Web Recruitment Enterprise Customers Free to Pay Prediction Based on PCA_BP: A Case Study

JIANG Guorui [1], **SI Xuefeng** [2]
1. School of Economics and Management, Beijing University of Technology, Beijing ,China
2. School of Economics and Management, Beijing University of Technology, Beijing ,China
1. e-mail jianggr@bjut.edu.cn, 2. e-mail sixuefeng@emails.bjut.edu.cn

Abstract: To the problem of web recruitment enterprise customers free to pay prediction, PCA_BP algorithm is proposed in this paper. Through extracted and analysis the enterprise customers online behavior log data to build data mart as prediction data source, The PCA method is used to data attribute reduce and then BP neural network algorithm is used for customers pay prediction. Further study shows that, PCA_BP is better than PCA_RBF to the prediction results. To describe the PCA_BP application, a well-known web recruitment site enterprise customers to pay prediction is discussed as a example and the results proved that compared with BP algorithm, the PCA_BP is more precise and efficient.

Keywords: web recruitment; web mining; PCA; BP neural network; pay prediction

基于 PCA_BP 神经网络的网络招聘企业客户付费预测实证研究

蒋国瑞 [1],**司学峰** [2]
1. 北京工业大学经济与管理学院，北京，中国，100124
2. 北京工业大学经济与管理学院，北京，中国，100124
1. e-mail jianggr@bjut.edu.cn, 2. e-mail sixuefeng@emails.bjut.edu.cn

【摘要】针对网络招聘企业客户由免费转为付费预测问题，本文提出了基于主成分分析(PCA)与 BP 神经网络相结合的组合建模方法，通过提取企业客户在线行为日志数据构建数据集市作为预测数据源，利用 PCA 方法剔除干扰信息、进行数据属性约减，应用 BP 神经网络算法进行客户付费预测，结果表明采用 PCA_BP 算法相比 BP 算法简化了网络拓扑结构、减少的训练次数，模型提升度也有明显改善。进一步研究表明，PCA_BP 比 PCA_RBF 有更好的预测效果。以某知名招聘网站为例描述了的方法应用，证明了 PCA_BP 方法的良好的预测效果。

【关键词】网络招聘；web 挖掘；主成分分析；BP 神经网络；付费预测

1 引言

网络招聘是传统人力资源管理与现代网络技术相结合的产物，网络招聘最早出现在美国，以其招聘范围广、招聘信息全、招聘方式便捷、时效性强等优点成为企业招聘的主要形式。随着网络的普及与发展，网络招聘市场规模巨大且增长迅速，网络招聘已经成为网络经济最成功的商业应用之一，据统计全球每天通过网络发布超过 2000 万条就业信息，3000 多万求职者通过网络发出求职简历。目前，网络招聘在欧美国家已经取代传统的印刷媒体的招聘广告，成为企业招聘的首选。在国内，网络招聘起步相对较晚，但发展较快，采用网络招聘的企业多集中在经济发达地区，而在中小城市特别是中小企业仍以常规的印刷媒体招聘广告或现场招聘的形式为主[1]。据艾瑞公司《2007-2008 中国网络招聘行业研究报告》显示（图 1），2007 年中国网络招聘市场规模达到 9.7 亿人民币，预计 2011 年中国网络招聘市场规模达到 26.3 亿，环比增长 27.7%。2007 年，前程无忧、中华英才网、智联招聘网等三大综合类招聘网站的招聘收入居市场前三甲，三家占中国网络招聘总收入的 70%，另外，行业细分与地方性招聘

978-1-935068-01-3 © 2009 SciRes.

网站，如中智英才网、卓越人才网、北京人才网等网络招聘收入也在不断增长。国内网络招聘方兴未艾，巨大的网络招聘市场与潜力也是专业网络招聘网站吸引国外风险投资的重要原因所在，竞争日益激烈的网络招聘市场在促使网络招聘网站提供更加专业化、个性化服务的同时，各大网站也在积极开拓新的业务与市场。另外，新的地方特色浓厚、行业细分明确类招聘网站不断涌现，这更加剧了网络招聘市场的激烈竞争态势。

