

基于支持向量机粒化的证券指数预测

陈孝全^{1,2}, 刘 波¹

(1. 华南师范大学 计算机学院, 广东 广州 510631;
2. 深圳博物馆网络中心, 广东 深圳 518026)

摘 要:为分析股票价格指数变化,文中提出一种采用近似支持向量机(PSVM)将金融时间序列数据进行模糊信息粒化的方法,并用此方法对上证指数数据进行回归分析预测。其实现过程是以 2008 年到 2013 年的上证综指数数据建立抛物型模糊粒子,运用近似支持向量机原理,采用交叉验证的方法对相关参数进行寻优,用优化参数对时间序列进行训练,并回归预测模糊粒子的三个参数来确定上证综指的走势变化。对于非线性难预测的股票指数,实验分析比较了实际数据与预测数据,证明具有较好的预测效果。

关键词:股票指数;近似支持向量机;模糊信息粒化;交叉验证;回归分析

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2015)04-0148-05

doi:10.3969/j.issn.1673-629X.2015.04.034

Stock Index Prediction Based on Support Vector Machines and Information Granulation

CHEN Xiao-quan^{1,2}, LIU Bo¹

(1. School of Computer Science, South China Normal University, Guangzhou 510631, China;
2. Information Centre of Shenzhen Museum, Shenzhen 518026, China)

Abstract:To analyze the changes of the stock index, put forward a financial data time series prediction method based on proximal support vector machine and fuzzy information granulation, and this method is used for the Shanghai composite index data regression analysis and forecast. The parabolic fuzzy particle model is established about the Shanghai composite index data set between 2008 and 2013. Use the theory of proximal support vector machine combined cross validation method to optimize the parameters, train the time series data with the optimized parameters, and three parameter of fuzzy particles regression prediction determine the variation trend of the Shanghai composite index. The stock index is nonlinear and hard to predict, the experiment analyzes and compares both the actual data and the predicted data, which proves that it has good prediction effect.

Key words:stock index; Proximal Support Vector Machine (PSVM); fuzzy information granulation; cross validation; regression analysis

0 引 言

近年,随着国家市场化脚步的加大,股票市场正在向着更深入更健康的方向发展。如何把握分析到市场的动向和趋势,是经济研究的重点。股票市场向来是国家经济的“晴雨表”,所以对股票市场的研究和预测也就成为了焦点。衡量中国股票市场的两大股指是上证综指和深证成指,研究这两大指数的走向,有效规避风险,就能扩大投资者的回报收益,实现资本的增值。股票指数是一个时间序列数据,根据其特点,对时间序列的分析研究可以采用的成熟模型很多,包括灰色理

论模型^[1]、ARMA 模型^[2]、GARCH 模型^[3-4]等,而从 BP 神经网络入手进行金融指数预测分析^[5-6]的思想也有学者在研究,特别近来首先由 Vapnik 等提出的支持向量机^[7-8]应用于金融领域是其一大大热点^[9-10]。而文中提出的对金融时间序列数据进行模糊信息粒化^[11-12]结合近似支持向量机的方法,能对中国证券市场股票指数的波动范围进行有效准确的预测。

1 信息粒化理论

信息粒化(Information Granulation)是由美国数学

家 L. A. Zadeh 教授于 1965 年提出,现在已经成为国内外计算机领域的研究热点。根据 Zadeh 教授的观点,信息粒化就是将一组相似的元素整合成一个对象去研究其性能特征,而这个对象就是信息粒,这种信息粒的命题表达方式为:

$$g \triangle (x \text{ is } G) \text{ is } \lambda \text{ 或者 } g \triangle x \text{ is } G$$

式中,论域 U 属于拓扑空间, $x \in U, G \subseteq U$, G 由隶属函数 μ_G 表达, λ 则表示可能性概率。通常设定 U 为实数集 $R(R^n)$, G 为 U 的凸模糊子集, λ 是单位区间的模糊子集。为达到良好效果,可以采用抛物线的模糊信息粒进行支持向量机的回归数据预测。抛物线的模糊粒子的隶属函数为:

$$A(x, c, a, b) = \begin{cases} 0, & x < a \\ 1 - \left(\frac{x-a}{c-a}\right)^2, & a \leq x \leq c \\ 1 - \left(\frac{b-x}{b-c}\right)^2, & c \leq x \leq b \\ 0, & x > b \end{cases} \quad (1)$$

