文章编号：1006－9348(2011)04－0115－04

支持向量机在电信客户流失预测中的应用研究

王观玉，郭 勇
(黔南民族师范学院计算机科学系，贵州 都匀 558001)

摘要：客户流失分析与预测是客户关系管理的重要内容。由于电信客户的特征呈高度非线性、严重冗余和高维数，传统方法无法消除数据之间冗余和捕获非线性规律，导致预测精度较低。为了提高电信客户流失预测精度，提出一种基于主成份分析（PCA)支持向量机（SVM)的电信客户流失预测方法（PCA－SVM)。首先利用主成分分析对原始数据进行特征降维，消除冗余，然后将得到的主成分作为非线性支持向量机的输入进行学习建模。对某电信公司客户流失数据进行了仿真，实验结果表明，PCA－SVM获得的命中率、覆盖率、准确率和提升系数远远高于其它预测方法。说明主成分分析结合支持向量机的数据挖掘方法具有很好的预测效果，为电信客户流失预测提供了一种新方法。

关键词：支持向量机；主成分分析；客户流失；预测
中图分类号：TP311.52 文献标识码：B

Study on Telecom Customer Leaving Prediction Based on Support Vector Machine

WANG Guan－yu, GUO－Yong
(Qiannan Normal College for Nationalities, Duyun Guizhou 558001, China)

ABSTRACT: Customer Leaving analysis and prediction is an important content of the customer relationship management. Features of Telecom Customer data are highly redundant and nolinear, therefore traditional method cannot eliminate data redundancy and draw the nonlinear rule, and the prediction accuracy is very low. In order to improve the accuracy of telecom customer leaving prediction, a new method is proposed based on principal component analysis (PCA) and support vector machine (SVM) in this paper. The original high dimensional is lowered by principal component analysis and principal components are determined. The low dimensional data sets are used as the inputs of support vector machine predictor. The experimental results of customer leaving prediction for a telecommunication carrier show that the PCA－SVM method is superior to traditional method in hit rate, covering rate, accuracy rate and lift coefficient. This research indicates that the data mining method of PCA－SVM has a good prediction effect and can work as a new method for customer leaving prediction.

KEYWORDS: Support vector machine (SVM); Principal component analysis (PCA); Customer leaving; Prediction

1 引言

随着通讯工具的日益普及，电信行业之间争取客户、扩大市场份额的竞争日益激烈。按照最新电信行业成本结构核算，流失一个已有客户的代价是发展一个新客户所带来利润的 5倍。因此在日趋饱和的客户市场中，如何预测客户的流失成为工作的重中之重[1]。

数据挖掘技术能够通过创建预测客户行为的模型，发现大量数据背后隐藏的重要信息，使营销变得更加准确而迅速[2]。目前针对客户流失问题，学者们提出了以下两类方法：第一类方法是传统分类方法，如决策树[3]、Logistic回归[4]、贝叶斯分类器和聚类分析[5]。该类方法主要特点可以对定类数据和连续性客户数据进行处理，对于所构建的模型有较强的可解释性。但是，该类方法在处理大规模、高维度、含有非线性关系、非正态分布、有时间顺序的客户数据时，不能保证所建模型的泛化能力；第二类方法是人工智能分类方法，如人工神经网络、自组织映射和进化学习算法。该类方法一定程度上能克服第一类方法面临的困难，不仅具有非线性映射能力和泛化能力，而且具有较好的鲁棒性和预测精度，但此类方法主要依靠经验风险最小化原则[6]，容易导致

收稿日期：2010－07－26

泛化能力下降且模型结构难以确定。这些不足极大地限制了上述方法在实际中的应用。因此，探索新的客户流失预测方法的研究工作仍方兴未艾[7]。

针对电信银行客户流失数据的特点，本文提出一种基于支持向量机的电信客户流失预测模型。通过主成分分析对高维度、非线性的电信客户数据进行降维和消除冗余处理，然后通过 10折交叉验证法对支持向量机参数进行优化，最后建立优化的预测模型。通过利用国内电信客户数据对其未来潜在的客户进行预测，与决策树、贝叶斯分类器和人工神经网络等方法进行了对比，本文提高出的方法精确度更高，预测速度也明显加快。

