Statistics and Application 统计学与应用, 2019, 8(3), 513-519
Published Online June 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.83058

Hans汉斯

Research on Renewal Rate of Vehicle Insurance Based on Relevance Analysis

Bixuan Li[1], Xinwei Wen[1], Zhuohui Zhong[1], Luming Shen[1,2]

[1]College of Information Science and Technology, Hunan Agricultural University, Changsha Hunan
[2]Agricultural Mathematical Model and Data Process Center, Hunan Agricultural University, Changsha Hunan
Email: math_forever@163.com

Received: Jun. 3rd, 2019; accepted: Jun. 18th, 2019; published: Jun. 25th, 2019

Abstract

This paper chooses the relevant characteristics of the customers who buy insurance as the research object, and based on the methods of correlation analysis and logistic regression, predicts the renewal rate of the people who buy insurance. Firstly, the data are preprocessed, and the available variables are transformed into virtual variables for correlation analysis. In this paper, two different types of data, discrete and continuous, are virtualized. For discrete data, discrete intervals or attributes can be directly transformed into "items". For continuous data, the idea of iteration dichotomy is introduced, which is based on support and confidence. The continuous attribute values are divided into optimal intervals, and a new item is created for each different attribute value pair to obtain the virtual variables of the continuous attribute. After quantifying the data obtained from association rules, the association analysis is carried out, and all other variables including "whether to renew insurance" variables in strong association rules are selected as the six main influencing factors: vehicle age, renewal year, insured age, new car purchase price, signing premium and insurance premium. The logistic regression model is used to obtain the relationship between these influencing factors and renewal rate. Then we predicted again to get renewal rate value. The fitting degree is 96.7%. Then Z statistics is constructed to locate the attributes of customers by statistical inference, so as to achieve accurate customer portraits.

Keywords

Auto Insurance Renewal, Iterative Bipartition, Association Analysis, Logistic Regression

基于关联分析对车险续保率的研究

李碧璇[1]，文欣薇[1]，钟卓辉[1]，沈陆明[1,2]

[1]湖南农业大学信息科学技术学院，湖南 长沙
[2]湖南农业大学农业数学建模与数据处理中心，湖南 长沙

文章引用：李碧璇，文欣薇，钟卓辉，沈陆明．基于关联分析对车险续保率的研究[J]．统计学与应用，2019，8(3): 513-519. DOI: 10.12677/sa.2019.83058

Email: math_forever@163.com

收稿日期：2019年6月3日；录用日期：2019年6月18日；发布日期：2019年6月25日

摘　要

本文选取购买保险的客户的相关特性作为研究对象，基于关联分析以及逻辑回归等方法，对百姓在购买保险方面的续保率进行预测。首先对数据进行预处理，将可用变量转化为虚拟变量以做关联分析，本文分别对离散与连续两种不同类型的数据进行虚拟化处理，针对离散型数据可直接将离散化区间或属性直接转化为“项”，针对连续型数据，本文引入迭代二划分的思想，基于支持度与置信度对连续属性值进行最优区间划分，为每个不同的属性值创建一个新的项来得到连续型属性的虚拟变量。将量化关联规则后得到的数据，对其进行关联分析，选取强关联规则中包含“是否续保”变量的其他所有变量：车龄，续保年，被保险人年龄，新车购置价，签单保费，三者险保费6种因素作为续保的主要影响因子，利用Logistic回归模型得到这些影响因素与续保率之间的关系，再预测得到续保率，其拟合度为96.7%。而后构建Z统计量，借助统计推断可为客户定位其属性，以实现精准的客户画像。

关键词

车险续保，迭代二划分，关联分析，逻辑回归

1. 引言

随着社会购车比例不断增加，汽车保险这一概念逐渐进入百姓视野。车险，即机动车辆保险，为分散机动车辆在行驶过程中可能发作的未知风险和损失的一种保障机制。中国目前的车险费率制度，大多数符合“从车主义”。即车险保费多少，主要取决于这辆车本身的各项情况，如车的购置价、座位数、排量、购车年限等，根据这些数据计算出一个基本的车险保费价格，再根据这辆车的上年理赔次数来打不同的折扣。这就导致了中国的车险定价模式非常的单调，相似情况的车型，保费也都差不多。可以预见未来车险行业的两大发展趋势：车险价格与驾驶行为密切相关[1]，未来的车险定价将逐渐转变为“从人主义”；同价位车型车险价格完全不同，未来中国车险业，同样的一款车，不同的人开，保费价格会完全不同。

