文章编号:1672-3961(2012)05-0053-06 DOI:10.6040/j.issn.1672-3961.2012.05.010

一种线性插补与自适应滑动窗口价格预测模型

朱全银[1],严云洋[1],周培[1],谷天峰[2]
(1.淮阴工学院计算机工程学院,江苏 淮安 223003;2. 河海大学水文水资源学院,江苏 南京 210098)

摘要:因为基于Web数据挖掘的商品价格预测的准确率都不高,所以为了提高价格预测的准确率,提出了一种基于线性插补与自适应滑动窗口的商品价格预测方法,给出了将线性数据插补方法与自适应滑动窗口结合的商品价格预测模型,并将该商品价格预测模型应用于手机与黄金价格的预测。实验结果表明,该预测模型获得了99%以上的预测准确率,提高了网页商品价格数据抽取的抗噪性能,解决了现有销售商只有历史销售价格数据没有基于多个销售商的预测价格问题,可以为商品的市场预测与分析提供依据。

关键词:商品价格;数据挖掘;预测模型;线性插补;自适应滑动窗口
中图分类号:TP391 **文献标志码**:A

Price forecasting model based on linear backfilling and adaptive sliding windows

ZHU Quan-yin[1], YAN Yun-yang[1], ZHOU Pei[1], GU Tian-feng[2]
(1. Faculty of Computer Engineering, Huaiyin Institute of Technology, Huaian 223003, China;
2. Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China)

Abstract: The accuracy rate of commodities price forecast based on Web mining is lower because of the network noise. In order to increase this accuracy rate, a novel price forecast method and a comprehensive price forecast model based on the linear backfilling and adaptive sliding windows algorithm were proposed. This comprehensive price forecast model was utilized in the commodities price forecast for cell phone and gold market. Experimental results showed that the mean absolute error of this proposed model could get more than 99 percent accuracy rate. In addition, the anti-noise performance of the webpage commodity price data extraction was improved. At the same time, this method could also solve the problem that the existing vendors only had the historical sales price data but did not have the forecasted price based on a plurality of vendors, which could also provide basis for the commodities market forecast and analysis.

Key words: commodity price; data mining; forecasting model; liner backfilling; adaptive sliding windows

0 引言

商品价格的预测方法是市场预测分析与商品生产销售决策的基础,是市场预测领域中的一个重要问题,在商品生产、销售等很多问题中起着关键作用。由于网络技术的发展与网络商店的普及,因此近年来,人们越来越重视对商品价格的预测方法的研究。商品价格的预测问题可以看作是基于时间序列的数据处理与数据分析问题,分为数据获取、数据处理与预测模型3个方面。股票市场、期货市场、电力市场等公开价格数据获取较为容易,用于价格预

收稿日期:2012-02-29
基金项目:国家星火计划资助项目(2011GA690190);淮安市科技计划资助项目(HAG2011052,HAG2010066,HAG2011045);江苏省青蓝工程资助项目;淮安市"533英才工程"资助项目
作者简介:朱全银(1966-),男,江苏盱眙人,教授,主要研究方向为智能信息处理、接口与通信. E-mail:hyitzqy@126.com

测的模型主要有最小二乘回归[1]、神经网络[2-6]、灰色马尔科夫链[7]、小波理论[8-9]和GM(1,1)模型[7]等。在以往研究Web数据挖掘与价格预测的基础上[10-18]，根据所挖掘的缺陷数据，研究基于缺陷数据的商品价格预测方法。通过滑动窗口的价格预测理论，给出了基于线性插补与自适应滑动窗口的商品价格预测模型，以网页挖掘的手机价格为预测对象，给出了实验结果，并且以纽约商品交易所的黄金交易价格数据再次验证所提出的价格预测模型的有效性。

1 滑动窗口算法

定义1 设周期在时间观测周期 t 内的时间序列为 $x_1, x_2, \cdots, x_t, \cdots, f_{t,1}$ 为 $t+1$ 即下一时刻的预测值；$f_{t,1}$ = 最新预测均值 $= x_t, x_{t-1}, \cdots, x_{t-N+1}$ 的平均值，N 为给定的参数，即预测窗口，又称步长；N 决定着预测精度，一般有实验数据根据经验获得。

