doi：10.11830/ISSN.1000-5013.201605022

电信客户流失的组合预测模型

余路[1,2]

（1. 西南大学 计算机与信息科技学院，重庆 北碚 400715；
2. 重庆涪陵广播电视大学 教务处，重庆 涪陵 408000）

摘要： 针对电信行业客户流失的问题，设计基于决策树 C5.0、BP 神经网络及 Logistic 回归算法的组合预测模型，并对某电信企业进行客户流失预测. 预测结果表明：与单一客户流失预测模型相比，组合预测模型命中准确率高，预测效果好，更能直观地显示出流失客户的基本特征.

关键词： 客户流失；预测模型；电信企业；决策树 C5.0；BP 神经网络；Logistic 回归算法

中图分类号： TP 311.5 **文献标志码：** A **文章编号：** 1000-5013(2016)05-0637-04

Combination Forecasting Model of Customer Churns in Telecom Industry

YU Lu[1,2]

（1. School of Computer and Information Science, Southwestern University, Chongqing 400715, China;
2. Teaching Affair Office, Chongqing Fuling Radio and television University, Chongqing 408000, China）

Abstract: According to telecommunication customer churn problem, the forecasting model based on decision tree C5.0, BP (back-propagation) neural network and logistic regression algorithm combination is designed, and according to orecasting of the customer churns in some telecom companies, the accuracy is higher and prediction effect is good in combination forecasting model compared to a single customer churn prediction model. It shows the basic features of the customer churn more directly.

Keywords: customer churn; forecasting model; telecom industry; decision tree C5.0; back-propagation neural network; logistic regression algorithm

电信市场的竞争愈来愈激烈，为使企业的利润最大化，各通信运营商都把争取更多的客户作为营销的最终目标. 但是随着竞争的不断加剧，客户流失成为各企业运营过程中面临的主要问题，不仅使市场份额减少，还会出现客户恶意离网产生欠费行为，增加了企业的运营成本，造成严重的经济损失[1]. 有分析称，开发一个新的客户比挽留一个老的客户所产生的成本高很多倍[2]. 因此，做好客户关系管理，防止客户流失是通信行业提升企业核心竞争力的有效手段. 针对以往客户关系管理过程中无法监控客户流失的问题，将数据挖掘技术应用到通信客户流失预警分析中，利用其强大的数据分析手段，建立客户消费特征等属性与客户流失可能性之间的关联模型，可实现对客户状态的实时监控. 因此，寻求一种有效的建模与评估方案是研究人员关注的重点[3-5]. 针对决策树、神经网络及逻辑回归 3 种单一算法的模型特点和预测效果，本文尝试建立一种基于 3 种算法的组合预测模型，并应用所建模型对某电信企业进行

收稿日期： 2016-06-20

通信作者： 余路(1972-)，男，讲师，博士，主要从事计算机数据库技术的研究. E-mail：flddyl@126.com.

基金项目： 重庆市自然科学技术研究项目(KJ131302)

客户流失预测,以验证模型的有效性.

1 数据挖掘理论

1.1 数据挖掘的定义

数据挖掘是利用数据分类算法在海量的、随机分布的数据中提取隐含在数据当中的,能为人们提供决策作用的信息的过程[6].数据挖掘包含两方面含义:一是能够处理海量数据;二是具有挖掘探索的能力.因强调从海量数据中获取信息的过程,所以数据挖掘技术更侧重于后者.

1.2 数据挖掘算法

1.2.1 *决策树分类方法* 决策树基于信息增益理论,通过分析样本中的数据挖掘其中的知识和规律,是目前应用最广泛的数据分类算法之一.决策树结构包含了若干个节点和分支,其中,节点表示某个属性上的测试,分支则表示测试的结果.常见的决策树算法有ID3,C4.5/C5.0等[7-9],主要用于事件的预测分析.决策树预测过程分两步进行:一是利用训练集建立并进化一棵决策树;二是测试各节点的属性值,对输入数据进行分类,用该类的属性值完成预测对象的估计.

