第 23 卷第 5 期
2006 年 10 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 23 No. 5
Oct. 2006

文章编号：1000−8152(2006)05−0805−05

快速增量加权支持向量机预测证券指数

李拥军[1,2]，奉国和[3]，齐德昱[1]

(1. 华南理工大学 计算科学与工程学院, 广东 广州 510640; 2. 广州市广播电视大学, 广东 广州 510260;
3. 华南师范大学 经济管理学院, 广东 广州 510006)

摘要: 传统支持向量机是对小样本提出, 对于大样本会出现训练速度慢、内存占用多等问题, 并且不具有增量学习性能.而常用的增量学习方法又会出现局部极小等问题. 本文阐述了一种改进的支持向量机算法(快速增量加权支持向量机算法)用于证券指数预测. 该算法先对指数样本做相空间重构, 再分解成若干个工作子集, 针对样本重要程度给出不同权重构建预测模型. 实验分析表明, 在泛化精度保持略好情况下, 训练速度明显提高.

关键词: 支持向量机; 增量学习; 证券指数预测; 相空间重构

中图分类号: TP18, O29　　**文献标识码**: A

Fast incremental weighted support vector machines for predicating stock index

LI Yong-jun[1,2]，FENG Guo-he[3]，QI De-yu[1]

(1. School of Computer Science and Engineering, South China University of Technology, Guangzhou Guangdong 510640, China;
2. Radio and Television Guangzhou University, Guangzhou Guangdong 510260, China;
3. College of Economics and Management, South China Normal University, Guangzhou Guangdong 510006, China)

Abstract: Traditional support vector machine (SVM) is effective only for small size of samples.When the size of sample is large, it exhibits a low training speed and a large required memory. Thus, it is not suitable for increment learning. Furthermore, traditional increment learning algorithms such as neural network have local minima only. To tackle this problem, a fast incremental weighted support vector machines for predicting the stock index is put forward. The algorithm model reconstructs the phase for the index, and then decomposes the sample space into subsets and gives different weights to them. Experimental results show that modified algorithm raises the training speed while maintaining the same precision.

Key words: support vector machine; incremental learning; stock index; phase-space reconstruction

1　引言(Introduction)

标准支持向量机的提出其初衷是为解决小样本机器学习而提出的一种学习模型, 其泛化精度不是依赖整个训练样本的, 而是训练样本的一个很小的子集, 该子集包含的信息等同与整个训练样本包含的信息. 标准支持向量机求解过程最后化为线性约束的凸二次规划问题, 因此解具有全局最优性和唯一性.如当训练样本数为n时, 该二次规划问题包含了$2n$个优化变量、1个等式约束、$4n$个线性不等式约束, 同时还涉及到$n \times n$维核函数矩阵的计算和矩阵与向量相乘计算[1], 所以求解的规模与样本数量有关. 对于样本量大训练, 其训练速度则会很慢, 训练时间与计算内存是大样本训练时所遇到的一个主要瓶颈. 而由于增量学习是一种样本量不断增加的过程, 这样不对传统的支持向量算法做改进, 是无法发挥支持向量的优势的, 萧嵘[2], S.Ruping[3], G.Fung等[4] 做过这方面的研究. 其中文献[2]提出的算法存在一些缺点, 如用户需要选择很多的参数, 而确定这些参数的值也是一个很棘手的问题.

本文提出一种改进的支持向量机算法—快速增量支持向量机学习算法(fast incremental weighted support vector machines, FIWSVM), 并用于证券指数预测, 实验结果表明, 该方法在保持泛化精度前提下, 提高了训练速度.

2　支持向量机回归理论(Support vector machine regression theory)

基于统计学习理论(statistical learning theory)的支持向量机给出了实际风险的上限, 并利用核函数

收稿日期: 2005−05−16; 收修改稿日期: 2006−03−29.
基金项目: 中国博士后科学基金资助项目 (2005037582); 广东省自然科学基金资助项目(05200300).

