应用因子分析和K-MEANS聚类的客户分群建模

彭 凯 秦永彬 许道云
(贵州大学计算机科学与信息学院 贵阳 550025)

摘 要 为挖掘存量用户的潜在数据业务使用需求,研究客户细分成为各电信运营商进行差异化营销所必须解决的问题。利用聚类算法提出了一种解决电信短信业务客户分群的应用模型。首先基于因子分析为复杂参数变量下的数据挖掘有效地减少了冗余字段,提高了模型构建的质量和效率,然后通过无监督的K-MEANS分群算法完成分群。经验证,该短信分群模型具备明显的特征差异性。2009年某西部通信企业应用该模型在数据业务差异化营销中取得了明显的效益。

关键词 增值业务,因子分析,短信渗透率,数据探索,数据训练,时间窗口
中图法分类号 TP312 **文献标识码** A

Customer Segmentation Modeling on Factor Analysis and K-MEANS Clustering

PENG Kai QING Yong-bin XU Dao-yun
(College of Computer Science and Engineering, Guizhou University, Guiyang 550025, China)

Abstract To develop customers' potential demands for data services, the research for customer segmentation has become a primitive work of telecommunications operators in order to run a differentiated users' marketing. Through the use of clustering algorithm, this paper presented a segmentation modeling for differentiating customers using short messaging services in telecommunications operators. Firstly, based on factor analysis, redundant properties were simplified in the complex data mining under variable parameters in order to improve the quality and efficiency of the modeling, and then the customer segmentation model was constructed through unsupervised clustering K-MEANS algorithm. It was verified that the SMS users have the obvious differentiation of characteristics by using the cluster model. In 2009, a western communications enterprise achieved significant benefits with application of the model in the differentiated data service marketing.

Keywords Value-added services, Factor analysis, Perm eability of short message service, Data exploration, Data training, Time interval

1 引言

移动增值业务经过近10年的发展逐渐得到壮大。2009年统计数据显示,各通信运营商非语音业务的收入占比已接近30%,这其中以点对点短信占比最高。目前该业务渗透率已经达到近70%,但近年来业务量发展逐渐开始趋于缓慢。为促进业务量进一步提升,刺激用户潜在需求,各运营商开始通过客户市场细分和提供精细化的精准营销以最低成本挖掘用户业务需求。如何有效地对短信客户进行分群以便实施差异化营销并降低成本,成为目前运营商需要关注的一个重要商业问题。通过分析,短信用户的以下几个特点决定了其用户分群的困难:在中国短信业务用户量大,因为地区、经济、文化等差异导致用户消费特性差异化比较明显;短信业务虽然已经发展近10年,但目前还没有有效的用户分群模型可供运营商学习参考;移动通信类用户特征属性值比较多,海量的数据导致通过专家法很难完成模型的搭建。

针对海量数据的分类问题,目前已经可以通过成熟的数据挖掘算法进行处理,包括分类算法和聚类算法都可以解决大数据量的分群。目前已公布文献提出了部分解决办法,主要有吴斌等人在文献[1]中将基于群体智能的聚类方法分析客户行为,通过选用由小到大的群体相似系数进行聚类,采用递归算法对不同消费特征的客户群体进行分类。梁静国等人在文献[2]中将模糊c均值聚类算法作为客户聚类的方法,为客户群的特征分析提供了量化依据,从而进行客户聚类。曲昭伟等人在文献[3]中利用遗传算法,通过构造模糊相异矩阵,对通信领域高价值客户进行聚类分析,实现了大客户的划分。郑国荣等人在文献[4]中利用基于密度的聚类方法获取高端消费模式的客户特征,并以此为基础了解顾客的消费模式,向其提供满意的产品和服务。吕巍等人在文献[5]中利用K-MEANS方法研究了中国移动市场顾客行为细分,为顾客的细分提出了一种比较适用的分析方法。陈治平在文献[6]中结合聚类算法的分析,将K-MEANS,SOM,BIRCH等聚类

到稿日期:2010-06-10 返修日期:2010-11-07 本文受国家自然科学基金项目(60863005,61011130038),贵州省省长基金(200404)资助。
彭 凯(1980—),男,博士生,工程师,主要研究方向为数据挖掘,E-mail:13908515671@139.com;秦永彬(1980—),男,博士生,讲师,CCF会员,主要研究方向为可计算分析、可信计算;许道云(1959—),男,教授,博士生导师,CCF高级会员,主要研究方向为计算复杂性、可计算分析。

方法应用于电信的客户细分案例分析，结果表明了该分析模型在客户市场细分中的应用有效性，并在此基础上，结合各聚类方法在各指标性能上所体现的区分度的差异，提出了一种聚类方法应用效果评估的方法。

