机器学习算法在股票走势预测中的应用

傅航聪[1,2]，张 伟[1,2]

（1. 上海理工大学 光电信息与计算机工程学院；2. 上海市现代光学系统重点实验室，上海 200093）

摘 要：结合K-近邻算法、支持向量机算法和时间序列算法的优点，整合其结果提出一种综合预测算法，并将其应用到沪深300指数的涨跌预测中。首先通过时间序列预测出股票未来一段时间内的走势图，然后结合K-近邻算法判断该走势图的总体涨跌趋势，最后将涨跌趋势量化作为一变量引入到支持向量机算法中，通过支持向量机算法预测最终的股票涨跌。该方法能够弥补3种算法的不足，能够更加准确地预测股市的涨跌趋势。

关键词：股票预测；K-近邻算法；支持向量机；时间序列

DOI：10.11907/rjdk.171549

中图分类号：TP312 **文献标识码**：A **文章编号**：1672-7800(2017)010-0031-04

The Application of Machine Learning Algorithms in Stock Movements Forecasting

FU Hang-cong[1,2], ZHANG Wei[1,2]

(1. *School of Optical Electronics and Computer Engineering, University of Shanghai for Science and Technology*; 2. *Shanghai Key Lab of Modern Optical SystemEngineering Research Center of Optical Instrument and System, Ministry of Education, Shanghai* 200093, *China*)

Abstract: This article is based on the K-nearest neighbor (KNN) algorithm, support vector machine (SVM) algorithm and time sequence algorithm's advantages and then integrate the results put forward a comprehensive prediction algorithm, and apply it to the outcome of the CSI 300 index prediction. First of all through the time series to predict stock over a period of time in the future charts, then concludes that the graph by combining KNN algorithm the overall trend of rise and fall of the up or down. Finally we use the fall trend of quantitative as a variable to beintroduced into the support vector machine (SVM) algorithm, by support vector machine (SVM) algorithm to predict the final stock price. In this paper, the proposed method can make up for the shortage of the three algorithms, can more accurately predict the trend of rise and fall of the stock market.

Key Words: stock; prediction; KNN; support vector machine (SVM); time series

0 引言

随着经济的快速发展，股票市场受到了投资者的普遍关注，掌握股市变化规律并预测其走势，一直是投资者和投资公司关注的热点，对于预测股市未来收益方法的探索也从来没有中断过。由于涉及到金融领域，因此最开始仅从金融工程、数理统计方面进行挖掘，探索相关方法，在文献[1]-[3]中提到了其中的部分金融模型与方法。然而，由于发现金融模型一定程度上无法满足人们对于预测准确性的要求，因此对于金融模型而言，更多地是建立在假设上。之后人们提出了机器算法模型，其实践性更强，适用范围更广，模型准确率更高[4-6]。至21世纪初期，金融市场发展迅速，同时也衍生出了更多金融产品。因此，单一的金融模型算法和机器学习算法已经满足不了市场需求，人们便将金融模型算法与机器学习算法结合起来，以达到更好的效果[7-8]。对于金融工程方面的股市量化预测分析，机器学习算法的优点[9]是能够最大程度地模拟对象的具体特征，另外在处理数据量及复杂度方面也有更大优势。另外一种预测手段是结合多种算法，能够一定程度上弥补单独算法存在的缺陷。

本文采用的即是机器学习和金融模型算法相结合的综合算法，其中包括K-近邻算法、支持向量机算法(SVM)、时间序列算法。K-近邻算法主要应用于分类方

收稿日期：2017－05－03

基金项目：国家自然科学基金项目(11502145)

作者简介：傅航聪(1994－)，男，浙江金华人，上海理工大学光电信息与计算机工程学院硕士研究生，研究方向为机器学习算法在金融市场的应用；张伟(1985－)，男，山东临沂人，博士，上海理工大学光电信息与计算机工程学院讲师，研究方向为最优控制、智能控制、飞行控制。

