



基于 K -means 和 MTLS-SVM 算法的生理参数监测系统

夏景明, 唐玲玲, 谈玲, 郑晗
(南京信息工程大学, 江苏 南京 210044)

摘要: 在非医模式的生理参数监测系统中, 对监测参数进行学习, 可以提高诊断和预测精度。针对多任务时间序列中存在的信息挖掘不充分、预测精度低等问题, 将机器学习中的监督和半监督学习方式结合起来对远程健康监护对象进行生理状况预测。该方法用 K -means 算法将相同类别的数据集群, 并使用多任务最小二乘支持向量机 (MTLS-SVM) 来训练历史数据来进行趋势预测。为了评估该方法的有效性, 将 MTLS-SVM 方法与 K -means、MTLS-SVM 方法比较, 实验结果表明该方法具有较高的预测精度。

关键词: 生理参数; 时间序列预测; K -means 聚类; 多任务学习

中图分类号: TN919

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017286

Biometric monitoring system based on K -means & MTLS-SVM algorithm

XIA Jingming, TANG Lingling, TAN Ling, ZHENG Han
Nanjing University of Information Science & Technology, Nanjing 210044, China

Abstract: In a nonmedical biometric monitoring system, the monitoring parameters are preceded with machine learning for precision promotion of diagnosis and prediction. Considering the problems of insufficient information mining and low prediction accuracy in multi task time series, both supervised and unsupervised machine learning techniques were applied to predict the physical condition of the remote health care. These techniques were K -means for clustering the similar group of data and MTLS-SVM model for training and testing historical data to perform a trend prediction. In order to evaluate the effectiveness of the method, the proposed method was compared with MTLS-SVM method. The experimental results show that the proposed method has higher prediction accuracy.

Key words: physiological parameter, time series prediction, K -means clustering, multi-task learning

1 引言

我国社会开始进入老年化阶段, 老龄化趋势明显, 患有慢性病的人群 (如患有心脏病和血压

不稳定人群) 也逐渐增长。老化性患者和慢性病患者需要进行长时间的观察与医治, 但由于条件的限制及其他因素, 这类人群通常是按期到医院复查诊治, 那么对病人的治疗效果会不佳, 除人

收稿日期: 2017-09-01; 修回日期: 2017-09-30

基金项目: 国家自然科学基金资助项目 (No.41505017); 江苏省自然科学基金资助项目 (No.BK20160951)

Foundation Items: The National Natural Science Foundation of China (No.41505017), The National Natural Science Foundation of Jiangsu Province of China (No.BK20160951)



力财力的浪费之外，对病人的生理和心理也都造成很大的影响。对于慢性病患者而言，他们的生命体征数据应该进行实时监测，否则会严重影响医生的准确诊断。还有一些患者生活不能自理，这就需要进行实时监护，以便得到更好的护理。随着嵌入式技术、无线通信技术的快速发展，基于无线传感网构建的远程健康监护系统在农村、社区医疗服务中得到广泛的关注和应用^[1,2]。

在生理参数监测预测方面，Palaniappan 等人^[3]提出了将决策树、朴素贝叶斯和神经网络应用于心脏病的预测，它促进了关于心脏病方面重要模式的建立。Tsipouras^[4]为冠状动脉疾病的诊断给出了一种基于模糊规则的决策支持系统。模糊模型是用来对参数进行优化，Patil 等人^[5]引入 *K*-means 集群算法从数据仓库提取适于心脏病发作的数据。Srinivas 等人^[6]提出数据挖掘方法来预测心脏疾病，使用病人的医学资料如年龄、性别、血压和血糖来预测病人患有心脏病的可能。Resul 等人^[7]提出通过神经网络集来预测，这种集成方法是通过结合现有的方法来创造新模式进行疾病预测。支持向量机（support vector machine, SVM）是由 Vapnik^[8]提出的一种监督的学习方式，SVM 主要用于时间序列的预测和分类。这些算法会出现训练时间过长、程序异常终止、预测精度不高的问题。