数据挖掘技术在国外已被应用于零售业的销售预测、金融业的客户信用分析、电信业的客户价值分析等方面。目前已经公布的国外文献中已经提及的客户分类方法有：Wayne Thompson 利用 SAS 工具基于聚类和决策树等算法对电信企业用户进行分类，以实现对客户多维度的分析、降低企业运营成本并提升客户服务质量[7]；Kim Su-Yeon 等人提出基于客户价值和生命周期理论，对通信公司的客户进行分类，并在此基础上提出价值提升的相关策略[8]；Hwang Hyunseok 等人对某通信行业的客户建立了一个价值评估模型，基于该模型将客户分为 3 个群：当前价值客户、潜在价值客户、忠诚客户，并针对不同类型的客户提出营销建议[9]；Rho Jae Jeung 提出引用决策树和聚类分析来分析网络交易数据，揭示了互联网渠道的客户行为，并进行了有效的客户分群，提出了有效的网络营销策略[10]。

而针对通信企业的短信客户分群，以上国内、外提出的相关解决方案主要存在以下几个问题：

(1)算法多以传统话音业务为主研究分群，而针对新业务的分群算法少有研究。

(2)部分文献主要关注算法性能的改进，而过于复杂的算法将导致实际商业操作过程中开发和维护成本增加。

(3)由于缺乏商业实践经验，文献中提出的算法没有给出如何进行挖掘前各种分析变量的选择方法，而变量选择成为模型最终能否真实反映市场现象的重要环节。

(4)大部分分类算法主要关注从客户价值角度来进行分群，对于从业务角度(特别是新业务)来进行客户分群几乎没有详细资料。

(5)分类前对输入算法中的各种分析变量的预处理也成为影响模型形成的重要因素，对于在模型质量和可操作性方面寻找一个平衡，已发布的文献中没有提出相关解决方案。

鉴于此，在商业实践中为了使项目成本控制和效益实现达到最好的平衡，我们提出了基于因子分析和 K-MEANS 聚类的短信客户群细分模型。首先采用基于因子分析为复杂参数变量下的数据挖掘有效地减少了冗余字段，提高了模型构建的质量和效率，然后通过无监督的 K-MEANS 分群算法完成分群。选择该方法的优势如下：

(1)聚类分析方法属于非监督型机器学习的数据挖掘方法，适用于较大数据样本和较多变量的分析任务。

(2)短信类业务作为新业务，没有可供参考的分群经验，选择有监督的分类算法将失去依据，所以我们选择没有监督的 K-MEANS 聚类算法就可以满足需求。

(3)通过因子分析，可有效减少数据挖掘过程中输入的冗余挖掘字段，可有效提高模型构建的质量和效率。

(4)由于 K-MEANS 算法可解释性较好，因此算法执行时不断迭代的过程便于业务人员的分析和观察，并根据业务经验决定最合适的迭代终止点。

(5)因子分析和 K-MEANS 在商业领域的灵活搭配应用，可有效减少操作可行性，降低操作成本。

2 数据预处理

点对点短信业务是一项运营了多年的业务，市场品牌认知度也非常高，因此在提升短信量和用户数方面，难度较大。我们将某运营商公司 2009 年 1 月份有发送点对点短信行为且到 2009 年 7 月份客户状态正常的用户作为研究的目标客户。为了确保模型的代表性，我们将利用该批目标用户 1 月份－3 月份的各属性指标的平均值进行分析。

在短信客户分群的研究中，往往需要对反映用户的多个变量进行大量的观测，收集大量数据，以便进行分析，寻找规律。多变量、大样本无疑会为研究提供丰富的信息，但也在一定程度上增加了数据采集的工作量。更重要的是在大多数情况下，许多变量之间可能存在相关性而增加了问题分析的复杂性，同时对分析带来不便。因此需要找到一种合理的方法来减少分析指标，并尽量减少原指标包含信息的损失，对所收集的资料作全面的分析。本次试验中，我们利用因子分析方法来进行降维处理。

假设该模型有观察变量 $X_1, X_2, \cdots, X_P$，每个变量可作如下分解[11]：

$$X_1 = A_{11} + A_{12} * S_1 + \cdots + A_{1n} * S_n$$

$$X_2 = A_{21} + A_{22} * S_1 + \cdots + A_{2n} * S_n$$

$$\cdots\cdots\cdots\cdots$$

$$X_m = A_{m1} + A_{m2} * S_1 + \cdots + A_{mn} * S_n$$

上式为因子模型，其中 $S_1, S_2, \cdots, S_n$ 叫做公共因子，它们是在各个变量中共同出现的因子。A_{ij} 为因子载荷，它是第 i 个变量在第 j 个主因子上的负荷，反映了第 i 个变量在第 j 个主因子上的相对重要性。因子分析的基本问题就是要确定因子载荷。利用因子载荷矩阵，可以得到各个因子的载荷值。试验中，根据业务专家建议，最终选择了信息量之和占总体信息量 85%以上的主因子[12]。

