Hans Journal of Data Mining 数据挖掘, 2019, 9(3), 69-80
Published Online July 2019 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2019.93009

Hans汉斯

Auto Insurance Renewal Forecast and Scheme Optimization Based on Customers' Characteristics

Kaibin Wu, Hanwen Deng, Xuchang He, Minmin Luo*

Department of Medical Engineering, School of Biomedical Engineering, Southern Medical University, Guangzhou Guangdong
Email: *22809017@qq.com

Received: Jun. 14th, 2019; accepted: Jun. 26th, 2019; published: Jul. 3rd, 2019

Abstract

This paper aims at the pre-judgement of the customers' renewal probability to the insurance company. Based on big data, this paper comprehensively uses the back-propagation neural network and variable weight combination to describe the images of the customers accurately, thus building a customer renewal probability model. Secondly, the L1 regularization is used to extract key features as the basis for classifying customers, thus making a customer classification. As a result, different preferential and welfare programs are designed for different types of customers to increase the renewal rate.

Keywords

BP Neural Network, L1 Regularization, Support Vector Machine, Big Data

基于客户特征的车险续保预测与方案优化

吴锴镔，邓汉闻，何旭昌，罗敏敏*

南方医科大学，生物医学工程学院，医学工程系，广东 广州
Email: *22809017@qq.com

收稿日期：2019年6月14日；录用日期：2019年6月26日；发布日期：2019年7月3日

摘　要

本文针对保险公司如何预判断客户续保概率这一问题，依托大数据的条件，综合运用反向传播BP神经网

*通讯作者。

文章引用：吴锴镔，邓汉闻，何旭昌，罗敏敏. 基于客户特征的车险续保预测与方案优化[J]. 数据挖掘，2019，9(3): 69-80. DOI: 10.12677/hjdm.2019.93009

络、变权组合等方法对客户精准画像，建立客户续保概率模型。其次，运用神经网络L1正则化提取关键特征作为划分客户类别的依据，实现客户的等级划分。由此，针对不同的客户设计不同的优惠和福利方案，提高客户的续保率。

关键词

BP神经网络，L1正则化，支持向量机，大数据

1. 引言

目前我国车险市场巨大，据统计，我国汽车保险行业的保险收入在 3000 亿元左右，并以逐年递增的趋势发展。我国在 2000 年前后经历了两次车险费率改革，但效果并不显著，各大车险公司纷纷通过恶性价格竞争和高成本营销手段赢得市场，导致车险市场急剧恶化，财险公司难以盈利。后来行业协会对整个行业进行调整，规定了行业基线，市场才有所好转。但与此同时，大部分车险费率的制定都主要遵从“从车主义”，导致目前中国车险定价模式单一，车险费率条款、产品同质化严重。“从车主义”费率模式系统整体比较简单，但其局限性也十分突出。根据交通统计资料的数据可知，发生交通意外超九成都是由于驾驶员因素造成[1]，故针对投保顾客的特征进行续保预测具有重要实用意义。

通过近年来数字化和数据挖掘技术的发展，可预见到未来的车险定价模式的转换。转变后的“从人主义”可以通过对驾驶人的驾驶经验、驾驶习惯来制定车险保费，使拥有不同驾驶习惯的驾驶员拥有属于自己独特的定费标准，这更符合保险中的公平互利原则，也使车险保费制度更具有多样性和竞争性。车险定费的“从人主义”需要大量的数据分析，通过对大量已有的客户数据行进分析，对客户的投保类型进行分类，并挖掘出车险客户潜在的投保规则，从而对不用驾驶习惯的投保人制定不同的投保方案。

然而，在中国保险运营研究领域中，目前业界对客户的数据仍然停留在了储存阶段，数据没有经处理分析，也并未具体应用到实际场合，2016 年唐俊虎等人[2]曾利用大数据所得的信息，对客户进行星级划分，但其准确率并不理想，同时也缺乏对提取到的客户特征进行说明。