本文设计了一个基于 *K*-means 和 MTL-SVM 算法的远程医疗监测护理系统。在生理监测的基础上，也需要对病人的状况进行预测，以便病患得到更加及时有效的护理，为此本文提出一种基于 *K*-means 和 MTL-SVM 的生理参数预测算法。

2 系统结构

针对慢性病（如心脏病）人以及生活不能自理的患者，本文设计了一种远程健康护理系统，让病人在设备不齐全、医疗条件受限的情况下得到更好的护理。系统通过监测端将数据转发到服务器端，并通过数据挖掘技术对数据进行处理之后传送到数据库专家系统，由医生对病患健康状况进行诊断。

硬件系统如图 1 所示，分为 3 个模块：信号采集模块、微控制器模块和传输数据模块。信号采集模块采集生理信号，由体温、心率、湿度传感器构成。体温和湿度是由 AM2320 数字温湿度传感器采集数据，采用单总线、标准 I^2C 两种通信方式；心率由一款用于脉搏心率测量的光电反射式模拟传感器（pulse sensor）采集，将其佩戴于手指、耳垂等处，通过导线连接可将采集到的模拟信号传输给监测终端用来转换为数字信号，再通过监测端简单计算后就可以得到心率数值。微控制器模块主要是处理和监测数据，监测模块

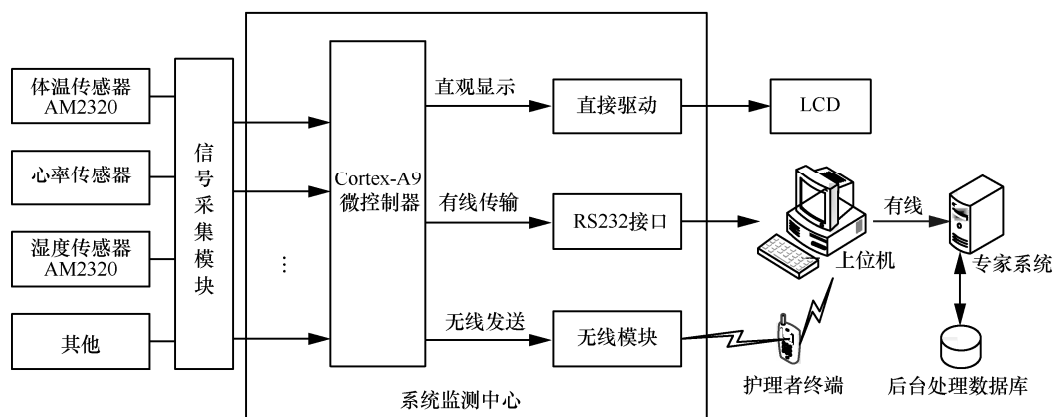


图 1 系统硬件设计

选用三星公司的 Exynos4412 作为主控芯片, 该芯片是以 Cortex-A9 作为架构的微处理器核心, 支持嵌入式 Linux 系统, 它是整个系统的核心, 控制着整个系统的工作流程。当它向采集信号模块发送命令时, 采集模块会发送采集到的数据给监测中心, 它会进行相应的处理, 之后通过不同的形式将数据显示出来。