定义2 x_t 为 t 时刻的实际值，$\hat{x}_t$ 为 t 时刻的预测值。

预测误差：

$$e_t = x_t - \hat{x}_t,\ t = 1, 2, \cdots, n。 \tag{1}$$

相对预测误差：

$$\tilde{e}_t = \frac{e_t}{x_t} = \frac{x_t - \hat{x}_t}{x_t},\ t = 1, 2, \cdots, n。 \tag{2}$$

绝对平均误差：

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} |e_t| = \frac{1}{n}\sum_{t=1}^{n} |x_t - \hat{x}_t|。 \tag{3}$$

绝对平均百分误差：

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \frac{|e_t|}{x_t} = \frac{1}{n}\sum_{t=1}^{n} \frac{|x_t - \hat{x}_t|}{x_t}。 \tag{4}$$

滑动窗口的预测理论为下一时刻的预测值

$$\hat{x}_{t+1} = f_{t,1} = \frac{x_t + x_{t-1} + x_{t-2} + \cdots + x_{t-(N-1)}}{N}。 \tag{5}$$

其观测周期 t 内的绝对平均误差为

$$\mathrm{MSE} = \frac{1}{t - N_{t-N+1}} \sum_{t-N+1}^{t} (x_t - \hat{x}_t)^2。 \tag{6}$$

以上算法是滑动窗口用于价格预测，但其只适用于数据正常情况，且为同一种类商品，将其用于有缺陷数据或不同品牌的商品时，具有很大的局限性。由于日常生活中的商品，如消费类产品，其每天的销售价格数据获取非常困难，采用基于网页的数据挖掘方法获取时，由于网络噪声的影响，会造成比较严重的数据缺陷问题。针对缺失数据的修补方法，常见的有均值替代法、Hedonic插补法[19]、Kriging插值法[20]、回归预测法等。除均值替代法外，其他算法计算复杂、效率低，而广大的销售商对不同消费种类商品市场预测分析与商品销售决策有迫切的需求，因此，需要找到一种能够对缺陷数据进行修补的方法和对不同种类商品价格的预测方法，以获得更高的预测准确率。

2 线性插补自适应滑动窗口预测模型

将线性插补方法与自适应滑动窗口预测方法结合，提供一种对缺陷数据且为不同品牌商品有效的价格预测方法，进而实现一种新的商品价格预测方法，以提高商品价格预测的准确率。

通过线性插值方法将网页挖取的数据进行预处理，在实现修补后的数据集上进行不同商品不同窗口下预测值的均方差分析，进而完成商品的市场价格预测。

一般理论上用于修补数据的方法中，最简单的是使用缺陷数据在时间上或空间上左右的值，求平均值后回填丢失之处，或者使用Hedonic插补法、Kriging插值法。但是对于商品价格数据的趋势性，及趋大、趋小或不变化，均值回填法带来的误差往往较大，Hedonic插补、Kriging插值都是计算复杂，效率低，用于商品价格预测时误差较大，在应用于价格预测分析上都存在较大的局限性。线性插补算法能够满足这样的需求。另外，由于基于滑动窗口的预测存在预测商品的单一性的缺点，建议采取自适应窗口法，针对不同种类或不同品牌，利用历史数据训练的方法，选取均方误差最小的预测值自适应调整窗口宽度。

具体的说，本研究提出的预测模型通过如下步骤实现线性插补与自适应滑动窗口的商品价格预测：

步骤A 抽取网页中商品的名称、型号、类型与价格数据，建立数据集 $X = \{A_1, A_2, \cdots, A_k\}$，设定需要预测价格的商品为 $A_i = \{x_1, x_2, \cdots, x_n\}$，$x_1, x_2, \cdots, x_n$ 指第 A_i 个商品从抽取的第1日至第 n 日的价格数据；查找 A_i 中异常数据，得到异常数据集 $B_j = \{b_1, b_2, \cdots b_m\}$，分别统计 B_j 中属于时间上连续的异常数据段，设共有 s 个日期连续的异常数据段，每段日期上连续异常的数据数量为 p；