2. 国内研究现状

目前国内对车险续保概率预测的研究并不多，特别是运用数据挖掘预测车险续保概率的研究。倪琪、刘骅飞、田雪颖(2011)利用某保险公司的车险续保数据，用逐步回归和数据拟合求出了车险保单各因素对续保率的影响因子，讨论了哪些因子对续保率有较大影响[3]。王钧等(2011)运用粗糙集理论(RS)产生规则和灰度关联度法，挖掘出车险保单数据中潜藏的续保规则[4]。本文采用 BP 神经网络，利用机器学习的方法对客户进行精准画像。

3. 问题提出

目前业界普遍对数据的挖掘应用不够，缺乏对客户信息的合理充分利用。本文通过运用神经网络对客户信息特征加以提取，提出有效方法以解决以下两个问题：

1) 问题一：对已有的大量车险保户信息进行处理，构建出精准客户画像模型，通过此模型得出影响

客户未来续保的相关因子，并给出这类客户的续保概率。

2) 问题二：通过得到的模型，在保证企业一定的盈利下，从影响车险保户续保的几大影响因素出发，为提高车险保户的续保率，制定出对不同类型的车险保户的优惠方案。

4. 实验思路

4.1. 问题的总体分析

给出客户的精准画像和续保概率是属于分类、预测问题，对于给出优惠方案，是一个约束性最优化问题。基于所得到的车险投保客户数据，运用 MATLAB 和 PYTHON 软件，通过神经网络训练，抽离出特征权重，从而对有明显差异的权重为特征进行分类。通过 BP 神经网络得出的权重表，通过 L1 正则化进行稀疏化，对相关性强对区间进行合并统计，从而得到针对不同种类客户的优惠方案，见图 1 所示。

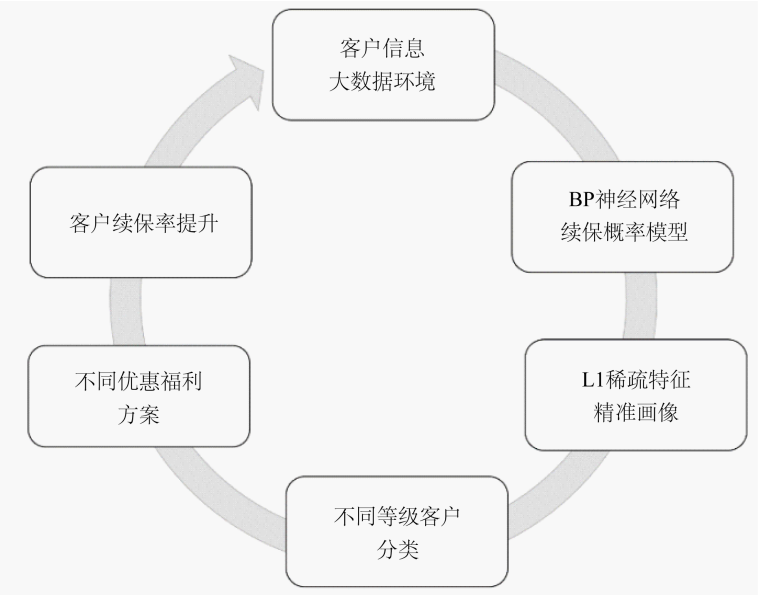

Figure 1. An overall analysis of the problem of increasing the probability of renewal
图 1. 提高续保概率问题的整体分析

4.2. 问题一解决

通过对数据的预处理，把保户信息数字化并忽略次要数据。将客户信息特征数据分为两种：将数值大小无关联的离散数字数据通过 ONE-HOT 进行编码，同时把数值大小有关联的连续数字数据信号进行归一化。

传统做法需要通过 PCA 降维，将关联度大的数据变量合成为相关性小的变量，从而达到减少变量的目的，同时也能得到影响续保权重的排序。本文使用深度学习 BP 神经网络训练投保用户已预处理过的数据信息，通过一层 Dense 层便可得到该批客户的影响续保权重的排序和续保概率。通过对比，发现 BP 神经网络的准确度与速度都高于 SVM，于是主要采用 BP 神经网络所得出的结果。

4.3. 问题二解决

通过已得到的 BP 神经网络续保概率模型，使用神经网络 L1 正则化对特性集合进行稀疏处理，找到

客户的关键特征，对特征权重中的签单保费和累积续保年进行拟合，再通过 Dense 层提取签单保费的影响因素，针对不同等级的客户实行不同的费用折扣，从而提高客户的续保率，把低等级保户转化成高等级保户。