基于以上人体生理参数, 设计了基于 Web 的生理监测系统, 并开发了图形化显示的应用程序, 病患监护界面图 2 所示。

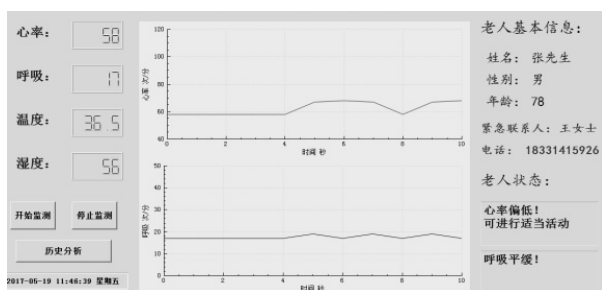


图2 系统软件设计

3 病患生理参数学习算法设计

由于监测系统是通过远程方式监测病患生理参数, 不能像医院那样及时解决病人问题, 所以进行病患的生理状况预测非常重要。例如对于患有心脏病的患者, 若不能及时处理就可能出现猝死现象, 若系统能及时预测病人在未来会出现的状况, 则能给护理者充分的时间准备解决方案, 不至于使护理者在出现问题时手足无措。此预测系统不仅能减小病人风险, 还能减轻护理者压力, 提高护理效率。

最小二乘支持向量机 (LS-SVM)^[9]是在二次损失函数下 SVM 的一种形式, 只对线性方程进行求解且求解非常迅速, 因此, 在时间序列预测中应用非常广泛^[10-12]。然而, 这些单任务学习算法在实际应用中经常遇到训练样本不充分的情况, 无法对模型进行充分训练, 常常会产生欠拟合现象。此外, 对任务间的相关性信息没有做进一步

的挖掘, 从输入样本中所获取的信息有限, 最终影响了模型的泛化性能。多任务学习通过利用相关任务的训练样本中所包含的信息进行归纳偏置, 对多个具有共享表达的任务进行并行学习, 每个任务的学习可以帮助其他任务学习得更好^[13], 因此, 多任务学习的研究受到广泛的关注。不像稳定数据, 时间序列数据的特征是数值是随着时间的变化而变化^[14], *K*-means 算法的原理是通过迭代的方法, 逐次更新各聚类中心的值, 直到得到最好的聚类结果。聚类的初始中心可以通过很多方式选择, 可以任意选择, 也可以依次选择。此外, 聚类的数量也是一个重要的参数, 它可以提前确定好, 也可以在处理过程中变化。

本文将采用多任务 LS-SVM 算法与 *K*-means 算法结合的方式, 该方法融合了半监督学习和监督学习算法。将输入值通过 *K*-means 算法进行聚类, 相近的样本被划分在同一个簇, 每个簇经过 MTLS-SVM 算法训练和测试得到预测的输出值。这种方法不仅改善了以往算法中序列信息挖掘不足、序列间相关性考虑欠佳等因素导致的预测精度低问题, 而且降低了支持向量机中的维度, 提高了计算速度。

3.1 基于 *K*-means 算法的学习过程

首先从 n 个数据对象中任意选择 k 个对象作为初始聚类中心, 对于所剩下的对象, 则根据它们与这些聚类中心的相似度 (距离), 将它们分别分配给与其最相似的聚类, 再计算每个所获新聚类的聚类中心; 不断重复这一过程直到标准测度函数开始收敛为止。

给出聚类 K 和数据子集 $T = \{t_1, t_2, \dots, t_n\}$, $t_i = (x_i, y_i)$ 。

K-means 算法的基本思想如下:

步骤 1 随机选择 K 作为初始质心点;

步骤 2 对于所剩下的对象, 则根据它们与这些聚类中心的相似度 (距离), 分别将它们分配给



与其最相似的聚类；

步骤 3 计算每个所获新聚类的聚类中心；

步骤 4 如果满足了标准，就停止步骤，否则转回步骤 2，直到条件满足。

停止条件如下：没有需要分配的任务到不同的簇，质心不再发生变化，或者均方差误差下降幅度很小，其计算式如下：

$$E = \sum_{k=1}^k \sum_{x \in c_k} d(x, m_k)^2 \quad (1)$$

其中， c_k 是第 k 个簇， m_k 是簇 c_k 的质心， $d(x, m_k)$ 是 x 和质心 m_k 之间的距离，在本文提出的方法中，欧氏距离是每个目标点到簇中心的距离：

$$d(x, m_k) = \|x, m_k\| = \sqrt{(x_1 - m_{k1})^2 + (x_2 - m_{k2})^2 + \dots + (x_n - m_{kn})^2} \quad (2)$$