2 电信客户流失预测原理

客户流失模型主要是对电信客户一定时间内流失与否的一种判断，其本质是一种分类问题，即将现有客户分为两类：有流失倾向的客户和无流失倾向的客户[8]。客户流失预测就是用样本数据库（其中既包含有一段时间内已流失的客户数据，也包含有未流失的客户数据）通过在一段时间内对客户的使用业务、缴费、服务满意情况和是否离网等因素的分析，建立起客户流失预测模型，找出影响不同客户群体是否会发生流失的规律性知识。然后可以利用这个模型对某一阶段所跟踪到的客户相关数据，分析其流失的概率。

在客户流失预测中，由于受多种因素的影响，常常不容易发现重要的特征，特征提取一般采用专家经验方法，其提取的特征往往带有主观性和猜测性，忽略了多个属性间的非线性相关性，从而导致预测效果不理想。为了消除特征提取的主观性和考虑属性间的非线性关系，需要从不同角度采用不同方法从客户流失有关数据中提取多种特征，并按照与预测（分类）有关的评价准则来挑选出最有效的特征子集，才能有效地捕捉电信客户流失变化规律，达到对电信客户流失进行准确的预测。

本文将能够快速降维、消除数据之间的高度冗余，提高对预测结果贡献大的特征的主成分分析方法和在数据处于非线性、高维度情况下分类较精确的优点支持向量机相结合进行电信客户流失预测建模。这样充分利用 PCA的特征提取能力和支持向量机优异非线性预测能力，来提高模型预测精度和加快预测速度。预测原理如图 1所示。

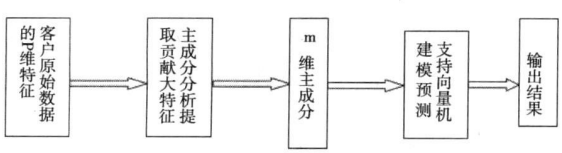


图 1 主成分的支持向量机客户流失预测原理图

3 基于 PCA— SVM的电信客户流失预测

3.1 主成分分析

主成分分析（Principal Component Analysis PCA）是一种统计分析技术，利用降低维度的思想在损失很少信息的前提下把多个指标转化为几个综合指标的多元统计方法，其中每个主成分为原始属性的线性组合且各个主成分之间线性无关。

设与电信客户流失数据相关的 P维随机变量 $X=(X_1, X_2, \cdots, X_p)^T$，设随机变量 X为 μ，协方差矩阵为Σ，Σ 的特征值为：

$$\begin{cases} Y_1 = \mu_{11}X_1 + \mu_{12}X_2 + \cdots + \mu_{1p}X_p \\ Y_2 = \mu_{21}X_1 + \mu_{22}X_2 + \cdots + \mu_{2p}X_p \\ \cdots\cdots \\ Y_p = \mu_{p1}X_1 + \mu_{p2}X_2 + \cdots + \mu_{pp}X_p \end{cases} \quad (1)$$

对此式加以约束：

1）$\mu_i' \mu_i = 1$，即 $\mu_{i1}^2 + \mu_{i2}^2 + \cdots + \mu_{ip}^2 = 1$（$i = 1, 2, \cdots, p$）；

2）Y_i 与 Y_j 与不相关；

3）Y_1 是 $X_1, X_2, \cdots, X_p$ 的一切满足原则的线性组合中方差最大者；是与不相关的所有线性组合中方差最大者，依次类推。

基于以上三原则决定的综合变量 $Y_1, Y_2, \cdots, Y_p$ 分别称为原始变量的第一、第二、…… 第 P个主成分。贡献率是指某个主成分提取的信息占总信息的比率，设第 K个主成分的贡献率 Y_k 的贡献率为：

$$Y_k = \lambda_k / \sum_{i=1}^{p} \lambda_i \quad (2)$$

其中前 m个主成分的贡献率之和为 $\sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{p} \lambda_i$ 称为 $Y_1, Y_2, \cdots, Y_m$ 的累计贡献率。实际应用中，考虑到既要保持原始数据的主要信息又能很好地进行属性约简，通常取 $70\% \leqslant \sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{p} \lambda_i \leqslant 90\%$，这样前 m个属性就可以作为支持向量机的输入。

3.2 支持向量机

随着统计学习理论的不断发展，使得支持向量机（Support Vector Machine SVM）方法于 1995年间应运而生，其是实现结构风险最小化的一个有效途径，具有良好的推广性和较好的分类精确性，已经成为继模式识别和神经网络研究之后机器学习领域新的研究热点[9]。

