

表7为改进的神经网络对股票市场涨跌的预测效果表,对比DMDI预测的结果可以明显看出,神经网络的预测正确率基本可以达到70%,训练精度可以达到98%[5-6]。本文提出的在线选择支持向量分类机的动态预测模型(DMDI)的预测效果比神经网络的预测效果要好得多——基本都保持在70%以上,好的时候可以达到80%多,甚至90%。但是其训练精度比神经网络方法的要低,基本在90%~98%之间,这正体现了支持向量机能够避免过学习的优越性,也从实际应用中说明了支持向量机方法比神经网络算法有更好的泛化能力,更适合用来作股票预测。

表7 BP神经网络的预测效果

实验	实际预报天数/d	正确率/%
1	17	64.7
2	19	68.1
3	21	71.4
4	17	69.3
5	16	66.7
6	23	69.2
7	14	62.5
8	21	66.7

5 结论

支持向量机法应用于具有非线性时间序列的经济预测问题的研究还处于初步阶段,比神经网络具有更好的拟合精度和泛化能力,体现了其可行性和优越性。但是,在股票市场预测中的应用还有很多

值得研究的:1)在具体应用过程中核函数的选取以及参数的选择都需要经验,还没有更好的选择方法,这点有待研究;2)对于股票数据来说,怎样选取输入向量的分量来降低输入空间的维数——特征选择也是很值得研究的一个方面,因为降低了输入空间的维数则可有效地降低期望风险的上界等。

参考文献:

[1] VAPNIK V N. The Nature of Statistical Learning Theory [M]. New York:Springer,1995.

[2] VAPNIK. Statistical Learning Theory Wiley [M]. New York:Springer,1998.

[3] 田翔,邓飞其. 精确在线支持向量回归在股指预测中的应用[J]. 计算机工程,2005,31(22):18-20.

[4] 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.

[5] 吴微,陈维强,刘波. 用 BP 神经网络预测股票市场的涨跌[J]. 大连理工大学学报,2001,41(1):9-15.

[6] 吴微,陈维强. 用于股市的 BP 算法的一些改进[J]. 大连理工大学学报,2001,41(5):518-522.

[7] 陶小龙. 基于支持向量机的股票预测[D]. 理学硕士学位论文. 北京:北京工业大学,2005.

Application of Online Selection Support Vector Classification in the Prediction of Ups and Downs in Stock Market

GUO Hui

(College of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047, China)

Abstract: Support Vector Machine (SVM) which is a new technology used in Data Mining. It is a new tool that accounts for the problems of the Machine Learning by the method of the optimization. Applying the support vector machine method in the research on the non-linear time series economic prediction problem is underway. It is more feasible and predominant than the Neural Networks algorithm in the extending ability and the tallying precision. After we studied the characteristics of the stock data and the rules of the stock market people, we put forward to a dynamic model which bases on the traditional support vector machine arithmetic. The model selects the training data online when we get the new data and then we modify the model each time base on the increased data in the aggregate. It is a dynamic model, so it can catch the real time change of the market. It make the prediction precision be improved comes to truth with the small workload as the cost. In this paper we use the support vector machine and the Time series dynamic model (DMDI) to predict the short-time and the medium-term ups and downs in the single stock and the holistic Shanghai stock market. We perform a large numbers of numerical experiments and compared with the results being got based on the methods of the BP neural networks and the static models which is not changed when the new data is got with the time going, and the prediction rightness probability is higher, and it is more feasible in the extending ability and the tallying precision through the actual application. In addition, It can also avoid the difficult problem——study of the training data excessively. The results show that the DMDI is more suitable for the forecasting the index time series of the stock market than the BP neural networks and the static models. The model we have proposed in this paper has more advantages in the prediction of the trends of the stock market than the conventional methods.

Key words: Support Vector Machine (SVM); stocks prediction; kernel function; dynamic selection

(责任编辑 游中胜)