

关于股票价格优化预测的建模仿真研究*

陈 远^{1,2}, 罗必辉¹, 蒋维琛¹, 孙宏伟³

(1.重庆大学 软件学院,重庆 400044;2.数字影视艺术理论与技术重庆市重点实验室,重庆 400044;
3.四川航天职业技术学院,四川 成都 610100)

摘要:股票数据具有非线性和含有大量噪声的特点,传统股票预测模型难以充分识别股票非线性特征以及降低噪声,导致预测精度不高.为了提高预测精度,去除冗余特征并加强特征的区分度,引入流形学习中的线性局部切空间排列算法,提出了一种新的支持向量回归机的股价预测优化模型.首先利用线性局部切空间排列算法对股票原始数据进行特征提取,然后采用支持向量回归机对提取到的特征和股票价格之间的非线性关系建模,并利用遗传算法优化支持向量回归机的参数,最终提高股票价格的预测精度.为证明模型的有效性,采用标准普尔 500 指数在 2012—2013 年、2014—2015 年 2 个时间段内的股票数据进行检验.实验证明,提出的模型相较于其他对比模型具有更高的预测精度,更强的泛化能力.

关键词:股票预测;遗传算法优化;线性局部切空间排列;支持向量回归机

中图分类号:TP 391.9 **文献标志码:**A **文章编号:**0258-7971(2016)04-0536-07

现如今越来越多人选择将资产投入股票交易市场,期望获得高额回报.但股票价格的变化具有非线性、时变性和高度不稳定性等特点,因此难以预测,使得广大股民在收获高收益时也要承担高风险.因此一个好的预测方法对于股民能否在收获高收益的同时尽量降低风险,减少亏损显得尤其重要.

股票预测的方法中传统的基本面分析法、技术分析法^[1]和时间序列分析法^[2]准确率低,不适合处理具有非线性特征的股票数据.为了提高预测精度,机器学习方法被运用到股票预测领域,其中人工神经网络(Artificial Neural Network, ANN)和支持向量机(Support Vector Machine, SVM)应用最为广泛.由于本身固有的非线性网络结构,ANN 算法非常适合用于捕捉具有非线性特征数据的内在结构.但该方法需要的训练集数据量大,容易陷入局部极小、且收敛速度慢,易造成过度拟合,导致预测结果失真.SVM 收敛速度快且具有全局最优解,正好克服了 BP 神经网络的缺点,而支持向量回归机(Support Vector Regression, SVR)是 SVM 在函数回归领域的应用,现已广泛被应用于股票预测.

由于单一的模型在预测股票时只能反映股票局部信息,往往效果不是最优.在此基础上, Gocken^[3]提出将和声搜索算法与 ANN 算法结合,克服了 ANN 模型过拟合和局部极小的缺点.吴春国等^[4]将正规正交分解方法与 RBF 神经网络结合,考虑了时间序列数据之间的时序关系,使得模型具有更好的预测精度和收敛稳定性. Talarposhti 等^[5]利用学习自动机群智能算法优化传统模糊时间序列的间隔,提出了指数模糊时间序列的组合方法,得到比传统模糊时间序列更好的预测效果.尽管组合模型在一定程度上弥补了单个模型对股票数据非线性特征刻画的不足,但是股票数据含有大量噪声和冗余信息,属性之间也有较强相关性,特征之间区分度不高,导致模型容易陷入过拟合. Shi^[6]在此基础上引入主成分分析(PCA)对股票数据进行预处理,并与因果神经网络相结合预测股票,缩短了训练时间,提高了预测精度. Kao 等^[7]则引入小波变换对股票数据进行特征提取,并利用多变量自适应回归(MARS)方法解决小波子序列的选择问题,最后与支持向量回归机(SVR)结合预测股票,

* 收稿日期:2016-02-22

基金项目:国家自然科学基金(61379158).

作者简介:陈 远(1976-),男,湖北人,博士,硕士生导师,主要研究领域为数字媒体、模式识别等.E-mail:455036550@qq.com.