支持向量机的回归算法的基本思想是通过一个非线性映射 Φ，将数据 x_i 映射到高维特征空间 F，并在这个空间进行线性回归。设定 K个样本数据集 $\{(x_i, y_i), i = 1, 2, \cdots, k\}$，k为样本个数，具体表现形式如下：

$$\omega^T \Phi(x) + b = 0 \quad (3)$$

其中，ω 为超平面的权值向量，b为偏置量。

本文采用 ε 不敏感损失函数（ε— insensitive cost function）支持向量机进行预测。ε 可用下式描述：

$$L_\varepsilon = \begin{cases} |f(x) - y| - \varepsilon & |f(x) - y| \geqslant \varepsilon \\ 0 & |f(x) - y| < \varepsilon \end{cases} \quad (4)$$

为了使训练集上获得的回归模型具有更好的推广能力，

不但要考虑经验风险的最小化，同时还要设法降低模型的复杂度。在这种理念指导下，SVM回归实际上就是求解一个优化问题：

$$\min_{w,b,\xi_i,\xi_i^*} = \frac{1}{2}\omega^T\omega + c\sum_{i=1}^{l}(\xi_i + \xi_i^*) \quad (5)$$

其中，ξ，ξ_i^* 为松弛变量。C表示模型复杂度和样本拟合精度之间的折衷，其值越大，拟合程度越高，这样相应支持向量回归估计函数为：

$$f(x) = \sum_{i=1}^{l}(a_i - a_i^*)k(x - x_i) + b \quad (6)$$

3.3 PCA—SVM的电信客户流失预测

本文利用 PCA—SVM模型进行电信客户流失预测，首先对影响电信客户的影响因子进行主成分分析，提取贡献率大的因子，对样本的属性进行维数处理，同时消除向量之间的冗余性和非线性关系，然后把所得到的主要影响因子作为支持向量机输入样本的主成分，利用支持向量机对训练集进行学习和训练，得到支持向量机预测模型，最后利用得到的预测模型对预测集进行预测，输出客户流失模型的预测值，其预测流程见图 2。

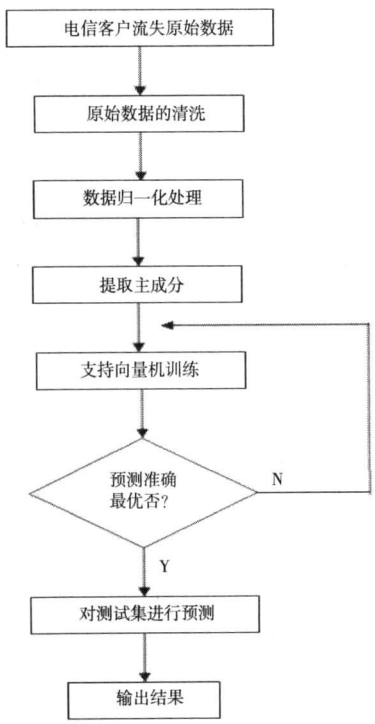


图 2 基于 PCA—SVM客户流失预测流程

4 仿真

4.1 原始数据及指标体系的构建

本文使用某市电信公司 2006年的真实客户数据来验证模型，包括 30个属性，40万条数据。根据经验以及相关文献选取影响客户流失的因素，包括客户的性别、年龄、入网时间等基本属性和客户每月消费金额以及每月通话时长（分）等属性作为预测指标体系。

4.2 时间窗口的确定

为了能够比较好的观测客户的行为变化，本课题的分析数据涵盖 2006年 2至 6月的数据，如图 3所示 6月份流失客户，选择 3、4、5月数据作为建模数据；而 7月份流失客户，选取 4、5、6月数据做为建模数据。其中 5月份，6月份流失客户数据作为训练数据；7月份流失客户数据作为测试数据。

