基于支持向量机的股票预测

张晨希,张燕平,张迎春,陈 洁,万 忠
(安徽大学 智能计算与信号处理重点实验室,安徽 合肥 230039)

摘 要:针对股票预测的特点,选择对上市公司股票走势有重要影响的相关数据进行测试。为了避免传统的预测算法(如BP算法)的一些弊端,使用可以避免这些弊端并且具有良好分类功能的支持向量机对该上市公司股票走势进行预测。测试表明预测的精度明显高于采用BP算法等传统神经网络分类方法的测试结果,预测达到了让人满意的效果。

关键词:股票;预测;支持向量机;数据
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2006)06-0035-03

Stock Prediction Based on Support Vector Machine

ZHANG Chen-xi, ZHANG Yan-ping, ZHANG Ying-chun, CHEN Jie, WAN Zhong
(Key Lab. of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039, China)

Abstract: According to the characteristics of the stock prediction, this paper selects the data that greatly influence the stock development trend of listed companies. In order to avoid the disadvantages of the traditional NN classification methods (e.g. BP algorithm), this paper uses the support vector machine(SVM) to predict the stock development trend of listed companies. The test shows that the accuracy of the prediction is obviously higher than traditional NN classification ways, such as BP algorithm and thus it has a satisfying result.

Key words: stock; prediction; support vector machine; data

0 引 言

股票市场是证券业和金融业必不可少的重要组成部分,受到投资者的普遍关注。有效的股票预测在金融投资领域占有重要地位,因此对股票价格进行分析和预测有着非常重大的理论意义和实践价值。由于股票受到政策、经济以及投资人心理等诸多复杂的因素的影响,而这些因素是没有确定规则的,因此近几年兴起了利用人工神经网络来对股票进行分析和研究的热潮[1,2]。在传统的几种利用人工神经网络进行股票预测的方法中,BP网络本身还存在隐层结构无规律可循、易陷于局部最小值等缺陷[1,3]。支持向量机(SVM)算法可有效地改善这些缺陷[4,5]。因此文中对股票数据进行分类分析,从而达到预测的目的。

传统的有关股票预测方面的研究所采用的数据大多直接来源于证券公司所公布的开盘价、收盘价、最高价、最低价、综合指数等交易日的数据,数据缺乏足够的透明性[6,7]。文中采用的数据则是通过对巨灵证券数据库产品3.0产品中117个报表中的海量数据中进行挖掘、提取,从而得到上市公司的净资产收益率、投资报酬率、销售报酬率、流动比率、留存盈利比例等16个数据。这些数据反映了该上市公司的自身经营情况,因此将对该上市公司股票走势有着决定性影响[8]。笔者正是力图对这些数据进行分析从而达到对该上市公司的股票进行预测的目的。

收稿日期:2005-09-23
基金项目:"九七三"计划国家重点基础研究(2004CB318108);国家自然科学基金(60475017, 60135010);安徽省自然科学基金(050420208)
作者简介:张晨希(1982-),男,安徽桐城人,硕士研究生,研究方向为智能算法及其应用;张燕平,教授,硕士生导师,研究方向为人工神经网络、智能算法及其应用。

1 支持向量机

SVM理论是在统计学习理论的基础上发展起来的。由于统计学习理论和SVM方法对有限样本情况下模式识别中的一些根本性的问题进行了系统的理论研究,很大程度上解决了以往的机器学习中模型的选择与过学习问题、非线性和维数灾难问题、局部极小点问题等,所以它们在20世纪90年代以来受到了很大的重视。SVM是针对二类模式识别问题而提出的。设训练样本集为(x_i, y_i),$y \in \{+1, -1\}$是类别标号($i = 1, \cdots, n, x \in R^d$)。对于线性可分情况,存在$(w, b)$,使$w \cdot x_i + b > 0, Px_i \in$ Class1, $w \cdot x_i + b < 0, Px_i \in$ Class2。分类的目的是寻求(w, b),使最优分类面满足分类间隔最大。为减少分类平面的重复,对(w, b)进行如下约束:

$$\min_{i=1,2,\cdots,n} = | w \cdot x_i + b | = 1$$

满足上式的超平面成为典型超平面。典型超平面到最近数据点的距离为$\frac{1}{\| w \|}$。这样求最优分类面的问题转化为求最优问题:

$$\min_{w,b} \frac{1}{2} \| w \|^2 \tag{1}$$

s.t. $y_i(w \cdot x_i + b) \geqslant 1, i = 1,2,\cdots,n$

定义 Lagrange 函数为：

$$L(w,b,a) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{n} a_i \{ y_i[(w \cdot x_i + b) - 1] \}, a_i > 0 \tag{2}$$

