文章编号:1001 - 9081(2005)03 - 0673 - 03

基于贝叶斯网络的推理在移动客户流失分析中的应用

叶　进[1],林士敏[2]

(1. 桂林电子工业学院 通信与信息工程系,广西 桂林 541004;
2. 广西师范大学 计算机科学系,广西 桂林 541004)

(Yejin@gliet.edu.cn)

摘　要:移动客户流失分析系统在数据挖掘的基础上,实现了客户流失模型的管理应用。其中关键的环节是根据先验知识的因果推理和基于贝叶斯网络的因果推理进行流失客户的分析,挖掘导致流失的因素,从而辅助市场经营人员制订相应的策略。实验说明,基于贝叶斯网络推理的知识可以不断修正先验知识,获得对客户流失等问题的正确认识。

关键词:因果推理;贝叶斯网络;机器学习;客户流失

中图分类号: TP181　**文献标识码**:A

Application of Bayesian network in subscriber churn analysis of wireless carriers

YE Jin[1], LIN Shi-min[2]

(1. *Department of Communication and Information Engineering, Guilin University of Electronic Technology, Guilin Guangxi* 541004, *China*;
2. *Department of Computer Science, Guangxi Normal University, Guilin Guangxi* 541004, *China*)

Abstract: The analysis system of subscriber churn in wireless carriers was constructed based on data mining. It used Bayesian network to build the subscriber churn model from existing business data. So the market people could make corresponding policy. Experiments prove that the priority knowledge can be corrected continually by inference information of Bayesian network, and help to solve the problem of subscriber churn.

Key words: causality; Bayesian network; machine learn; subscriber churn

0　引言

移动通信用户的客户流失是一个长期以来困扰全球移动电话运营商的难题。在欧洲,每年有 35% ~50% 的客户流失;而获取一个新客户的平均成本超过了 $ 700,这几乎相当于一个客户 5 年内给公司带来的净利润。这种情况直接导致客户回报率的下降。[1]根据流失的客户和没有流失的客户性质和消费行为,进行推理分析,建立客户流失预测模型,分析哪些客户的流失概率较大,流失客户的消费行为如何,正在成为移动公司面临的重要课题。

"移动客户流失分析系统"是一个智能信息分析系统,它从用户资料、账单、详单等业务数据中提取相关的信息,通过流失客户的分析,发掘导致流失的因素,辅助市场经营人员制订相应的销售策略,同时为挽留客户提供决策依据。

整个系统的业务流程如图 1 所示,把数据挖掘得到的知识和市场的经验、客服的信息结合起来,应用于数据库中的数据进行流失客户的预测、分析,对确认有流失倾向的客户根据不同的情况进行预警处理,包括套餐资费调整、服务方式更改、竞争对手调查等,同时将预警客户名单通过公司的数据交换平台下发给各个业务分区,进行摸底跟踪,实施关怀工程。

其中流失客户的预测和分析是两个关键的环节,需要选择合适的推理分析方法,获取有用的模型和知识应用于系统中,才能进行科学的辅助决策。

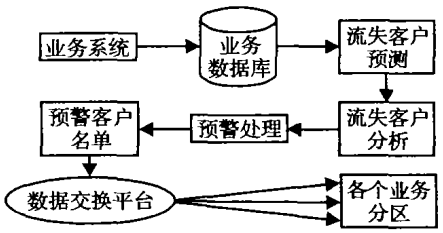


图 1　移动客户流失分析系统业务流程

本文介绍了根据先验知识的因果推理和基于贝叶斯网络的推理在该过程中的应用,实验结果说明,基于贝叶斯网络的推理是较优的。

1　因果推理与贝叶斯网络

在信息处理自动化得到不断提高后,人们已经逐步开始了对思维自动化的思考。而实现思维自动化的关键问题之一,就是如何有效表达和解决不确定性问题。

贝叶斯网络(Bayesian Network)[2,3]又称为置信网络(Brief Network),是基于概率分析、图论的一种不确定性知识的表达和推理的模型。从直观上讲,Bayesian 网络表现为一

收稿日期:2004 - 08 - 29;**修订日期**:2004 - 11 - 08

基金项目:清华大学智能技术与系统国家重点实验室开放课题资助(99002);广西教育厅学科软环境建设项目(D20327)

作者简介:叶进(1970 -),女,江苏泰兴人,讲师,主要研究方向:数据挖掘、机器学习、数据库技术;　林士敏(1941 -),男,广西贵港人,教授,主要研究机器学习、知识发现.