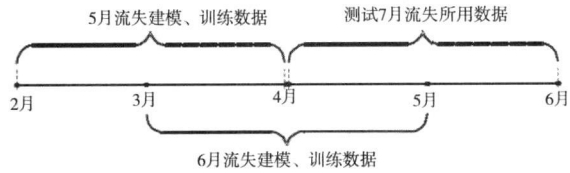


图 3 客户流失中的时间窗口

4.3 数据清洗

原始数据中不可避免地存在着一些空缺值、噪声数据、不正确数据等“脏”数据，它们可能使建模过程陷入混乱，导致不可靠的输出，因此需要进行数据清洗。数据记录中有些字段的值为空，对这样的记录要做相应的处理。如果样本数据库中的数据量巨大，可以用 SQL语句将空值记录给删除掉，不会对后面的分析产生任何影响。如果样本数据有限，可考虑将这些字段为空的值取非空记录的平均值、或赋最常见的值、或根据值的分布情况赋其它的值。还存在着一些数据属于噪音或异常情况，某些属性的取值没有落在相应的区间之内，如盗打行为，使得月通话时长变化太大。一方面根据业务知识对这些属性设置一个正常的取值区间，也可以根据该属性的均值和方差设置一个区间，落在区间之外的记录称为异常或噪音，将这类记录从样本数据库中删除。

4.4 数据的归一化处理

属性的取值区间变化较大，在进行属性相关性分析时，取值较大的属性比取值小的属性会产生更大的影响。为了使分析不受到取值范围不同的影响，在分析前对数据进行标准化处理，使它们都处于相似的区间。本文采用最小—最大规范化的方法，其计算公式如下：

$$v' = \frac{v - \min X}{\max X - \min X} \quad (7)$$

其中，$\min X$和 $\max X$分别为属性 X的最小和最大值，这样属性 X的值映射到[0，1]区间。

4.5 提取主成分

本文利用 DPS6.5对影响因子进行相主成份分析，将其按贡献率大小依次排序主成分分析结果见表 1从表 1可知，前面 12个主成份贡献率已经达到了 85.517这样，主成分 m=12其它的成分可作为噪声项不纳入训练集的输入变量中，将 12主成分作为支持向量机样本的属性。

表 1 主成分的贡献率表

No	特征值	百分率	累计百分率
1	7.19298	39.961	39.961
2	2.77413	15.41391	55.375
3	1.04041	4.88966	60.265
4	0.93183	5.17535	65.440
5	0.87843	4.87987	70.320
6	0.78409	4.35566	74.675
7	0.51175	2.84088	77.516
8	0.39071	2.1716	79.688
9	0.37113	2.06124	81.749
10	0.27857	1.54593	83.295
11	0.22072	1.22642	84.522
12	0.17889	0.99591	85.517
13	0.14596	0.81168	86.329
14	0.10858	0.60431	86.933
15	0.07921	0.43966	87.373
16	0.05963	0.33464	87.708
17	0.04806	0.26789	87.976
18	0.02403	0.13439	88.110
	……		
30	0.00012	0.000025	100

4.6 结果与分析

支持向量机选择径向基核函数，利用 10折交叉验证算法进行参数优化，然后选择 12个主成分进行训练建预测。为了验证本文提出的模型的优劣性，参比模型为：决策树、BP神经网络贝叶斯网络及不进行主成分分析的支持向量机（SVM）。模型的评价指标为模型准确率、命中率、覆盖率和提升系数，各模型预测分析结果比较见表 3。

表 2 预测流失计算指标

样本中客户状态	预测流失	预测非流失失
实际流失	A	B
实际非流失	C	D

$$\text{模型准确率} = \frac{(A+D)}{(A+B+C+D)} \tag{8}$$

$$\text{命中率} = \frac{A}{(A+C)} \tag{9}$$

$$\text{覆盖率} = \frac{A}{(A+B)} \tag{10}$$

$$\text{提升系数} = \frac{\text{命中率}}{\text{测试数据中的客户流失率}} \tag{11}$$

表 3 各模型预测结果

模型	准确率	命中率	覆盖率	提升系数
决策树	0.881	0.612	0.405	4.219
BPNN	0.950	0.902	0.281	6.564
贝叶斯网络	0.926	0.767	0.262	5.732
SVM	0.947	0.853	0.520	6.893
PCA— SVM	0.956	0.895	0.567	7.655

从表 3评比指标结果来看，本文提出的 PCA— SVM模型预测准确率比所有的对比模型都要好，其他指标均也具有一定的优势。在表 3中的 BPNN覆盖率为 0.281，由此可知，该方法出现了过拟合现象。而 SVM能解决非线性、高维和局部极小问题，且分类面简单、泛化能力强、拟合精度高，该模型较好地解决了决策树、神经网络的一些缺点。另外，相比普通的 SVM，PCA— SVM可大大降低原始属性集的维数，减少特征选择的代价，简化预测模型的结构，缩短预测模型的学习时间和提高预测效果，降维使模型具有更高的精确性和泛化性。