2.1 数据准备

表 1 属性因子载荷得分

公共因子	变量名	因子载荷
Factor 1	点对点短信上行条数	0.951
	工作日点对点短信上行条数	0.949
	短信费用	0.939
	白天点对点短信上行条数	0.923
	点对点短信总收发条数	0.916
	工作日白天点对点短信上行条数	0.913
	节假日点对点短信上行条数	0.901
	点对点短信网内上行条数	0.9
	工作日晚上点对点短信上行条数	0.883
	发送密友点对点短信条数	0.871
	节假日晚上点对点短信上行条数	0.87
	晚上点对点短信上行条数	0.865
	节假日白天点对点短信上行条数	0.856
	点对点短信网外上行条数	0.669
	语音密友发送短信条数	0.485
Factor 2	点对点短信下行条数	0.969
	点对点短信网外下行条数	0.969
	工作日点对点短信下行条数	0.955
	晚上点对点短信下行条数	0.936
	工作日晚上点对点短信下行条数	0.932
	节假日点对点短信下行条数	0.927
	白天点对点短信下行条数	0.907
	节假日晚上点对点短信下行条数	0.9
	工作日白天点对点短信下行条数	0.875
	节假日白天点对点短信下行条数	0.851

续表

公共因子	变量名	因子载荷
Factor 3	普通语音呼叫密友通话时长	0.891
	普通语音呼叫密友通话次数	0.88
	普通语音呼叫密友通话时长占总普通语音呼叫时长比例	0.85
	普通语音呼叫密友通话次数占总普通语音呼叫次数比例	0.813
	普通语音主被叫密友通话时长	0.738
	普通语音主被叫密友通话次数	0.735
	普通语音主被叫密友通话时长占总普通语音主被叫时长比例	0.644
	普通语音主被叫密友通话次数占总普通语音主被叫次数比例	0.631
	发送密友点对点短信条数占总发送条数的比例	0.476
Factor 4	点对点短信上行关联号码数	0.828
	点对点短信总收发关联号码数	0.812
	点对点短信网内上行关联号码数	0.808
	点对点短信网外上行关联号码数	0.628
Factor 5	梦网短信总收发条数	0.994
	梦网短信下行条数	0.994
	新业务费用	0.991
	消费水平	0.869
Factor 6	点对点短信上行关联号码数占总收发关联号码数的比例	0.94
	上行条数占总收发条数的比例	0.928
	工作日点对点短信上行条数占总上行条数比例	0.751
	点对点短信网内上行关联号码数占总上行关联号码数的比例	0.67
Factor 7	网外上行条数占总上行条数的比例	0.944
	点对点短信网外上行关联号码数占总上行关联号码数的比例	0.933
	网内上行条数占总上行条数的比例	-0.71
Factor 8	gprs 闲时流量	0.904
	gprs 忙时流量	0.891
Factor 9	WAP 使用时长	0.908
	WAP 使用次数	0.807
	WAP 平均上网时间	0.74
Factor 10	长途通话费用	0.738
	漫游通话费用	0.612
	本地通话费用	0.611
Factor 11	普通语音接听密友通话时长占总普通语音接通时长比例	0.856
	普通语音接听密友通话次数占总普通语音接通次数比例	0.84
	普通语音接听密友通话次数	0.813
	普通语音接听密友通话时长	0.81
Factor 12	工作日点对点短信下行条数占总下行条数比例	0.778
	白天点对点短信下行条数占总下行条数比例	0.751
	点对点短信网外下行关联号码数	0.698
	点对点短信下行关联号码数	0.698
	节假日点对点短信下行条数占总下行条数比例	0.696
	晚上点对点短信下行条数占总下行条数比例	0.68
Factor 13	晚上点对点短信上行条数占总上行条数比例	0.795
	白天点对点短信上行条数占总上行条数比例	0.77
Factor 14	节假日上行条数占比	0.927
Factor 15	点对点短信网内下行条数	0.815
	点对点短信网内下行关联号码数	0.803
Factor 16	语音密友发送条数占比	0.74
	语音密友中有发送短信号码数	0.69
	语音密友号码数(除去固话)	0.68
	语音密友中有发送短信号码数占比	0.54
Factor 17	点对点短信平均上行次数	0.715
	点对点短信平均次数	0.708
	点对点短信密友平均上行次数	0.504
Factor 18	gprs 平均上网时间	0.877
	gprs 使用时长	0.859
	gprs 使用次数	0.434

通过专家法，我们采用 2009 年 1 月份发过短信的客户在 2009 年 1 月—3 月的数据来生成数据分析和探索的宽表。宽表中主要包含以下类型的变量：点对点短信信息(发送条数、发送网内条数等)、账务(ARPU、短信费用、新业务费用等)、通话行为(本地、漫游话次数等)、交往圈(网内、网外)。初始细分变量主要用来将客户进行分群，最后得到的各个变量的细分变量的因子分析结果如表 1 所列。

