数据仓库和数据挖掘技术在保险公司中的应用

杨 杉[1],何 跃[2]
(1.四川大学 锦城学院,四川 成都 611731;
2.四川大学 工商管理学院,四川 成都 610064)

摘 要:随着我国保险市场的开放,我国保险业的垄断格局被打破,竞争也日趋激烈。保险业作为传统数据处理密集型行业之一,已经积累的大量业务数据。如果能够根据保险公司的实际情况,构建数据仓库平台,利用数据挖掘技术挖掘其中蕴涵的知识和信息,就能有效地制定市场策略,以及时把握市场机会。结合A人寿保险公司的实际情况,详细设计和实现了A人寿保险公司的数据仓库,接着以该数据仓库为数据源,分别实现了客户流失挖掘模型和客户理赔风险模型,利用直观的图表方式将数据挖掘的结果展示出来。最后给出了模型的验证与评价方法,得出了有价值的结论,可以为保险公司的决策层提供参考。

关键词:数据挖掘;人寿保险;客户流失;客户理赔
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2011)06-0157-04

Application of Data Warehouse and Data Mining to Life Insurance Company

YANG Shan[1], HE Yue[2]
(1. Jincheng College of Sichuan University, Chengdu 611731, China;
2. Business Management College, Sichuan University, Chengdu 610064, China)

Abstract: With the openning of domestic insurance market, foreign insurance companies entered, which break the monopolization of domestic insurance. The competition between insurance companies is intense. Insurance is a traditional industry which faces lots of business data everyday. If the insurance companies can construct data warehouse according to the actual situation and scoop out information from it, then they can efficiently make marketing tactics and seize opportulities. Firstly, designed the data warehouse of A Life Insurance industry detailedly. Secondly took the data warehouse as data source, desinged and realised models of customer churn prediction and customer compensate risk prediction. Finally, verified and evaluated the methods and gave out valuable conclusions which can provide reference to the management levels of Life Insurance companies.

Key words: data mining; life insurance; customer churn; customer claim

0 引 言

随着外资公司不断涌入中国保险市场,国内保险市场的竞争愈发激烈,给中国寿险带来了巨大冲击。假如保险公司无法有效利用积累的内部业务数据,就很难获取有价值的信息或规律,也就很难把握市场机会,规避市场风险。

中国人寿保险股份有限公司A分公司(简称"A人寿保险公司")是文中研究的具体实例。该公司内部使用的计算机业务处理系统已大大提高了工作效率,但这些数据并没有系统集中,而是分散在不同地点的业务系统和不同地点的计算机中。如果能够利用数据仓库[1]和数据挖掘[2]技术,将这些零散的数据集中起来,并挖掘出有价值的信息,将大大提高管理层的决策能力,从而提供更好的产品和服务,赢得更多客户。

收稿日期:2010-10-28;**修回日期**:2011-01-30
基金项目:国家自然科学基金资助项目(70771067)
作者简介:杨 杉(1983-),女,四川成都人,硕士,研究方向为数据挖掘、管理信息系统、决策技术;何 跃,博士,副教授,研究方向为管理信息系统、数据挖掘、决策支持系统。

1 保险公司数据仓库的设计

保险公司分支机构较多,地域分布较广,数据仓库的数据来源于总部及各个地理位置分散的分公司。

1.1 数据仓库的概念模型设计

在本保险业数据仓库系统中,主要是分析与客户相关的信息,因此只关心与客户密切相关的四个主题:客户个人信息、承保信息(新投保信息、续保信息)、退保信息和理赔信息。

1.2 数据仓库的逻辑模型设计

数据仓库中广泛使用的多维模型主要有星型模

型[3](Star Schema)和雪花模型[4](Snowflake Schema)两种,文中采用"星型模型"。将退保主题和理赔主题作为中心的星型数据模型,由一个事实表、五个维表组成。事实表中的每条记录含有指向每个维表的指针,从而将多维数据连接起来,如图1、图2所示。