得到比传统小波方法更高的预测精度.苏治^[8]引入核主成分分析(KPCA),结合SVR模型预测股票收益率用于选股,并利用遗传算法(GA)优化SVR模型的训练参数.PCA方法通过将数据样本集线性映射到低维空间来实现降维的效果,但只能提取到数据二阶的统计信息,忽略了数据点之间的非线性相关性.而KPCA虽然提取的是非线性相关的特征,但其核矩阵的大小与数据样本大小的平方成正比,在数据样本数量很大的情况下造成核矩阵过大,从而导致计算效率低下.除此之外,这些传统特征提取方法只能发现全局的欧式结构,不能进一步发掘数据隐藏的模式.流行学习方法将欧式空间扩展至流形空间,并将数据样本集看成嵌入在高维空间中的一个低维流形,通过将原始数据样本点投影到低维特征空间并同时保留原始数据的几何结构,使得其十分擅长发现具有非线性特征的高维数据集中固有的结构模式^[9-10].本文在此基础上,将股票数据集看成一个潜在的非线性流形,在股票预测领域第一次引入流形学习方法中的线性局部切空间排列算法(LLTSA)对股票数据集进行降维去噪,并与遗传算法优化的SVR相结合,提出了一种新的股价预测优化模型.相对流形学习方法中的传统特征提取算法而言,LLTSA具有更加优异的聚类性能,使得提取的特征更加具有区分度和代表性^[11],从而提高模型的预测精度.

1 股票价格预测原理

假设影响股票价格波动的一组因素为 X ,且与股票真实价格 Y 有对应关系 $Y = \{X_1, X_2, \dots, X_n\}$,则股票预测任务为对 X 和 Y 之间的非线性关系建模,预测模型的公式如下:

$$Y' = f(X_1, X_2, \dots, X_n). \quad (1)$$

预测的最终目标是使预测值 Y' 尽可能接近实际值 Y ,从而实现对未来股票价格的预测,以此作为投资者进行交易决策的参考.

由于股票价格波动受到政治、经济、军事、上市公司自身盈利情况、投资者心理等多方面因素的影响,具有时变性、非线性和不稳定等特点,传统的预测方法主要依靠个人经验,可靠性不足且效率低下.人工神经网络算法又容易陷入局部极小点且收敛速度慢,导致预测精度难以满足投资者的需求.支持向量机基于结构风险最小化原则,非常适合处理小样本非线性问题,且有全局最优解,能用于对

股票数据的非线性特征进行建模.然而股票历史价格数据和影响股价波动的因素等都含有大量噪声和冗余,因此需要对原始数据进行降噪和提取特征,加强特征区分度,从而提高预测精度.本文引入线性局部切空间排列算法对股票原始数据进行特征提取,采用支持向量回归机对提取到的特征和股票价格之间的非线性关系建模,并利用遗传算法优化支持向量回归机的参数,最终提高股票价格的预测精度.

2 股票价格预测模型

2.1 线性局部切空间排列算法 线性局部切空间排列(LLTSA)是一种基于流形学习的数据集特征提取与降维方法,由局部切空间排列(LTSA)算法发展而来.LLTSA相比LTSA增加了PCA预处理数据的步骤,以此降低原始数据的噪声,随后通过逼近每个数据点邻域的切空间来描述数据集的局部几何结构,再将这些局部切空间进行全局排列,得到整体低维嵌入坐标,完成高维原始数据到低维空间的线性映射,从而实现数据集的降维操作.其基本原理为:

对于 R^m 空间中含噪声的股票数据集 $S = [S_1, S_2, \dots, S_n]$,LLTSA算法的目标是找到一个转换矩阵 A ,将 S 映射到低维空间 R^d 中,得到降维后的样本集 $O = [O_1, O_2, \dots, O_n]$,转换表达式如下所示:

$$O = A^T S H_n, d < m. \quad (2)$$

求解转换矩阵 A 的步骤为:

步骤1 PCA投影.将数据集 S 投影到PCA主子空间并舍弃一些次要成分.为了更清晰地表达,仍然采用 S 表示PCA子空间中的数据集, A_{PCA} 表示PCA的转换矩阵.经过PCA的预处理,能够帮助降低数据集噪声.

步骤2 求取数据点邻域.对 S 中的每一个数据点 $S_i, i = 1, \dots, n$,根据距离决定其 k 个邻域 $S_{ij}, j = 1, \dots, k$.得到包含邻域信息的数据集 $S, S_i = [S_{i1}, S_{i2}, \dots, S_{ik}]$.