为简化描述,模糊粒子 P 代替模糊概念 $G, P = A(x, a, b, c) = A(x), a, b, c$ 为模糊粒子化的三个参数,分别是模糊集下界 a , 上界 b 和模糊粒子集的核 c 。

2 支持向量机

2.1 标准支持向量机

支持向量机(Support Vector Machine, SVM)建立在统计学习理论的 VC 维理论和结构风险最小原理上,事实上对于超平面 $H: \omega \cdot x + b = 0$ 分割样本,可以构造条件极值 $\begin{cases} \min \|\omega\|^2/2 \\ \text{s.t. } y_i(\omega \cdot x_i + b) - 1 \geq 0 \end{cases}$,再根据拉格朗

日函数求解,求解过程中引入松弛变量、惩罚参数等概念。在线性情况下,得到分类决策函数为 $f(x) = \text{sgn} \left\{ \sum_{i=1}^m a_i^* y_i x_i \cdot x + b^* \right\}$; 在非线性情况下,根据 Mercer 条件,分类决策函数为 $f(x) = \text{sgn} \left\{ \sum_{i=1}^l a_i^* y_i K(x_i \cdot x) + b^* \right\}$ 。其中:拉格朗日乘子 a_i 求出的解就是支持向量, b^* 是分类的阈值。

2.2 近似支持向量机

为了提高分类效率, Mangasarian 等在支持向量机的基础上提出近似支持向量机^[13-14](PSVM)。以下从最优化理论出发, PSVM 模型转化成如下的求解约束。

$$\begin{cases} \min_{(\omega, b, \eta) \in R^{n+1+m}} \frac{1}{2} \nu \|\eta\|^2 + \frac{1}{2} (\omega^T \omega + b^2) \\ \text{s.t. } D(A\omega - eb) + \eta = e \end{cases} \quad (2)$$

式中, A 为 $m \times n$ 训练集; ν 为权重因子; $\eta \in R^m$ 为松弛变量; $\omega \in R^n$ 为分界面的法向量; b 为偏置向量; D 是一个 $m \times m$ 的对角矩阵, 矩阵元素 $D_{ii} \in \{1, -$

$1\}$; e 为 m 阶都是 1 的向量。

如果在线性情况下的 PSVM, 将 ν 换成对角矩阵 V , 根据 KKT 条件, 取拉格朗日梯度为 0 求解式(2)可得拉格朗日乘子 $\lambda = (V^{-1} + D(AA^T + ee^T)D)^{-1}e$, 分类决策函数为:

$$x^T \omega - b \begin{cases} > 0 \text{ 则 } X \in A_+ \\ < 0 \text{ 则 } X \in A_- \\ = 0 \text{ 则 } X \in A_+ \text{ 或 } X \in A_- \end{cases} \quad (3)$$

其中, A_+ 为正训练样本集合; A_- 为负训练样本集合。

如果是在非线性情况下的 PSVM, 同理先将 ν 换成对角矩阵 V , 再用高斯核 $K(A, A^T)$ 取代 AA^T 对式(2)进行推导运算, 求解优化条件方程可得拉格朗日乘子 $\lambda = (V^{-1} + D(KK^T + ee^T)D)^{-1}e$, 分类决策函数为:

$$(K(x^T, A^T)K(A, A^T) + e^T)D\lambda \begin{cases} > 0 \text{ 则 } X \in A_+ \\ < 0 \text{ 则 } X \in A_- \\ = 0 \text{ 则 } X \in A_+ \text{ 或 } X \in A_- \end{cases} \quad (4)$$