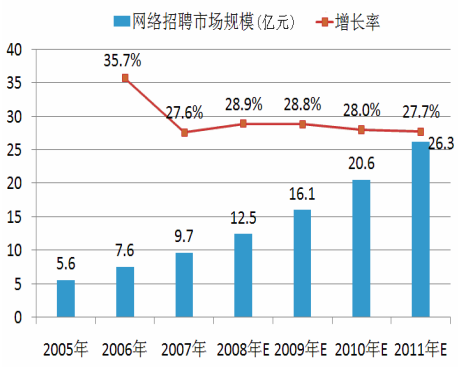

图 1. 2005-2011 年中国网络招聘市场规模及预测

招聘网站在运营初期或实力雄厚的综合类网络招聘网站在开拓新市场时，为了占领市场、获得高点击量、吸引企业发布职位、增加求职者在线注册量等，往往对求职者提供免费服务的同时，在一定的期限内对企业客户也采取免费服务的方式。根据企业客户在免费期间的招聘效果对其进行持续跟进与客户关系维护，适时推出付费产品服务，以促使企业客户由免费向付费转换。然而企业客户能否在免费期内进行付费涉及诸多因素，不但涉及到企业的规模、性质、招聘需求还涉及发布职位、招聘效果、及客户招聘页面浏览量等诸多因素。本文案例为国内某知名综合类招聘网站在开拓地级市业务时遇到的问题，由于未对免费企业客户在线行为进行深入分析与挖掘，导致企业客户有免费到付费的转换率比较低，而如果企业用户在免费期内的黄金时间不能付费，后续的付费概率就更小，导致公司投资回报率很低，市场开拓困难重重。

为了提高企业客户付费转换率，在公司人、财、物资源相对有限的情况下，必须对网络招聘企业客户基本特征及在线行为进行分析挖掘，对企业客户是否付费做出预测[2-5]。BP 神经网络是一种有效的预测技术与方法，然而，若 BP 前端输入特征过多，不但会使网络拓扑结构复杂化，而且会降低网络的训练次数和预测效果。本文提出了采用主成分分析 PCA 与 BP 神经网络相结合的方法对客户付费行为进行预测。首先对企业客户在线行为日志数据集进行预处理，建立影响客户付费的关键性能指标 KPI，然后结合 CRM 系统数据库中的客户基本信息及产品数据库中付费客户信息形成客户预测数据源，再通过 PCA 方法对源数据进行处理，降低维度并删除冗余信息，最后，采用 BP 算法进行客户付费行为进行预测建模。

2 算法原理

2.1 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 $X_1, X_2, \cdots X_p$ 为招聘企业付费相关的 P 个特征信息，记 $X=(X_1, X_2, \cdots X_p)^T$，$\mu$ 为 X 的均值，Σ 为 X 的协方差矩阵，Σ 的特征值为：$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$，相应的正交单位化特征向量为：$e_1, e_2 \cdots e_p$。对 X 进行线性变换后形成新的综合变量 Y：$Y_i = e_i^T X = \sum_{j=1}^{p} e_{ij} X_j$

其中：称 Y_i 为第 i 个主成分，

Y_i与Y_j线性无关$(i \neq j; i, j = 1, 2, \cdots p)$；

Y_1为$X_1, X_2, \cdots X_p$中线性组合中方差最大者；

Y_2为$X_1, X_2, \cdots X_p$中线性组合中方差次最大者，且与Y线性无关；

以此类推。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 $\lambda_k \Big/ \sum_{i=1}^{p} \lambda_i$。其中前 m 个主成分的贡献率之和为

$\sum_{i=1}^{m}\lambda_i \Big/ \sum_{i=1}^{p}\lambda_i$ 称为 $Y_1, Y_2, \cdots Y_m$ 的累计贡献率。实际应用中，考虑到即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 $80\% \le \sum_{i=1}^{m}\lambda_i \Big/ \sum_{i=1}^{p}\lambda_i \le 90\%$ 。

2.2 BP 神经网络

BP 神经网络(Back-propagation Neutral Network)是一种高效的目标分类器，由输入层、隐藏层及输出层构成，各层间通过神经元进行连接，采用梯度下降算法调节连接权值以最佳拟合输入-输出对组成的训练集合，被广泛应用于模式分类、语音识别、预测建模等多个领域[6-7]。