3.2 基于 K-means 与 MTL-SVM 结合的预测模型

本文提出的基于 K-means 与 MTL-SVM 结合的时间序列预测模型，是将输入值通过 K-means 算法进行聚类，相近的样本被划分在同一个簇，每个簇经过 MTL-SVM 算法训练和测试得到预测的输出值。算法包括数据预处理、任务构造和模型训练 3 个部分，如图 3 所示。

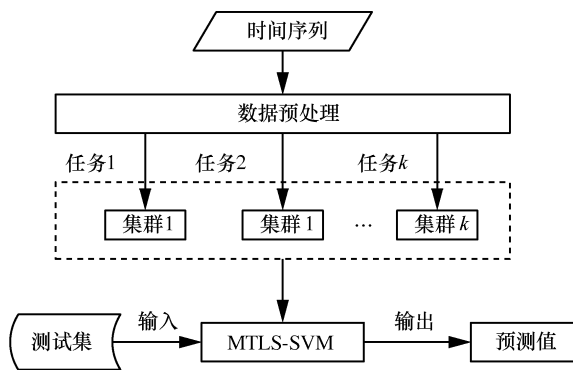


图 3 K-means 与 MTL-SVM 时间序列预测模型

(1) 数据预处理

设时间序列 $\{y_1, y_2, \dots, y_n, y_{n+1}, \dots, y_t\}$ 为等间隔采样所得，其中， t 表示当前时刻，时间序列预测就是根据历史采样值估计 $t+h$ 时刻的值，即 y'_{t+h}

如式 (3)：

$$y'_{t+h} = f(y_t, y_{t-1}, \dots, y_{t-n+1}) \quad (3)$$

其中， h 为预测步长，若 $h=1$ ，则为单步预测，若 $h>1$ ，则为多步预测。 f 为预测模型， n 为时延。将采集到的观测值 $\{y_1, y_2, \dots, y_n, y_{n+1}, \dots, y_t\}$ 通过 K-means 算法聚类构造数据集样本。

(2) 任务构造

将多任务学习方法用于时间序列预测的本质在于相邻时间点的密切相关性，传统的单任务学习方法使用前 n 个时刻的观测值预测 $t+h(h \geq 1)$ 时刻的值，模型输出为单值。而使用多任务学习方法可以将 $t+1, t+2, \dots, t+k$ 一起作为预测输出，以起到归纳偏置的作用，提高预测精度。时间序列中构造的单任务与多任务如图 4 所示。

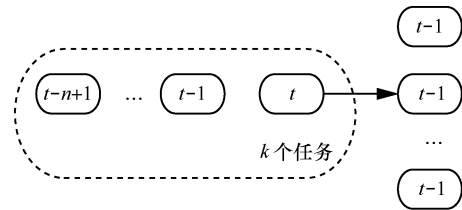


图 4 时间序列中多任务构造

采用多任务学习方法只需构造一个多输出预测模型，得到 k 个任务的输出值。根据时间序列预测目标，选取适当的 k 个任务，并构造每个任务相应的训练集。

(3) 模型训练

将 k 个学习任务相对应的训练集同时训练 MTL-SVM 模型。由于这些任务同时对相邻的时刻值进行预测，因此，它们之间是平行关系而且有紧密的相关性。在训练的过程中，任务之间能够相互牵制起到归纳偏置的作用，从而提高模型的泛化性能。基于 MTL-SVM 的时间序列预测步骤如下。