将线性不可分转化为特征空间线性可分, 最后化为求解一个线性约束的凸二次规划求解问题.

给定训练集$G=\{(x_i,y_i)\}_{i=1}^n$, 其中$x_i\in\mathbb{R}, y\in\mathbb{R}$, 确定一个基于训练集$G$的函数

$$f(x)=<\omega\cdot\phi(x)>+b \tag{1}$$

来逼近未知的实际函数. 其中: $<\cdot>$表示在高维特征空间Ω中的内积, $\phi: x\to\Omega, b\in\mathbb{R}$.考虑误差, 引入松弛变量$\xi_i^+\geqslant 0, \xi_i^-\geqslant 0$, 同时给定损失函数

$$L_\varepsilon(f(x)-y)=\max\{0,|f(x)-y|-\varepsilon\}. \tag{2}$$

问题化为在约束条件下

$$\begin{cases} y_i-<\omega\cdot\phi(x)>-b\leqslant\varepsilon+\xi_i^+,\\ <\omega\cdot\phi(x)>+b-y_i\leqslant\varepsilon+\xi_i^-,\\ \xi_i^+\geqslant 0,\\ \xi_i^-\geqslant 0. \end{cases} \tag{3}$$

求

$$\phi(\omega,\xi)=\frac{1}{2}||\omega||^2+C\sum_i(\xi_i^++\xi_i^-) \tag{4}$$

最小值, 其中C为常数. 引入拉格郎日乘子构造拉格郎日泛函, 得到原问题的对偶问题(dual problem)[1]

$$\begin{aligned}&\max_{\alpha,\alpha^*}W(\alpha,\alpha^*)=\\ &\max_{\alpha,\alpha^*}-\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)k(x_i,x_j)+\\ &\qquad\sum_{i=1}^n[\alpha_i(y_i-\varepsilon)-\alpha_i^*(y_i+\varepsilon)].\end{aligned} \tag{5}$$

所要求的回归方程为

$$\begin{aligned}f(x)=&<\omega\cdot\phi(x)>+b=\\ &\sum_{i=1}^n(\alpha_i^*-\alpha_i)k(x_i,x)+b^*.\end{aligned} \tag{6}$$

其中$k(x,y)=<\phi(x)\cdot\phi(y)>$是满足Mercer条件[1]的核函数. 文中采用RBF核函数: $k(x,y)=\exp(-1/\sigma^2(x-y)^2)$, σ为RBF核的宽度. 由于求解最后化为一个线性约束的凸QP问题, 所以解具有全局最优性和唯一性. 在式(6)中$\alpha_i-\alpha_i^*\neq 0$所对应的样本称为支持向量(support vector, SV), 这样的样本数一般占整个样本数的比例很少, 也就是说决策函数的构造是由这少数的样本来决定的. 这样在不影响精度的情况下, 如果用支持向量取代原来的训练样本进行学习, 可以极大地减少训练时间和内存的占有率.

3 支持向量增量学习模型及其算法(Support vector machine learning model and algorithm)

3.1 支持向量增量学习问题(Incremental support vector machine learning)

设初始训练样本集为G, 增量训练样本集为I, 增量学习问题就是研究针对新增训练样本如何有效地学习下步需要的模型. 传统的方法是新增训练样本加入时, 历史训练结果舍弃, 而将初始训练样本和新增训练样本重新构成新训练样本集来进行学习. 但用传统支持向量算法来进行这种学习效率比较低, 而且它不适应于在线学习, 因为其涉及的是一个$n\times n$的海森矩阵的计算, 随着样本量n的增加, 训练速度会越来越慢.