步骤3 求取局部坐标.通过线性逼近每个数据点及其邻域对应的切空间,得到数据点投影在切空间的局部坐标.计算公式如下:

$$\arg \min_{\Theta, Q} \|S_i H_k - Q_i \Theta_i\|^2, \quad (3)$$

其中 $H_k = \frac{I - ee^T}{k}$ 为中心化矩阵, Q 为切空间的正交

基矩阵, $\Theta_i = [\theta_{i1}, \theta_{i2}, \dots, \theta_{ik}]$, θ_{ij} 为基于正交基 Q_i 的局部坐标, 式中下标表示求取欧几里得 2 范数, 上标表示求取范数的平方, 则该式转换为 $\Theta_i = Q_i^T S_i H_k$, 最优的 Q_i 已证明可由矩阵 $S_i H_k$ 的 d 个最大特征值对应的特征向量构成。

步骤 4 整合局部信息. 将对局部坐标进行全局排列, 即找到一个局部排列转换矩阵 L_i , 将所有数据点对应的切空间局部坐标 Θ_i 映射到低维全局空间 $O_i = [O_{i1}, O_{i2}, \dots, O_{ik}]$, 并使得误差和最小, 其表达式如下:

$$\arg \min_{O_i, L_i} \sum_i \|O_i H_k - L_i \Theta_i\|_2^2, \quad (4)$$

其中最优的 $L_i = O_i H_k \Theta_i^+$, Θ_i^+ 为 Θ_i 的 Moore - Penrose 广义逆. 若指定 D_i 为 0 ~ 1 选择矩阵, 则有 $OD_i = O_i$, 其中 $D = [D_1, D_2, \dots, D_n]$, 则公式(4)可转换为:

$$\arg \min_O \|ODW\|_F^2 = \arg \min_O \text{tr}(ODWW^T D^T O^T), \quad (5)$$

其中 $W = \text{diag}(W_1, W_2, \dots, W_n)$, $W_i = H_k(I - \Theta_i^+ \Theta_i)$, 且将矩阵的 Frobenius 范数平方转化为求对称矩阵的迹. 为了唯一确定 O , 限制 $OO^T = I_d$, 最后考虑到映射关系 $O = A^T S H_n$, 则可将目标函数(5)转换为广义特征值的求解:

$$SH_n B H_n S^T \alpha = \lambda SH_n S^T \alpha, \quad (6)$$

其中排列矩阵 $B = DW W^T D^T$.

步骤 5 求得转换矩阵. 对公式(6) 求取其最小的 d 个广义特征值对应的广义特征向量 $\alpha_1, \alpha_2, \dots, \alpha_d$, 则得到 $A_{\text{LLTSA}} = [\alpha_1, \alpha_2, \dots, \alpha_d]$, 最终转换矩阵为 $A = A_{\text{PCA}} A_{\text{LLTSA}}$.

通过转换矩阵 A , 可将原始高维股票数据集映射到低维流形空间, 根据 $O = A^T S H_n$ 即可得到降维后的数据集 O . LLTSA 算法提取出股票数据潜藏的非线性子流形结构, 并且由于其优异的聚类性能, 使得数据点的特征之间更加独立, 增加了特征的区分度^[11]. 将得到的低维特征集作为 GA 优化的 SVR 输入预测股票, 具有更好的预测效果。

2.2 遗传算法优化的支持向量回归机 支持向量回归机(SVR) 的基本原理为: 给定输入数据集 $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$, 其中 x_i 为数据样本值且 $x_i \in R^d$, $y_i \in R$ 为输出值, SVR 的工作是将数据样本非线性映射到高维特征空间中, 映射函数为 $\Phi(x)$, 并在这个空间中建立数据样本点的线性回归函数, 则解决该线性回归问题就对应于解决了原

空间非线性回归问题. 且为了避免计算映射到高维空间中的训练和测试样本点的点积, 引入核函数, 最终的回归函数如下:

$$f(x) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) K(x_i, x_j) + b, \quad (7)$$