输入：时间序列 $y = \{y_1, y_2, \dots, y_n, y_{n+1}, \dots, y_t\}$

输出：未来时刻的估计值 y'_{t+h}

步骤 1 数据预处理，将原始时间序列 y 经过

表 1 k 取 2~5 时的算法精确度

	$k=2$			$k=3$			$k=4$			$k=5$		
	数据集 1	数据集 2	数据集 3	数据集 1	数据集 2	数据集 3	数据集 1	数据集 2	数据集 3	数据集 1	数据集 2	数据集 3
心电情况	82.5%	78.1%	79.4%	80.6%	79.3%	79.8%	80.2%	77.9%	78.9%	79.7%	77.2%	77.4%
血压情况	78.7%	80.2%	79.7%	77.9%	80.8%	80.2%	76.5%	79.1%	77.8%	77.6%	79.2%	79.1%

表 2 基于 K-means+MTLS-SVM 算法的数据趋势分析

	下降			平缓			上升		
	数据集 1	数据集 2	数据集 3	数据集 1	数据集 2	数据集 3	数据集 1	数据集 2	数据集 3
心电情况	15	13	11	18	16	19	17	21	20
血压情况	10	12	14	18	15	17	22	23	19

K-means 算法聚类后, 根据预测需要确定不同的 h , 并构造数据集样本。

步骤 2 根据目标估计值构造 k 个学习任务 $t+1, t+2, \dots, t+k$, $k \leq h$ 针对每个任务对应的输出 y'_{t+h} , 分别从数据集通过 K-means 算法聚类, k 个类组成训练集。

步骤 3 将 k 个任务的训练集同时训练 MTLS-SVM 模型, 得到最终的预测模型。

步骤 4 将 k 个任务相应的测试样本输入训练好的 MTLS-SVM 模型, 可同时得到 k 个估计值, 最后从中选择需要的估计值 y'_{t+h} 。

选取需要的估计值之后, 即可以与实测数据进行对比, 从而得出本文算法预测的精确情况。

4 实验结果与分析

为了验证本文算法的有效性, 得到精确的实验结果, 以安徽省安庆市宿松县中心医院 180 名心脏病人和高血压患者作为实验对象, 在 2017 年 2 月 13—20 日经过数据收集和数据处理过程之后, 用 K-means 算法来实现聚类算法。本实验中, 需要选择合适的 k 值, 以便达到更高的精确度, 表 1、表 2 分别是 k 取 2、3 和 4、5 时, 通过不同子集测试后的结果。

数据集 1、2、3 分别使用的样本值个数为 203、421、548。从表 1 可以看出, 当数据集为小样本时, k 取 2 时预测精确度最高, 当数据集为较大样本 k

取 3 时预测精确度最高, 其中, KMs.MTLS-SVM 表示使用了 K-means 聚类算法。从数据集 1、数据集 2、数据集 3 中选取 50 个样本做模型测试, 显示监测病患当天的心电和血压情况, 分为上升、下降和平缓, 表 3 为模型测试的结果。

表 3 算法预测精度和误差比较

	误差	MTLS-SVM	基于 K-means+ MTLS-SVM	精度 提高
心电情况	RMSE	4.36	3.46	20%
	MAE	3.68	2.87	22%
血压情况	RMSE	4.03	3.25	19%
	MAE	3.79	3.07	19%

除了要得到本文算法预测的精确度, 还需要比较两种算法的精确度和预测误差, 误差指标采用均方根误差 (RMSE) 与平均绝对误差 (MAE), 如式 (4) 和式 (5) 所示, 式中, y_i 表示真实值, y'_i 表示预测值, N_t 表示预测点数, 表 4 为两种算法的精确度和预测误差比较, 可以看出, 本文算法精确度更高, 误差更小。