步骤B 当 s 的值为0时，直接执行步骤F，当 s 的值不为0时重复执行步骤C到步骤E；

步骤C 设第 s 个日期连续的异常数据段中的数据在 A 中的位置为 $\{x_i, x_{i+1}, \cdots, b_{i+p-1}\}$；

步骤 D　求 $\Delta p=\frac{x_{i+m}-x_{i-1}}{p}$和 $x_{i+k}=\sum_{k=0}^{p-1}(x_{i-1}+(k+1)\Delta p)$;

步骤 E　$s=s-1$,返回步骤 B;

步骤 F　针对不同的商品,A_i 经过步骤 B 到步骤 E 后可以得到插补后的数据集 $\hat{X}=\{\hat{A}_1,\hat{A}_2,\cdots,\hat{A}_k\}$,设 $\hat{A}_i=\{\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_n\}$;

步骤 G　设定商品价格观测窗口宽度为 L,定义 $f_{L,1}$ 为观测窗口后一天的预测值,定义用于预测的滑动窗口宽度为 $N_{L,r}$;

步骤 H　选取预先设定的不同的 r 值,求

$$\tilde{x}_{r,n+1}=\frac{\hat{x}_n+\hat{x}_{n-1}+\hat{x}_{n-2}+\cdots+\hat{x}_{n-(N_{L,r}-1)}}{N_{L,r}};$$

步骤 I　计算不同 r 值的均方误差,$\mathrm{MSE}_{L,r}=\frac{1}{L-N_{L,r}}\sum_{t-N_{L,r}+1}^{t}(\hat{x}_i-\tilde{x}_{r,n+1})^2$,找到 $\mathrm{MSE}_{L,r}$ 最小时的 r 值和 $\tilde{x}_{r,n+1}$;

步骤 J　第 $n+1$ 天的预测值

$$\tilde{x}_{n+1}=\tilde{x}_{r,n+1}\big|_{(\mathrm{MSE}_{L,r})_{\min}};$$

步骤 K　重复步骤 B 到步骤 J,可以得到数据集 X 中所有商品的预测值。

步骤 A 到步骤 K 参数使用说明:

步骤 A 中所述抽取网页中商品的名称、型号、类型与价格数据是指利用任意 Web 数据抽取算法,抽取商品在网页上显示的名称、型号、类型与价格数据。其中 $x_1,x_2,\cdots,x_n$ 可以是第 A_i 个商品从一个网页中抽取的第 1 日至第 n 日的价格数据,也可以是从多个网页中抽取的第 1 日至第 n 日的平均价格数据;步骤 A 中异常数据集 B_j 的数据数量不大于 A_i 的总数据量的 10%。

步骤 B 到步骤 E 是针对任意一个商品在一个网页中不同日期的价格数据的插补。

步骤 G 到步骤 J 是针对任意一个商品在一个网页中不同日期的价格数据的预测值,或多个网页中不同日期的平均值价格数据的预测值。

步骤 G 中观测窗口宽度 L 的取值一般为 3 个月,设为 13 周,91 d。

步骤 H 中预先设定的不同的 r 值一般为 3 d,5 d,7 d,10 d,15 d 和 30 d。

相比现有技术的各种价格预测方法,本研究提出的价格预测模型对网页商品的价格数据挖掘,通过线性插补方法,将缺陷数据修复,提高对商品价格预测的准确率,改变了现有价格预测方法只能对完整无缺陷数据的预测,提高了数据挖掘的抗噪性能,通过自适应滑动窗口,自动选取不同商品的最佳滑动窗口值,可以达到更高的商品价格预测准确率。