3 证券指数预测算法

在证券市场上,既有稳定性,也有波动性。由于传统相关理论对证券市场预测皆存在缺陷,文中将以一种新的角度和方法对证券市场进行预测。为实现预测,首先以模糊集理论为指导建立模糊信息粒化模型,并用近似支持向量机的思想,将样本数据通过非线性变换,将低维映射到高维特征空间,在高维特征空间进行数据拟合,并做出回归预测,下面给出具体算法步骤描述:

Step1: 对数据进行初始化: 训练数据集 A , 模糊集分类矩阵 MH , 向量 $sv = []$ (支持向量初始化为 0), 类标 D , 高斯核参数 σ , 所需支持向量个数 nSV , 每次迭代次数 q 。

Step2: 读取证券数据文件, 对上证或者深证的时间序列数据进行预处理, 构造模糊粒子, 计算模糊参数 a, b, c 。

Step3: 为更好地应用 SVM 的性质, 取得良好预测效果, 利用高斯函数构造特征权矩阵 $V_i = \sum_{j=1}^N \exp(-\frac{\|x_i - x_j\|}{d/4})$ 。其中, $d \in [0.1, 0.4]$ 为邻域直径, N 表示 x_i 以 d 为直径的邻域点数量, 计算每个 V_i 的值, 以便代入目标函数

$\begin{cases} \min_{(\omega, b, \eta) \in R^{n+1+m}} \left\| \left(\frac{1}{2}V\right)^{1/2} \eta \right\|^2 + \frac{1}{2} (\omega^T \omega + b^2) \\ \text{s.t. } D(A\omega - eb) + \eta = e \end{cases}$ 进行数据处理。

Step4:利用拉格朗日函数和 KKT 条件求取优化参数,因为 PSVM 对比标准 SVM,其约束条件由不等式变成等式,可以采用步长迭代的方式从大量数据中选取某点使目标值最小,产生优化分类器。先进行外部循环 for $i = 1$ to nSV ,循环内部再分三小步:第一步从 A 中任意选出不包括支持向量 sv 中已经有数据行的 q 行数据,记做 A_1, A_2, \dots, A_q ;第二步再进行一个嵌套循环 for $j = 1$ to q ,for 的内部执行 $sv[] = A_j + sv[]$; $K = K(A, sv[])$,其中 $sv[] = A_j + sv[]$ 表示将 A_j 行加入到向量 sv 中。用 A_j 更新后对应的核矩阵 K ,代入目标函数为 $z_j = \|(\frac{1}{2}V)^{1/2}\eta\|^2 + \frac{1}{2}(\omega^T\omega + b^2)$ 的约束求解,再根据 PSVM 的决策函数,即上面推导的式(3)和式(4)进行寻优运算和分类判断;第三步使泛函取得最小值 $j^* = \arg \min_j(z_j)$,可得到相应的 $\eta, \omega^{(j^*)}, b^{(j^*)}$ 。

Step5:根据 j^* 求得最优分类面的支持向量 sv 、相应的法向量 ω 和偏置向量 b ,相关数据保存到模糊集分类矩阵 MH 中。

Step6:根据模糊集模型,训练数据,并对参数 a, b, c 做出回归预测。

4 实验结果与分析

4.1 数据平台

中国证券市场上证综指 2007 年曾一度冲高 6 000 多点后迅速回落,现文中提取以 2008 年 10 月 28 日的阶段最低点 1 664.93 点为起始点,到 2013 年 11 月 22 日为结束共 1 231 条上证综指的日线数据为基本数据,以上证指数每天的开盘价、收盘价、最高价、最低价建立粒化模型,按照文中的算法进行数据处理和分析,通过计算模型的参数实现对时间序列的拟合,对上证综指点数的变化做出预测。

4.2 实验运行与结果

一般情况,上证和深证一周交易五天,可将模糊粒子大小设定为 5,方便为研究上证交易周 K 线数据提供參考,其实只要对研究目标有需要,模糊粒子也可以设定为其他值,如 10,20,365 等,模糊粒子粒化度大小用 gd 表示。图 1 和图 2 分别是 $gd = 5$ 和 $gd = 10$ 时上证综指日线粒化图。