5 结束语

针对当前电信客户流失数据的高维、非线性特点以及传统方法的不足，本文把主成分法和支持向量机有机的结合，首先通过主成分分析法对电信客户样本数据有效降维和消除重复处理，然后利用支持向量机在数据处于小样本、非线性及维度高情况下分类较精确的优点，建立了基于支持向量的电信客户流失预测模型。该模型较好地解决了决策树、神经网络和贝叶斯网络等方法的一些缺陷，同时相对于标准的 SVM，又能够降维消除噪声，使模型具有更高的精确性和泛化性。实例证明，基于主成分分析的支持向量机电信客户流失模型具有广泛的应用前景和应用价值。然而，由于支持向量机核函数还没有统一的选择标准，因此，下一步的研究主要是针对选择更合理的支持向量机核函数类型和它们的参数。同时，对于其他行业的客户流失预测应用也是未来的研究方向。

参考文献：

[1] 罗布 马蒂森. 电信业客户流失管理—电信管理精选译丛[M]. 北京：人民邮电出版社，2006.

[2] 贾琳，李明. 基于数据挖掘的电信客户流失模型的建立与实现[J]. 计算机工程与应用，2004，40(4)：185—187.

[3] 盛昭瀚，柳炳祥. 客户流失危机分析的决策树方法[J]. 管理科学学报，2005，8(2)：20—25.

[4] 郭明，郑惠莉，卢毓伟. 基于贝叶斯网络的客户流失分析[J]. 南京邮电大学学报(自然科学版)，2005，25(5)：79—83.

[5] 李艳美，张卓奎. 基于贝叶斯网络的数据挖掘方法[J]. 计算机仿真，2009，26(12)：27—31.

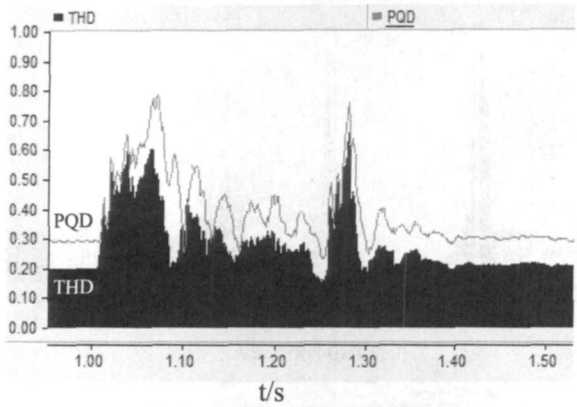
[6] 叶良. 数据挖掘技术在证券客户关系中的应用[J]. 计算机仿真，2008，25(2)：87—389.

[7] 赵宇，李兵，李秀. 基于改进支持向量机的客户流失分析研究[J]. 计算机集成制造系统，2007，13(1)：202—207.

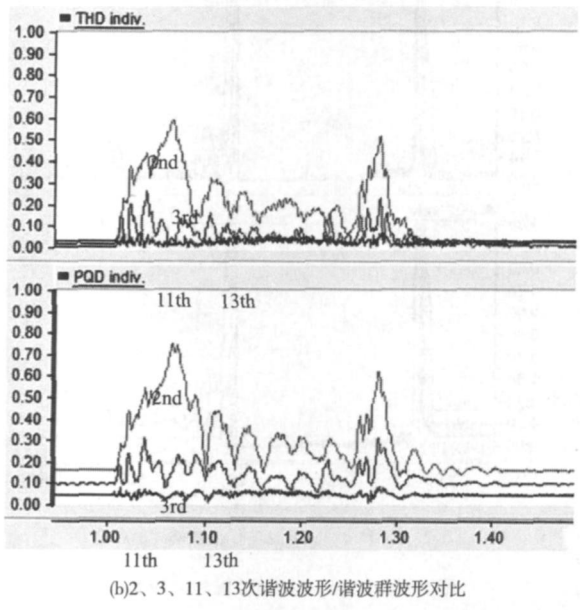
[8] Lemmensa C Croux. Bagging and boosting classification trees to predict churn[J]. Journal of Marketing Research，2005，43(2)：276—286.