面，通过近邻算法将相似的样本归为一类；SVM 支持向量机算法可以有效解决神经网络无法避免的局部最小化问题，而且对于小样本容量、非线性及高维数模式，以及在克服维数过大和过度拟合学习方面具有优势[9]；时间序列算法能够很好地展示一定时间内事物的发展变化趋势与规律，从而对未来的变化进行有效预测。

本文将通过沪深 300 数据，结合 K-近邻算法、支持向量机算法(SVM)、时间序列算法的优点，以达到最好的股指预测效果。本文首先结合近邻算法和时间序列算法，得到一个未来趋势的涨跌变量，通过图像处理将图片转换成二维码数字，用近邻算法分析识别时间序列预测的曲线图是涨还是跌；然后将该变量输入到支持向量机算法中，得到最终的综合算法；最后与单一的 SVM 算法、时间序列算法进行比较，得到最终的准确率。

1 理论基础

1.1 KNN 算法

在 KNN 算法中，样本集里的每个样本对应着自己的分类标签。在计算机中输入没有分类标签的数据，通过与样本集的比对，提取与样本中特征最接近的分类标签。通常是提取前面 K 个样本，选择在 K 个样本中出现次数最多的分类作为新数据的分类，也即是要检验的样本分类[10-11]。

1.2 支持向量机(SVM)

支持向量机(SVM)技术有着坚实的统计学理论基础，并在许多实际应用中展示了很好的实践效果。SVM 很大程度上解决了以往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[12]。

SVM 是针对包含 N 个训练样本的二元分类问题。在文献[13]中，每个样本表示一个二元组$(X_{y_i}, y_i)(i=1,2,\cdots,N)$，其中 $X_i=(x_{i1},x_{i2},\cdots,x_{id})^T$ 对应于第 i 个样本的属性集。为方便计算，令 $y_i\in\{-1,1\}$ 表示它的类标号。一个线性分类器的决策边界可以写成如下形式：

$$W\cdot X+b=0 \tag{1}$$

SVM 最终的学习任务可以描述为以下被约束的优化问题：

$$\min_w=\frac{\|w\|^2}{2} \tag{2}$$

其受限于：

$$y_i(W\cdot X_i+b)\geqslant 1, i=1,2,\cdots,N \tag{3}$$

为了防止对模型过分拟合，可以引入正值的松弛变量 ξ 来实现放松不等式约束，以适应非线性可分数据，如下所示：

$$W\cdot X_i+b\geqslant 1-\xi_i \quad \text{如果 } y_i=1 \tag{4}$$

$$W\cdot X_i+b\leqslant -1+\xi_i \quad \text{如果 } y_i=-1 \tag{5}$$

其中，$\forall\,\xi_i>0$。

为了避免造成误分训练实例，必须修改目标函数，以惩罚那些松弛变量很大的决策边界，其目标函数如下：

$$f(w)=\frac{\|w\|^2}{2}+C(\sum_{i=1}^{N}\xi_i)^k \tag{6}$$

其中 C 和 k 是用户指定的参数，表示对误分训练实例的惩罚，为了简化该问题，假定 $k=1$。参数 C 可以根据模型在确认集上的性能进行选择。

因此，被约束的优化问题拉格朗日函数可以表示为以下形式：

$$L_p=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i-\sum_{i=1}^{N}\lambda_i\{y_i(W\cdot X_i+b)-1+\xi_i\}-\sum_{i=1}^{N}\mu_i\xi_i \tag{7}$$