2.2 描述变量

在初始挖掘阶段，为了有效地描述各个群体的特征，我们提出如表 2 所列的描述变量(该标量在进行分群时不做分群变量)。

表 2 描述变量表

变量类型	描述变量	变量类型	描述变量	变量类型	描述变量
语音呼叫时长	普通语音通话时长	语音呼叫次数	普通语音通话次数	属性特征	品牌
	主叫通话时长		主叫通话次数		性别
	被叫通话时长		被叫通话次数		VIP 客户标志
	呼叫转移时长		呼叫转移次数		在网时长
	IP 通话主叫时长		IP 通话次数		年龄
	非漫游通话时长		非漫游通话次数		短信套餐类型
	漫游通话时长		漫游通话次数		套餐短信剩余条数(包月条数－发送条数)
	网内通话时长		网内通话次数		ARPU
	网外通话时长		网外通话次数		优惠金额
	白天通话时长		白天通话次数		是否飞信活跃用户
	晚上通话时长		晚上通话次数		飞信活跃天数
	闲时通话次数		闲时通话次数		
	忙时通话次数		忙时通话次数		普通语音通话费用
	闲时通话时长比例		闲时通话时长比例		IP 通话费用
	非漫游通主叫话时长		非漫游主叫通话次数	消费情况	网内通话费用
	非漫游通被叫话时长		非漫游被叫通话次数		网外通话费用
	非漫游市话时长		非漫游市话次数		白天通话费用
	非漫游长途时长		非漫游长途次数		晚上通话费用
	漫游通主叫话时长		漫游主叫通话次数		省内漫游费用
	漫游通被叫话时长		漫游被叫通话次数		省际漫游费用
	长途通话主叫时长		长途通话主叫次数		国际漫游费用(包含港澳台)
	长途通话被叫时长		长途通话被叫次数		
	网内市话通话时长		网内市话通话次数		语音联系号码数
	网内长途通话时长		网内长途通话次数	交往圈	主叫号码数
	网内主叫通话时长		网内主叫通话次数		被叫号码数
	网内被叫通话时长		网内被叫通话次数		网内联系号码数
	网外主叫通话时长		网外主叫通话次数		网外联系号码数
	网外被叫通话时长		网外被叫通话次数		网内联系号码数占比
	白天长途通话时长		白天长途通话次数		网外联系号码数占比
	晚上长途通话时长		晚上长途通话次数		
	网内通话次数/总通话次数				
	网外通话次数/总通话次数				
	网内通话时长/总通话时长				
	网外通话时长/总通话时长				

根据因子分析结果对变量进行整合，最后形成的用于模型的特征变量如下(共 16 个)：消费水平；点对点短信上行条数；工作日点对点短信上行条数占总上行条数比例；晚上点对点短信上行条数占总上行条数比例；网外上行条数占总上行条数的比例；短信费用；新业务费用；短信上行交往圈；数据业

务费用占比；梦网短信条数；GPRS+WAP 时长；语音交往圈；本地通话时长；本地通话次数；漫游/长途通话时长；漫游/长途通话次数。

3 模型构建

通过探索型数据分析，把所有数据准备好之后，就可以选用适当的数据挖掘工具及数据挖掘技术来建立短信分群分析模型。其中的处理过程描述如下。

(1)汇总合并：将各个基表中数据汇总合并到总表中，再从总表中取 3 个月的数据平均后生成新的变量。并生成派生变量，把这些数据都汇总到宽表中，作为模型训练和验证的样本数据集。

(2)数据探索(EDA)[13]：值分析(Value Analysis)、统计分析(Statistic)、数值分析(Histogram)。

(3)数据训练[14]：使用 K-MEANS 算法训练个人客户分群分析模型。在模型训练之前，必须对样本数据进行 Z-SCORE 转换，以消除量纲对模型训练的影响。

(4)K-MEANS 算法是最常用的聚类算法之一。它是把对象集合 D 划分成一组聚类 $\{C_1,C_2,C_3,\cdots,C_K\}$，这里 $\bigcup_{i=1}^{k} C_i = D$，其中 K 是要得到的聚类个数。

K-MEANS 算法的过程如下。

(1)假设要聚成 K 个类。由人为决定 K 个类中心 $Z_1(1),Z_2(1),\cdots,Z_k(1)$。

(2)在第 k 次叠代中，样本集 $\{Z\}$ 用如下方法分类：

对所有 $i=1,2,\cdots,K,i\neq j$，若 $||Z-Z_j(k)||<||Z-Z_i(k)||$，则 $Z\in S_j(k)$。

(3)令由文献[2]得到的 $S_j(k)$ 的新的类中心为 $Z_j(k+1)$，令 $J_j=\sum_{Z\in S_j(k)}||Z-Z_j(k+1)||^2$ 最小，$j=1,2,\cdots,K$，则 $Z_j(k+1)=\frac{1}{N_j}\sum_{Z\in S_j(k)}Z$。$N_j$ 是 $S_j(k)$ 中的样本数。对于所有的 $j=1,2,\cdots,K$，若 $Z_j(k+1)=Z_j(k)$，则终止。否则 goto (2)[2]。