其中 α_i, α_i^* 为拉格朗日算子. 本文采用 RBF 作为核函数. 则在 SVR 模型中, 需要确定的参数有惩罚因子 C 、核函数的核参数 γ 和误差范围 ε , 这 3 个参数的选取好坏直接决定了模型的效果, 而确定参数的方法一般凭借交叉实验和经验进行优化搜索, 效率低下且不一定找得到全局最优参数, 故本文采用遗传算法(GA)^[12] 对这 3 个参数进行寻优, 其算法步骤为:

步骤 1 随机生成初始种群, 采用二进制编码方案对其进行编码。

步骤 2 按照适应度函数评估种群中每条染色体的适应度大小. 本文采用均方误差作为适应度函数, 将每条染色体解码得到 3 个参数并训练 SVR 模型, 将预测值 y_i^* 与实际值 y_i 对比, 按照公式(8) 计算适应度函数值的大小. 均方误差值越小代表误差越小, 适应度越好。

$$\text{fitness} = \frac{1}{n} \sum_{i=1}^n (y_i - y_i^*)^2. \quad (8)$$

步骤 3 去掉适应度大的个体, 运用选择、交叉和突变算子对余下的个体进行遗传操作, 形成新的种群。

步骤 4 判断适应度大小, 若满足设定最优值或达到迭代次数上限, 则停止迭代, 适应度最优个体即为所求; 若不满足以上 2 个条件, 则重复迭代步骤 2 直到满足条件为止。

最后将得到的适应度最优个体进行解码操作, 即可得到最优的 C, γ, ε . 将 SVR 的参数设置为所得最优值, 则可得到优化的 SVR。

2.3 LLTSA-GA-SVR 股票预测流程 运用本文提出的 LLTSA-GA-SVR 模型进行预测的过程如下:

(1) 数据预处理: 将数据集每一个属性值都变换到 $[0, 1]$ 范围中, 公式为:

$$s' = \frac{(s - s_{\min})}{(s_{\max} - s_{\min})}, \quad (9)$$

其中 s 为原始股票某一属性值, $s_{\max}$ 和 $s_{\min}$ 分别为该属性所有值当中的最大值和最小值. s' 为标准化之后的该属性值。

- (2) LLTSA 提取股票特征,降低数据噪声.
- (3) 利用染色体解码得到的参数训练 SVR.
- (4) 对适应度进行评估,运用遗传算子对剩余个体进行遗传操作,形成新的种群.
- (5) 判断种群中每个个体适应度大小,若满足设定最优值或迭代次数达到上限,则继续执行;若不满意则返回步骤(3)重新执行.
- (6) 最优参数输入 SVR 得到优化 SVR 模型预测股票.输出股票预测值.综上所述,LLTSA-GA-SVR 预测股票价格的流程如图 1 所示.

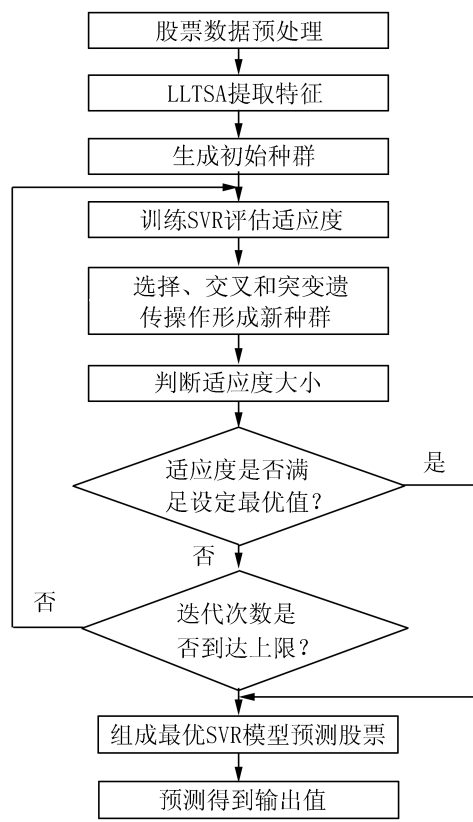


图 1 LLTSA-GA-SVR 模型预测股票流程
Fig.1 Forecasting flow of LLTSA-GA-SVR model

3 仿真实例

3.1 数据来源 本文的实验采用 Matlab R2015a 作为仿真平台,利用 libsvm 工具箱和菲尔德大学开发的 GATBX 工具箱编程实现 GA 参数优化的 SVR 模型.选取标准普尔指数分别在 2012-01-03 到 2013-12-31、2014-01-02 到 2015-12-31 这两个时间段内股票交易数据,该数据从东方财富网的数据中心提取得到.数据内容包括股票每天的开盘价、最高价、最低价、收盘价和成交量,成交额这 6