3　实验分析

为了验证所提出的预测模型的有效性,利用从 2011 年 8 月 1 日至 2011 年 11 月 1 日从不同网页中抽取的 5 种不同商品的价格数据,采取人为丢失数据(见表 1)的方法来验证基于线性插补与自适应滑动窗口的价格预测方法的有效性,先求得每一种商品的周价格平均值,其中选取人为丢弃的商品数据都为价格正好发生变动的日期,保证了数据具有较好的典型性,并进行 2011 年 11 月 1 日的价格预测,最后求的平均误差为 0.81%(见图 1),而如果没有人为丢失数据,其预测的平均误差为 0.82%(见图 2),即修补数据预测的准确率比利用原始数据预测的准确率反而提高了 0.01%。

表 1　人为丢失数据表
Table 1　Abandoned data

手机型号	人为丢失数据						修补的数据						时间
Sony EricssonMT15i	2 388	2 388	2 388	2 680	2 680	2 680	2 436.7	2 485.4	2 534.1	2 582.8	2 631.5	2 680.2	10.15 ~ 10.20
Samsung S5670	1 299	1 299	1 299	1 299	1 299	—	1 299.0	1 299.0	1 299.0	1 299.0	1 299.0	—	10.18 ~ 10.22
Nokia E72i	1 879	1 788	1 788	1 788	1 788	1 788	1 863.8	1 848.6	1 833.4	1 818.2	1 803.0	1 787.8	10.15 ~ 10.20
HTC A510c	1 879	1 879	1 798	1 798	1 799	—	1 863.0	1 847.0	1 831.0	1 815.0	1 799.0	—	10.19 ~ 10.23
Motorola ME511	999	999	999	999	999	—	1 079.0	1 059.0	1 039.0	1 019.0	999.0	—	10.16 ~ 10.20

为了比较平均值修补法和线性插补法对预测准确率的影响,采用相同时间区间不同商品的价格数据进行实验,实验结果显示采用原始数据预测的误差率为 0.82%,采用线性插补预测的误差率为 0.81%,而采用均值修补法预测的误差率为 2.74%(图 3)。实验结论是线性插补预测的误差率比均值修补法预测的误差率下降了 1.66%。图 4 给出了二种不同插补方法的商品价格预测误差的比较。