BP 神经网络是一种有监督的学习算法，模型的学习训练由两个环节组成，输入层信号的正向转播和误差信号的方向转播。在正向传播过程中，数据由输入层经隐藏层逐层先前传播至输出层，通过输出层神经元的实际输出值与目标输出进行对比，如果达不到预期输出，则通过 delta 法则反向传播调整各层神经元的连接权，通过反复训练直到输出层神经元的输出结果与目标值小于最小期望误差。

2.3 RBF 神经网络

径向基函数(RBF)是一种以函数逼近理论为基础而构造的一类前向网络,RBF 的基本结构分为三层：输入层、RBF 层、输出层。拟合与插值是函数逼近理论的重要组成部分，插值函数又成基函数，插值问题表述为：对于一个包含 N 个不同点的集合 $\{x_i \in R^n \mid i = 1, 2, \cdots, N\}$ 和相应的 N 个实数的一个集合 $\{d_i \in R^1 \mid i = 1, 2, \cdots, N\}$，

寻找一个函数 $F: R^n \to R^1$ 满足插值条件

$F(x_i) = d_i$ 。RBF 技术就是要寻找一个函数具有形式：$F(x) = \sum_{i=1}^{N} w_i \Phi(\|x - x_i\|)$，其中 $\{\Phi(\|x - x_i\| \mid i = 1, 2, \cdots, N\}$ 是 N 个任意函数的集合，称为径向基函数。本文中径向基函数采用高斯函数 $\exp\left[-\|x - c_i\| / 2\sigma_i^2\right]$ $i = 1, 2, \cdots, m$ 。

其中，x是n维输入向量；

c_i 是第i个基函数的中心；

σ_i是第i个感知的变量，决定了该基函数围绕中心点的宽度；m 是单元个数。

在 RBF 网络中，输入层到隐藏层的映射为非线性的，即隐藏层中神经元作用的函数为非线性函数，而隐藏层到输出层则是线性的，能够把不易处理的非线性问题线性化。

3 实证研究

3.1 案例背景及开发工具

本案例为某知名综合类人才招聘网站，规模收入处于国内领先位置。在开拓地级市业务过程中，首先通过电话营销的方式向当地中小企业客户免费推广其网络招聘平台，在企业免费试用期间，通过持续跟进客户进行网络招聘产品促销以使免费客户转换为付费客户。然而，在客户跟进过程中，面临的主要问题是无法判断客户是否会付费，无法对客户付费进行有深入的分析与预测，导致客户跟进比较盲目，客户付费转换率比较低，公司投资回报率较差。 针对客户在线行为通过分析其日志数据，对价值客户进行识别并展开有针对性的营销策略是一种有益的探索[8]。

本文采用 SQL Server 2005 中的 Integration Services 工具通过提取公司 CRM 数据库中的客户基本信息、客户在线行为日志数据及销售数据库中客户信息建立数据集市作为预测的数据源，预测模型采用数据挖掘工具 Clementine 中的 PCA 与神经网络工具箱[9-10]。

3.2 数据获取

从数据集市中抽取 2007 年 3 月到 2007 年 11 间共计 10131 家客户数据，其中已经转换为付费客户 724 家，仍在免费期内客户 7896 家，免费期终止未转换为付费的客户 1509 家。数据字段的选择通过与业内人士的反复论证，把客户付费预测模型数据源字段归为客户基本信息、客户活跃信息，发布职位信息，职位浏览、收到简历信息等 5 大类别共计 33 个属性。一方面，客户在线活跃度高，招聘需求旺盛，招聘效果好容易付费。其次，客户付费并不一定有实际的招聘需求，而是通过招

978-1-935068-01-3 © 2009 SciRes.

聘网站进行广告宣传。另外，通过免费期内已付费用户分析发现，在半年的试用期内，客户能否付费跟其已经试用的时间也有很大关系如图 3 所示，付费客户在试用期第一个月付费比率为 18%，第二个月占 38%，而在试用期的最后一个月仅为 2%。