种指标,且以前 5 天的 6 种指标数据作为输入变量,以第 6 天的收盘价作为输出的预测对象.以此类推得到的样本集数据共 492(2012—2013 年)和 494(2014—2015 年)条数据,且本文相较于其他论文中只选取极少量股票数据(一般为 10~15 条)作为测试数据集而言,选取最后 76 组数据作为测试集,这样更能反映出模型的泛化能力.

3.2 参数设置 根据反复实验,本文的 GA 算法的最大遗传代数设为 100,种群数为 20,个体染色体长度为 20,规定 5 折交叉验证.SVR 模型设置损失函数取值限制在 $[0.01,1]$,惩罚因子取值范围限制在 $[0,100]$,核函数选择径向基核函数,核函数参数取值范围限制在 $[0,100]$.

3.3 评价指标 本文选用均方误差 MSE 作为预测结果的评价指标.该值越大,则说明预测值与实际值差距越大,反之则说明预测值与实际值差距越小,拟合效果越好.

3.4 结果与分析 作为实验的对比,本文实现了特征提取方法中的主成分分析法 (PCA)、核主成分分析法 (KPCA)、概率主成分分析法 (PPCA)、Stochastic Proximity Embedding (SPE),再将这些特征提取过程与 GA-SVR 模型组合进行预测,与本文提出的模型效果进行对比评估,其中 2014—2015 年数据的预测效果分别以图和表格的形式展示如表 1 和图 2.另外选取多元线性回归 (MLR)、BP 神经网络 (BPNN) 和 SVR 与 LLTSA-GA-SVR 模型进行对比,其在 2014—2015 年的预测效果如表 2 和图 3 所示.

由图 2 可以看出除了 KPCA 算法之外,其他所有特征提取算法在标准普尔指数 2014 到 2015 的预测效果都相对比较好,拟合效果比较接近,而其中 LLTSA 算法的拟合效果最优.由表 2 可看出 LLTSA-GA-SVR 组合模型在预测集局部的预测精度不是最优,但从图 3 可以很明显地看出整个预测集中组合模型的预测精度最高,效果最好.

为了更加具体地了解各个模型的预测效果,采用评价指标均方误差来评估,结果如表 3、表 4 所示.

由表 3、表 4 可以看出,LLTSA-GA-SVR 模型的均方误差在 2012—2013 年、2014—2015 年 2 个时间段里都取得了相对较低的均方误差,说明本文提出的模型具有较好的泛化能力,具有更精确的预测效果.

表 1 标准普尔 500 2014—2015 年各特征提取方法组合模型预测值

Tab.1 Forecasting values of each feature extraction integrated method used in &SP500 index from the year 2014 to 2015					
实际值	PCA	KPCA	PPCA	SPE	LLTSA
1978.1	1954.1	1955.5	1989.2	1955.0	1956.5
1995.3	1960.4	1969.0	1971.9	1971.6	1967.8
1990.2	1978.8	1995.4	1997.2	1991.2	1985.1
1958.0	1994.4	2011.0	2000.9	1998.6	1995.2
1967.0	1982.9	1976.8	1941.7	1960.6	1961.8
...	...	...	...	...	...
2061.0	2031.9	2049.5	2039.5	2054.0	2055.6
2056.5	2051.3	2036.3	2076.6	2057.0	2053.0
2078.4	2048.5	2043.2	2055.1	2052.4	2068.2
2063.4	2095.4	2041.3	2079.7	2076.9	2074.8
2043.9	2061.4	2049.0	2047.9	2066.5	2067.9