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_t} (y_i - y'_i)^2}{N_t}} \quad (4)$$

$$MAE = \frac{1}{N_t} \sum_{i=1}^{N_t} |y_i - y'_i| \quad (5)$$

为了验证本文算法的可行性, 采用病患心电和血压的收缩压样本数据来进行模型训练, 图 5

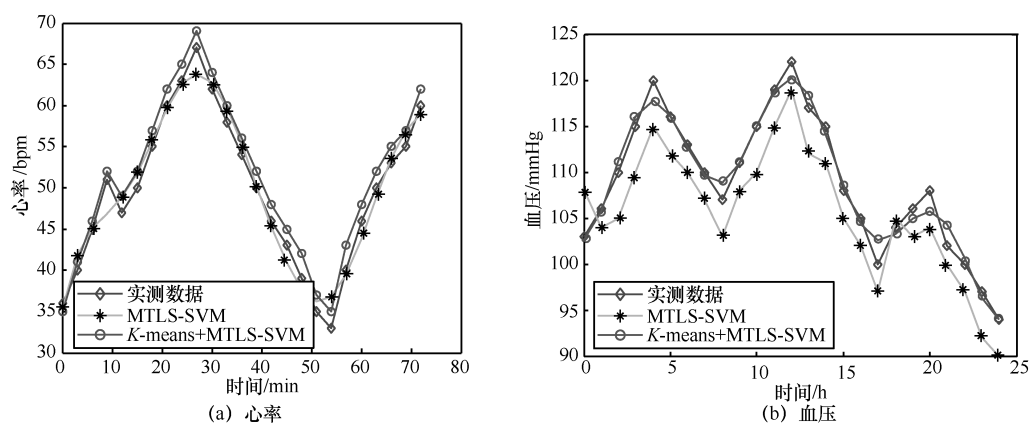


图5 病患1 预测结果对比

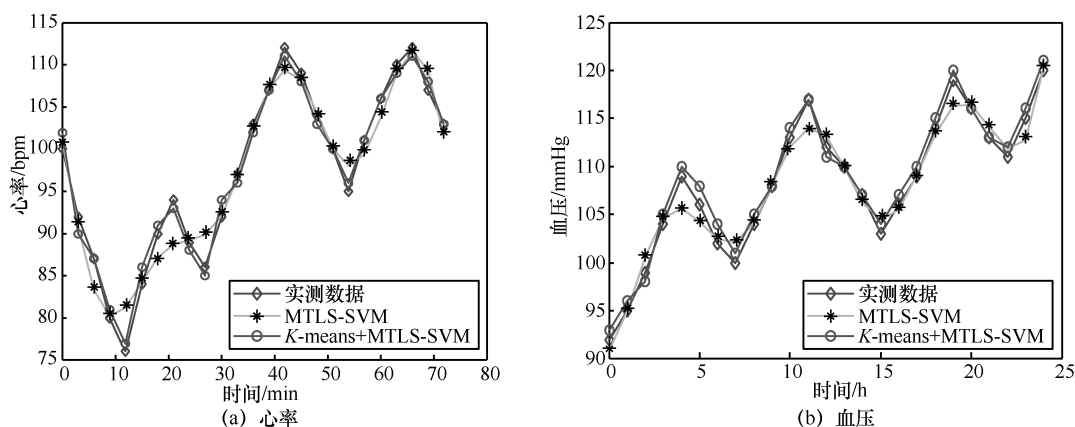


图6 病患2 预测结果对比

和图6为病患1和病患2的生理参数数据经过两种算法学习之后得到的对比结果,实验结果表明,本文算法更精确,预测值更逼近实测值,从而体现了本文方法的有效性。

5 结束语

本文提出一种将 K -means 算法与多任务 LS-SVM 算法结合的预测算法,应用于人体生理参数监测系统中。首先通过 K -means 算法将数据聚类,然后用多任务 LS-SVM 算法进行模型训练,该方法不仅改善了以往算法中序列信息挖掘不足、序列间的相关性考虑欠佳等因素导致的预测精度低问题,在病患的生理参数预测,如心电和血压数据预测中得到了更精确的结果,改进了远程护理不及时、不能有效应对病