为研究证券指数走势,文中将上证指数每天的开盘、收盘、最高、最低、成交量、成交额等影响股指走势的数据作为输入向量,并将这些数据做成粒化模型。粒化模型采用抛物型,其中参数 a, b, c 分别对应模糊粒化数据图的 low、high 和 avg 值,分别表示日线数据波动变化范围的最小值、最大值和平均值。在数据处理上,首先为保证程序运行时的快速收敛,必须对上证数据进行归一化处理。文中将数据归一化到区间

$[160, 360]$,另外因为径向基函数(RBF)对非线性、高维数据具有良好的适应性,所以支持向量机选取 RBF 核,即 $K(x, y) = \exp(-\gamma \|x - y\|^2)$,核参数 γ 通过交叉验证法获取,即通过将参数的取值范围分成一系列小区间,然后再通过迭代求值的方法。通过这种方法获得核参数 γ 、惩罚因子 C 和其他参数。获得参数后,可计算对应的支持向量 sv 和对应的支持向量个数 nSV 和 q 值,然后可对粒化模型的参数进行预测。图 3 ~ 5 分别是对上证综指进行抛物线粒化的三个参数下

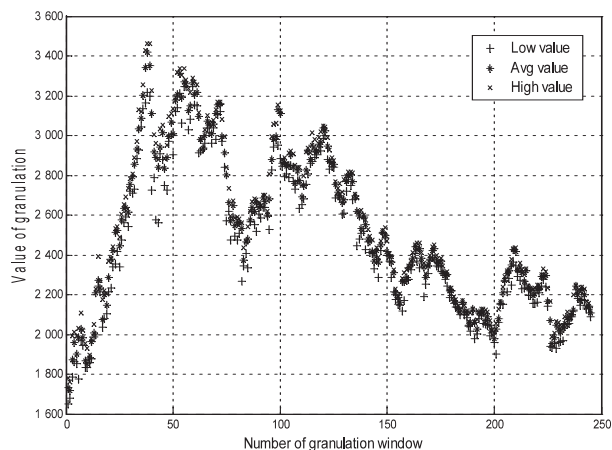


图 1 上证综指日线粒化图($gd = 5$)

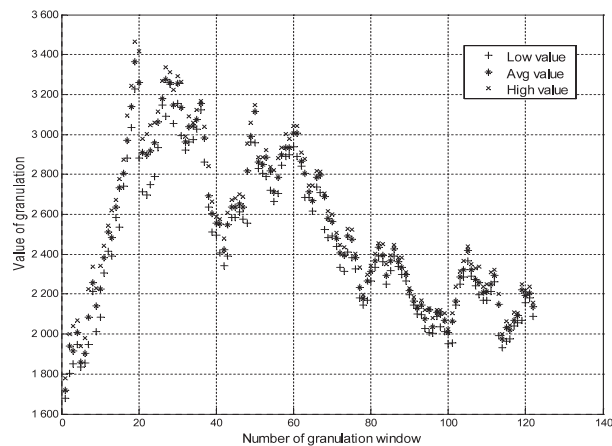


图 2 上证综指日线粒化图($gd = 10$)