其中 $||Z-Z_j(k)||$ 和 $||Z-Z_i(k)||$ 为计算某点到某群中心点的距离，这里使用比较常用的欧氏 EUCLIEAN 距离计算方法：

$$d_{ij}=\sum_{k=1}^{m}|x_{ik}-x_{jk}|^2$$

对于该过程解释为：首先从数据集中随机选取 K 点作为初始聚类中心，然后计算各个样本到聚类中心的距离，把样本归到离它最近的那个聚类中心所在的类。计算新形成的每一个聚类的数据对象的平均值，得到新的聚类中心。如果相邻两次的聚类中心没有任何变化，说明样本调整结束，聚类准则函数 Jc 已经收敛。本算法的一个特点是在每次迭代中都要考察每个样本的分类是否正确。若不正确，就要调整。在全部样本调整完后，再修改聚类中心，进入下一次迭代。如果在一次迭代算法中，所有的样本都被正确分类，则不会有调整，聚类中心不会再有变化。

4 建模结果

我们运用 1 月份—3 月份的数据作为训练集，用细分变量对目标客户进行分群，然后根据得到的分群结果通过分析“描述变量”对模型进行优化和调整。

构建本次模型选择的最大迭代次数为 500 次。按照以下方式设定该分析的参数(见表 3)。

表 3 模型结果参数表

Clustering Algorithm	K-MEANS
Convergence Criterion 收敛标准	0.001
Maximum Iterations 最大迭代次数	500

在迭代过程中，邀请业务专家确认每一次迭代结果。如果满足业务需求，将提前人工停止迭代。当进行到第 321 次迭代后认为该结果比较理想，因此进行人工停止处理。对于准则函数 Jc，我们设置的收敛值为 0.001，这是一般商业分群的基本需求。最后通过聚类获得了 11 个有效分群模型，编号为 1 到 11。并根据用户的属性集特征将群分为 3 大类，分别为低端短信客户群(包括群 1,2,3,4,5)、中端短信客户群(包括群 6,7,8,9)和高端短信客户群(包括群 10,11)。图 1 和图 2 为各个群的特征变量的统计特征值(描述变量仅作为业务分析辅助，不作为分群依据)，我们可以通过这些值验证分群的效果。

变量性质	变量名称	低端用户群				
		群 1	群 2	群 3	群 4	群 5
描述变量	人数	615693	727114	1069657	967129	633349
	人数占比	15.05%	13.42%	20.11%	17.85%	15.38%
	消费水平	47.43	47.65	54.62	56.88	116.66
建模变量	点对点短信上行条数	4.46	13.55	27.59	38.53	34.2
	工作日点对点短信上行条数占总上行条数比例	0.19	0.45	0.63	0.66	0.58
	晚上点对点短信上行条数占总上行条数比例	0.22	0.4	0.61	0.89	0.53
	网外上行条数占总上行条数的比例	0.13	0.46	0.16	0.88	0.3
	短信费用	1.28	2.38	3.92	5.34	5.64
	新业务费用	6.6	6.97	7.67	8.91	14.29
	短信上行交往圈	1.62	3.8	6.22	6.81	9.8
	数据业务费用占比	0.16	0.16	0.15	0.17	0.13
	梦网短信条数	8.02	7.92	9.54	11.55	20.96
	GPRS+WAP 时长	21.72	37.76	55.09	69.07	83.21
	语音交往圈	39.05	39.08	41.57	43.6	99.73
	本地通话时长	10660.13	11320.7	12768.51	14540.2	31459.7
	本地通话次数	155.36	152.66	171.96	181.68	456.86
	漫游/长途通话时长	983.17	1151.72	1360.71	1540.25	3873.84
	漫游/长途通话次数	8.09	8.59	10.28	10.6	31.65

图 1 细分变量表(低端用户群)

变量性质	变量名称	中端用户群				高端用户群	
		群 6	群 7	群 8	群 9	群 10	群 11
描述变量	人数	212403	95586	26171	124111	451661	76091
	人数占比	3.92%	1.76%	0.48%	2.29%	8.34%	1.40%
	消费水平	201.61	281.84	111.97	131.77	99.72	196.51
建模变量	点对点短信上行条数	71.99	97.64	149.06	160.47	224.22	[illegible]
	工作日点对点短信上行条数占总上行条数比例	0.59	0.6	0.63	0.65	0.64	0.65
	晚上点对点短信上行条数占总上行条数比例	0.52	0.54	0.59	0.64	0.61	0.61
	网外上行条数占总上行条数的比例	0.29	0.32	0.4	0.43	0.41	0.35
	短信费用	11.8	14.06	21.39	18.61	29.74	74.97
	新业务费用	30.18	31.53	35.11	13.36	17.31	[illegible]
	短信上行交往圈	20.58	19.37	21.8	20.84	29.08	[illegible]
	数据业务费用占比	0.17	0.14	0.31	0.11	0.18	0.4
	梦网短信条数	39.06	30.26	28.9	17.63	23.76	65.74
	GPRS+WAP 时长	172.77	248.21	11738.55	201.35	394.03	627.32
	语音交往圈	185.36	125.99	61.89	85.96	60.69	94.04
	本地通话时长	[illegible]	28274.9	25650.02	[illegible]	24911.3	45066.74
	本地通话次数	[illegible]	356.48	280.32	[illegible]	264.17	454.01
	漫游/长途通话时长	7348.27	47916.7	5630.9	2963.66	3664.14	[illegible]
	漫游/长途通话次数	71.43	390.72	35.15	16.44	25.5	[illegible]