[9] 邓乃扬，田英杰. 数据挖掘中的新方法—支持向量机[M]. 北京：科学出版社，2004. 25—120.

(下转第 312页)



(a) THD/PQD波形对比



(b)2、3、11、13次谐波波形/谐波群波形对比

图 5 逆变侧三相接地故障时间段波形

电网技术, 2007, 31(18): 43—47.

[3] A E Emanuel, J A Orr, D Cyganski, E M Gulachenski. A Survey of Harmonic Voltage and Currents at Distribution Substations[J]. IEEE Transaction on Power Delivery, 1991, 6(4): 1883—1890.

[4] 张宇辉, 金国彬, 李天云. 用对称分量法检测间谐波背景下的基波分量[J]. 电网技术, 2006, 30(4): 31—35.

[5] 沈睿佼, 杨洪耕, 吴昊. 基于奇异值总体最小二乘法的间谐波估计算法[J]. 电网技术, 2006, 30(23): 45—49.

[6] 邓淑娟, 等. 新型工业变流系统间谐波与次谐波特性分析[J]. 电网技术, 2009, 33(5): 28—32.

[7] 李晶, 等. 一种用于电力系统谐波与间谐波分析的超分辨率算法[J]. 中国电机工程学报, 2006, 26(15): 35—39.

[8] 李季, 等. 一种滤波换相换流器工作机理与稳态模型[J]. 电工技术学报, 2008, 23(8): 53—59.

[9] 许加柱, 等. 自耦补偿与谐波屏蔽换流变压器的接线方案和原理研究[J]. 电工技术学报, 2006, 21(9): 44—50.

[10] 罗隆福, 等. 基于新型换流变压器的谐波治理研究[J]. 高压电器, 2006, 42(2): 96—98.

[11] 张学武, 闫萍. 电力系统谐波分析及仿真研究[J]. 计算机仿真, 2005, 22(9): 195—200.

[12] F J Harris. On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform[C]. Proceedings of the IEEE, 1978, 66(1): 549—557.

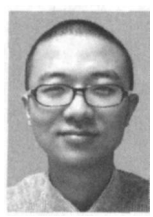
[13] 柯勇, 陶以彬, 王世华. 间谐波检测的 FFT算法改进和 DSP实现[J]. 北京科技大学学报, 2008, 30(10): 1194—1199.

[14] S Chin, N Moo Yong, P Chang Phil, P Mok. A Digital Measurement Scheme for Time—Varying Transient Harmonics[J]. IEEE Transactions on Power Delivery, 1995, 10(2): 588—594.

[15] A E Emanuel, J A Orr, D Cyganski, E M Gulachenski. A Survey of Harmonic Voltage and Currents at Distribution Substations[J]. IEEE Transactions on Power Delivery, 1991, 6(4): 1883—1890.

[16] A Domijan, G T Heydt, A P S Meliopoulos. Directions of Research on Electric Power Quality[J]. IEEE Trans. on Power Delivery, 1993, 8(1): 429—436.

[作者简介]



潘 侃(1985—), 男(汉族), 湖南郴州人, 硕士研究生, 主要研究领域为基于新型换流变压器的直流输电系统谐波抑制与无功补偿等。

罗隆福(1962—), 男(汉族), 湖南人, 教授, 博士生导师, 中国电机工程学会高级会员, 主要研究领域为现代电器设备的设计和优化、新型换流变压器的研制和高压直流输电新理论等研究工作。

许加柱(1980—), 男(汉族), 安徽人, 副教授, 主要研究领域为电器设备的设计优化及仿真研究, 自耦补偿与谐波屏蔽换流变压器的研制及对应的高压直流输电新理论研究工作。

李 勇(1982—), 男(汉族), 河南人, 博士研究生, 主要研究领域为基于新型换流变压器的直流输电系统新理论及谐波抑制与无功补偿研究。

(上接第 118页)

[作者简介]

王观玉(1964—), 女(汉族), 贵州瓮安人, 硕士, 副教授, 主要研究方向: 信息管理及信息技术应用, 信息技术教育。

郭 勇(1971—), 男(汉族), 贵州贵定人, 硕士, 副教授, 主要研究方向: 信息系统开发, 数据库及网络应用技术。