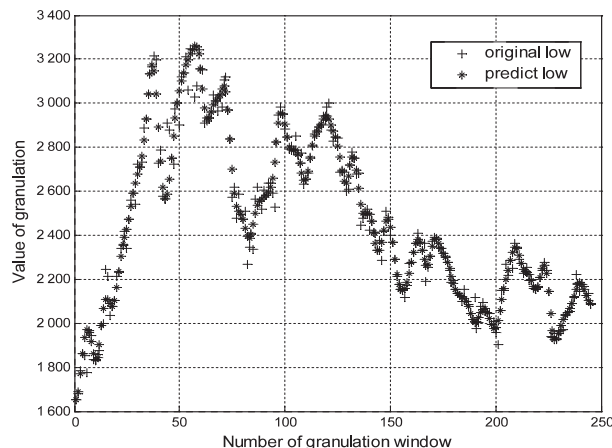


图 3 上证综指 low 的原始值和预测值

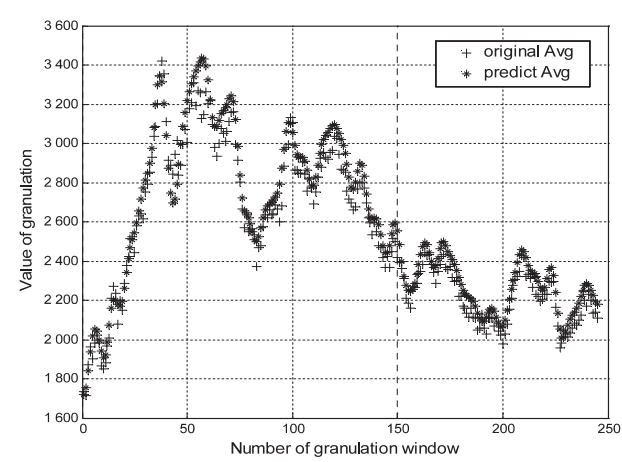


图4 上证综指 avg 的原始值和预测值

界 low、均值 avg 和上界 high 的回归图像,图中用符号“+”表示原始数据值,用符号“*”表示预测值。三个参数的窗口粒化度 gd=5。

4.3 实验预测与分析

对于粒化的三个参数,进行回归预测,可得到 low = 2 153.3, avg = 2 230.5, high = 2 272.8, 具体数据如表

表1 上证指数真实数据和预测数据表

波动范围	计算结果	交易日期	开盘	收盘	最高	最低
实际变化	(low, avg, high) = (2 107.03, 2 182.3, 2 213.97)	2013/11/18	2 147.37	2 197.22	2 198.32	2 143.77
		2013/11/19	2 198.32	2 193.13	2 203.30	2 186.11
		2013/11/20	2 201.47	2 206.61	2 207.11	2 186.64
		2013/11/21	2 197.01	2 205.77	2 206.53	2 177.74
		2013/11/22	2 206.47	2 196.38	2 211.07	2 188.77
预测变化	(low, avg, high) = (2 153.3, 2 230.5, 2 272.8)	2013/11/25	2 186.06	2 186.12	2 209.15	2 181.42
		2013/11/26	2 184.33	2 183.07	2 192.58	2 176.18
		2013/11/27	2 181.73	2 201.07	2 207.62	2 176.85
		2013/11/28	2 204.38	2 219.37	2 234.39	2 202.62
		2013/11/29	2 221.62	2 220.50	2 224.94	2 212.26

通过对下界 low 参数的回归预测,可得均方误差 MSE = 136.582,核参数 $\gamma = 0.062\ 5$,惩罚因子 $C = 128$ 。当 gd = 5 时,训练和预测所需时间为 217.191 3 s;当 gd = 10 时,训练和预测所需时间为 96.103 4 s。这是因为当 gd 越大,所产生的支持向量个数 nSV 越小,训练和预测所需要的时间就越少,即本实验时间复杂度 $T(\text{gd}) = T(n) \propto \frac{1}{\text{nSV}}$ 。以 low 为例计算预测与实际的误差,绘制成图表如图 6 所示。

从图中可看出误差主要集中在纵坐标轴上的 0 值附近。通过计算可得综合误差率 $p = \frac{1}{n} \sum_{i=1}^n |x_i - y_i| / x_i = 6.38\%$,其中 x_i, y_i 表示粒化的真实值和预测值。而从最近的预测更得到 low 的误差率 $p = 1 - 2\ 153.3 / 2\ 176.18 = 1.1\%$, high 的误差率 $p =$