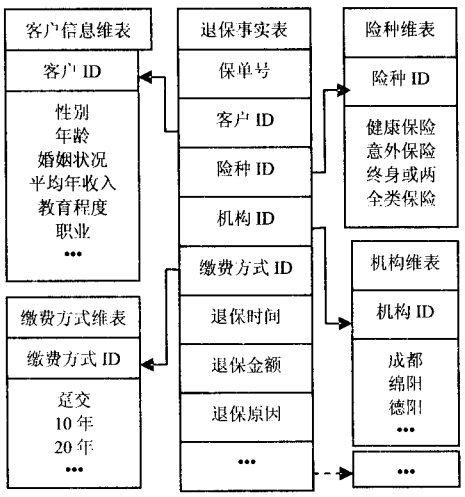

图1 以退保信息为主题的星形模型

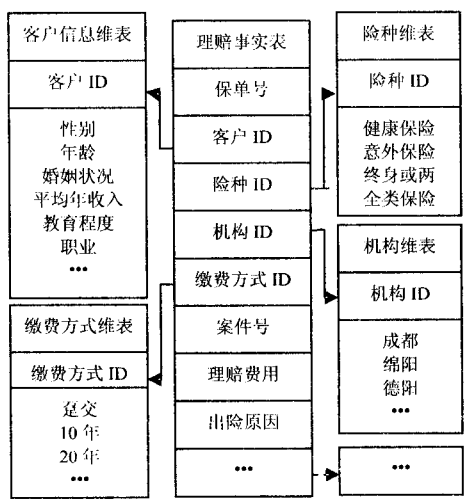

图2 以理赔信息为主题的星形模型

2 保险公司数据挖掘实证研究

在本章中,将利用数据仓库系统,从中抽样提取A人寿保险公司业务数据(包括退保信息、理赔信息),经过对数据的预处理,来分别建立客户流失[5]、客户理赔风险[6]模型,并应用数据挖掘工具Clementine[7]进行了模型的验证。

2.1 数据处理流程

A人寿保险公司的业务数据在建模之前,必须经过一系列的处理流程,分别包括:选取目标数据集和数据预处理过程。

2.1.1 选取目标数据集

客户流失分析所需字段组成数据集"lost_set":客户号(client_id)、性别(sex)、年龄(age)、婚姻状况(marial_status)、收入(income)、教育程度(education)、职业(occupation)、机构(agent_name)、险种(product_name)、缴费方式(paytype_name)、总保费(total_premium)、退保金额(quit_money)、退保原因(quit_reason)。

客户理赔风险分析所需字段组成的数据集"conpensate_risk_set":客户号(client_id)、性别(sex)、年龄(age)、婚姻状况(marial_status)、收入(income)、教育程度(education)、职业(occupation)、机构(agent_name)、险种(product_name)、理赔金额(conpensate_money)

文中选取了2007.01.01-2008.11.30时间段内的6万条数据,分别是:3.3万条退保数据以及相应的客户信息数据、2.7万条理赔客户数据以及相应的客户信息数据,这些数据均从设计好的保险数据仓库中提取,以便为数据挖掘提供分析数据源。

2.1.2 数据预处理

保险公司采用SQL Server2000来建立数据仓库,数据库表的属性大部分采用varchar类型。数据挖掘算法在处理这些数据的时候速度比较慢而且资源消耗比较大,为了解决这个问题,在建立模型之前,需要对所选择的建模属性值进行数字离散化[8]。进行客户流失分析和客户理赔风险分析时,只需将各个字段的连续值离散化为"0,1,2,…"或"F,M"这类集合。

2.2 模型验证与评价

2.2.1 客户流失模型

利用Clementine工具中C5.0决策树算法[9]对客户流失进行特征分析,随机选取现有3.3万条数据的66%作为训练集,剩下34%作为验证集。将客户退保原因分为"经济原因退保-0"和"险种或服务不理想-1"两类,作为输出属性,承保人的性别、年龄、婚姻状况、收入、教育程度、职业、承保险种作为输入属性,从而分析哪些属性值导致两类退保原因的出现。