除此之外，目前我国机动车辆保费占全部财产保险业务的高比例在全世界都是罕见的。但事实上，真正能够做到高续保率的保险公司并不占多数。客户留存率低、无法积累长期客户、销售成本和费用居高不下成为了多数以车险为主要业务收入的财险公司的一大难题。所谓续保率就是当年到期的客户中续保客户所占的比重。一直以来，续保都被认为是保险公司业务流程中最为重要的一个环节。续保状况对保险公司及整个汽车保险业都有着重大的影响。

针对上述现象，本文利用关联分析以及构造统计量的思想对客户进行精准画像，给出客户的续保概率。

2. 数据预处理

2.1. 数据来源与模型假设

本文选取购买保险的客户的相关特性作为研究对象，比如：车龄、签单保费、新车置购价等，数据

来源于第十二届认证杯数学中国数学建模网络挑战赛。为了简化模型过程，本文做出以下假设：

1) 假设国家新出台的有关保险政策，短期内不会影响个人的决策意愿；

2) 假设近期 NCD 系数不会发生改变。

2.2. 数据清理与数据表示

由于部分客户的部分属性数据存在空缺、不一致的现象，并且表达形式多样化，不利于进一步的数据分析。为提高分析结果的精度和有效性，本文对数据进行预处理，分以下两步进行：

数据清理：处理样本数据中存在的空缺数据或奇异值数据，对于有缺失属性值的样本进行删除，以及对每个属性的奇异值用领域专家认定的相应的缺省值来代替。去除数据冗余度，如将属性为立案件数和使用性质分别以是否曾经立案和是否运营车辆为基准划分 0~1 变量。

数据表示：原始数据的形式类型不统一，因此需要对数据进行离散化和概念分层，要点包括：对离散型数据 0~1 化；对连续型数据离散化，通过线性的比例变换映射到某一区间，最后根据属性值的范围划分为N个不相交的区间，用编号代替每个区间属性的值。

3. 量化关联规则

针对连续型的属性变量，若将其映射到二元属性集中，将更好地适用于关联分析模型。为此，本文给出一种基于支持度与置信度对连续属性值的最优划分区间的算法。现给出支持度与置信度的定义：

$$SP(A \to B) = \frac{\sigma(A \cup B)}{\sigma(A)}$$

$$CF(A \to B) = \frac{\sigma(A \cup B)}{\sigma(A)}$$

其中 SP 表示支持度，CF 表示置信度，σ 表示对集合内的元素进行计数，A 与 B 为两个互不相交的项集。

属性离散化的一个关键参数是用于划分每个属性的区间个数，一般情况下，该参数根据模型的算法复杂度以及经验值而定。若区间间隔取值过宽，则可能因为缺乏置信度而丢失某些特征；若区间间隔取值过窄，则可能因为缺乏支持度而丢失某些特征。本文考虑采取迭代二划分的思想，对第一个区间的右端点在有效数据范围内进行搜索，每次搜索得到一个 2 × 2 支持度矩阵与一个 2 × 2 置信度矩阵，对这两个矩阵中对应位置的一对元素值满足仅当其支持度超过 5%，并且它的置信度超过 65%的情形进行统计，得到满足该条件下统计得到的数量最多相应的搜索区间作为其最优划分区间，依次类推，当划分的最后一个区间段内的数据统计值小于之前所有区间段内数据统计值的平均值时，停止迭代[2]。

根据上述思想，本文中针对续保年、车龄、被保险人年龄、签单保费、立案件数、已决赔款、新车置购价 7 个连续变量分别取 5、7、6、5、2、2、6 作为其划分区间个数，其离散化效果图如图 1 所示：

由图 1 可知：基于支持度与置信度对连续属性值的最优划分区间在一定程度上接近于等频离散化属性值的区间，它们的划分区间都集中分布于数据量密集处。由上图也可以看出，签单保费与新车置购价在价格中下游档次有大量购买现象，而在高价格区段购买量少，该现象与当今社会消费结构相符。