要克服这种困难, 则需要充分运用支持向量的特点, 因为在构造决策函数时, 除了少数样本(支持向量)有效外, 其余都是冗余的, 其对决策函数的构造不起作用. 这个思想启发笔者在进行增量学习时, 不需要全部样本参与训练, 而只选取支持向量即可. 文献[5]指出在SVM增量学习算法中两个主要关注的问题是:

第一, 如何利用历史训练结果让再次训练预测函数时更快?

第二, 如何在不损失预测精度的前提下抛弃样本点?

3.2 加权回归支持向量机(Weighted support vector machine)

在式(4)中对超出ε管道的样本实行相同的惩罚C, 但在一些应用比如股票价格、证券指数等这样的时间序列数据中, 近期数据显得比远期数据重要, 近期数据错误造成的影响远比远期数据要大, 而体现这种重要性常用的手段就是给与不同的权值. 这样在上述优化问题中, 可以采取不同的惩罚因子C, 以得到更准确的回归估计. 将式(4)改写为

$$\Phi(\omega,\xi)=\frac{1}{2}||\omega||^2+C\sum_i s_i(\xi_i^++\xi_i^-). \tag{7}$$

s_i为加权系数, 式(7)的约束条件与式(4) 相同, 得到对偶最优化问题

$$\begin{aligned}&\max_{\alpha,\alpha^*}W(\alpha,\alpha^*)=\\ &\max_{\alpha,\alpha^*}-\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)k(x_i,x_j)+\\ &\qquad\sum_{i=1}^n[\alpha_i(y_i-\varepsilon)-\alpha_i^*(y_i+\varepsilon)].\end{aligned} \tag{8}$$

约束条件

$$\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0,\ 0\leqslant\alpha_i\leqslant Cs_i,$$
$$0\leqslant\alpha_i^*\leqslant Cs_i,\ i=1,2,\cdots,n. \tag{9}$$

在本文中采用Logsitic指数权函数[6], 即

$$s_i=\frac{1}{1+\exp(a-2ai/n)}. \tag{10}$$

其中α为控制上升速率的参数.s_i随着α不同取值而产生不同的权值变化曲线, 如图 1 所示, 其中实线、虚线、“*” 线、“+” 线分别对应r的值为0.1, 1, 10, 100. 不同的α值反映了权值上升的速率是不同的, 在实际的问题中可根据具体情况给出α的值.

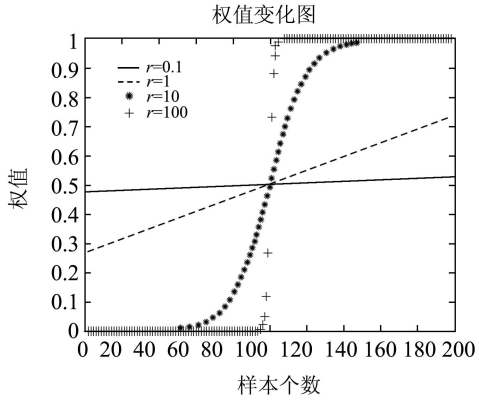

图 1　权值变化图

Fig. 1　Curve of changed weight

在以后3.3算法的第2步和第3步运用3.2节提出的加权回归支持向量机SV_{new}, 实际上就是根据式(8)(9) 来训练新工作集SV_{new}. 同时新工作集中不同样本依据Logsitic函数即式(10)给出不同权值.

3.3　快速增量加权支持向量机学习模型与算法(Fast incremental weighted support vector machines model and algorithm)

支持向量集描述了整个训练集的特点, 而且在绝大多数情况下, 训练集中的支持向量数目只占训练集的很少一部分, 即$N_{\text{sv}}<<N_{\text{train}}$(其中: N_{sv}为支持向量数, N_{train}为训练样本数). 因此可以使用支持向量集取代训练集进行函数逼近学习, 使得在不影响逼近精度的情况下极大地减少了训练时间和内存开销.