令

$$\frac{5L(w,b,a)}{5w} = w - \sum_{i=1}^{n} a_i y_i x_i = 0 \tag{3}$$

$$\frac{5L(w,b,a)}{5b} = \sum_{i=1}^{n} a_i y_i = 0 \tag{4}$$

若 a_i^3 为最优解，则

$$w^3 = \sum_{i}^{n} a_i^3 y_i x_i \tag{5}$$

将式(3) 和式(4) 代入式(2)，得

$$F(a) = \sum_{i=1}^{n} a_i - \frac{1}{2} \| w^3 \|^2 = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j (x_i x_j)$$

这样，式(1) 的优化问题转化为对偶问题：

$$\min Q(a) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j (x_i x_j) - \sum_{i=1}^{n} a_i$$

$$\text{s.t.} \quad a_i \geqslant 0, i = 1,2,\cdots,n \tag{6}$$

$$\sum_{i=1}^{n} y_i a_i = 0$$

通过对式(6) 的求解，得到最优分类函数为：

$$f(x) = \text{sign}\{ \sum_{i=1}^{n} a_i^3 y_j (x_i \cdot x) + b^3 \} \tag{7}$$

式中 b^3 的求解可通过任选一支持向量，由式(1) 的约束方程(此时取等号) 求出。根据 KT 条件，对于大多数样本而言，$a^i = 0$。对应 $a^i \neq 0$ 的样本称为支持向量(Support Vector,SV)。由支持向量集决定的分类面和由全体样本集决定的分类面是等价的。对于线性不可分情况，可采用增加松弛变量或采用合适的核函数，通过非线性变换将低维的输入空间映射到高维的特征空间，然后在这个新空间中求取最优线性分类面。目前得到研究的核函数主要有 3 类[5]，包括：

阶次为 d 的多项式核函数：

$$K(x,x_i) = [(x,x_i) + 1]^d$$

径向基函数型核函数：

$$K(x,x_i) = \exp\{ - \frac{| x - x_i |^2}{\sigma^2} \}$$

神经网络 S 形函数核函数：

$$K(x,x_i) = \tanh(\Omega(x,x_i) + c)$$

引入核函数 $K(x,x_i)$ 代替最优分类面中的点积，则优化方程变为：

$$\min Q(a) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j (x_i x_j) - \sum_{i=1}^{n} a_i$$

$$\text{s.t.} \ 0 \leqslant a_i \leqslant c, \ i = 1,2,\cdots,n \tag{8}$$

$$\sum_{i=1}^{n} y_i a_i = 0$$

最优分类面方程为：

$$\sum_{i=1}^{n} y_i a_i = 0 \tag{9}$$

任选一支持向量 x_i，由式(10) 求得 b：

$$y_j(\sum_{i=1}^{n} a_i y_i K(x_j x_i + b) - 1) = 0 \tag{10}$$

2 支持向量机的测试结果及其分析

在股票的若干种数据中，每股收益是每个股民最关心的部分，也是股民投资股票的动机所在。故在选取每股收益作为分类标准的同时，选取了每股净资产、股利支付率、每股股利、净资产收益率、留存盈利比例、流动比率、速动比率、负债比率、长期负债比率、应收账款比率、存货周转率、销售报酬率、净利润率、投资报酬率、净值报酬率这 15 个对每股收益影响最大的数据作为输入样本的属性[8]。表 1 是对巨灵证券数据库产品 3.0 产品中的上市公司数据进行整理分析后所得到的样本的基本概况。

表 1 样本概况

样本个数	6942
样本属性个数	15
决策属性类别	4
样本分布情况	
第一类	509
第二类	133
第三类	4008
第四类	2292

2.1 测试结果

利用支持向量机(SVM)对样本进行交叉训练测试，即将样本平均分成 10 组，其中 9 组用来学习，1 组用来测试，循环交叉。分类测试的结果如表 2 所示。

表 2 试验结果(一)

分类识别率/%	训练时间/s	测试时间/s
70.086	10.642	0.4095

利用 BP 算法对样本进行分类测试的结果如表 3 所示。

表 3 试验结果(二)

分类识别率/%	训练时间/s	测试时间/s	神经元个数
58.582	10.35	0.01	4
57.736	11.667	0.01	5
57.736	13.469	0.01	6

利用 RBF 算法对样本进行分类测试的结果如表 4 所示。

表 4 试验结果(三)

分类识别率/%	训练时间/s	测试时间/s
66.934	11.63	1.1406

从测试结果中，可以发现在训练时间、测试时间和 BP 算法相差不多的情况下，通过支持向量机所取得的分类识别率明显高于 BP 算法和 RBF 算法的分类识别率，为了更

一步验证支持向量机在分类识别率的优越性,将表 1 样本中的决策属性类别由四类改成三类后进行测试,其测试结果如表 5,6 所示。

表 5　十交叉训练试验结果(四)

分类识别率/%	训练时间/s	测试时间/s
98.652	2.9612	0.023

表 6　BP 算法试验结果(五)