针对连续型变量离散化后区间，本文考虑将划分后区间转化为“项”，这种类型的变换可以通过为每个不同的属性值对创建一个新的项来实现。针对其他分类属性以及对称二元属性，本文直接将其转化为“项”。

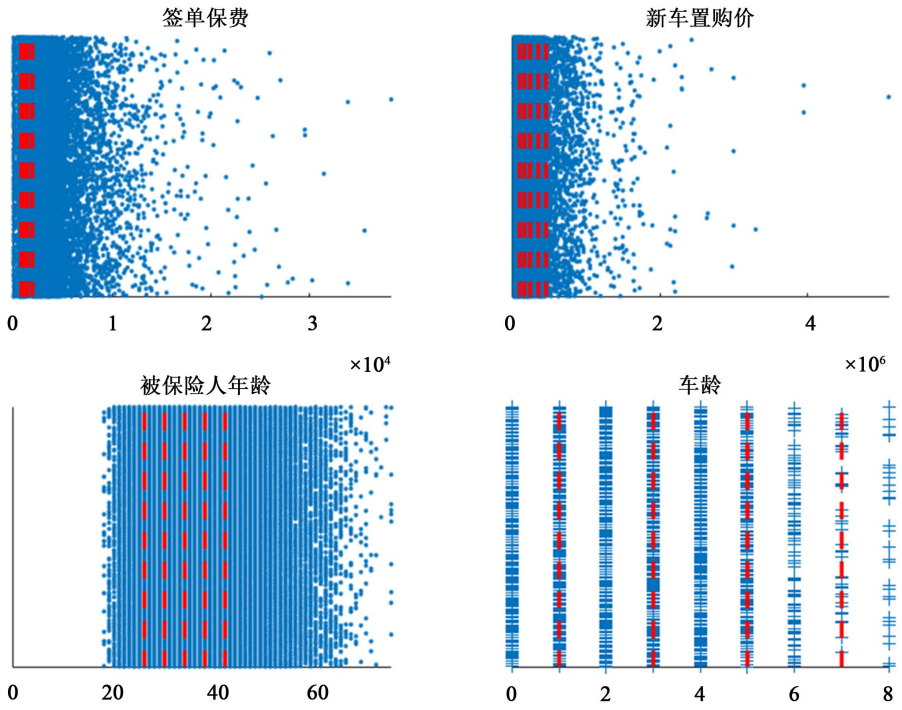

Figure 1. Distribution of discrete data
图 1. 离散化数据分布图

4. 基于关联分析的客户画像

- **关联分析提取变量**

为实现对客户进行精准画像，本文利用 Apriori 算法[3]，以逐层计算的思想，来简化频繁项集的构造，即首先对数据预处理中所得到的 60 个变量构建候选 1-项集，以此类推，从频繁 1-项集到最长频繁项集，每次遍历项集格中的一层，即经历每一次迭代后，新的候选项集由前一次迭代发现的频繁项集产生，然后对每个候选的支持度进行计数，并与最小支持度阈值进行比较，最终得到其最长频繁项集，过程如图 2 所示。

候选 1-项集

项集	计数
续保	17691
转保	49744
……	……
性别	16668
……	……
立案	1579

候选 2-项集

项集	计数
续保，机构	119
转保，机构	189
……	……
投保车损，续保年	25323

候选 3-项集

项集						计数
车龄	续保年	被保险人年龄	新车购置价	签单保费	三者险保费	3580

Figure 2. Generates frequent item sets using Apriori algorithm
图 2. 使用 Apriori 算法产生频繁项集

图 2 中，蓝色底纹代表因支持度低而被删除的项集。在候选集的选取的过程中，会产生一些不必要的或者重复的候选集，该现象直接影响算法的复杂度。本文采用 $F_{k-1}\times F_k$ 方法，即每一个频繁 k-项集都是由一对频繁 $(k-1)$-项集合并而成，其中它们的前 $k-2$ 个项都相同。事实上，它确保每个频繁项集中的项以字典序存储，每个频繁 $(k-1)$-项集只用字典序比该项集中所有的项都大的频繁项进行扩展。