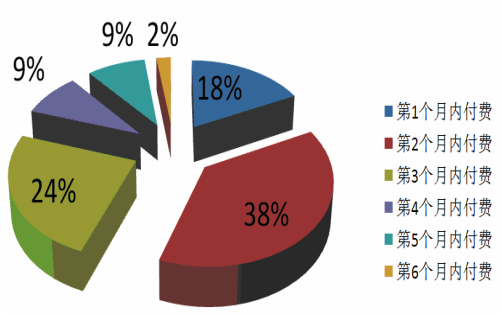

图 2：免费期内客户付费时间分布

客户基本信息来源于招聘网站在前期推广过程中获取的客户数据，包括公司名称，企业规模，企业性质，注册资金，成立时间等；而其余四类信息则涉及客户在线行为日志数据，数据的获取要从网络日志数据库中对客户在线行为数据进行抽取、转换、加载即 ETL 过程形成数据集市[11]，构建与客户付费预测密切相关的 KPI 指标，包括客户激活到现在的时间（月），客户平均登陆系统次数，发布职位总数，发布职位总次数，登陆系统次数，收到简历数量，客户发布职位被浏览次数等如表 1 所示。

表 1. 客户预测指标

信息类别	数据属性
客户基本信息	企业性质，成立时间，注册资金，企业规模，客户类型(免费客户、付费客户)等
客户活跃信息	客户激活距当前时间(天)，最近登入时间，登陆系统次数，登陆系统频率(天)等
发布职位信息	发布职位总数数，最近 2 周发布职位数，职位更新次数，平均发布职位数/次等
职位浏览信息	职位总浏览量，最近 2 周职位浏览量，职位最大浏览量，职位平均浏览量等
收到简历信息	收到简历总数，最近 2 周收到简历数，职位收到简历最大数量，职位平均收到简历数等

3.3 数据预处理

采用主成分分析法对预测数据源进行预处理，计算得到源数据对应的特征根 $\lambda_1,\lambda_2,\lambda_3,\cdots\lambda_{33}$ 为 18.9785, 6.1634, 3.7032, 1.904, 1.0031, 0.7034,0.5738, 0.1135, ⋯，0.0001，经计算 5 个特征较大值的方差贡献率已经达到 83.7%，把特征向量与原始样本数据相乘得到的主成分数据源如表 2 所示。

表 2. 预测源数据主成分值

编号	PC1	PC2	PC3	PC4	PC5
1	-0.7591	-0.5544	-0.2030	1.7442	-0.1575
2	0.6869	-0.2795	-0.3218	0.4019	-1.1948
3	-1.2914	1.1571	1.0390	-0.6803	0.4122
4	-1.2142	1.5826	0.8404	0.3614	1.7704
⋮	⋮	⋮	⋮	⋮	⋮
	0.3075	-0.0401	-0.4904	0.2385	0.9507
	-0.2583	0.6949	-1.6279	-0.7793	0.3424
	0.5960	0.0750	1.0696	-1.9256	-1.0416
⋮	⋮	⋮	⋮	⋮	⋮
10130	-0.5372	-0.0297	0.5346	1.6847	-0.7787
10131	-0.4268	-0.6512	0.1406	-0.3916	0.2559

3.4 预测建模

预测建模过程如图 3 所示，步骤如下：

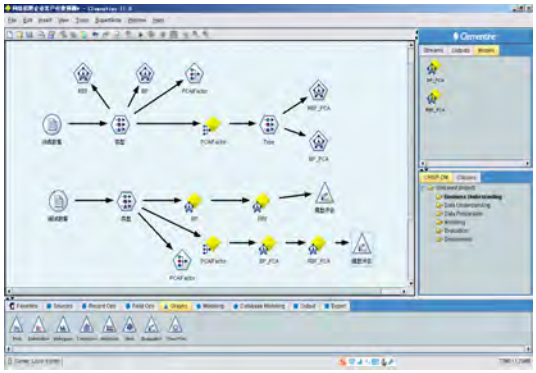

图 3：预测建模

1) 把源数据按 7:3 比例划分为训练集与测试集
2) 把训练数据依次通过 BP 神经网络，RBF 网络建模建立 BP 及 RBF 模型，将训练数据通过 PCA 处理后建立 PCA_BP 及 PCA_RBF 模型
3) 通过测试数据评估 BP 神经网络及 RBF 网络模型，将通过 PCA 处理后的数据评估 PCA_BP 及 PCA_RBF 模型