个赋值的复杂因果关系网络图,网络中的每一个节点表示一个变量,即一个事件。各变量之间的弧表示事件发生的直接因果关系。贝叶斯网络 $G = \langle S, P \rangle$ 由网络的拓扑结构 S 和局部概率分布的集合 P 两部分组成。S 表示节点变量之间的因果联系,P 代表用于量化网络的一组参数,包括边缘概率和条件概率,表达原因对结果的作用程度[2,3]。

利用 Bayesian 网络进行推理的前提是从原始数据中构造 Bayesian 网络模型,实际上是对原始数据进行数据挖掘:首先找出最符合原始数据的定性的网络结构,然后根据网络结构中的因果关系,计算节点间的条件概率。

Bayesian 网络的推理原理基于 Bayesian 定理:

$$p(B_i \mid D) = \frac{p(D, B_i)}{p(D)} = \frac{p(B_i)p(D \mid B_i)}{p(D)}$$

其推理过程实质就是概率计算。Bayesian 网络的推理主要有以下三种形式[4]:

因果推理 原因推知结论——由顶向下的推理。已知移动的原因(证据),求出在该原因的情况下结果发生的概率。

诊断推理 结论推知原因——由底向上的推理。目的是在已知结果时,找出产生该结果的原因[4]。

支持推理 提供解释以支持所发生的现象。目的是对原因之间的相互影响进行分析[5]。

在"移动客户流失分析系统"中,进行客户流失的趋势预测,正是 Bayesian 网络的因果推理的任务。

BN Toolkit (BNT)[6] 是 Kevin P. Murphy 基于 Matlab 语言开发的关于叶斯网络学习的软件包,提供了许多贝叶斯网络学习的底层基础函数库,支持多种类型的节点(概率分布)、精确推理和近似推理、参数学习和结构学习、静态模型和动态模型。BNT 是个完全免费的软件包,其代码完全公开,系统的可扩展性良好。因此,我们选择 BNT 作为 Bayesian 网络的学习和推理的实验工具。

2 推理机制在客户流失分析中的应用

2.1 数据样本的组建

数据样本是数据挖掘过程的基本组成部分。客户的历史行为数据中隐含着大量与流失相关的行为模式,数据样本必须围绕移动市场分析得到的流失变量来组建。由移动公司提供的 10000 条原始数据经过清洗、处理异常等工作,获取了 1250条符合要求的候选数据。

1) 选取网络节点

表 1 流失变量与状态的关系

A	B	C	D	E	R
0.43	0.97	0	0.09	0	1
0.23	0.92	0	0.07	0	1
0.04	0.42	0	0.07	0	1
…	…	…	…	…	…

结合移动市场得到的先验知识,我们选取了下列流失变量:上月话费下降比率、本月话费下降比率、是否呼叫转移到网外、与其他移动公司用户通话比例、是否拨打××××(其他移动公司服务热线电话),分别标记为 A(fee_rate1),B(fee_rate2),C(call_remove),D(call_rate_union) 和 E(call_1001),即为贝叶斯网络模型中的节点,得到的候选数据集如表 1,其中 R(is_churn) 标记为该样本的状态值,"1"表示在用用户,"0"表示离网用户。

2) 样本数据离散化

在贝叶斯网络中使用分类型变量,因此要利用本移动公司客户的固有特点进行离散化转换。结合对历史数据进行 OLAP(在线分析处理)的结果,将数据进行处理如表 2。

表 2 对表 1 离散化处理的结果

A	B	C	D	E	R
mid	low	no	low	no	no
high	low	no	low	no	no
high	mid	no	low	no	no
…	…	…	…	…	…

例如对"本月话费下降比率"这一变量,根据 OLAP 的初步分析,一般用户如果低于 30% 是在用的正常波动,如果高于 60% 是离网的预警标志。因此将变量 A 分 3 层(<30%,30% ~60%,>60%),分别标记为 low,mid 和 high。

3) 数据格式转换

根据 Matlab 的特点,将数据集转化为矩阵的形式,而且所有属性的取值全部依次编号为 1,2,3。

表 2 对应的矩阵为:
$$\begin{bmatrix} 2 & 3 & 2 & 2 & 2 & 1 \\ 1 & 3 & 2 & 2 & 2 & 1 \\ 1 & 2 & 2 & 2 & 2 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