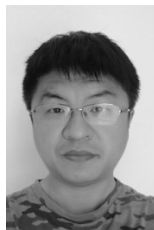
患突发状况的不足,为医疗远程监护的发展起到了一定的推动作用。如何快速有效地优化模型参数,并进一步提高模型的泛化性是下一步的研究方向。

参考文献:

- [1] 褚航,曾碧. 非接触式心跳监控系统——面向社区医疗服务的物联网应用系统研究与开发[J]. 计算机系统应用, 2012, 21(8): 233-235.
CHU H, ZENG B. Non contact heartbeat monitoring system—research and development of internet of things application system for community medical service[J]. Computer System & Applications, 2012, 21(8): 233-235.
- [2] 谈玲,张隼. 交互式多生物特征识别技术在电子商务中的应用[J]. 电信科学, 2015, 31(10): 124-129.
TAN L, ZHANG L. Application of interactive multi-biometrics recognition on E-commerce[J]. Telecommunications Science, 2015, 31(10): 124-129.
- [3] PALANIAPPAN S, AWANG R. Intelligent heart disease predic-

- tion system using data mining techniques[C]//International Conference on Computer Systems & Applications, March 30-April 3, 2008, Doha, Qatar. New Jersey: IEEE Press, 2008: 108-115.
- [4] TSIPOURAS M G, EXARCHOS T P, FOTIADIS D I, et al. Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling[J]. IEEE Transactions on Information Technology in Biomedicine, 2008, 12(4): 447-458.
- [5] PATIL S B, KUMARASWAMY Y S. Intelligent and effective heart attack prediction system using data mining and artificial neural network[J]. European Journal of Scientific Research, 2009, 31(4): 642-656.
- [6] SRINIVAS K, KAVITHA R B K, GOVRDHAN A. Applications of data mining techniques in healthcare and prediction of heart attacks[J]. International Journal on Computer Science and Engineering, 2010, 2(2): 250-255.
- [7] RESUL D, TURKOGLU I, SENGUR A. Effective diagnosis of heart disease through neural networks ensembles[J]. International Journal of Expert Systems with Applications, 2009(36): 7675-7680.
- [8] VAPNIK V. The nature of statistical learning theory[M]. New York: Springer, 1995.
- [9] SAPANKEVYCH N I, SANKAR R. Time series prediction using support vector machines: a survey[J]. IEEE Computational Intelligence Magazine, 2009, 4(2): 24-38.
- [10] KAYTEZ F, TAPLAMACIOGLU M C, CAM E, et al. Forecasting electricity consumption: a comparison of regression analysis, neural networks and least squares support vector machines[J]. International Journal of Electrical Power & Energy Systems, 2015(67): 431-438.
- [11] VAN G T, SUYKENS J K, BAESTAENS D E, et al. Financial time series prediction using least squares support vector machines within the evidence framework[J]. IEEE Transactions on Neural Networks, 2001, 12(4): 809-821.
- [12] YE M Y, WANG X D. Chaotic time series prediction using least squares support vector machines[J]. Chinese Physics, 2004, 13(4): 454.
- [13] CARUANA R. Multitask learning[M]. New York: Springer, 1998: 95-133.
- [14] HAN J W, KAMBER M. Data mining: concepts and techniques: 2nd edition[J]. San Francisco: Morgan Kaufman Publishers, 2006.

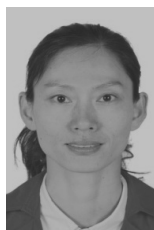
[作者简介]



夏景明 (1980-), 男, 博士, 南京信息工程大学电子与信息工程学院副教授, 主要研究方向为信号处理。



唐玲玲 (1992-), 女, 南京信息工程大学电子与信息工程学院硕士生, 主要研究方向为智能仪器。



谈玲 (1979-), 女, 博士, 南京信息工程大学计算机与软件学院副教授, 主要研究方向为物联网大数据及气候变化。

郑晗 (1994-), 男, 南京信息工程大学计算机与软件学院硕士生, 主要研究方向为物联网。