其中，前两项是需要最小化的目标函数，第 3 项表示与松弛变量相关的不等式约束，而最后一项要求 ξ_i 的值是非负结果。此外，利用如下的 KKT 条件，可以将不等式约束变换成等式约束：

$$\xi_i\geqslant 0,\lambda_i\geqslant 0,\mu_i\geqslant 0 \tag{8}$$

$$\lambda_i\{y_i(W\cdot X_i+b)-1+\xi_i\}=0 \tag{9}$$

$$\mu_i\xi_i=0 \tag{10}$$

令 L 关于 W、b 和 ξ_i 的一阶倒数为零，则可得到如下公式：

$$W=\sum_{i=1}^{N}\lambda_i y_i x_{ij} \tag{11}$$

$$\sum_{i=1}^{N}\lambda_i y_i=0 \tag{12}$$

$$\lambda_i+\mu_i=C \tag{13}$$

将公式代入到拉格朗日函数中，得到如下对偶拉格朗日函数：

$$L_D=\sum_{i=1}^{N}\lambda_i-\frac{1}{2}\sum_{i=1}^{N}\lambda_i\lambda_j y_i y_j x_i \tag{14}$$

另外在公式中 λ_i 不应该超过 C，因此其被限制在 $0\leqslant\lambda_i\leqslant C$。然后使用二次规划技术，对对偶问题的数值求解，得到拉格朗日乘子 λ_i。可以将这些乘子代入公式和 KKT 条件中，从而得到决策边界的参数。

另外最后结果可能涉及到空间向量对的点积，该运算相当复杂，可能导致维数灾难问题，因此这里引入了核函数[14]。

1.3 时间序列

由于股票价格随着时间不断发生变换，而且有着非常明显的时间先后顺序。这种按照时间序列排列起来的一系列观测值称为时间序列，因此可以认为时间序列包含了某一个或某几个统计指标特征。一个时间序列里一般包含 4 种信息：长期趋势、循环、季节变换、不规则变换。常用的时间序列模型主要有自回归模型(AR)、移动平均模型(MA)、自回归移动平均模型(ARMA)、齐次非平稳模型(ARIMA)等。时间序列模型又分为传统时间序列模型和现代时间序列模型。传统时间序列模型把时间序列看成是长期趋势、季节变动、循环变换、不规则变换的复合体。现代时间序列模型则将时间序列看作一个随机概率过程，ARIMA 模型算法即是这类模型的代表[15]。

下面将结合 3 种单独的算法得到最终的综合算法。

2　综合算法预测模型

2.1　算法模型概括

综合算法预测模型主要分为 3 部分，首先在相同的数据条件下分别计算出支持向量机算法、时间序列算法与综合算法的准确率，然后对其进行比较。具体步骤为：①首先对数据进行标准化处理；②经过归一化处理，分别计算出在支持向量机算法、时间序列算法下的准确率；③按照相同像素大小的图片，随机截取 20 个间隔为 30 天的沪深 300 历史涨跌图作为样本集。将 20 个曲线图分为两类，即涨和跌，每一类的样本数量为 10；④按照支持向量机选取的变量日期，画出时间序列未来 30 天的预测曲线图，按照相同的像素大小截取预测曲线；⑤通过近邻算法进行比对，识别出时间序列未来 30 天的曲线图是涨还是跌，并将涨定义为 1，跌定义为－1；⑥将通过近邻算法得到的结果作为其中一个支持向量机变量输入其中，最后计算出综合算法的预测准确率；⑦比较支持向量机算法、时间序列算法和综合算法的准确率，得到最后的结论。

2.2　综合算法原理

本文中的综合算法用到了图像处理技术以及时间序列算法、近邻算法、支持向量机算法。

对于图像处理技术而言，因为在股票市场某一阶段并不能通过肉眼准确地判断涨跌，因此将图像处理技术和机器学习算法相结合能够避免肉眼带来的局限性，同时还避免了重复工作带来的错误率。通过 MATLAB 软件对彩色图片进行二值化处理，使图片只显示黑白两色。通过使用 MATLAB 里的 imread()函数读取需要操作的图片，然后调用 MATLAB 中的二值化处理函数使图片转换为二值化图片，并且输出二值化图片的二进制数据。二值图像也称为黑白图像，因为其中的每个像素都只有黑白两色，并且没有中间的过渡颜色，其中输出的二进制矩阵中的 0 和 1 也代表了二值图像的黑白值，其中黑色像素块代表 0，白色像素块代表 1。