1)挖掘结果(如图3所示)。

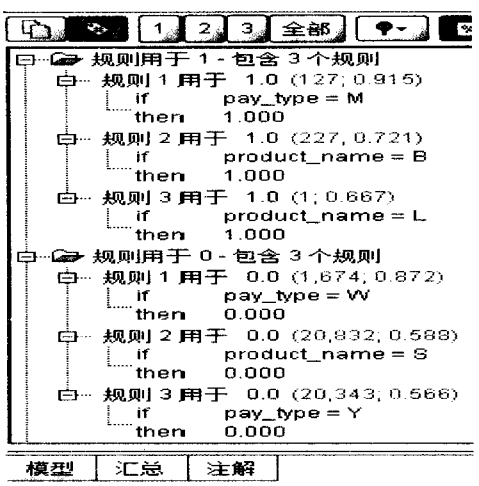

图3 客户流失挖掘规则集模型

从图3可以看出:共产生6个规则,其中规则1、规则2是有效的规则集。从以上规则集中可以看出:客户选择的缴费方式是影响客户退保的一个关键因素,同时客户选择的险种也是影响客户退保的主要因素。

2)模型的验证和评估。

把用于分析的数据分成两部分,随机抽取60%作为训练集,剩下40%作为验证集。训练集与验证集的收益表如图4、图5所示。

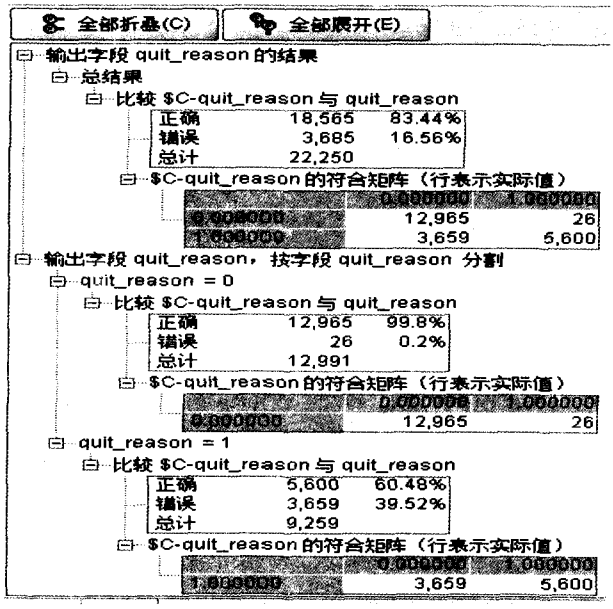

图4　客户流失挖掘训练集收益表

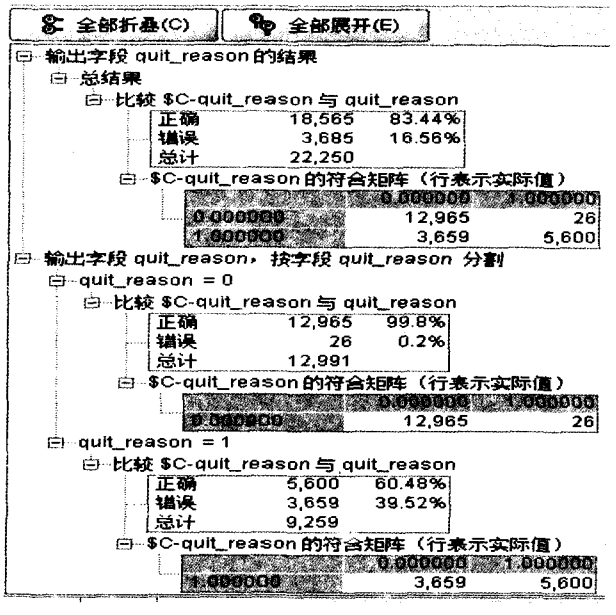

图5　客户流失挖掘验证集收益表

从上图可以看出,在训练集和验证集上的错误率分别是16.56%和16.68%,即正确率分别是83.44%和83.32%,这说明本研究中客户流失特征模型的正确率是比较令人满意的。

2.2.2　客户理赔风险模型

利用Clementine工具中C&R Tree算法[10]对客户理赔风险特征进行分析。将现有2.7万条数据分成两部分,其中,随机选取的60%作为训练集[11],剩下40%作为验证集[12]。

客户风险可以分为"低风险-0"和"高风险-1"两类,作为输出属性;险种、年龄、性别、婚姻状况作为输入属性。从而分析哪些属性值导致客户理赔高风险,哪些属性值导致客户理赔低风险。