训练样本集$G=\{(x_i,y_i)\}_{i=1}^{n}$, 其中$x_i\in\mathbb{R}^n, y\in\mathbb{R}, G_1,G_2,\cdots,G_m$为训练工作子集, 且$G_1\cup G_2\cup\cdots\cup G_m=G, G_i\cap G_j=\varnothing$, $i\neq j, SV_1,SV_2,\cdots,SV_m$为在各个$G_i$上产生的支持向量集, 将这些数据顺次并成一个新的训练集$SV_{\text{new}}=SV_1\cup SV_2\cup\cdots\cup SV_m$. 在新的训练集$SV_{\text{new}}$上进行加权回归支持向量机(weighted support vector machines, WSVM), 得到下步要预测的决策函数[15]. 当新训练样本集I加入时, 在I上抽取支持向量得到集合I_{sv}, 舍弃最远的样本子集G_{m}, 将剩余工作子集的支持向量和I_{sv}重新组成训练集$SV_1\cup\cdots\cup SV_{m-1}\cup I_{\text{sv}}$, 在其上构造加权支持向量机作为下步要预测的决策函数. 每增加新训练样本集时重复进行该步骤, 进行增量学习, 这样用于模型学习的训练样本始终是最近的m个工作子集. 该算法步骤主要思想简单描述如下:

第1步　将初始训练样本集分解为m个互不相交的工作子集, 对每个工作子集G_i 抽取支持向量SV_i, 将$\sum_{i=1}^{m}SV_i$ 集成一个新训练集SV_{new};

第2步　运用3.2节提出的加权回归支持向量机SV_{new}, 得到下步要预测的模型;

第3步　增量训练样本集I加入时, 抽取支持向量得到集I_{sv}; 同时去掉离新训练样本最远的工作子集合G_m, 将I_{sv} 及$SV_1,SV_2,\cdots,SV_{m-1}$依次重编为$SV_1,SV_2,\cdots,SV_m$组成新的初始训练样本集$SV_{\text{new}}$, 运用3.2节提出的加权回归支持向量机训练$SV_{\text{new}}$, 得到下步要预测的模型;

第4步　重复第3步, 不断进行增量学习.

针对大样本支持向量机问题, Boser和Vapnik提出了分块算法(chunking algorithm)[7], Osuna等提出分解算法(decomposition algorithm)基本框架[8,9], Jachims在上述分解算法的基础上做了几点重要改进[10], 同时利用该方法实现的SVM^{light}是设计SVM分类器的重要软件. Platt在分解算法基础上提出了贯序最小化算法(sequential minimal optimization, SMO)[11], 该算法是分解算法的一种特殊情况. 这些算法在一定程度上解决了大规模样本需要大内存的的困难, 但同时也出现了一定的不足, 即这些算法为了寻找最优化解需要反复地迭代, 会出现收敛速度较慢的情况. 本文提出的模型在保证泛化精度前提下, 极大加快了训练速度.

该算法的完备性由算法的分解过程可以看出, 该分解始终是有限个工作子集, 同时引入适当权重, 最远工作子集抽取的支持向量对预测起的作用趋于零, 去掉与不去掉在同一时间窗口里效果趋于相同, 关于增量学习的完备性在文献[12, 13] 里有详细说明.

4 基于快速增量加权支持向量机证券指数预测(Stock index predicting based on FI-WSVM)

4.1 证券指数预测相空间重构(Phase-space reconstruction of stock index)

证券指数是一混沌时间序列, 本来对于混沌时间序列是没办法去预测的, 但笔者运用Takens定理[14]通过选择合适的参数m,τ, 将混沌时间序列重构相空间为

$$\{x(n\tau)|x(n\tau)=[x((n-1)\tau),x((n-2)\tau),\cdots,$$
$$x((n-m)\tau)],n=1,2,\cdots\}. \quad (11)$$