图 2 细分变量表(中、高端用户群)

从图 1 和图 2 中可以看到，每一个群都有明显区隔于其它群的特征变量值(用灰色标注)，因此我们认为本次通过聚类完成的分群有效。借助描述变量，我们将分群用户的特征进行归纳，归纳后的群特征如表 4 所列。

表 4　群特征表

短信发送量	分组号	群特征
低	群 1：节假日短信活跃客户群	短信发送量非常低，节假日短信所占总发送量 80%以上
	群 2：短信间歇活跃客户群	发送短信量偏低，客户需求波动较大，有的月份发送较多，有的月份没有发送
	群 3：低端普通短信客户群	发送短信量偏低，但是客户每月都有发送，发送条数较少
	群 4：短信网外联系紧密群	发送短信量偏低，但是网外短信量多，占比达到 66%
	群 5：中价值(ARPU)低短信客户群	发送短信量偏低，但是此群的客户价值属于中端水平，人均水平达到了 116 元
中	群 6：高价值本地客户群	发送短信量中等，客户较为偏好语音，价值较高，人均 ARPU 达到 201 元
	群 7：高价值移动客户群	发送短信量中等，客户漫游多，价值较高，人均 ARPU 达到 281 元
	群 8：新技术偏好客户群	发送短信量中等偏上，此部分客户非常偏好 GPRS 和 WAP
	群 9：中价值(ARPU)本地联系紧密群	发送短信量中等偏上，本地语音非常高
高	群 10：偏好短信客户群	短信发送量很高，ARPU 一般，人均 ARPU 仅有 100 元，但是发送短信量达到 224 条
	群 11：热衷短信客户群	发送短信量非常高，梦网短信，ARPU 也很高，语音中等

5　模型效果检验

5.1　模型分群效果

由于本模型采用的 K-MEANS 聚类算法属于无监督学习算法，因此我们依据各个群的强势特征变量的数据分布，并结合实际应用，对模型进行评估。短信提升模型应用的最终目的，就是提升用户的短信量，故各分群对短信量维度的区分能力是十分重要的。以下依次为群 1 到群 11 的发送短信量的分布图(见图 3 到图 13)。

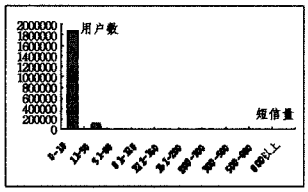

图 3　群 1 短信用户数分布情况

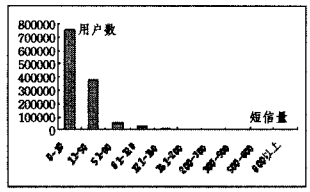

图 4　群 2 短信用户数分布情况

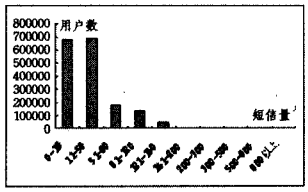

图 5　群 3 短信用户数分布情况

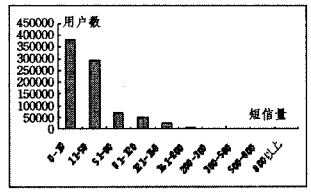

图 6　群 4 短信用户数分布情况

从图 3 到图 13 可以看出，群 1 到群 5 的短信量集中在 0～10 条和 10～50 条上。尤其群 1，大多数用户都是 0～10 条。这 5 个群可视为低端短信发送集群。从群 6 开始到群 9，发送短信超过 120 条的用户比例逐步增高，但没有占到绝对的比例，可视为中端短信发送集群。在群 10 和群 11，短信发送量大的用户占了绝对数目，即为高端短信发送集群。纵向比较，群 1 的短信发送量主要集中在0～10 条间，群 11 的短信发送量集中在 500～800 条间，差异明显。