按照相同的像素大小截取涨跌图片，得到沪深 300 涨跌图如图 1 所示。图中随机从样本集中挑选了 2 张截取的图片，一张代表上涨趋势，一张代表下跌趋势。然后将两张图对应的黑色方格计为 1，白色方格计为 0。算法中按照相同像素大小的图片随机截取 20 个间隔为 30 天的沪深 300 历史涨跌图作为样本集，将 20 个曲线图分为两类，即涨和跌，每一类的样本数量为 10。然后将这 20 个样本集图片转化为二值图片后，再通过程序以二进制矩阵的数据形式输出，并且贴上“涨”、“跌”两类标签分别保存。

通过时间序列算法预测未来 30 天的股市涨跌，然后再截取相同像素的涨跌图片，进行二值化并将其转换为二进制数据矩阵。相对于其它算法，时间序列算法具有能反应未来股市涨跌趋势的功能，并且能够最大限度地把握股市周期性。之后通过截取时间序列预测图，将其转化成二进制数据，作为 K-近邻算法的测试集。

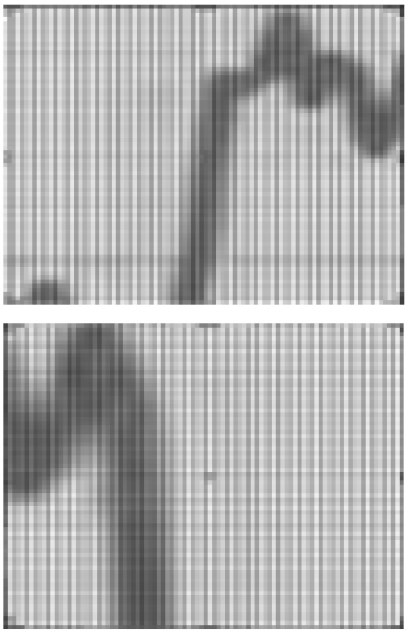

图 1　沪深 300 涨跌图

之后通过 K-近邻算法训练样本集，然后输入测试集判断该组二进制矩阵数据是涨还是跌。将涨的状态定义为 1，将跌的状态定义为－1，储存作为支持向量机的输入量之一。通过 K-近邻算法计算测试集和样本集数组之间的距离，以此区分测试集表示的状态是涨还是跌，然后将得到的实验结果作为支持向量机的输入变量。同时采用支持向量机的其它变量，即沪深 300 在 2015～2016 年间循环每前 3 天的日收盘价和日交易量，预测第 4 天的股指涨跌比率[16]，因为在这样的条件下得到的预测结果更准确，实验效果更好。得到了所需的变量数据后，分别对其进行归一化处理，使实验结果更加准确。

因此对于 SVM 算法而言，主要有 3 个输入参数，第一个参数是通过时间序列预测的涨跌数据，第二个是提前 1、2、3 天的日收盘价，第三个是提前 1、2、3 天的日交易量。将训练集单日的涨跌状态定义为 1 或－1。然后通过算法计算测试集，通过测试集得到的结果计算该综合算法的最终准确率。

3　实验结果及讨论

3.1　图像识别处理

在综合算法中，首先需要通过 MATLAB 软件对图片进行二值化处理，得到只有黑白像素的图片，然后得到二值图片通过二进制表示的矩阵数据，其中黑色表示“1”，白色表示“0”。其最终的实验结果如图 2 所示，图中包含了彩色图片与二值图片。

3.2　SVM 算法仿真结果

本文通过使用沪深 300 指数 2015～2016 年的数据输入向量为提前 1、2、3 天的日收盘价和日交易量来预测第 4 天的股指涨跌比率[16]。其过程通过 SPSS Modeler 来实