由此得到了 1250 条样本数据集作为模型数据集(包括训练数据集、测试数据集)来构建模型。

2.2 Bayesian 网络的建模

为了比较先验知识与 Bayesian 方法获得的知识,我们在实验中从 2 个不同的角度来建立 Bayesian 网络的模型。

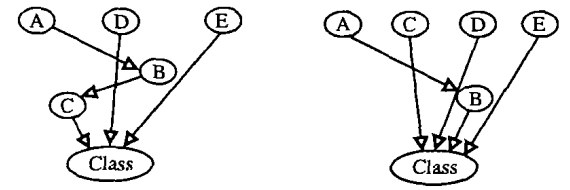
一方面,把先验知识的因果关系直接表达为 Bayesian 网络结构[7],然后采用训练数据进行该网络的参数学习。图 2 描述了两种先验认识:图 2(a)表示 C(呼叫转移到网外),D(与其他移动公司用户通话比例)和 E(拨打其他移动公司热线××××)是导致客户流失的直接因素,而 C 是由 B(本月话费下降)和 A(上月话费下降)引起的;图 2(b)表示的因果关系与图 2(a)的区别在于把两个月话费连续下降(A -> B)当作导致客户流失的直接因素。

另一方面,通过机器学习来获得 Bayesian 网络结构。在 Bayesian 学习方法中实用性很高的一种是朴素 Bayesian 网络,如图 2(c)所示,它基于如下假设:给定目标值时属性值之间相互条件独立,即:$p(X) = \prod_{i=1}^{n} p(x_i \mid pa_i)$。

朴素 Bayesian 方法不需要搜索,只是简单地计算训练样本中不同数据组合的出现频率。更为精确的学习方法是搜索所有可能的假设结构,从中选择一个"好的"模型。即:定义一个随机变量 B_i,表示网络结构的不确定性,并赋予先验概率分布 $p(B_i)$,然后计算后验概率分布 $p(B_i \mid D)$,选择后验概率最大的结构作为学习结果。在无约束多项分布、参数独立、采用 Dirichlet 先验和数据完整的假设前提下,对 $p(D \mid B_i)$ 的计算公式[8] 如下:

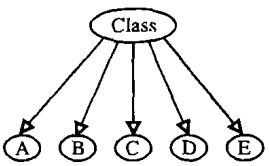
$$p(D \mid B_s) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[\frac{\Gamma(a'_{ij})}{\Gamma(a'_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(a'_{ijk} + N_{ijk})}{\Gamma(a'_{ijk})} \right]$$

在一般情况下,N 个变量可能的网络结构数目大于以 N 为指数的函数。学习完全的贝叶斯网络的结构是 NP 难问题,若干研究者的工作表明,使用贪心搜索法选择单个好的模型通常会得到准确的预测。在此采用 Cooper 和 Herskovits[8] 提出的 K2 算法进行结构学习,其基本思想是从一个空网络开始,根据事先确定的节点次序,选择使后验结构概率最大的节点作为该节点的父节点,依次遍历完所有的节点,逐步为每一个变量添加最佳父节点。算法中限制变量的最大父节点数目为 2 来进行优化。学习结果如图 2(d)所示,影响客户流失的因素是 D(与其他移动公司用户通话比例)和 E(拨打其他移动公司热线××××),与 C(呼叫转移到网外)无关,同时客户流失倾向将导致 A(上月话费下降)和 B(前月话费下降),这个结果可以帮助市场人员重新修正原有的先验知识。

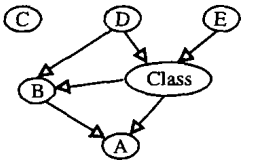


(a) 基于先验知识的网络结构一

(b) 基于先验知识的网络结构二



(c) 朴素Bayesian网络结构



(d) 选择Bayesian网络结构

图2　推理学习的网络结构

确定贝叶斯网络的结构后,就可以进行参数学习了,即计算每个变量在贝叶斯网络中的条件概率分布,作为下一步推理的依据。

2.3　Bayesian 网络的因果推理

利用网络结构性质和条件独立性的关系已经设计出计算后验概率 $p(B \mid D)$ 的有效方法,如:基于消息传播的算法、基于子团的算法以及基于邻接树的算法等。这些方法虽然有着不同的计算思路,但是计算结果是一致的,区别主要在于计算速度。本次实验采用 BNT 工具包中的联合树推理引擎来进行精确推理,随机抽取样本:$X = (A(fee\_rate1) = 3, B(fee\_rate2) = 2, C(call\_remove) = 0, D(call\_union) = 1, E(call\_1001) = 1)$,推理结果:$p(X \mid R(is\_churn) = yes) = 0.8427$,$p(X \mid R(is\_churn) = no) = 0.1573$,因此认为该样本所标识的用户流失的可能性较大。

