基于 WEKA 平台的移动客户流量消费分析

戴 琳,张 悦,韦 玉,景子倩,张 沫,宫 婧
(南京邮电大学 理学院,江苏 南京 210000)

摘 要:随着移动互联网的飞速发展,手机网民规模迅速扩张,作为移动互联网关键环节的中国移动正面临着这一机遇与挑战;如何根据用户的业务使用情况,对移动客户流量消费进行分析是增加业务收入、提高用户满意度的重要研究课题。文中主要研究了基于 WEKA 平台的移动客户流量消费分析。首先,进行客户群与客户发展趋势的细分,对用户业务数据进行特征选择、数据清洗以及数据类型转换的预处理。其次,以客户群作为添加属性,以客户发展趋势作为目标属性,基于 WEKA 平台的决策树算法对预处理后的业务数据进行分析,建立手机上网用户的决策树模型。最后,根据移动公司提供的 2 万条客户业务数据对模型进行验证。结果表明,当样本数在 10 000 至 20 000 时,模型有很好的分类预测效果,能够挖掘出潜在的高流量用户,从而达到精确营销的目的。

关键词:移动客户流量消费;WEKA;决策树;分类预测;精确营销
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2016)01-0115-04
doi:10.3969/j.issn.1673-629X.2016.01.024

Analysis of Mobile Customer Traffic Consumption Based on WEKA Platform

DAI Lin, ZHANG Yue, WEI Yu, JING Zi-qian, ZHANG Mo, GONG Jing
(School of Science, Nanjing University of Posts and Telecommunications, Nanjing 210000, China)

Abstract: With the rapid development of mobile Internet, mobile Internet users scale expands rapidly, the China mobile, as the key link of the mobile Internet, is facing the opportunities and challenges. How to analyze the mobile client traffic consumption according to the user's business is important research subject to increase revenue and improve customer satisfaction. The analysis of the mobile customer traffic consumption based on WEKA platform is studied. Firstly, subdivide the development trend of customer base and customer, selecting the user business data feature, cleaning the data and converting the data types. Secondly, adding customers as property, development trend of the customer as the target attribute, analyze business data after pretreatment based on the decision tree algorithm on WEKA platform, mobile Internet users of the decision tree model is established. Lastly, verify this model according to the mobile 20 000 customer business data provided by the company. The results show that the model has good classification prediction effect when the number of samples is from 10 000 to 15 000, able to dig out the potential high flow users so as to achieve the purpose of precise marketing.

Key words: mobile customer traffic consumption; WEKA; decision tree; classification prediction; precise marketing

0 引 言

现如今,移动通信流量业务的发展变得高速化多样化,经营竞争环境愈演愈烈,对该行业的服务需求提出了更高、更新的要求。流量时代客户的流量消费行为具有更大的弹性和更大的粘性。而移动通信业流量业务的爆炸性增长也成为移动运营商必须面对的问题。利用数据挖掘[1-2]在这些海量数据背后及时发现有用的知识,提高流量信息利用率,满足客户需求,实现精细化营销[3]变得十分重要。如何尽量满足客户对流量的多样需求,如今对移动通信业具有革命性的意义。

一直以来,国内外学者致力于改进决策树算

收稿日期:2015-04-14 **修回日期**:2015-07-16 **网络出版时间**:2016-01-04
基金项目:国家自然科学基金资助项目(61373135);江苏省高校自然科学研究重大项目(12KJA52003);南京邮电大学大学生科技创新训练计划(STITP)(XYB2014154)
作者简介:戴 琳(1994-),男,研究方向为数据挖掘与大数据分析;张 沫,讲师,研究方向为分布式计算和数据挖掘;宫 婧,副教授,研究生导师,研究方向为数据挖掘、模式识别、智能算法等。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160104.1510.042.html

法[4-5]来对客户进行分类,从而预测潜在的高流量用户[6]。实践表明,算法的改进确实提高了模型的效果,但是,改进算法毕竟只是一方面,若能从其他方面双管齐下,必然会取得意想不到的效果。文中的创新点在于先对客户群进行细分,并添加客户群作为客户的属性,最后建立手机上网用户的决策树模型。

1 数据预处理

数据预处理[7-8]的效果会直接影响到模型的性能与分类预测的效果。一方面,通过对数据格式和内容的调整,可以使建立的模型更准确、简单且便于理解;另一方面,可以降低学习算法的时间和空间复杂度。文中先将客户群与客户发展趋势作为客户新衍生出的属性,然后基于新数据进行数据的微处理,包括特征选择[9]、数据清洗[10]以及数据类型的转换。