为了比较两种算法预测的精确度，采用提升度进行评

价。提升度 Lift 是数据挖掘中评价模型性能的常用方法，用于比较应用模型与不用模型时，预测能力提高的倍数。以本文而言，

$lift = (\text{付费客户数}/\text{样本数量})/(\text{付费客户数}/\text{客户总数})$，

lift 图的纵轴表示 lift 值，对应点的横轴表示样本数量占样本的比值，如果模型起初有比较高提升度值，随着抽样比例的增加逐渐降为 1 则说明模型预测性能好。图 4 表明，PCA_BP 的提升度相比 BP 神经网络有明显提高，且比 PCA_RBF 网络预测效果好，而 PCA 应用于 RBF 模型相比 BP，提升度并不明显改进。

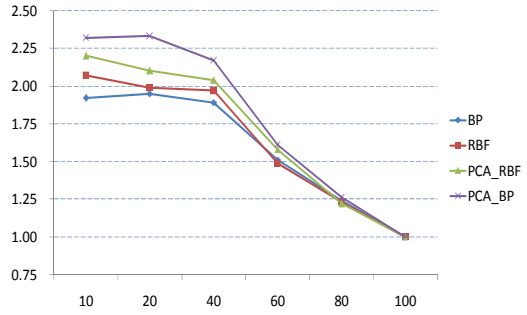

图 4. 预测性能评估

4 结束语

本文分析了我国网络招聘市场特点和前景，针对招聘网站业务推广中，对免费企业客户后期跟进效果差、客户付费率低，客户流失严重的问题，通过基于数据挖掘技术分析客户在线行，对目标客户通过基于 PCA_BP 神经网络进行预测和识别，取得了良好的预测效果。同时，通过对比分析 BP 神经网络与 RBF 网络的预测效果发现，PCA 技术应用于 BP 神经网络后预测效果有明显提高，而应用于 RBF 网络则无明显提高。另外，基于 PCA_BP 神经网络技术预测模型的提升度也比 PCA_RBF 提升度高。对于网络招聘企业客户行为的分析研究有一定的借鉴和知道意义。

致谢

本文得到国家自然科学基金会的支持
(基金号：70639002)

References (参考文献)

[1] 熊军.网络招聘的应用研究[J].科技管理研究,2006.11:153−155
Xiong jun. web recruitment application research [J] Science Management Research, 2006.11:153−155.

[2] Berson A, Smith K, Thearing K. Building data mining applications for CRM[M]. Mc Graw-Hill, NewYork, 2000, 207−211.

[3] Reichheld F F, Sasser WE. Zero defections: quality comes to service [J]. Harvard Business Review, 1990, 68 (5), 105−111.

[4] Rosset S., Neumann E. Integrating Customer Value Considerations into Predictive Modeling [C].
Third IEEE International Conference on Data Mining, 2003

[5] M Kitayama, R Matsubara, Y lzui. Application of data mining to customer profile analysis in the power electric industry[C]. 2002 IEEE Power Engineering Society Winter Meeting, 2002, 632−634.
M Kantardzie, AN SRIVASTAVA

[6] JIN Xue-xiang, ZHANG Yi, YAO Dan-ya. Simultaneously Prediction of Network Traffic Flow Based on PCA- SVR[C]. Lecture Notes on Computer Science. Springer-Verlag, 2007, 4492: 1022−1031.

[7] 杨静, 毛宗源. 基于 PCA 和神经网络的识别方法研究[J].计算机工程与应用,2007,43(25):246−248.
Yang Jing,Mao Song Yuan.Recognition method based on principal component analysis and neural network[J].Computer Engineering and Applications, 2007,43(25):246−248.

[8] 郭岩等.网络日志规模分析和用户兴趣挖掘[J].计算机学报,2005,28(9):1483−1496
Guo Yan et al.. Analyzing Scale of Web Logs and Mining Users' Interests [J]. Chinese Journal of Computers, 2005, 28 (9): 1483−1496.

[9] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data (2000).SIGKDD Explorations, Vol. 1, Issue 2, 2000.

[10] Alex Buchner and Maurice D Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. SIGMOD Record, 27 (4): 54−61, 1998.

[11] BAO Yubin; SONG Jie; LENG Fangling; WANG Daling.Study and Implementation of a New SQL-Based ETL Approach [J]. Wuhan University Journal of Natural Sciences. 2007, 12 (5): 804−808.