依据Takens理论, 在一定条件下, 对几乎所有的τ和满足特定条件的m, 存在一个光滑映射$f:\mathbb{R}^m\to\mathbb{R}$, 使得

$$x(n\tau)=$$
$$f(x((n-1)\tau),x((n-2)\tau),\cdots,x((n-m)\tau)). \quad (12)$$

实验中采用2002年1月4日~2004年4月13日的深圳成指、上证180各543个数据点, 根据证券市场及证券指数的特点, 在式(14)中取$\tau=1,m=5$, 通过相空间重构, 问题转化为估计下面的动态系统:

$$x_t=f(x_{t-1},x_{t-2},\cdots,x_{t-5}). \quad (13)$$

其中x_t为在时间t的股票指数. 这样一来, 整个系统的输入就是时间t前5个历史数据, 而输出则是时间t时的值.

4.2 试验设计(Experiment design)

在数据进行模拟前先对数据预处理, 数据归一化, 归一化的数据经过时间序列相空间重构后, 整个样本数为538个, 将数据集分成11个工作子集, 其中前10个工作子集的样本数均为50, 第11个工作子集样本数为38. 对于增量学习, 先以前8个工作子集抽取的支持向量作为初始训练样本集, 以后每次加入增量训练样本集, 删除最前面的工作集, 从增量训练集中抽取支持向量并与余下的工作子集的支持向量组成新的训练集, 这样就每次用来学习的训练样本量维持在一个很小的数目, 大大加快了学习速度.同时设计对比计算, 即传统支持向量学习, 初始训练集为前8个工作子集, 以后每加入一增量训练集, 则将最前面的一个工作子集删除, 剩余工作子集与增量集组成新的训练集, 也就是说保持最近的m个工作子集构造下步的SVM预测模型.

4.3 实验结果(Experiment results)

实验中采用均方误差MSE作预测精度的评价标准, 表中T为训练时间(单位为: s), SVs为支持向量数. 增量学习栏中第1列In1, In2, In3分别为第1次、第2次、第3次增量训练样本数同时其也作为上次的测试样本数, S为参与学习的训练样本数(即工作子集抽取的支持向量数); 传统学习栏中第1列Test1, Test2, Test3为用于测试的样本数, S为参与训练样本数. 计算结果如表2所示:

表 1 实验结果比较

Table 1 Results comparison of experiment

	增量加权支持向量学习		传统支持向量学习	
深圳成指	In1=Test1=50 S=59	MSE=0.0049 SVs=35, T=6.8	Test1=50 S=400	MSE=0.0052 SVs=53, T=452.3
	In2=Test2=50 S=39	MSE=0.0542 SVs=20, T=6.3	Test2=50 S=400	MSE=0.0649 SVs=37, T=458.3
	In3=Test3=38 S=41	MSE=0.0074 SVs=20, T=6.3	Test3=38 S=400	MSE=0.0093 SVs=41, T=455.6
上证180	In1=Test1=50 S=132	MSE=0.0045 SVs=81, T=21.1	Test1=50 S=400	MSE=0.0054 SVs=106, T=445.6
	In2=Test2=50 S=111	MSE=0.0092 SVs=56, T=12.8	Test2=50 S=400	MSE=0.0097 SVs=81, T=417.7
	In3=Test3=38 S=118	MSE=0.0070 SVs=68, T=15.4	Test3=38 S=400	MSE=0.0071 SVs=93, T=424.8

从两种算法的计算结果可以看出, 增量学习泛化精度略高, 训练速度与传统方法相比是数量级的提高. 同时, 增量学习比传统方法支持向量的数目少, 这样由式(6)可知, 在一定程度上压缩了支持向量数, 加快了预测时间.

5　结论(Conclusion)

支持向量机近年来成为机器学习研究的一个热点, 但它不支持增量学习. 本文中针对时间序列提出了一种增量的支持向量学习算法, 将大的训练集合先分成若干个小的训练工作集, 每个工作集上抽取支持向量, 同时根据数据重要程度不同给出不同的权值, 当增量训练样本集加入时, 淘汰最不重要的工作子集, 剩下的与增量训练样本构成新的训练集用于下步模型学习. 实验表明该算法与传统增量学习相比不但泛化精度提高了, 训练速度速度也极大地加快了.