根据先验原理得到：车龄、续保年、被保险人年龄、新车购置价、签单保费、三者险保费 6 个变量与是否续保属于强关联规则。

- **客户精准画像**

考虑量化关联规则 $A\to t:\mu$，其中 A 是频繁项集，t 是连续的目标属性，而 μ 是被 A 覆盖的事物 t 的平均值。此外，设 μ' 是未被 A 覆盖的事物 t 的平均值。目标是检测 μ 和 μ' 之差是否大于客户指定的某个阈值 Δ。在这种情况下，原假设 $H_0:u'=u+\Delta$，而备择假设是 $H_1:\mu'>\mu+\Delta$，统计量为：

$$Z=\frac{\mu'-\mu-\Delta}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \tag{1}$$

其中，n_1 是支持 A 的事务个数，n_2 是不支持 A 的事务个数，s_1 是支持 A 的事务的 t 的标准差，而 s_2 是不支持 A 的事务的 t 的标准差。如果有 $Z>Z_\alpha$，则拒绝原假设，我们可以认为该量化关联规则是有趣的[4]。

由于量化关联规则可以用来推断总体的统计性质，根据关联分析以及得到的与是否续保存在强关联规则的 6 个变量，可以提取如下形式的量化关联规则：

{0 < 车龄 < 5，0 < 续保年 < 1，被保险人年龄 > 34，103,701 < 新车置购价 < 245,001，签单保费 > 1920，400,000 < 三者保险额 < 800,000，是否续保 = 是} → 年龄：均值 = 3.91

{0 < 车龄 < 5，0 < 续保年 < 1，被保险人年龄 > 34，103,701 < 新车置购价 < 245,001，签单保费 < 807，签单保费 > 1920，400,000 < 三者保险额 < 800,000，是否续保 = 是} → 年龄：均值 = 36.39

综上，为实现对每一客户的精准画像，只需找到量化关联规则，并将其基本统计信息带入 Z，得出其值，即可通过统计决断判别客户的各类属性。

5. 续保率预测

为实现对每一客户的续保率预测，本文选取 Logistic 回归模型[5]。Logistic 回归是针对因变量为定性变量、自变量为分类变量的一种解决方案，根据关联分析模型，选取离散化后的车龄，续保年，被保险人年龄，新车购置价，签单保费，三者险保费 6 种因素为续保的主要影响因素，以是否续保为因变量，建立了影响汽车续保定性评价的分组数据 Logistic 回归方程。

Step 1：构建 Logistic 回归模型

对于每一张保单来说，发生续保即为 $y=1$ 的情况，不发生续保就是 $y=0$ 的情况。各续保保单 Y 的概率函数为：

$$P(Y=y)=p_i^y(1-p_i)^{1-y},\ y=0\text{或}1 \tag{2}$$

$$p_i=\frac{\exp(\beta_0+\beta_1x_{i1}+\cdots+\beta_kx_{ik})}{1+\exp(\beta_0+\beta_1x_{i1}+\cdots+\beta_kx_{ik})} \tag{3}$$

其中 $X=(x_1,\cdots,x_k)$ 是影响续保发生与否的因素，β_i 是 x_i 的回归系数，代表影响因素 x_i 对续保的影响程度。

现考虑 n 个保单类别，所有保单的对数似然函数为：

$$l(p_1,\cdots,p_n;y)=\sum_{i=1}^{n}\left[y\log p_i+(1-y)\log(1-p_i)\right] \tag{4}$$

进一步假设续保率是解释变量的函数，即：

$$g(p_i)=\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik} \tag{5}$$

其中 g 为连接函数，通常选用 logit 函数，即有：

$$\operatorname{logit} p_i=\log\left(\frac{p_i}{1-p_i}\right)=\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik} \tag{6}$$

这就是续保模型的一般形式。

Step 2：参数估计及显著性检验

根据 Logistic 回归，可以得出所选变量的系数(如表 1)，对每一辆汽车是否续保的相关数据进行分析和总结，来获得续保率和车龄、续保年、被保险人年龄、新车购置价、签单保费、三者险保费等多种因素之间线性和非线性的定量定性关系。