1.2.2 *神经网络分类方法* 作为一种人脑思想仿真的数据分析模式,神经网络以海量数据并行处理和计算为基础,用于描述认知,决策等智能控制行为.典型的神经网络的模型结构包括输入层、隐含层和输出层,由若干神经元连接而成,如图1所示.BP神经网络是应用最广泛的神经网络算法,其输出表达式[10-11]为

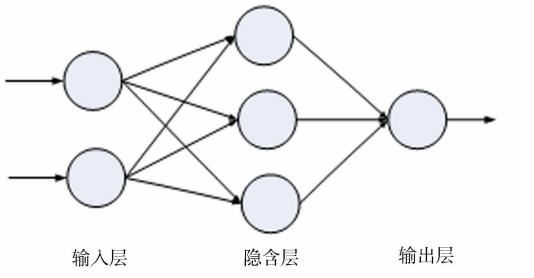

图1 神经网络模型结构

Fig. 1 Model structure of neural network

$$H = f_i(\sum w_{i,j}x_i + \theta_j). \tag{1}$$

式(1)中:$w_{i,j}$为连接权系数;f_j为激励函数;θ_j为神经元的阈值;x_i为神经元的输入.

BP神经网络采用有师学习方的式进行训练,能够实现任何复杂非线性映射的功能,其训练过程以输出误差最小为原则,逐层修正各连接权系数和阈值,其训练过程如图2所示.

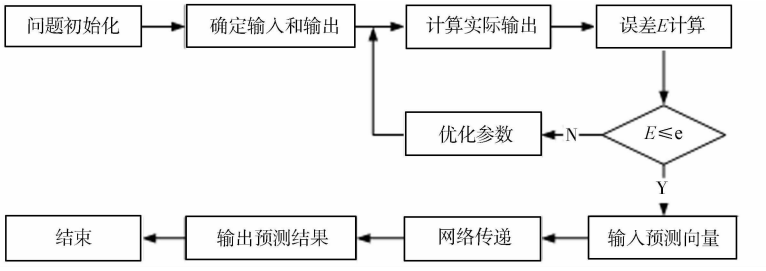

图2 神经网络的训练过程

Fig. 2 Training process of neural network

1.2.3 *逻辑回归分类方法* 逻辑回归的思想来源于多元线性回归,与多元回归连续性变量不同,逻辑回归的因变量是非连续性的变量.逻辑回归主要用来预测某种情况下事件发生的概率,一般用于处理二值型因变量,一般用"1"或"0"代表预测结果[12-13].

设事件发生的影响因素为m个变量,用向量$\boldsymbol{X}'=(X_1,X_2,X_3,\cdots,X_m)$表示;根据观测量相对于某事件发生的概率为条件概率,用$P(Y=1|x)=p$表示,则逻辑回归的模型可表示为

$$P(Y=1\mid x)=\frac{1}{1+\mathrm{e}^{-g(x)}}. \tag{2}$$

2 多组合预测模型的建立与评价

2.1 组合预测模型的建立

针对典型分类算法的特点,在开放式数据挖掘工具Clementine中建立基于决策树、神经网络及逻辑回归算法的组合客户流失模型,构造Lagrange函数[14]为

$$L(\alpha_1,\alpha_2,\alpha_3)=\sum_{i=1}^{3}[(\alpha_1x_i+\alpha_2y_i+\alpha_3z_i-x_i)^2+(\alpha_1x_i+\alpha_2y_i+\alpha_3z_i-y_i)^2+$$

$$(\alpha_1 x_i+\alpha_2 y_i+\alpha_3 z_i-z_i)^2+\lambda(\alpha_1 x_i+\alpha_2 y_i+\alpha_3 z_i-1)]. \tag{3}$$

式(3)中：x_i，y_i，z_i 分别为 C5.0，BP 和 Logistic 回归算法单一模型的预测值；λ 为 Lagrange 算子；α_k 为组合预测的权重系数，且 $k=1,2,3$.