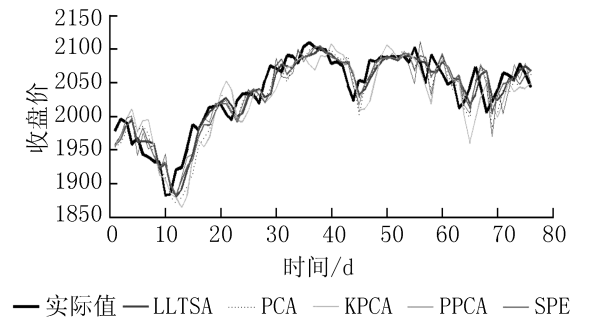


图 2 2014—2015 年各特征提取算法组合模型预测结果

Fig.2 Forecasting results of each feature extraction integrated method from the year 2014 to 2015

表 2 标准普尔 500 2014—2015 年各模型预测值

Tab.2 Forecasting values of each model for &SP500 from the year 2014 to 2015				
实际值	MLR	BPNN	SVR	LLTSAGASVR
1978.1	1978.1	1951.9	1955.0	1956.5
1995.3	2009.6	1957.1	1961.8	1967.8
1990.2	2003.4	1965.4	1971.5	1985.1
1958.0	1981.6	1973.4	1977.5	1995.2
1967.0	1994.1	1967.4	1961.5	1961.8
...	...	...	...	...
2061.0	2093.5	2040.1	2053.0	2055.6
2056.5	2059.2	2043.8	2059.1	2053.0
2078.4	2090.1	2040.7	2047.5	2068.2
2063.4	2092.9	2049.0	2053.8	2074.8
2043.9	2065.1	2052.1	2055.2	2067.9

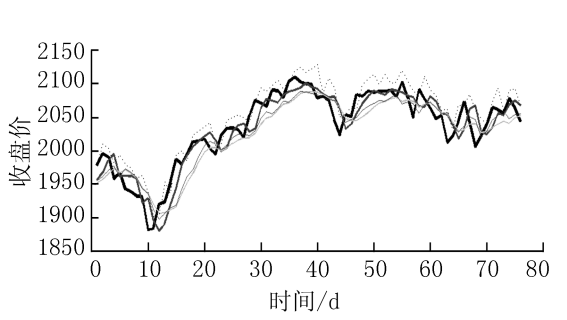


图 3 2014—2015 年各模型预测结果

Fig.3 Forecasting results of each model from the year 2014 to 2015

表 3 各特征提取算法与 GA-SVR 组合模型的均方误差值(%)

Tab.3 Mean squared error of each feature extraction method integrated with GA-SVR					
时间	PCA	KPCA	PPCA	SPE	LLTSA
2012—2013	0.13	0.78	0.25	0.13	0.13
2014—2015	0.36	0.79	0.43	0.36	0.32

表 4 各模型的均方误差值(%)

Tab.4 Mean squared error of each model				
时间	MLR	BPNN	SVR	LLTSAGASVR
2012—2013	0.71	0.55	0.35	0.13
2014—2015	0.62	0.56	0.48	0.32

4 结束语

本文引入线性切空间排列算法(LLTSA),并与遗传算法优化的支持向量回归机(SVR)相结合,提出了一种新的股票价格优化预测模型.该模型既克服了 ANN 泛化能力不强,容易陷入局部极小等缺点,也解决了 SVR 参数选择困难的难题,而且采用 LLTSA 进行数据特征提取与降噪操作,具有更好的预测精度和泛化能力.本文选取了标准普尔指数在 2012—2013 年和 2014—2015 年 2 个时间段的股票数据,将 LLTSA 与目前常用的几种特征提取算法与 GA-SVR 模型融合构建了共 5 种股票预测模型,并选择 3 种经典的股票预测模型多元线性回归、BP 神经网络和 SVR 共同进行仿真实验,最后对预测的结果进行对比.实验证明 LLTSA-GA-SVR 算法在不同的时间段、不同特征提取方法和不同预测模型中都具有比较好的预测结果和较强的泛化能力.因此该模型对股票交易决策具有较大的指导意义,可用来为股民或证券机构提供有价值的参考数据,具有广泛的应用前景.

