研究简报

PCR-RBF-SVM预测模型在财政数据中的应用

王 喆[1], 王有力[2], 孙雯雯[1], 吕 巍[1]
(1. 吉林大学 计算机科学与技术学院, 长春 130012; 2. 吉林吉信通信咨询设计有限公司, 长春 130012)

摘要: 通过使用支持向量机算法将主成分回归的线性预测结果和径向基神经网络的非线性预测结果相结合, 提出一种新的预测模型, 该模型提高了预测精度, 解决了预测方式单一的问题. 将新预测模型应用于财政数据预测结果表明, 与传统主成分回归和径向基神经网络方法相比, 该模型预测效果更好.

关键词: 主成分回归; 径向基神经网络; 支持向量机; 预测
中图分类号: TP399 **文献标志码**: A **文章编号**: 1671-5489(2012)01-0111-03

Application of Prediction Model Based on PCR-RBF-SVM to Finance Data

WANG Zhe[1], WANG You-li[2], SUN Wen-wen[1], LÜ Wei[1]
(1. *College of Computer Science and Technology, Jilin University, Changchun* 130012, *China*;
2. *Jilin Jixin Communications Consulting and Design Co., Ltd, Changchun* 130012, *China*)

Abstract: On the basis of support vector machine algorithm and the result of principal component regression of the linear prediction and radial basis function neural network of the non-linear prediction, a new forecasting model was proposed by which one can effectively improve the prediction accuracy and solve the problem of single prediction. Application of the new prediction model to the prediction of finance data showed that compared with the traditional principal component regression and radial basis neural network method, the new model has better effect and practical significance in prediction.

Key words: principal component regression; radial basis neural network; support vector machine; prediction

随着经济的快速发展, 财政数据越来越繁琐, 冗余信息较多, 因此从繁杂的信息中提取出有用的信息, 并针对数据进行有效预测具有重要意义. 目前, 预测算法主要有线性预测[1]、神经网络预测[2]、支持向量机预测[3]、决策树预测[4]、灰色预测[5]、朴素贝叶斯预测[6]、时间序列预测[7]和遗传算法预测[8]等. 本文提出一种基于主成分回归[9]、径向基神经网络[10]和支持向量机[11]的综合预测算法, 将3种算法的优势结合, 用于预测数据.

主成分回归、径向基神经网络和支持向量机3种预测方法的特点如下:

1) 主成分回归算法适用于属性较多的数据, 且这些属性满足: ① 数据之间具有相关性, 可以提取主成分; ② 数据满足一定的线性关系, 即目标属性与相关属性间满足某些特定的线性关系式.

2) 径向基神经网络适用于非线性的数据关系表达式, 给定输入输出, 神经网络通过权值系数即可学习数据属性间的关系, 从而逼近任意的函数. 如果数据复杂, 同时目标属性与相关属性间满足一定

收稿日期: 2011-04-22.

作者简介: 王 喆(1974—), 男, 汉族, 博士, 副教授, 从事数据挖掘和商务智能的研究, E-mail: wz2000@jlu.edu.cn. 通讯作者: 吕 巍(1971—), 男, 汉族, 硕士, 高级工程师, 从事计算机应用的研究, E-mail: lvwei@jlu.edu.cn.

基金项目: 国家自然科学基金(批准号: 60673099; 60873146)和吉林省科技发展计划重点项目(批准号: 20090304).

的非线性关系，则可使用径向基神经网络预测.

3）支持向量机适用于小样本、非线性及高维模式下的数据.

主成分回归用于处理线性问题，而径向基神经网络用于处理非线性的预测. 当一组样本数据满足的关系既有线性关系，又有非线性关系时，则使用支持向量机. 本文通过支持向量机，将主成分回归得到的线性预测结果和径向基神经网络得到的非线性预测结果作为支持向量机的数据进行计算，从而得到将主成分回归和径向基神经网络预测相结合的结果.