将式(3)对组合权重系数求偏导数，得到 α_k 的极值 $\alpha_k{}^*$，令 φ_i^* 为第 i 个预测对象的组合预测值，则其表达式为 $\varphi_i^*=\alpha_1^* x_i+\alpha_1^* y_i+\alpha_1^* z_i$. 利用组合模型预测有以下 5 个步骤.

步骤 1　将预处理后得到的数据集进行划分：文中的划分比例为训练集占 60%，测试集占 40%.

步骤 2　选用决策树 C5.0、BP 神经网络及逻辑回归 3 个基本分类模型分别对训练集进行建模.

步骤 3　将测试集中的样本数据带入前面建好的模型中进行预测，得到预测分析结果.

步骤 4　分别将 3 种单一模型的预测结果带入构造好的 Lagrange 函数，得到多算法组合预测模型的权重系数，从而建立组合预测模型.

步骤 5　计算预测结果.

基于 Lagrange 函数的多算法组合模型的预测流程，如图 3 所示.

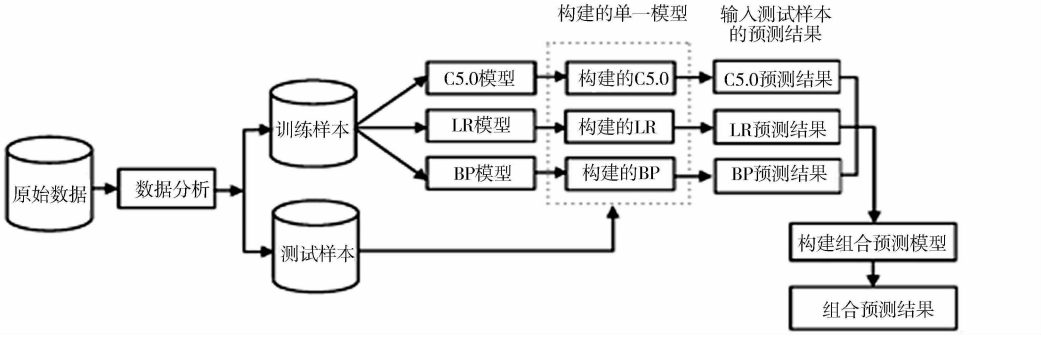

图 3　组合模型预测流程

Fig. 3　Forecasting process of combination model

2.2　模型评价

分别采用单一算法模型与多算法组合模型进行预测分析，预测结果如表 1 所示. 由于论文篇幅的限制，这里仅列出其中的 10 个预测结果. 为进一步分析不同模型算法的预测准确程度，对表 1 中各预测结果进行统计，结果表明：C5.0 模型的命中率为 88.95%；LR 模型的命中率为 87.38%；BP 模型的命中率为 87.11%；组合模型的命中率为 92.07%.

基于 Lagrange 的多算法组合预测模型集合了各单一模型的预测优势，大大提高了客户流失的预测命中率，达到 92.07%，比单一模型的预测命中率提升了近 5%.

假设某电信运营企业的流失客户数为 300 000 个，根据各模型算法的预测命中率进行计算，分别得到单一模型和组合模型预测客户流失的数量及误判率，如表 2 所示. 表 2 中：m 为流失数量；n 为误判人数；η 为误差率；w 为误判损失.

表 1　不同模型算法的客户流失预测结果

Tab. 1　Customer churn prediction results for different model algorithms

编号	流失标志	各模型预测结果			
		C5.0 模型	LR 模型	BP 模型	组合模型
100001	0.0	0.0	0.0	0.0	0.0
100002	0.0	0.0	0.0	0.0	0.0
100003	0.0	0.0	0.0	0.0	0.0
100004	1.0	0.0	1.0	1.0	1.0
100005	0.0	0.0	0.0	0.0	0.0
100006	0.0	0.0	0.0	0.0	0.0
100007	0.0	0.0	0.0	0.0	0.0
100008	1.0	1.0	1.0	1.0	1.0
100009	0.0	0.0	0.0	0.0	0.0
100010	0.0	0.0	0.0	0.0	0.0