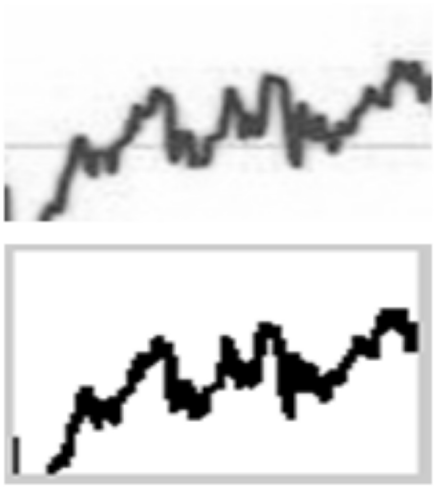

图 2 图像处理结果

现最终的算法，并且将数据的 60％作为训练集，剩下的 40％作为测试集。将训练集单日的涨跌状态定义为 1 或 -1，其最终的预测准确率如表 1 所示。因为只是单一的 SVM 算法，对核函数非常敏感而且更加适合处理二类问题，并且里面没有整合股市的周期性规律变换。因此 SVM 仿真预测股票涨跌的率结果并不是非常理想，约 67.5％。

表 1 SPSS Modeler SVM 仿真结果对比

分区	1_训练	占比(％)	2_测试	占比(％)
正确	230	89.15	116	67.52
错误	28	10.85	56	32.84
总计	258		172	

3.3 时间序列仿真结果

时间序列算法在进行运算前，首先要对原型图进行平稳化，并且去周期化。在 SPSS Modeler 里会自动建模，选择最佳的一阶差分次数和一阶季节差分次数。本文通过输入沪深 300 在 2015 年的收盘价、开盘价交易量来预测最后的股市。本文没有考虑引进季节差分，通过软件自动得到 ARIMA(1,0,2)模型。

图 3 为按日为周期时间序列训练预测结果图。由于股票交易存在周末与节假日，对于以日为周期，需要重新编排日期，因此跳过了周末，相比于按月为周期，其最终日期会提前。由图可见，时间序列的训练效果较好。从表 2 分析结果中也能看到，其相关性为 99.1％左右。

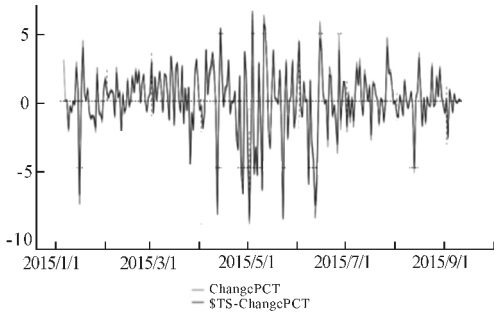

图 3 按日为周期时间序列训练预测结果

最终预测结果的准确率为71.4％，虽然训练集表现

表 2 时间序列训练结果

最小误差	最大误差	平均误差	绝对平均误差	标准差	线性相关	发生率
-1.216	1.749	0.032	0.215	0.339	0.991	249

效果较好，但预测集的效果不太令人满意。而且这里的准确率指的是涨跌概率，并没有精确到涨跌比率。其中可能的原因是模型没有加入季节性，而且参数的输入可能过少，不能完全反映股市的整体现状。对于时间序列算法而言，完全依据历史数据，对于未来的预测范围将缺乏一定准确性，并且其中的随机性因素也较大。但是时间序列算法的优势是可以大致研判未来股市的总体周期趋势。因此，时间序列算法还需要与其它算法相结合，才能更好地提高预测准确率。

3.4 综合算法仿真结果

按照综合算法原理，随机选取 20 张涨跌图，分别转换为二进制矩阵数据，并且分为两类，分别是涨和跌。然后通过时间序列预测后面一个月的图，最后一个月即是预测的图像，然后不断往后推一天，通过 SPSS 软件保存预测得到的未来一个月涨跌图。然后通过图像处理得到二值图像，并且输出保存每张图像的二进制矩阵数据。SVM 算法每一组数据都对应预测一个月未来涨跌的图像。