参考文献(References):

[1] VAPNIK V N.*The Nature of Statistical Learning Theory*[M]. New York: Springer, 1999.

[2] 萧嵘, 王继成, 孙正兴, 等. 一种SVM 增量学习算法α-isvm[J]. 软件学报, 2001, 12(12): 1818 – 1824.
(XIAO Rong, WANG Jicheng, SUN Zhengxing, et al. An incremental SVM Learning algorithm α-isvm[J]. *J of Software*, 2001, 12(12): 1818 – 1824.)

[3] RUPING S. Incremental learning with support vector machines[C]//*Proc of Int Conf on Data Mining*. San Jose: IEEE Computer Science Press , 2001: 641 – 642.

[4] FUNG G, MANGASARIAN O L. Incremental support vector machine classification[C]//*Proc of the Second SIAM Int Conf on Date Mining SDM*. Arlington, Vinginia: IEEE Computer Science Press, 2002: 135 – 145.

[5] XIAO R, WANG J, ZHANG F . An approach to incremental SVM learning algorism[C]// *Proc of the 12th Int Conference on Tools with Artificial Intelligence*.New York: IEEE Computer Science Press, 2000: 268 – 273.

[6] FRANCIS E H T, CAO L J. Modified support vector machines in financial time series forecasting[J]. *Neural Computing*, 2002, 48: 847 – 861.

[7] BOSER B E, GUYON I M, VAPNILK V N. A training algorithm for optimal margin classifiers[C] //*Proc of the Fifth Annual Workshop on Computational Learning Theroy*. Pittsburgh, PA: ACM Press, 1992: 144 – 152.

[8] OSUNA E, FREUND R, GIROSI F. An improved training algorithm for support vector machines[C]//*Proc of Int Conf on Newral and Signal Procossing*. New York: IEEE Computer Science Press, 1997: 276 – 285.

[9] OSUNA E, FREUND R, GIROSI F. *Suppport Vector Machines:Training and Application*[M]. Cambridge, MA: Massachusetts Institute of Technology, AILab, 1997.

[10] JOACHIMS T. Making large-scale support vector machines learning practical[C]//*Advances in Kernel Methods-Support Vector Learning*1999. Cambridge, MA: MIT Press, 1999: 169 – 184.

[11] PLATT J. Fast training of support vector machines using sequential minimal optimization [C]//*Advances in Kernel Methods-Support Vector Learning*. Cambridge, MA: MIT Press, 1999: 185 – 208.

[12] CAUWENBERGHS G, POGGIOT. Incremental and decremental support vector machine[J]. *Advances in Neural Information Processing Systems*, 2001, 13(8): 409 – 415.

[13] MA J S, THEILER J, PERKINS S. Accurate on-line support vector regression[J]. *Neural Computation*. 2003, 15(11): 2683 – 2704.

[14] TAKENS F. Detecting strange attractor in turbulence[J]. *Lecture Notes in Mathematics*, 1981: 898(2) :361 – 381.

[15] 奉国和, 朱思铭. 基于支持向量机的分解合作加权算法及其应用[J]. 计算机科学, 2005, 32(4): 91 – 93.
(FENG Guohe, ZHU Siming. Decomposition-cooperation weighted support vector machines and it'a application[J]. *Computer Science*, 2005, 32(4): 91 – 93.)

作者简介:

李拥军　(1968—), 男, 博士, 副教授, 主要研究领域为计算机网络与人工智能, E-mail:Liyi@scut.edu.cn ;

奉国和　(1972—), 男, 博士, 讲师, 主要研究领域为人工智能、机器学习、数据挖掘;

齐德昱　(1959—), 男, 博士生导师, 教授, 研究方向为分布式系统.