表2 组流失量预测及误判率对比结果

Tab. 2 Comparison results for prediction and erroneous judgement rate of group loss amount

模型	m	n	η/%	w/万元
实际流失	300 000	—	—	—
C5.0模型	265 498	34 502	11.5	103.5
LR模型	268 652	31 345	10.4	94.0
BP模型	274 583	25 417	8.5	76.3
组合模型	283 564	16 436	5.5	49.3

由表2可知:在客户流失数量的预测中,多算法组合模型的误判人数明显减小,预测误差率仅为实际数量的5.5%;设每个人的月均消费为30元,那么由组合模型所造成的误判损失也将大大降低,仅是单一模型预测损失的一半左右.由此可见,与单一客户流失预测模型相比,基于Lagrange的多算法组合模型预测效果好,可有效预测客户流失和流失倾向,达到预测期望,企业可针对预测结果制定相应的避免客户流失的对策.

3 结束语

客户流失是通信行业运行过程中常见的问题,直接影响到运营商的企业效益.数据挖掘可以根据客户信息、消费行为等历史数据判断客户流失的可能性,避免因营销手段的盲目性造成的成本浪费.对决策树C5.0、BP神经网络和Lagrange回归算法3种典型数据分类方法进行分析,针对单一模型客户流失预测建模的特点,建立了基于Lagrange函数的组合预测模型.预测结果表明:所建立的组合模型对电信客户流失预测命中率大幅提高,预测效果好,能有效获取客户的流失倾向,使电信企业营销方案的制定更具针对性.

参考文献:

[1] 夏国恩.客户流失预测的现状与发展研究[J].计算机应用研究,2010,27(2):151-153.

[2] 张线媚.数据挖掘在电信行业客户流失预测中的应用[J].微型机与应用,2015,34(15):99-102.

[3] 刘光远,苑森淼,董立岩.数据挖掘方法在用户流失预测分析中的应用[J].计算机工程与应用,2007,43(9):154-156.

[4] 郭俊芳,周生宝.基于联合决策树的客户流失预测模型设计[J].计算机与现代化,2010(5):5-7.

[5] 尹婷,覃锡忠,贾振红,等.基于WEKA的客户流失预测研究[J].激光杂志,2013,34(5):44-46.

[6] 仲继.电信企业客户流失预测模型研究[D].西安:西安科技大学,2011:21-22.

[7] 张晓滨,高峰,黄慧.基于客户细分的客户流失预测研究[J].计算机工程与设计,2009,30(24):5755-5758.

[8] 王晓华.电信数据挖掘的数据质量评估技术研究[D].杭州:浙江大学,2010:7-10.

[9] 潘大胜,屈迟文.一种改进ID3型决策树挖掘算法[J].华侨大学学报(自然科学版),2016,37(1):71-73.

[10] CONG H E, REN Lihong, DING Yongsheng. Performance prediction of carbon fiber protofilament based on SA-GA-SVR [J]. Journal of Donghua University,2014,31(2):92-97.

[11] 李爱群,乔晗,王汝传,等.基于分布式混合数据挖掘的电信客户流失分析[J].计算机技术与发展,2010,20(10):43-46.

[12] 朱龙.利润约束的关联规则挖掘算法[J].华侨大学学报(自然科学版),2015,36(9):522-526.

[13] THANGAPARVATHI B,ANANDHAVALLI D,SHALINIE S M. A high speed decision tree classifier algorithm for huge dataste[C]// IEEE-International Conference on Recent Trends in Information Technology. [S. l.]:IEEE Press,2011,10(6):695-700.

[14] 迟准.电信运营企业客户流失预测与评价研究[D].哈尔滨:哈尔滨工程大学,2013:73-74.

(**责任编辑**:陈志贤 **英文审校**:吴逢铁)