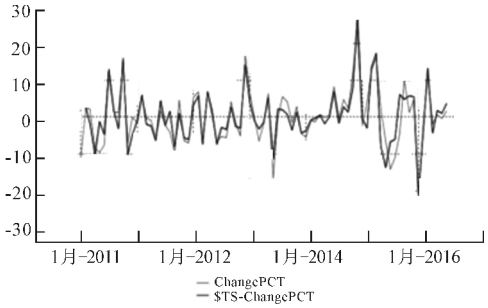

图 4 按月为周期的时间序列结果

得到 KNN 算法所需的样本集和测试集，便可以调用 KNN 算法判断每一个测试集最终的类别是涨还是跌，然后将其保存作为 SVM 算法的输入变量。最后支持向量机算法需要输入的参数便是 1、2、3 日的日收盘价和日交易量，结合时间序列算法和 KNN 算法预测其未来一个月的涨跌趋势，并和之前的 SVM 算法一样，将其 60％的数据作为训练集，剩下的 40％作为测试集，只是在其中添加了一组变量参数，总体框架并没有改变。综合算法的最终仿真结果如表 2 所示。

表 3 通过 SPSS Modeler 综合算法仿真结果对比

分区	1_训练	占比(％)	2_测试	占比(％)
正确	244	94.57	131	76.16
错误	14	5.43	41	23.84
总计	258		172	

表 3 得到了综合算法、单独 SVM 算法以及单独时间序列算法的准确率。然后对 3 种算法进行比较，结果如表 4 所示。

（下转第 46 页）

Complex Sciences. Springer Berlin Heidelberg,2009:657-670.
[4] LUONG HIEP,HUYNH TIN,GAUCH SUSAN,et al. Exploiting social networks for publication venue recommendations[C]. Proceedings of the International Conference on Knowledge Discovery and Information Retrieval,KDIR, Barcelona,2012:239-245.
[5] MEDVET ERIC,BARTOLI ALBERTO,PICCININ GIULIO. Publication venue recommendation based on paper abstract[C]. Proceedings of International Conference on Tools with Artificial Intelligence,ICTAI,Limassol Cyprus,2014:1004-1010.
[6] ANAS ALZOGHBI, VICTOR ANTHONY ARRASCUE AYALA, PETER MFISCHER, et al. PubRec: recommending publications based on publicly available meta-data[C]. Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB, Trier, Germany,2015:11-18.
[7] TRAN HUNG NGHIEP, HUYNH TIN, HOANG KIEM. A potential approach to overcome in scientific publication recommendation[C]. Proceedings of 2015 IEEE International Conference on Knowledge and Systems Engineering, Ho Chi Minh City, 2015: 310-313.
[8] HUYNH TIN, NGUYENTRAC-THUC, TRAN HUNG-NGHIEP. Exploiting social relations to recommend scientific publications [J]. Lecture Notes in Computer Science,2016,9795:182-192.
[9] 徐镇.基于垂直搜索引擎的论文投稿推荐系统研究[D].兰州:兰州大学,2010.
[10] 秦臻.学术社会网络建模和学术资源推荐方法研究[D].北京:北京邮电大学,2015.
[11] 王亮,张绍武.基于主题模型的文本挖掘的研究[D].大连:大连理工大学,2015.
[12] PORTER BW, BARESS E R, HOLTE R. Concept learning and heuristic classification in weak theory domains[J]. Artificial Intelligence,1989,45:229-263.
[13] QUINLAN J R. Induction of decision trees[J]. Machine learning, 1986(1):81-106.
[14] 王小巍,蒋玉明.决策树 ID3 算法的分析与改进[J].计算机工程与设计,2011,32(9):3069-3076.
[15] 黄爱辉,陈湘涛.决策树 ID3 算法的改进[J].计算机工程与科学,2009,31(6):109-111.
[16] 刘任任,欧阳建权.多值逻辑函数结构理论研究[M].北京:科学出版社,2010:2-20.
[17] 陶维丽.科技期刊的综合评价比较研究[D].武汉:华中师范大学.2012:1-30.
[18] 孟晓明,陈慧萍,张涛.基于 WEKA 平台的 Web 事务聚类算法的研究[J].计算机工程与设计,2009,30(6):1332-1334.