1)挖掘结果(如图6所示)。

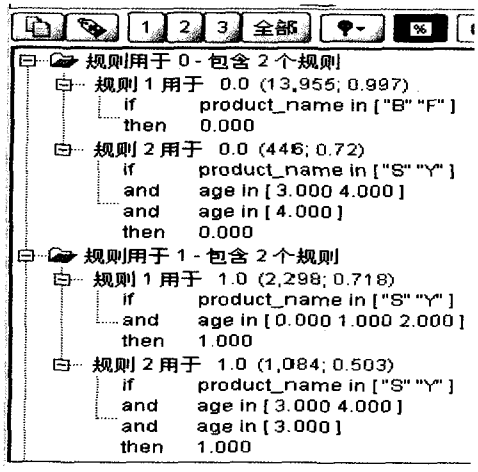

图6　客户理赔风险挖掘规则集模型

从上图结果中可以看出:共产生4个规则,其中规则1是有效的规则集。

从规则1可以看出:客户选择的险种是影响客户理赔风险的一个关键因素,同时客户的年龄也是影响客户理赔风险的一个主要因素。

2)模型的验证和评估。

将现有数据分成两部分,其中,随机选择60%作为训练集,剩下的30%作为验证集,训练集与验证集的收益表如图7、图8所示。

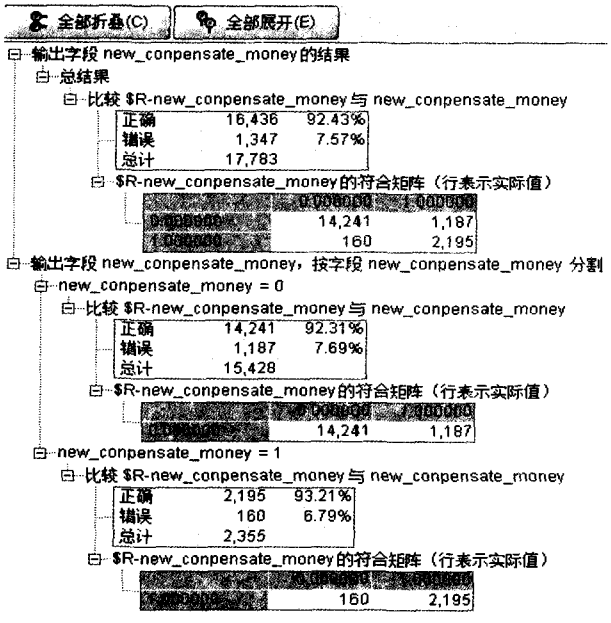

图7　客户理赔风险挖掘训练集收益表

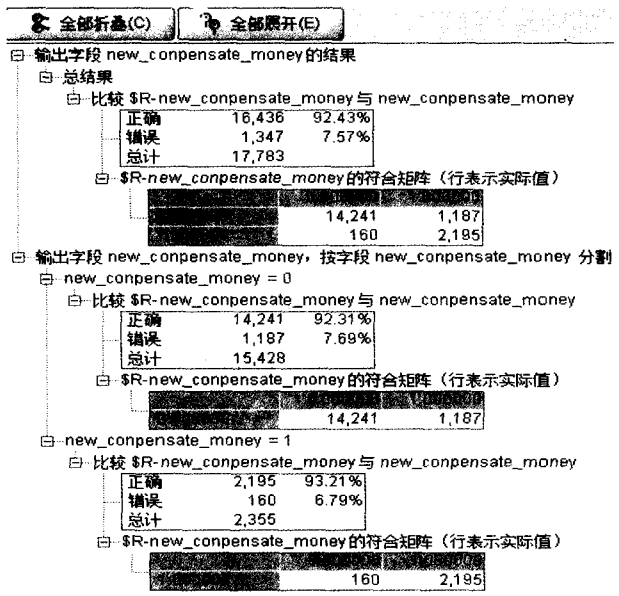

图8 客户理赔风险挖掘验证集收益表

从上图可以看出，在训练集和验证集上的错误率分别是7.57%和7.65%，即正确率分别是92.43%和92.21%，这说明本研究中客户理赔风险特征模型的正确率是非常令人满意的。

3 结束语

通过客户流失模型和客户理赔风险模型，从中得出了客户流失的特征、客户理赔风险大小的特征等结论。在面向分析的数据仓库的基础上，可以利用数据挖掘技术来设计险种绑定销售，制定挽留客户的措施，并且控制保险公司理赔风险，为保险产品定价提供合理依据。

参考文献：

[1] 闫娜娜，刘 锋，李锡娟，等. 支持CRM分析的数据仓库多维启动模型[J]. 计算机技术与发展，2008，18(5)：21-22.