Table 1. Regression coefficient analysis
表 1. 回归系数分析

变量	离散化变量	回归系数	标准差	Sig.
常数项	常数项	−2.8932	0.0121	0.020
车龄	车龄(0, 1]	3.0608	0.047	0.000
	车龄(1, 3]	1.8422	0.048	0.000
	车龄(3, 5]	1.386	0.05	0.011
续保年	续保年(0, 1]	−2.4409	0.045	0.045
被保险人年龄	被保险人年龄(34, +]	0.3832	0.025	0.032
新车购置价	新车购置价(103,701, 169,501]	0.1279	0.029	0.023
	新车购置价(169,501, 245,001]	0.291	0.037	0.020
签单保费	签单保费(−0.807]	0.5228	0.045	0.000
	签单保费(1920, +]	0.377	0.031	0.000
三者险保额	三者险保额(400,000, 800,000]	−0.1472	0.033	0.002

通过进行回归方程显著性检验可知：$R^2=0.973$，说明模型具有很强的解释力；由回归系数显著性检验，可以得到公式(3)中每个变量的系数，由上表可知，车龄对 p_i 的影响最大，其次是续保年和签单保费。且由各变量参数估计及其统计检验可知，各系数统计量检验的效果显著(Sig. < 0.05)。以上充分说明了模型的可行性。

Step 3：模型续保率预测

根据 Logistic 回归预报出每一保单续保的情况，预测得到其平均续保率为 19.02%。

6. 模型检验

本文将每一保单预测续保的情况与样本数据中实际续保情况进行拟合，经计算得到其拟合度高达 96.7%。为了更加清晰地体现其拟合程度，图 3 给出前 100 个样本序列拟合情况[6]。

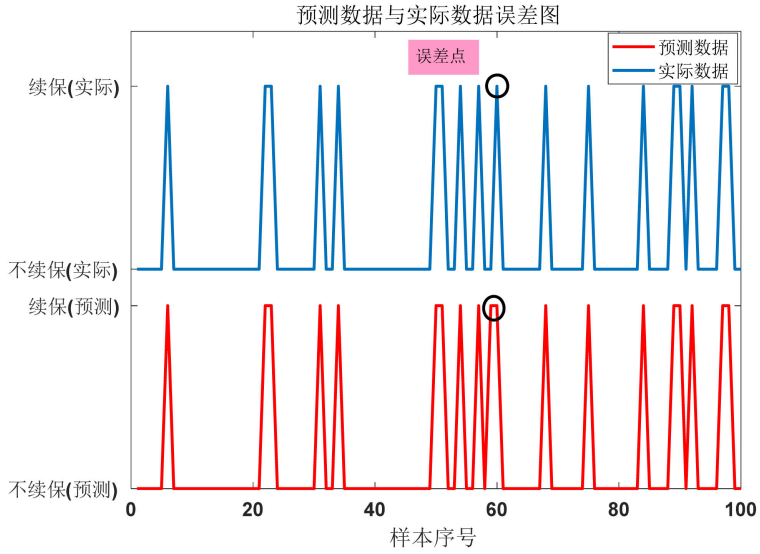

Figure 3. Error maps of predicted and actual data (the first 100 samples)
图 3. 预测数据与实际数据误差图(前 100 个样本)

基金项目

本文由湖南农业大学大学生创新性实验计划项目(SCX1802)资助。

参考文献

[1] 杨子江, 王野, 马天诣. 影响汽车保险续保率的因素分析[J]. 企业研究, 2011(10): 107.

[2] 倪琪, 刘骅飞, 田雪颖. 车险续保率影响因素模型[J]. 企业研究, 2011(10): 112-113.

[3] 顾茜. 我国金融业与三次产业关联分析——基于 2012 年中国投入产出表[J]. 中国集体经济, 2019(12): 12-13.

[4] 温桂国. 浅谈财产保险公司车商业务续保困境及思考[J]. 商业经济, 2018(5): 137-138.

[5] 陆芳园. 基于 Logistic 模型的中小企业信用风险管理研究[D]: [硕士学位论文]. 成都: 西南财经大学, 2011.

[6] 卓金武. MATLAB 在数学建模中的应用[M]. 北京: 北京航空航天大学出版社, 2011.