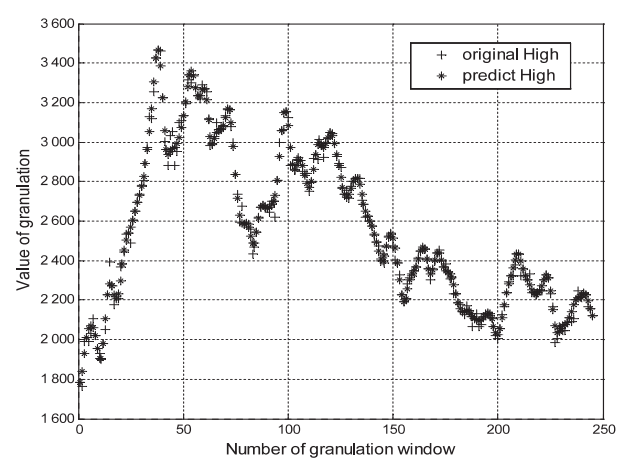


图5 上证综指 high 的原始值和预测值

1 所示。通过验证可知,在 2013 年 11 月份预测的 5 个交易日的相关指数变化范围与实际范围接近,而且上证指数的变化趋势是向上运行,预测结果可以为投资提供数据指导,在接近低点的时候买入,在接近高点的时候卖出。

1.7%,鉴于证券市场的复杂化,从误差率的值分析为一个可接受的预测结果。

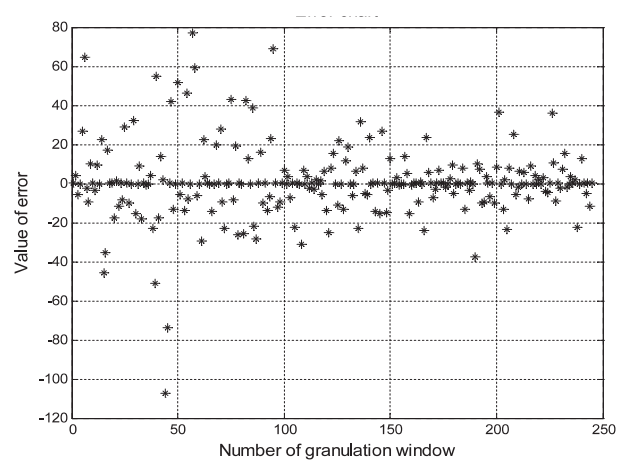


图6 low 预测与实际误差关系表

5 结束语

文中基于 2008 年 10 月至 2013 年 11 月的上证指数数据,研究关于证券指数的走势问题,提出一种将支持向量机和信息粒化理论相结合的回归预测方法模型。该方法能对下一个市场交易周期的指数波动范围进行预测,包括最高点、最低点和平均值,对比当天上证指数的实际运行数据,预测结果具有投资准确性。虽然该方法具有投资指导意义,但也存在增加可研究性和改进的地方,如模糊粒子类型的选择,高斯核参数和惩罚系数的选取都会影响预测值,此外中国证券市场指数的变化是非线性的,不但受一个国家经济运行和政策的影响,也受国际政治经济环境的制约,所以为达到综合有效的分析,其研究往往需要结合经济层面的其他各种因素和技术指标,如指数平滑移动平均线 MACD,指数平均数 EXPMA,心理线 PSY 等,去判断指数和市场的走向,以达到有效投资、实现扩大收益的目的。

参考文献:

- [1] 刘 星,迟建新,宿成建,等. 股票价格指数灰色系统预测与分析[J]. 数量经济技术经济研究,2003(8):128-131.
- [2] 吴朝阳. 改进的灰色模型与 ARMA 模型的股指预测[J]. 智能系统学报,2010,5(3):277-281.
- [3] 徐 枫. 股票价格预测的 GARCH 模型[J]. 统计与决策,2006(18):107-109.
- [4] 莫 扬,刘剑初. 上证指数的巨幅波动和单整类型的实证

检验[J]. 上海金融,2013(3):78-84.