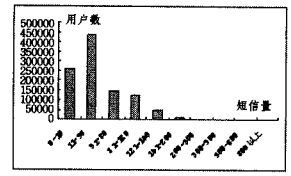

图 7　群 5 短信用户数分布情况

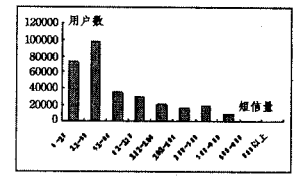

图 8　群 6 短信用户数分布情况

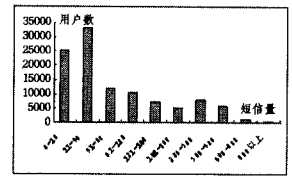

图 9　群 7 短信用户数分布情况

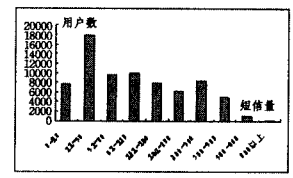

图 10　群 8 短信用户数分布情况

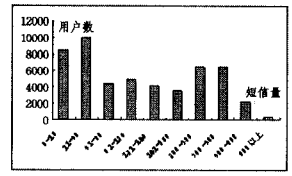

图 11　群 9 短信用户数分布情况

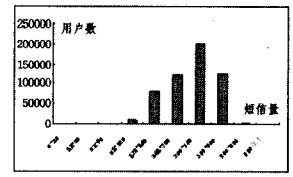

图 12　群 10 短信用户数分布情况

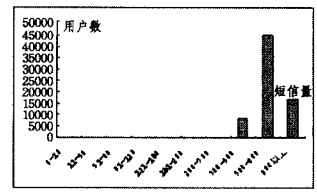

图 13　群 11 短信用户数分布情况

由上述可知，短信发送量在分群时起到了非常重要的作用。现有分群方法体现了该变量的重要性。对于 3 个集群内部各分群，可以通过它们的强势变量进行划分。本模型中各群的分群变量强弱差异各有不同，从而可以很好地把各个群区分出来。可以通过气泡图同时对比 2 个强势变量来检验分群效果。本模型采用了 16 个分群变量，两两变量组合，有 120 个组合，即最多可以产生 120 张汽泡图。在本评估文档中，只抽取 2 张气泡图进行展现，其它类似，不一一列举。

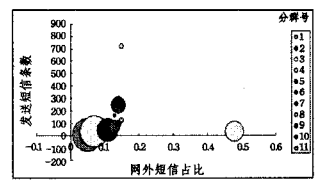

图 14　网外短信条数和发送短信量的组合汽泡图

由图 14 可知，各个分群的强势特征各有差异，分群效果理想。

5.2　模型实际运用效果

根据分群结果，2009 年 10 月到 12 月期间，我们在西部某通信公司针对模型进行了实际应用。即以节日为契机，通过开展“短信送祝福”系列活动(包括以推荐有礼、感恩有礼、贺岁有礼和多发有礼为主题的营销活动)来提升点对点短信人均发送条数，增加用户粘性。该活动取得了明显的经济效益，同时从实践中验证了模型的分群效果。本次活动参与用户数达到 194213，直接提升经济效益 430 万元。活动中随机

(下转第 198 页)

[10] 周玉华，张冠宇，史开泉. P-集合与双信息规律生成[J]. 数学的实践与认识，2010，40(13)：71-80
[11] Zhang Guanyu, Li Enzhong. Information gene and its information knock-out/knock in [J]. An International Journal Advances in Systems Science and Applications, 2010, 10(2): 267-275
[12] Zhang Li, Cui Yuquan. Outer P-sets and data internal-recovery [J]. An International Journal Advances in Systems Science and Applications, 2010, 10(2): 229-236
[13] Liu Jiqin. P-probabilities and its applications [J]. An International Journal Advances in Systems Science and Applications, 2010, 10(2): 237-244
[14] Liu Hongkang, Li Yuling. P-sets and its P-separation theorem [J]. An International Journal Advances in Systems Science and Applications, 2010, 10(2): 245-251
[15] 张飞，陈萍，张丽. P-集合的 P-分离与应用[J]. 山东大学学报：理学版，2010，45(3)：18-22
[16] 于秀清. $P_{(\rho,\sigma)}$-集合与它的随机特性[J]. 计算机科学，2010，37(9)：218-221
[17] 于秀清. P-集合与数据挖掘-还原[J]. 河南师范大学学报：自然科学版，2010，38(5)：52-55
[18] 于秀清. P-集合的动态特性[J]. 计算机工程与应用，2010，46(18)：45-48
[19] 于秀清. P-集合的识别与筛选[J]. 山东大学学报：理学版，2010，45(1)：94-98
[20] 罗承忠. 模糊集引论[M]. 北京：北京师范大学出版社，2005：13-18
[21] Dubois D, Prade H. Measuring properties of fuzzy sets: A general technique and its use in fuzzy query evaluation [J]. Fuzzy Sets and Systems, 1990, 38: 137-152
[22] Esogbue A O, Theologidu M, Guo Kejiao. On the application of fuzzy sets theory to the optimal flood control problem arising in water resources systems [J]. Fuzzy Sets and Systems, 1992, 48(2): 155-172
[23] Chen Degang, He Qiang, Wang Xizhao. Fuzzy rough set based support vector machines [J]. Fuzzy Sets and Systems, 2010, 161(4): 596-607
[24] Kuncheva L I. Fuzzy rough set: Application to feature selection [J]. Fuzzy Sets and Systems, 1992, 51(2): 147-153
[25] Atanassova L C. Remark on the cardinality of the intuitionistic fuzzy sets [J]. Fuzzy Sets and Systems, 1995, 75(3): 399-400
[26] Yang Xinmin. Some properties of convex fuzzy sets [J]. Fuzzy Sets and Systems, 1995, 72(1): 129-132
[27] Bigand A, Colot O. Fuzzy filter based on interval-valued fuzzy sets for image filtering [J]. Fuzzy Sets and Systems, 2010, 161(1): 96-117
[28] Chen Shui-li, Cheng Ji-shu. θ-Convergence of nets of L-fuzzy sets and its applications [J]. Fuzzy Sets and Systems, 1997, 86(2): 235-240
[29] Radecki T. A theoretical background for applying fuzzy set theory in information retrieval [J]. Fuzzy Sets and Systems, 1983, 10(1): 169-183
[30] Wygralak M. Questions of cardinality of finite fuzzy sets [J]. Fuzzy Sets and Systems, 1999, 102(2): 185-210