1 PCR-RBF-SVM 预测模型

PCR-RBF-SVM 模型的流程如图 1 所示. 步骤如下：

1）将数据样本进行主成分回归(PCR)，得到线性预测结果. 主成分回归是主成分分析和线性回归相结合的算法. 主成分分析先将原始数据标准化，再计算主成分贡献率和累计贡献率，最后计算主成分得分. 多元线性回归模型用于研究目标变量和多个自变量间的多元线性关系，将得到的主成分与目标变量建立线性关系即可得到主成分回归的预测结果.

图 1 PCR-RBF-SVM 模型流程

Fig. 1 Flow chart of PCR-RBF-SVM model

2）通过径向基神经网络(RBF)对数据样本进行目标值的预测，得到预测结果. 径向基神经网络是一种局部逼近网络，它能以任意精度逼近任一连续函数. 本文使用的径向基神经网络包含输入层、隐含层和输出层，隐含层的节点数为样本数 n，隐含层的传输函数为有辐射状作用的高斯函数：

$$u_i = \exp[-(x - c_i)^{\mathrm{T}}(x - c_i)/(2\sigma_i^2)], \quad i = 1,2,\cdots,n, \tag{1}$$

其中：u_i 表示第 i 个隐节点的输出；σ_i 表示第 i 个隐节点的标准化常数. 误差函数采用均方误差：

$$\mathrm{MSE} = \frac{1}{n}\sum_{p=1}^{n}\sum_{k=1}^{l}(T_k^p - Y_k^p)^2, \tag{2}$$

其中：T_k 表示第 k 个样本的目标量；Y_k 表示第 k 个样本的预测量.

3）将 1)和 2)的预测结果组成新数据，利用支持向量机(SVM)算法得到新预测结果. 核函数取高斯函数：

$$K(x_i,x_j) = \exp(-\|x_i - x_j\|^2/\sigma^2). \tag{3}$$

通过上述步骤，即可得到 PCR-RBF-SVM 预测模型的结果. 该结果综合利用了线性预测与非线性预测相结合的优势，使支持向量机将主成分回归和神经网络相结合，针对数据进行预测，可达到一定的预测精度.

2 PCR-RBF-SVM 模型在经济数据中的预测

本文使用的经济数据为《中国统计年鉴》中的财政支出数据，通过多年的数据样本，研究财政支出数据中多个地区各项支出与该地区预算间的关系，建立 PCR-RBF-SVM 预测模型. 数据共 31 个样本，即全国 31 个省、直辖市、自治区行政单位，其中，每个样本有 14 个财政支出指标和 1 个财政预算支出数值，通过研究 1984 年～2009 年的数据样本，将 2009 年的预算作为预测值，与 2009 年的真实预算值相比.

2.1 主成分回归 先将财政数据进行主成分分析，将 14 个支出的指标降维到 5 个主成分，列于表 1.

表 1 主成分的贡献率

Table 1 Contribution rate of principal component

主成分	特征值	贡献率/%	累积贡献率/%
1	4.326	33.065	33.065
2	3.529	22.089	55.154
3	2.368	15.698	70.843
4	1.925	10.354	81.097
5	1.026	8.268	89.465

由表 1 可见，5 个主成分特征值均大于 1，且反映原始数据信息的 89.465%，所以将 14 维属性降

到 5 维. 将 31 个样本 14 个财政指标的数据转化为具有 5 个主成分的数据，再建立财政预算支持 Y 与 5 个主成分间的多元线性方程式：

$$Y = 0.982X_1 + 0.683X_2 - 0.325X_3 + 0.298X_4 - 0.268X_5, \quad (4)$$

通过计算，可得到主成分回归的预测结果.

2.2　RBF 神经网络预测　将 31 个样本的 14 个财政指标作为 RBF 神经网络的输入端，再将 31 个样本下一年的预算支出作为 RBF 神经网络的输出端，即神经网络通过学习该年份的财政支出与下一年的财政预算间的关系，建立神经网络模型.