参考文献:

- [1] 王泳.股票投资短期预测的多模匹配识别算法[J].计算机应用,2014(S2):180-183.
WANG Y. Short-term prediction of stock investment based on multi-pattern matching algorithm[J]. Journal of Computer Applications, 2014(S2):180-183.
- [2] ARIYO A A, ADEWUMI A O, AYO C K. Stock price prediction using the ARIMA model[C]. UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, 2014:106-112.
- [3] GÖÇKEN M, ÖZÇALICI M, BORU A, et al. Integrating metaheuristics and artificial neural Networks for improved stock price prediction[J]. Expert Systems with Applications, 2015, 44(C):320-331.
- [4] 吴春国,朱世钊,汪秉宏,等.用于非线性时间序列预测的 POD-RBF 神经网络[J].小型微型计算机系统, 2013, 34(8):1 848-1 851.
WU C G, ZHU S Z, WANG B H, et al. POD-based radial basis function neural network for nonlinear time series Prediction[J]. Journal of Chinese Computer Systems, 2013, 34(8):1 848-1 851.
- [5] TALARPOSHI F M, SADAIE H J, ENAYATIFAR R, et al. Stock market forecasting by using a hybrid model of exponential fuzzy time series[J]. International Journal of Approximate Reasoning, 2015, 70:79-98.
- [6] SHI H, LIU X. Application on stock price prediction of Elman neural networks based on principal component analysis method[C]. 11th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2014:411-414.
- [7] KAO L J, CHIU C C, LU C J, et al. A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting[J]. Decision Support Systems, 2013, 54(3):1 228-1 244.
- [8] 苏治,傅晓媛.核主成分遗传算法与 SVR 选股模型改进[J].统计研究, 2013, 30(5):54-62.
SU Z, FU X Y. Kernel principal component genetic algorithm and Improved SVR stock selection model[J]. Statistical Research, 2013, 30(5):54-62.
- [9] ZHANG T, YANG J, ZHAO D, et al. Letters: Linear local tangent space alignment and application to face recognition[J]. Neurocomputing, 2007, 70(7-9):1 547-1 553.
- [10] 薛巍,王正群,徐春林,等.线性局部切空间排列的传播半监督学习方法[J].计算机应用研究, 2014(8):2 334-2 337.
XUE W, WANG Z Q, XU C L, et al. Semi-supervised dimensionality reduction based on linear local tangent space alignment and label propagation[J]. Application Research of Computers, 2014(8):2 334-2 337.
- [11] LI F, TANG B, YANG R. Rotating machine fault diagnosis using dimension reduction with linear local tangent space alignment[J]. Measurement, 2013, 46(8):2 525-2 539.
- [12] 闵文文,梅端,代婷婷,等.基于遗传算法 SVM 的基因表达谱数据分析[J].云南大学学报:自然科学版, 2013, 35(4):441-446.
MIN W W, MEI D, DAI T T, et al. Tumour classification and information genes discovery from microarray data using generic algorithms based on SVM[J]. Journal of Yunnan University: Natural Sciences Edition, 2013, 35(4):441-446.

Modeling and simulation of optimized stock price prediction

CHEN Yuan^{1,2}, LUO Bi-hui¹, JIANG Wei-chen¹, SUN Hong-wei³

(1.College of Software Engineering,Chongqing University,Chongqing 400044,China;

2.Chongqing Key Laboratory of Digital Film and Television Art Theory and Technology,Chongqing 400044,China;

3.Sichuan Aerospace Polytechnic,Chengdu,610100,China)

Abstract:Stock data has the characteristics of nonlinearity and lots of noise.Traditional stock prediction models cannot sufficiently recognize the nonlinear features and reduce noise in stock data,which leads to low prediction accuracy.In order to improve the prediction accuracy,remove redundant attributes and enhance discriminability of features,this study cites linear local tangent space alignment from manifold learning and proposes a novel stock price prediction model based on optimized support vector regression.Firstly,linear local tangent space alignment algorithm is used to extract features of original stock data.Then support vector regression is utilized to model the nonlinear relationship between those features and stock close price.Meantime genetic algorithm is used to optimize parameters of support vector regression.Finally,the prediction accuracy is improved.To validate the effectiveness of the model,S&P 500 index stock data from the year 2012 to 2013,and the year 2014 to 2015 is adopted.The experiments prove that the proposed model outperforms others and has higher accuracy and more powerful generalization ability.

Key words:stock price prediction;genetic algorithms;linear local tangent space alignment;support vector regression