(上接第 158 页)

选择了 10%的分群用户不参与活动，最后与参与活动的相比，发现通过模型得到的分群结果比不分群的用户在营销活动参与率上提高了近 6 倍。

结束语 本文通过分布图分析和汽泡图分析可以看出，分群区分特征非常明显，因此 K-MEANS 对于新业务分群效果较好。模型可以直接作为运营商今后短信业务客户群细分的标准，采用具备监督的分类算法可以更显著地完成分群工作。通过因子分析，可以很好地提升模型开发的效率。K-MEANS 方法迭代过程可以通过专家根据实际情况停止迭代次数，以达到最好的迭代效果。分群结果可以直接被应用，成为运营商的标准，减少再次挖掘的成本。2009 年 10 月到 12 月期间，我们在西部某通信公司针对模型进行了实际应用，直接提升经济效益 430 万元。通过比较发现，通过模型得到的分群结果比传统随机抽取结果在营销活动参与率上提高近 6 倍。

总体而言，本文所提出的分群模型能够较为准确地反映某公司的短信业务的细分效果。但由于 K-MEANS 算法本身的局限性导致模型仍然具备提升的空间。例如，算法在第一次迭代中，对中心点的选择是随机进行的，因此可能导致模型结构受到影响。另外，判断算法是否终止的标准可以是多个因素决定：目标函数值不再下降；两次迭代得到的聚点相同；两次迭代得到的划分相同；达到最大迭代次数。组合优化的问题是 NP 完全问题，解决这些问题可以通过爬山算法。但为了节省时间，我们这里通过人为经验决定，而理论上可以采用启发式算法，其更能提升模型的理论精度。

参 考 文 献

[1] 吴斌，郑毅，傅伟鹏，等. 一种基于群体智能的客户行为分析算法[J]. 计算机学报，2003，26(8)：913-918
[2] 梁静国，张亚光，戈华. CRM 中的模糊 c 均值(FCM)客户聚类算法研究[J]. 哈尔滨工业大学学报，2004，25(2)：257-260
[3] 曲昭伟，郑岩，吕廷杰. 基于聚类实现客户行为分析[J]. 东北师大学报：自然科学版，2006，38(2)：19-21
[4] 郑国荣，张郑礼，郭鹏，等. 聚类分析在电信消费模式中的应用[J]. 重庆大学学报：自然科学版，2006，29(4)：119-121
[5] 吕巍，蒋波，陈沽. 基于 K-Means 算法的中国移动市场顾客行为细分策略研究[J]. 管理学报，2005，2(1)：80-84
[6] 陈治平，胡宇舟，顾学道. 聚类算法在电信客户细分中的应用研究[J]. 计算机应用，2007，9081(10)：2569-2577
[7] Glendon C, Thompson W. Understanding Your Customer: Segmentation Techniques for Gaining Customer Insight and Predicting Risk in the Telecom Industry [R/OL]. http://www2.sas.com/proceedings/forum2008/TOC.html
[8] Kim S-Y. Customer segmentation and strategy development based on customer lifetime value: A case study[J]. Expert Systems with Applications, 2006, 31: 101-107
[9] Hwang H, Jung T. An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry [J]. Expert Systems with Applications, 2004, 26(2): 181-188
[10] Rho J J. Internet Customer Segmentation Using Web Log Data [J]. Journal of Business & Economics Research, 2004, 2(11)
[11] 王芳. 主成分分析与因子分析的异同比较以及应用[J]. 统计教育，2003(5)：14-17
[12] 于秀林，任雪松. 多元统计分析[M]. 北京：中国统计出版社，1999：5-100
[13] Kantardzic M. Data Mining Concepts, Models, Methods, and Algorithms[M]. [S. l.]: IEEE Press Publishing House, 2002
[14] Dunhan M H. 数据挖掘教程[M]. 郭崇慧，田凤占，译. 北京：清华大学出版社，2005