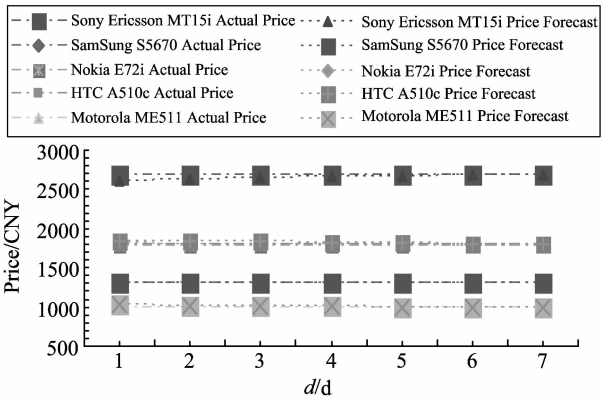

图 1 线性插补后的预测误差

Fig. 1 Forecasting MAE after using liner backfilling

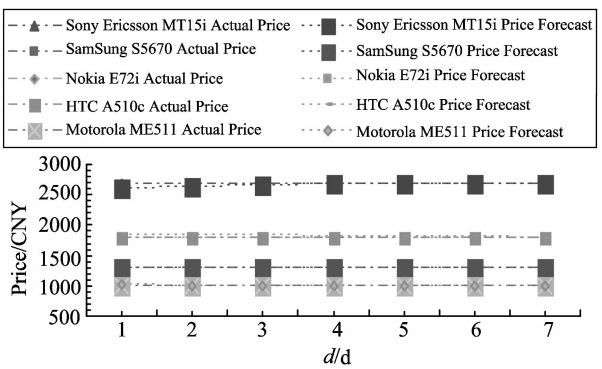

图 2 原始数据的预测误差

Fig. 2 Forecasting MAE using original data

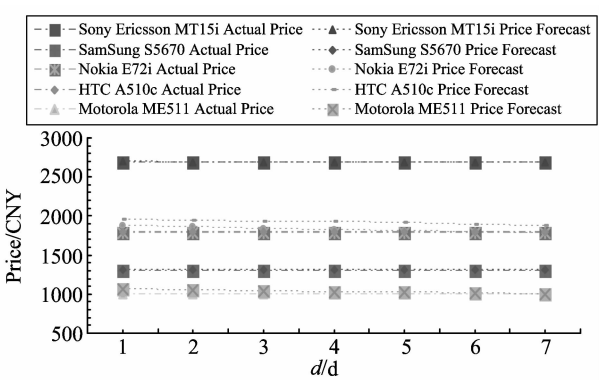

图 3 平均值插补的预测误差

Fig. 3 Forecasting MAE using average backfilling data

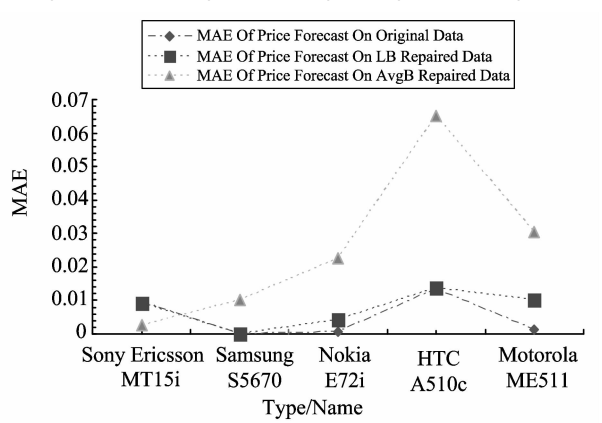

图 4 线性插补、均值插补与原始值的预测误差比较

Fig. 4 Comparisons of forecasting MAE with liner backfilling, average backfilling and original data

采用自适应滑动窗口与定值滑动窗口进行实验结果进行比较,结果是:在利用 2011 年 8 月 1 日至 2011 年 10 月 30 日的 2 个不同网站抽取的数据预测 11 月 1 日一款手机商品飞利浦 V816 价格预测中,采用定值为 7 d 的滑动窗口的预测的误差率为 0.82%,而采用自适应滑动窗口后,求得的最佳预测窗口值为 5 天,其预测的误差率为 0.19%;在利用 2011 年 11 月 1 日至 27 日的纽约商品交易所黄金交易价格数据进行 11 月 28 日的价格预测中,若采用 7 d 的定值滑动窗口的预测误差率为 1.20%,而采用自适应滑动窗口后,求得的最佳预测窗口值为 10 d,其预测的误差率为 0.49%。实验结果表明采用自适应滑动窗口的误差率比采用定值滑动窗口预测的误差率平均下降了 0.63%(见表 2)。

表 2 不同窗口值的黄金价格预测误差

Table 2 Forecasting MAE of gold price using different *N*

数 值	窗口值 10	窗口值 7	窗口值 5	窗口值 3
预测值	1 707.90	1 695.62	1 693.87	1 691.98
真实值	1 716.25	1 716.25	1 716.25	1 716.25
误差	0.004 865	0.012 020	0.013 040	0.014 141

由于网络的噪声等原因,在利用网页抽取价格数据的过程中,在每天抽取大量的数据的情况下,无法做到人工的数据确认修补,为了更进一步说明本研究提出的预测模型的实际应用价值,利用抽取的商品价格数据所进行的修补后预测价格与实际价格的比较,选择了网页抽取价格数据中存在缺陷的 5 种商品价格进行实验,实验结果是缺陷数据下经过线性插补后,采用自适应窗口预测的平均准确率达到 99.47%。见图 5。