分类识别率/%	训练时间/s	测试时间/s	神经元个数
92.407	3.565	0.03	4
92.352	3.875	0.04	5
92.407	4.4415	0.04	6

2.2　结果分析

从上面的测试结果可以得到如下结论:

(1) 采用支持向量机进行股票预测的准确率在决策属性类别为四类的时候明显高于 BP 算法和 RBF 算法,在决策属性的类别为三类的时候虽然 BP 算法取得了较好的分类识别率,但支持向量机仍然比 BP 算法的分类识别率高出了将近 7 个百分点。因此,相比传统的利用神经网络来预测股票的几种方法来说,支持向量机在预测精度上明显提高。

(2) 从表 2 和表 5 的对比中发现,随着决策属性类别的增加,利用支持向量机所获得分类识别率也有所降低,但仍然优于同等情况下的 BP 算法和 RBF 算法的测试结果,虽然在预测每股收益具体数值的性能有待提高,但对股票是否盈利的预测取得了很好的效果。因此支持向量机进行股票预测有着良好的可行性和应用性。

(3) 股票市场受到政策、经济以及人为因素的干扰和影响,从而大大增加预测的难度,要想非常准确地进行股票预测并进入实用阶段,仍然有很长的路要走。但是通过测试可以发现利用对上市公司本身经营情况的分析可以排除掉部分人为因素的干扰,达到较好的预测效果。

3　结束语

利用支持向量机并采用对上市公司股票走势有着重要影响的相关数据来达到对股票进行预测的目的,从测试上来看预测精度明显高于传统的神经网络的预测方法,提高了分类的准确率。所采用的支持向量机克服了 BP 算法固有的缺陷,如学习过程收敛速度慢、网络性能差、可能存在局部极小值等。虽然股票市场非常复杂,但是如果能够对反应上市公司经营状况的数据进行合理分析,从而提取影响该上市公司股票的关键数据,再使用支持向量机对股票走势进行预测,仍然可以得到令人满意的结果。

参考文献:

[1] Baba N, Kozaki M. An Intelligent Forecasting System of Stock Price Using Neural Networks[A]. In Proceedings of IJCNN [C]. Los Alamitos: IEEE PRESS, 1992. 652 - 657.

[2] Raymond S, Lee T. iJADE Stock Advisor: An Intelligent Agent Based Stock Prediction System Using Hybrid RBF Recurrent Network[J]. IEEE Transactions on Systems, Man, and Cybernetics - part A: Systems and Humans, 2004, 34(3): 421 - 428.

[3] 张　铃,张　钹. 人工神经网络理论及应用[M]. 杭州:浙江科技出版社,1995.

[4] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2000,26(1):32 - 41.

[5] Burgers B C. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121 - 167.

[6] 吴　微,陈维强,刘　波. 用 BP 神经网络预测股票市场涨跌[J]. 大连理工大学学报,2001(1):11 - 15.

[7] 张　建,陈　勇,何永保. 人工神经网络之股票预测[J]. 计算机工程,1997(2):52 - 55.

[8] 财政部注册会计师考试委员会办公室. 财务成本管理[M]. 北京:经济科学出版社,2003.

(上接第 34 页)

地节约网络带宽、降低网络负载。使用组播也存在一些缺点,如一些防火墙和路由器会阻塞组播的消息,同时还存在其它 IP 组播的障碍,比如个人防火墙、子网路由器。不过 JXTA 还支持其它多种发现机制,在实际应用中,可以通过多种发现的机制的结合应用来达到更好的目的[9]。

参考文献:

[1] Milojicie D S, Kalogeraki V, Lukose R, et al. Peer to Peer Computing[R]. Palo Alto: HP Laboratories, 2002. 2 - 6.

[2] Bawa M, Cooper B F, Crespo A. Peer - to - Peer Research at Stanford[R]. USA: Stanford University, 2003.

[3] 吴胜浩,钟亦平,张世永. 用 IP 组播发现同位体发现机制[J]. 计算机工程,2004,30(3):119 - 121.

[4] 刘　波. IP 组播通信机制及其实现[J]. 计算机工程,2003,29(6):131 - 133.

[5] Johnson V, Johnson M. How IP Multicast Works[EB/OL]. http://ipmulticast. com/community/whitepapers/howipmcworks. pdf, 2003 - 11 - 26.

[6] Moore D, Hebeler J. 对等网[M]. 苏　忠,战晓雷等译. 北京:清华大学出版社,2003.

[7] Krikorian R. HelloJXTA! [EB/OL] http://www. onjava. com/pub/a/onjava/2001/04/25/jxta. html? page=1. 2001.

[8] Oaks S, Traversat B, Li Gong. JXTA 技术手册[M]. 技　桥译. 北京:清华大学出版社,2004.

[9] 许　斌. JXTA—Java P2P 网络编程技术[M]. 北京:清华大学出版社,2003.