3　结果分析

表 1　基于图 2 所示网络结构的推理结果分析

Bayesian 网络	推理方法	TP	TN	FP	FN	检测率(%)	误检率(%)	总体准确率(%)
图2(a)	联合树	541	397	301	25	95.58	43.12	74.21
图2(b)	联合树	660	354	182	68	91.67	33.96	80.22
图2(c)	联合树	732	359	110	63	92.08	23.45	86.31
图2(d)	联合树	797	328	45	94	89.45	12.06	89.00

为了分析预测准确率,采用 5 叠交叉验证(5- fold Cross Validation)来测试分类结果:将初始数据集随机划分成 5 个互不相交的子集 $S_1, S_2, \cdots, S_5$,即 $S_1 \cap S_2 \cap \cdots \cap S_5 = \emptyset$,每个子集的大小基本相同。学习和测试分别进行 5 次。在第 i 次迭代,S_i 用作测试集,其余的子集都用于训练分类器。取 5 次迭代正确分类数除以初始数据中的样本总数的平均准确率作为最终评估的结果,见表 1。

TP (True Position):正确肯定的数目,将流失的客户预测为流失的数目;

TN (True Negatives):正确否定的数目,将正常的客户预测为正常的数目;

FP (False Positives):错误肯定的数目,将正常的客户预测为流失的数目;

FN (False Negatives):错误否定的数目,将流失的客户预测为正常的数目;

检测率:TP/(TP + FN);

误检率:FP/(FP + TN);

总体准确率:(TP + TN)/(TP + FN + FP + TN)。

从各项指标来看,对客户流失的先验认识中,图 2(a)所示的知识不如图 2(b)的准确;通过机器学习推理的结果普遍比使用先验知识推理的结果好;使用选择 Bayesian 网络的推理结果具有较高的可信度。

4　结语

基于贝叶斯网络的推理方法可以定期分析业务数据,将客户流失倾向的概率值由大到小排序,最终导出前面若干部分的客户名单作为目标变量提交给市场部。在客户流失模型建立的过程中还可加入客户属性资料、合同资料、呼叫模式和付费数据等更多变量进行学习和推理,进一步分析出更加全面的客户流失原因及对策。

由于 Bayesian 方法可以综合先验信息和后验信息,既可避免只使用先验信息可能带来的主观偏见,和缺乏样本信息时的大量盲目搜索与计算,也可避免只使用后验信息带来的噪音影响[3]。因此,在具有概率统计特征的数据挖掘和知识发现的领域中,得到了广泛的应用。

参考文献:

[1] 段云峰, 吴唯宁, 李剑, 等. 数据仓库及其在电信领域中的应用[M]. 北京: 电子工业出版社, 2003.

[2] HECKERMAN D. A Tutorial on Learning With Bayesian Networks [EB/OL]. http://www.accai.org, 1995.

[3] 林士敏, 王双成, 陆玉昌. Bayesian 方法的计算学习机制和问题求解[J]. 清华大学学报(自然科学版), 2000, 40(9): 61-64.

[4] 胡玉胜, 涂序彦, 崔晓瑜, 等. 基于贝叶斯网络的不确定性知识的推理方法[J]. 计算机集成制造技术, 2001, 7(12): 65-68.

[5] TIAN J, PEARL J. Causal Discovery from Changes[A]. Proceedings of the 17th Conference on Uncerntainty in Artificial Intelligence[C], 2001.

[6] MURPHY KP. The Bayes Net Toolbox for Matlab[J]. Computing Science and Statistics, 2001, 33: 331-351.

[7] LIU ZQ. Causation, Bayesian Network, and Cognitive Maps[J]. ACTA AUTOMATICA SINICA, 2001, 27(7): 552-566.

[8] COOPER G, HERSKOVITS E. A Bayesian method for the induction of probabilistic networks from data[J]. Machine Learning, 1992, 9(4): 309-347.