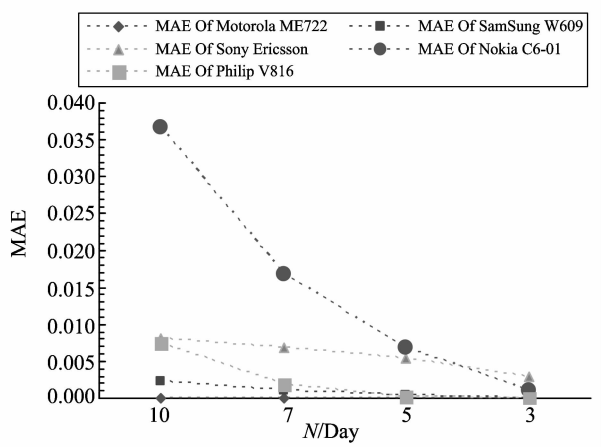

图 5 不同窗口值线性插补下 5 款手机的预测误差比较

Fig. 5 Comparisons of forecasting MAE with different *N* of five type's cell phone

4　结论

本研究提出了将线性插补方法结合自适应滑动窗口预测的新的商品价格预测方法,并将该商品价格预测方法应用于手机、黄金等商品价格的预测,解决了现有销售商只有历史销售价格数据没有基于多个销售商的预测价格问题,同时还提高了网页商品价格数据抽取的抗噪性能。

参考文献:

[1] 王永杰,王礼沅,张恒喜,等. 基于递阶偏最小二乘回归的飞机采购价格预测[J]. 火力与指挥控制,2010, 35(10):98-101.
WANG Yongjie, WANG Liyuan, ZHANG Hengxi, et al. The purchasing price prediction of aircraft based on hierarchical partial least squares regression[J]. Fire Control & Command Control, 2010, 35(10):98-101.

[2] 常松,何建敏. 基于小波包和神经网络的股票价格预测模型[J]. 中国管理科学, 2001, 9(5):8-15
CHANG Song, HE Jianmin. Stock price forecasting model basedon wavelet packet and neural network [J]. Chinese Journal of Management Science, 2001, 9(5):8-15.

[3] 张鸿彦,林辉. 应用混合神经网络和遗传算法的期权价格预测模型[J]. 管理工程学报, 2009, 23(1):59-62.
ZHANG Hongyan, LIN Hui. Option price forecasting model by applying hybrid neural network and genetic algorithm [J]. Journal of Industrial Engineering and Engineering Management, 2009, 23(1):59-62.

[4] LEE S H, LIME J S. Forecasting exchange rate by weighted average defuzzification based on NEWFM [C]// 6th IEEE International Conference on Industrial Informatics. Daejeon: Institute of Electrical and Electronics Engineers Inc., 2008:1036-1041.

[5] WANG Hua, LIU Bingxiang, CHENG Xiang, et al. An exchange rate forecasting method based on probabilistic neural network[C]// International Conference on Electronic and Mechanical Engineering and Information Technology. Harbin: IEEE Computer Society, 2011:3124-3126.

[6] YANG Hengli, LIN Hanchou. Applying EMD-based neural network to forecast NTD/USD exchange rate[C]// 7th International Conference on Networked Computing and Advanced Information Management. Daejeon: IEEE Computer Society, 2011:352-357.

[7] WU Hong, CHEN Fuzhong. Chinese exchange rate forecasting based on the application of grey system DGM(2, 1) model in post-crisis era [C]// 3th International Conference on Information Management, Innovation Management and Industrial Engineering. Kunming: IEEE Computer Society, 2010: 592-595.

[8] HADAVANDI E, GHANBARI A, ABBASIAN N S. Developing a time series model based on particle swarm optimization for gold price forecasting [C]// 3th International Conference on Business Intelligence and Financial Engineering. Hong Kong: IEEE Computer Society, 2010:337-340

[9] LIU Fanyong. The hybrid prediction model of CNY/USD exchange rate based on wavelet and support vector regression[C]// 2nd International Conference on Advanced Computer Control. Shenyang: IEEE Computer Society, 2010:561-565.

[10] 王红艳,朱全银,严云洋,等. 商品价格数据的两种WEB 挖掘算法比较[J]. 微电子学与计算机, 2011, 28(19):168-172
WANG Hongyan, ZHU Quanyin, YAN Yunyang, et al. Compare two Web mining algorithm for commodity price [J]. Microelectronics & Computer, 2011, 28(19):168-172.