2.3　建立 PCR-RBF-SVM 预测模型　将主成分回归和神经网络预测得到的结果相结合，作为支持向量机的数据端，从而将两种结果进行结合，得到 31 个样本的预测值.

2.4　3 种预测方法的比较　分别使用主成分回归、径向基神经网络和 PCR-RBF-SVM 预测模型分别对 31 个样本的 2009 年财政预算支出进行预测，得到预测结果与真实值间的差别，得出均方误差 MSE，结果列于表 2. 由表 2 可见，基于主成分的径向基神经网络均方误差(MSE)比其他方法提高了精度，误差减小了一个数量级.

表 2　3 种预测方法的均方误差

Table 2　Mean square errors produced by different methods

预测方法	主成分回归	径向基神经网络	PCR-RBF-SVM 预测模型
预测均方误差(MSE)	0.689 6	0.358 4	0.078 9

综上所述，本文将财政支出数据进行主成分回归，得到了关于主成分的线性预测值；将财政支出数据通过径向基神经网络进行预测，得到了非线性的预测值；通过支持向量机将主成分回归和径向基神经网络的预测值相连接，即将线性预测和非线性预测相结合，得到了 PCR-RBF-SVM 预测模型，该模型结合主成分回归、神经网络和支持支持向量机用于预测，取得了较好的预测效果.

参　考　文　献

[1] XU Shan-qing, MENG Qing-ping, RONG Yong-hua, et al. A Theoretical Model on Solvus Line Prediction of Film and Its Application in Nanogranular Al-Cu System [J]. Journal of Shanghai Jiaotong University: Science, 2007, 12(3): 341-346.

[2] Khoshhal A, Dakhel A A, Etemadi A, et al. Artificial Neural Network Modeling of Apple Drying Process [J]. Journal of Food Process Engineering, 2010, 33(Suppl 1): 298-313.

[3] Lopez-Meyer P, Schuckers S, Makeyev O, et al. Detection of Periods of Food Intake Using Support Vector Machines [C]//32nd Annual International Conference of the IEEE EMBS. Washington DC: IEEE Press, 2010: 1004-1007.

[4] LI Tun, LIU Gong-shen. Stock Price's Prediction with Decision Tree [J]. Applied Mechanics and Materials, 2011, 48/49: 1116-1121.

[5] Wu Y G, Huang G F. Motion Vector Generation for Video Coding by Gray Prediction [J]. IET Computer Vision, 2011, 5(1): 14-22.

[6] Hofwing M, Strömberg N. D-Optimality of Non-regular Design Spaces by Using a Bayesian Modification and a Hybrid Method [J]. Structural and Multidisciplinary Optimization, 2010, 42(1): 73-88.

[7] YANG Yun, CHEN Ke. Time Series Clustering via RPCL Network Ensemble with Different Representations [J]. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, 2011, 41(2): 190-199.

[8] Song C H, Jiang Y, Liao W F. Solution of TSP Based on the Improved GA [C]//Proceedings of the 2nd International Conference on Modelling and Simulation. Nanchester: Academic Press, 2009, 8: 382-387.

[9] Lipponen J A, Tarvainen M P, Laitinen T, et al. A Principal Component Regression Approach for Estimation of Ventricular Repolarization Characteristics [J]. IEEE Transactions on Biomedical Engineering, 2010, 57(5): 1062-1069.

[10] XIAO Xue-zhong, HUANG Hua, MA Li-zhuang. RBF Network-Based Temporal Color Morphing [J]. Computer Animation and Virtual Worlds, 2010, 21(3/4): 289-296.

[11] ZHOU Shui-sheng, LIU Hong-wei, YE Feng, et al. A New Iterative Algorithm Training SVM [J]. Optimization Methods and Software, 2009, 24(6): 913-932.

(责任编辑：韩　啸)