- [5] 杨小平. 基于主成分与 BP 神经网络的股票价格预测分析[J]. 统计与决策,2004(12):42-43.
- [6] 孙 彬,李铁克,王柏琳. 基于股票市场灵敏度分析的神经网络预测模型[J]. 计算机工程与应用,2011,47(1):26-31.
- [7] Vapnik V N. The nature of statistical learning theory[M]. New York:Springer-Verlag,1995.
- [8] 孙名松,张立新,杜春燕. 增量支持向量机算法研究[J]. 计算机技术与发展,2011,21(5):40-43.
- [9] 辛治运,顾 明. 基于最小二乘支持向量机的复杂金融时间序列预测[J]. 清华大学学报:自然科学版,2008,48(7):1147-1149.
- [10] 刘道文,樊明智. 基于支持向量机股票价格指数建模及预测[J]. 统计与决策,2013(2):76-78.
- [11] Zadeh L A. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic[J]. Fuzzy Sets and Systems,1997,90(2):111-127.
- [12] Song Qiang, Chissom B S. Fuzzy time series and its models[J]. Fuzzy Sets and Systems,1993,54(3):269-277.
- [13] Fung G, Mangasarian O. Proximal support vector machine classifiers[C]//Proceeding of the 7th ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, CA, USA: ACM,2001:77-86.
- [14] Agarwal D K. Shrinkage estimator generalizations of proximal support vector machines[C]//Proceedings of the 8th international conference on knowledge discovery and data mining. Edmonton, Canada: [s. n.], 2002:173-183.

(上接第 147 页)


- 技术,2009,33(8):1-7.
- [2] 钟 金,郑睿敏,杨卫红,等. 建设信息时代的智能电网[J]. 电网技术,2009,33(13):12-18.
- [3] Wang Wenye, Xu Yi, Khanna M. A survey on the communication architectures in smart grid[J]. Computer Networks,2011,55(15):3604-3629.
- [4] 余贻鑫. 智能电网的技术组成和实现顺序[J]. 南方电网技术,2009,3(2):1-5.
- [5] 赵鸿图,周京阳,于尔铿. 支撑高效需求响应的高级量测体系[J]. 电网技术,2010,34(9):13-20.
- [6] 刘 念,张建华. 互动用电方式下的信息安全风险与安全需求分析[J]. 电力系统自动化,2011,35(2):79-83.
- [7] 王思彤,周 晖,袁瑞铭,等. 智能电表的概念及应用[J]. 电网技术,2010,34(4):17-23.
- [8] Kushalnagar N, Montenegro G. IPv6 over low-power wireless personal area networks: assumptions, problem statement, and goals[S/OL]. [2013-01-28]. <http://tools.ietf.org/html/rfc4919>.

- [9] Yan Ye, Qian Yi, Sharif H. A secure and reliable in-network collaborative communication scheme for advanced metering infrastructure in smart grid[C]//Proc of 2011 IEEE wireless communications and networking conference. Cancun, Quintana Roo; IEEE,2011:909-914.
- [10] Berthier R, Sanders W H, Khurana H. Intrusion detection for advanced metering infrastructures: requirements and architectural directions[C]//2010 first IEEE international conference on smart grid communications. Gaithersburg, MD; IEEE,2010:350-355.
- [11] 苏 忠,林 闯,封富君,等. 无线传感器网络密钥管理的方案和协议[J]. 软件学报,2007,18(5):1218-1231.
- [12] 国家电网公司. 国家电网智能化规划总报告[R]. 出版地不详:国家电网公司,2010.
- [13] 中电联. 2012 年经济形势与电力发展分析预测会[C]. 出版地不详:中电联,2012.
- [14] 国务院节能减排十二五规划[R]. 出版地不详:出版者不详,2012.

基于支持向量机粒化的证券指数预测

作者：[陈孝全](#)，[刘波](#)，[CHEN Xiao-quan](#)，[LIU Bo](#)

作者单位：[陈孝全, CHEN Xiao-quan\(华南师范大学 计算机学院, 广东 广州 510631; 深圳博物馆网络中心, 广东 深圳 518026\)](#)，[刘波, LIU Bo\(华南师范大学 计算机学院, 广东 广州, 510631\)](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2015(4)

引用本文格式：[陈孝全](#). [刘波](#). [CHEN Xiao-quan](#). [LIU Bo](#) [基于支持向量机粒化的证券指数预测](#)[期刊论文]-[计算机技术与发展](#) 2015(4)