[11] ZHU Quanyin, YAN Yunyang, DING Jin, et al. The commodities price extracting for shop online[C]// International Conference on Future Information Technology and Management Engineering. Changzhou: IEEE Computer Society, 2010(2):317-320.

[12] ZHU Quanyin, YAN Yunyang, DING Jin, et al. The case study for price extracting of mobile phone sell online[C]// IEEE 2nd International Conference on Software Engineering and Service Science. Beijing: IEEE Computer Society, 2011:281-295.

[13] ZHU Quanyin, CAO Sunqun, DING Jin, et al. Research on the price forecast without complete data based on Web mining[C]// 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science. Wuxi: IEEE Computer Society, 2011:120-123.

[14] ZHU Quanyin, ZHOU Hong, YAN Yunyang, et al. Commodities price dynamic trend analysis based on Web mining[C]// 3rd International Conference on Multimedia Information Networking and Security. Shanghai: IEEE Computer Society, 2011:524-527.

[15] ZHU Quanyin, CAO Sunqun, ZHOU Pei, et al. Integrated price forecast based on dichotomy backfilling and disturbance factor algorithm[J]. International Review on Computers and Software, 2011, 6(6):1089-1093.

[16] ZHU Quanyin, ZHOU Hong, YAN Yunyang, et al. Exchange rate forecasting based on adaptive sliding window

and RBF neural network[J]. International Review on Computers and Software, 2011, 6(7):1290-1296.

[17] DING Jin, ZHU Quanyin, ZHOU Lujiang, et al. Research on the new products discovery based on Web mining [C]// 3rd International Conference on Multimedia Information Networking and Security. Shanghai: IEEE Computer Society, 2011:528-532.

[18] DENG Jianping, CAO Fengwen, ZHU Quanyin, et al. The Web data extracting and application for shop online based on commodities classified [J]. Communications in Computer and Information Science, 2011, 234(4): 189-197.

[19] 高艳云. 质量调整的价格指数编制中 Hedonic 插补法的应用[J]. 数理统计与管理, 2010, 29(6):1077-1083.
GAO Yanyun. The application of Hedonic imputation method in the quality-adjusted price index [J]. Journal of Applied Statistics and Management, 2010, 29(6): 1077-1083.

[20] 邹海翔,乐阳,李清泉,等. 基于 Kriging 插值的无检测器路段交通数据插补方法[J]. 交通运输工程学报, 2011, 11(3):118-126.
ZOU Haixiang, YUE Yang, LI Qingquan, et al. Traffic data interpolation method of non-detection road link based on Kriging interpolation [J]. Journal of Traffic and Transportation Engineering, 2011, 11(3): 118-126.

(编辑:陈燕)

(上接第 52 页)

[18] 许宏科, 房建武, 文常保. 基于亮度与火焰区域边缘颜色分布的火焰检测[J]. 计算机应用研究, 2010, 27(9):3581-3584.
XU Hongke, FANG Jianwu, WEN Changbao. Fire detection based on brightness and color distribution of fire edge regions[J]. Application Research of Computers, 2010, 27(9):3581-3584.

[19] TOREYIN B U, DEDEOGLU Y. Computer vision based method for real-time fire and flame detection[J]. Pattern Recognition Letters, 2006(27):49-58.

[20] 王俊明, 杨永跃, 付贵权. 多判据图像型火灾探测系统的研究[J]. 工业控制计算机, 2008, 21(2): 50-51.
WANG Junming, YANG Yongyue, FU Guiquan. Research of multi-recognition fire image detecting system [J]. Industrial Control Computer, 2008, 21(2):50-51.

[21] 肖靓. 基于支持向量机的图像分类研究[D]. 上海:同济大学,2006.
XIAO Liang. Research on SVM-based image classification[D]. Shanghai:Tongji University, 2006.

(编辑:陈斌)