(责任编辑:孙 娟)

(上接第 34 页)

表 4 综合算法实验结果比较

	1	2	3	4	5	6	7	…	准确率(%)
SVM 算法	涨	涨	跌	跌	涨	涨	跌	…	67.5
时间序列	涨	跌	跌	涨	涨	跌	涨	…	71.4
综合算法	涨	跌	跌	涨	涨	跌	涨	…	77.6
正确集	涨	涨	跌	涨	跌	跌	跌	…	100

从实验结果可以看出,单独的 SVM 算法和时间序列算法得到的准确率并不是很高,普遍只有 70%左右,而综合算法的准确率接近 80%。因此,相比于单一算法,综合算法的效果更好。因为综合算法结合了时间序列和 SVM 算法的优点,使 SVM 算法能够考虑到时间序列上的时间特性比如季节性等,又结合了 SVM 算法的优点,即其自身的错误率较低,而且计算开销小,适合运行大批量数据,得到的结果也更加直观具体。

4 结语

本文结合 K-近邻算法、支持向量机算法和时间序列算法的优点,提出了一种综合预测算法,并将其应用到沪深 300 指数的涨跌预测中,取得了较好效果。然而,综合算法虽然相比于单一算法,准确率有改善,但是未来提升的空间还有很大,需要通过发掘探索,不断组合一些更高效的算法,以得到更高的准确率。

参考文献:

[1] 何永沛. ARMA 模型参数估计算法改进及在股票预测中的应用[J].重庆工学院学报:自然科学版,2009(2):6-8.
[2] 于志军,杨善林.基于误差校正的 GARCH 股票价格预测模型[J].中国管理科学,2013(S1):28-32.
[3] 费时龙,任洪光.多重马氏链模型在股市预测中的应用[J].德州学院学报,2016(8):28-35.
[4] 王领,胡扬.基于 C4.5 决策树的股票数据挖掘[J].计算机与现代化,2015(10):38-51.
[5] 张鹏.基于 SVR 的股市预测与择时研究[J].重庆文理学院学报,2016(3):148-155.
[6] 孙海波,王丽敏.引入趋势因子的 BP 模型在股市预测中应用[J].统计与决策,2015(19):87-89.
[7] 李雁.基于 ARIMA 模型和神经网络模型的股票价格预测[J].金融商务,2014(4):77-80.
[8] 刘海玥,白艳萍.时间序列模型和神经网络模型在股票预测中的分析[J].管理科学,2011(4):22-31.
[9] 金得宝.基于支持向量机的股市预测研究[D].杭州:浙江大学出版社,2010:2-10.
[10] PETER HARRINGTON.机器学习实战[M].北京:人民邮电出版社,2013:15-20.
[11] 王波,程福云. KNN 算法在股票预测中的应用[J]. 武汉:科技创业月刊,2015(16):10-15.
[12] 张晨希,张燕平.基于支持向量机的股票预测[J].计算机技术与发展,2006,16(6):34-35.
[13] PANG NING TAN, MICHAEL STEINBACH.数据挖掘导论[M].北京:人民邮电出版社,2011:157-170.
[14] BURGERS B C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery ,1998,2(2):121-167.
[15] 张文彤,董伟. SPSS 统计分析高级教程[M].北京:高等教育出版社,2013:393-395.
[16] 林利敏.基于支持向量机的股价短期预测研究[D].杭州:杭州电子科技大学,2010:45-50.

(责任编辑:黄 健)