1.1 客户群的细分

分析客户业务数据,对其进一步处理得到客户群的细分,将其分为四类,如图 1 所示。

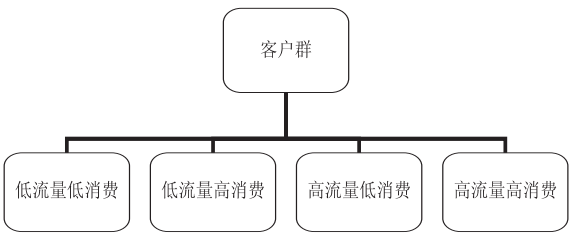

图 1 客户群细分

为将客户群划分为以上四类,文中定义了客户群阈值作为细分客户群的标准。

(1)客户群阈值的定义。

客户群阈值:移动互联网用户属于哪一类客户群的分界值。文中给出了两大标准:客户的月平均使用流量和客户的月平均消费额。

(2)客户群阈值的确定。

文中通过对客户的月平均使用流量和客户月平均消费额进行分析,给出了各种客户群的判断阈值,如表 1 所示。

表 1 移动用户客户群阈值的判定

客户群	月平均使用流量/MB	月平均消费额/元
低流量低消费	0 ~ 100	0 ~ 50
低流量高消费	0 ~ 100	50 以上
高流量低消费	100 以上	0 ~ 50
高流量高消费	100 以上	50 以上

(3)客户群的应用。

文中将每个客户进行归类,把客户所属客户群作为其添加属性,为建立决策树模型打下基础。

1.2 客户发展趋势的细分

文中为挖掘潜在的高流量用户,定性地将客户发展趋势细分为三类,如图 2 所示。

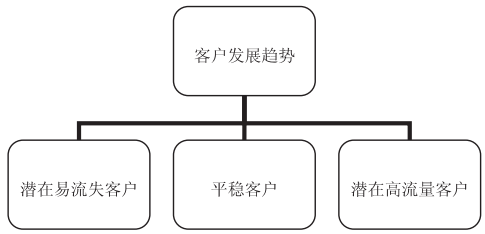

图 2 客户发展趋势细分

为反映客户发展趋势,文中利用客户连续三个月的流量消费情况衍生出流量变化率(BHL)这一属性,并且将客户发展趋势阈值作为图 2 细分的标准。

(1)流量变化率(BHL)的定义。

$$BHL=\frac{当前月份的流量-近三个月的平均流量}{近三个月的平均流量}$$

(2)客户发展趋势阈值的确定。

有关客户发展趋势阈值,由于跟客户群阈值类似,在此不做赘述。各类客户发展趋势阈值判定如表 2 所示。

表 2 客户发展趋势阈值判定

客户发展趋势	BHL
潜在易流失客户	<0
平稳客户	0 ~ 0.2
潜在高流量客户	>0.2

(3)客户发展趋势的应用。

分析每个客户所属的发展趋势,以其作为目标属性,建立决策树模型,能够挖掘出潜在的高流量客户。

1.3 客户业务数据的预处理

(1)特征选择。

特征选择的效果会直接影响到分类模型的性能。通过特征选择,可以减少样本的维度,大大减少计算量,降低时间和空间复杂度,简化学习模型。针对该样本数据集,处理方法如下:

①对于类别值唯一或者类别值众多的特征予以删除,例如地域(该样本数据集针对某地市,所以地域唯一)、用户 ID(类别值众多)等特征。

②利用 spss 对特征之间的相关性进行分析,删除一些与目标特征相关性小的特征,例如通话费、通话时间等与 GPRS 通信流量无关。

(2)数据清洗。

数据清洗的目的是补全数据、处理缺失数据、除去噪声及改进不协调的数据。由于客户业务数据样本较大,文中直接对含缺失值或者含异常数据的样本进行删除。针对该样本数据集,处理方法如下:

①由于该样本数据集样本众多,对于含缺失值的

样本直接删除。

②对于含异常数据的样本直接删除,例如年龄里小于 0 的样本。

③对于已经离网或停机的样本删除。

(3)数据类型转换。

由于原始数据保存在 excel 中,为了能在 WEKA 中打开,必须将原始数据保存为 arff 格式文件。具体方法是:将 excel 的原始数据另存为 csv 文件格式,再在 WEKA 中打开,最后保存为 arff 格式。