[2] 于红蕾，华庆一，刘燕玲，等. 数据仓库在电信统计分析中的应用[J]. 计算机技术与发展，2007，17(8)：59-60.

[3] 张 宁，贾自艳，史忠植. 数据仓库中ETL技术的研究[J]. 计算机工程与应用，2002(24)：214-215.

[4] 连立贵，金 凤，蔡家媚. 数据仓库中的数据提取[J]. 计算机工程，2001(9)：61-62.

[5] Lingand R，Yen D C. Customer Relationship Management：An Analysis Framework and Implementation Strategies[J]. Journal of Computer Information System，2001(3)：82-97.

[6] Ruggieri S. Efficient C4.5[J]. IEEE Transactions on knowledge and Data Engineering，2002，14(2)：438-444.

[7] 梅 强，张冬茉. 数据挖掘在保险分析中的应用[J]. 计算机工程，2004(12)：37-38.

[8] 桂现才，彭 宏；王小华. 基于决策树的保险客户流失分析[J]. 计算机工程与设计，2005(8)：59-60.

[9] 马建红，王万森. 基于数据仓库的保险管理系统的设计与实现[J]. 微机发展，2004，14(7)：64-66.

[10] 梁 循. 数据挖掘：建模、算法、应用和系统[J]. 计算机技术与发展，2006，16(1)：86-87.

[11] 王爱平，王占凤，陶嗣干，等. 数据挖掘中常用关联规则挖掘算法[J]. 计算机技术与发展，2010，20(4)：17-20.

[12] 姚毓才，王本年. 数据挖掘工具的分类与挖掘[J]. 计算机技术与发展，2006，16(8)：25-27.

(上接第156页)

计算简单，易于实现，适用于主旋律只布在一个音轨的音乐。通过对50多首MIDI音乐进行分析统计，92.8%都能准确地提取出主音轨。结果令人满意，同时也表明了该方法的有效性及可行性，从而为音乐灯光表演方案辅助设计系统的构建进行了很好前提准备。

参考文献：

[1] Liu Li，Cai Junwei，Wang Lei，et al. Melody Extraction from Polyphonic MIDI Files Based on Melody Similarity [C]//2008 International Symposium on Information Science and Engineering (ISISE '08). [s. l.]：[s. n.]，2008：232-235.

[2] 赵 芳，吴亚栋，宿继奎. 基于音轨特征量的多音轨MIDI主旋律抽取方法[J]. 计算机工程，2007，33(2)：165-167.

[3] 冯国杰，王吉军. 基于分层次聚类的MIDI音乐主旋律提取方法[J]. 计算机工程与应用，2009，45(26)：233-235.

[4] 叶 霖，李雄飞，刘丽娟，等. 一种有效识别MIDI文件中主旋律音轨的方法[J]. 计算机应用与软件，2010，27(1)：48-50.

[5] 金 毅，黄 敏. 基于旋律的音乐检索研究——旋律特征的表达和提取[J]. 信息检索技术，2003，4：49-51.

[6] 杨 军. MIDI消息和标准MIDI文件格式剖析及应用[J]. 中南民族大学学报(自然科学版)，2009，22(Sup)：62-64.

[7] 刘嘉欣. 嵌入式MIDI文件格式解析设计与实现[J]. 微计算机信息，2006，22(11-2)：66-67.

[8] 秦 丹. 利用C#从MIDI文件中获取音乐旋律[J]. 电脑知识与技术，2009(7)：4281-4284.

[9] 彭 琼，支 琤. 计算机自动识别音乐情感的关键技术研究[J]. 电声基础，2008，32(4)：35-38.

[10] Zhu Bin. Music Features Recognition and its Application in National Music Protection[C]//7th International Conference on Computer-Aided Industrial Design and Conceptual Design，2006(CAID-CD '06). [s. l.]：[s. n.]，2006：1-6.

[11] Li Jiangtao，Yang Xiaohong，Chen Qingcai. MIDI melody extraction based on improved neural network[C]//2009 International Conference on Machine Learning and Cybernetics. [s. l.]：[s. n.]，2009：1133 - 1138.

[12] 孙即祥. 现代模式识别[M]. 长沙：国防科技大学出版社，2001：46-70.