其次,基于 WEKA 的 J48 算法[11]对数据类型的要求,文中将数值属性转换为分类属性,如表 3 所示。

表 3　分类属性的定义

符号	含义	符号	含义
A_1	青壮年	V_1	一级 VIP
A_2	中老年	V_2	二级 VIP
A_3	老年	T_1	潜在易流失客户
M	男	T_2	平稳客户
W	女	T_3	潜在高流量客户
B_1	神州行	H_{11}	低流量低消费
B_2	动感地带	H_{12}	低流量高消费客户
B_3	全球通	H_{21}	高流量低消费客户
V_0	非 VIP	H_{22}	高流量高消费客户

2　手机上网用户决策树模型的建立

文中对移动客户流量消费进行分析,重点建立对潜在高流量用户的预测模型。而根据各类算法的优缺点,选择解释比较方便的决策树进行建模。

决策树是对数据进行分类,以此达到预测的目的。WEKA 中的 J48 算法就是决策树 C4.5 算法[12-13],其核心算法是 ID3 算法[14]。ID3 算法是以信息论为基础,以信息熵和信息增益度为衡量标准,从而实现对数据的归纳分类。而 C4.5 算法继承了 ID3 算法的优点,并在以下几方面对 ID3 算法进行了改进:

(1)用信息增益率来选择属性,克服了用信息增益选择属性时偏向选择取值多的属性的不足;

(2)在树构造过程中进行剪枝;

(3)能够完成对连续属性的离散化处理;

(4)能够对不完整数据进行处理。

J48 算法具有产生的分类规则易于理解、准确率较高的优点。因此,基于 WEKA 平台的 J48 算法对数据预处理后的业务数据进行分析。得到的决策树模型如图 3 所示。

依照建好的决策树模型,沿决策树从上到下遍历,在每个节点都会遇到一个测试,对每个节点上问题的不同的测试输出导致不同的分支,最后会到达一个叶子节点。这个过程就是利用决策树进行分类的过程,利用若干个变量来判断所属的类别,从而预测客户在未来的发展趋势,判断其是否为潜在的高流量用户,以此实现精确营销的目的。

图 3　决策树模型

3　模型的理解

由图 3 所建立的决策树模型,可以得到以下重要结论:

(1)客户所属客户群是决策树模型的根节点,因此客户群属性是信息增益值最大的特征属性,即决定客户发展趋势最重要的特征属性。

(2)高流量低消费客户群是潜在的高流量客户,低流量低消费客户群是潜在易流失客户。

(3)高流量高消费客户群中未办理 VIP 服务且年纪较轻的客户是潜在的高流量客户,未办理 VIP 服务而年纪较大的客户是平稳客户;办理一级 VIP 服务中年纪较轻的客户是平稳客户,而办理一级 VIP 服务中年纪较大的客户是潜在易流失客户;办理二级 VIP 服务中男性属于潜在易流失客户,而办理二级 VIP 服务中女性属于平稳客户。

(4)低流量高消费客户中使用神州行服务的客户属于潜在易流失客户,而使用全球通和动感地带服务的客户属于潜在高流量客户。

4　模型的验证

对于以上所建立的决策树模型,文中根据移动公

司提供的 2 万条客户业务数据对模型进行验证。采取的方法是:随机抽取 1 000、2 000、5 000、10 000、20 000 条客户业务数据来预测潜在高流量客户,并将预测结果与实际结果进行比较,得到预测值与实际值的比值,从而验证模型的准确性。结果如图 4 所示。

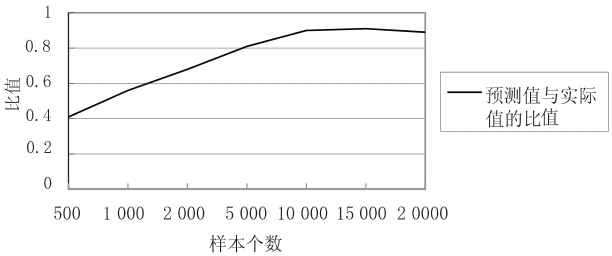

图 4 潜在高流量用户的预测检验

从图 4 中可以看出,当样本数<1 000 时,由于偶然性大,预测值与实际值的比值小于 0.5,说明预测效果并不好。当样本数在 1 000 至 10 000 时,预测值与实际值的比值越来越大,并逐渐接近于 1,说明预测效果越来越好。当样本数在 10 000 到 20 000 之间时,预测值与实际值的比值趋于稳定并最接近于 1,说明预测效果最好。但是当样本数大于 20 000 后,模型的效果有略微下降趋势。

综上,样本数在 10 000 至 20 000 之间时,模型的预测效果较好,从而验证了模型的准确性。

5 结束语

文中通过对移动客户业务数据的预处理包括客户群的细分,建立了手机上网用户的决策树模型,并通过大量的测试数据对模型进行验证与评估,最后发现样本数据在 10 000 到 20 000 之间时预测效果较好。这说明该方法对于分类与预测潜在的高流量用户有较大的改进,从而能更好地为移动运营商适时推荐套餐,实现精确营销提供决策支持。但由于该样本数据集包含客户基本特征有限,例如客户学历、职业等特征的缺少,文中所研究的内容还有待更进一步的深入。

参考文献:

[1] 严霄凤,张德馨. 大数据研究[J]. 计算机技术与发展,2013,23(4):168-172.

[2] 范 明,孟小峰. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2001.

[3] 陈志竞,梁伯瀚. 数据挖掘助力精细化流量经营[J]. 电信科学,2012,28(7):1-5.

[4] 徐 鹏,林 森. 基于 C4.5 决策树的流量分类方法[J]. 软件学报,2009,20(10):2692-2704.

[5] Han Hui, Mao Feng, Wang Wenyuan. Review of recent development in decision tree algorithm in data mining[J]. Application Research of Computers, 2004, 21(12): 5-8.

[6] 黄潇聪. 手机上网零流量用户"破零"模型的研究与应用[J]. 电信科学,2013(S2):26-29.

[7] 董 艳. 数据预处理方法在移动通信行业中的应用[J]. 计算机技术与发展,2010,20(11):225-228.

[8] Fayyad U M. Data mining and knowledge discovery: making sense out of data[J]. IEEE Expert-Intelligent Systems & Their Applications, 1996, 11(5): 20-25.

[9] 张 靖. 面向高维小样本数据的分类特征选择算法研究[D]. 合肥:合肥工业大学,2014.

[10] 郭志懋,周傲英. 数据质量和数据清洗研究综述[J]. 软件学报,2002,13(11):2076-2082.

[11] 赵 蕊. 基于 WEKA 平台的决策树算法设计与实现[D]. 长沙:中南大学,2007.

[12] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques[C]//Proc of SIGMETRICS. Banff: ACM, 2005: 50-60.

[13] Moore A W, Papagiannaki K. Toward the accurate identification of network applications[C]//Proc of LNCS. Heidelberg: Springer-Verlag, 2005: 41-54.

[14] 李 霞. ID3 分类算法在银行客户流失中的应用研究[J]. 计算机技术与发展,2009,19(3):158-160.

(上接第 114 页)

[12] Fanelli G, Dantone M, Gall J, et al. Random forests for real time 3D face analysis[J]. International Journal of Computer Vision, 2013, 101(3): 437-458.

[13] Ma B, Chai X, Wang T. A novel feature descriptor based on biologically inspired feature for head pose estimation[J]. Neurocomputing, 2013, 115: 1-10.

[14] Geng X, Xia Y. Head pose estimation based on multivariate label distribution[C]//Proc of IEEE conf on computer vision and pattern recognition. Columbus, Ohio: IEEE, 2014: 1837-1842.

[15] Murphy-Chutorian E, Trivedi M M. Head pose estimation in computer vision: a survey[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(4): 607-626.

[16] Geng X, Smith-Miles K, Zhou Z H. Facial age estimation by learning from label distributions[C]//Proc of 24th AAAI conf on artificial intelligence. Atlanta: [s. n.], 2010: 451-456.

[17] Geng X, Yin C, Zhou Z H. Facial age estimation by learning from label distributions[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35(10): 2401-2412.

[18] Felzenszwalb P F, Girshick R B, McAllester D A, et al. Object detection with discriminatively trained part-based models[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

基于 WEKA 平台的移动客户流量消费分析

作者： 戴琳， 张悦， 韦玉， 景子倩， 张沫， 宫婧， DAI Lin， ZHANG Yue， WEI Yu， JING Zi-qian， ZHANG Mo， GONG Jing

作者单位： 南京邮电大学 理学院,江苏 南京,210000

刊名： 计算机技术与发展 ISTIC

英文刊名：

年，卷(期)： 2016(1)