5. 实验过程

5.1. 数据描述

本文采用某财险公司的保单数据，数据总数 65,000 条，原始数据特征有 27 个。其中某些数据存在缺失，缺失在大量保单数据中属于正常现象，后续会对此进行处理。

模型输入为已投保客户个人信息数据特征，其中包括：投保人性别、投保人年龄、保单时长、保单性质、客户类别、车龄、新车购置价、险种、风险类别、NCD (无赔款优待系数)、风险类别等车辆保险基本信息。

这些信息普遍在客户购买保险时便有所记录，从而确保了模型的可推广性。

5.2. 预处理

1) 我们首先利用 EXCEL 软件，将所有的中文信息按照一定规则编译成数字信息和英文，便于导入 MATLAB 进行处理，也就得到一个纯数字化的客户信息数据。

2) 将被保人性别、客户类别这类按一定规则数字化过后没有大小强度意义的离散数值进行 ONE-HOT 编码。通过 ONE-HOT 编码，把大权重拆分成了若干个小权重，这既对数据的特征进行了扩充，降低了异常值对模型的影响，增强了模型的稳定性，也提升了模型的非线性能力。

3) 将签单保费、年龄这类有大小意义的连续数值和离散数值进行归一化处理。再清洗数据中混杂的坏死值，用平均数填充数据的缺损值。这样即可得到一份从 0 到 1 的数字信息数据方便作为输入。

5.3. 模型建立

5.3.1. 模型假设

1) 假设所给的保户信息都正确；

2) 假设社会在一段时间内保费制度不发生改变；

3) 假设保户在一段时间内按照自己过去的习惯进行投保。

5.3.2. 模型简介

1) 主成分分析法：利用降维的思想，把多指标转化为比较数量的几个综合指标(即主成分)，降维后的数据能够反映原始变量的 N%信息。本文我们定义要求反映 99%的效果。

最后，我们得出 12 个提取后的特征对顾客实现精准画像。这种方法虽然在引进多方面变量的同时将复杂因素归结为几个主成分，使问题简单化，同时得到的结果更加科学有效，但对复杂的模型数据并不适用，尤其是非线性回归模型。

我们为了全面、系统地分析问题，我们必须考虑众多影响因素，也就是特征或称为指标，在多元统计分析中也称为变量。本文选用一层 Dense 层的 BP 神经网络对客户续保概率进行预测。

2) 支持向量机由 Vapnik 提出，通过控制超平面的间隔度量和核技巧能够解决线性和非线性分类问题。最大间隔分类器是支持向量机的一种，是通过在特征空间找到一个超平面将不同类别分割开，因此最大间隔分类器只能适用于线性可分的二分类问题。最大间隔分类器需要在保证 2 类样本无错误的分开的同时，使得 2 个类别的分类间隔最大，见图 2 所示。

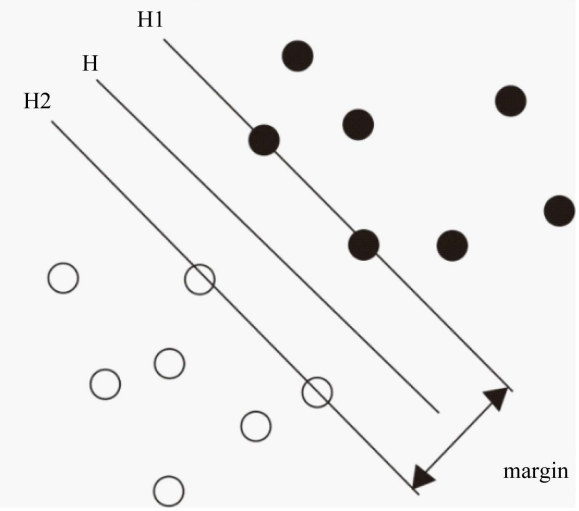

Figure 2. Finding hyperplane in support vector machine
图 2. 支持向量机寻找超平面

3) BP 神经网络：深度学习中的人工神经网络是由大量简单处理单元相互连接构成的高度并行的非线性系统，具有大规模并行性处理特征。大量简单处理单元的并行活动使网络呈现出丰富的功能并具有较快的运行速度。

BP 神经网络中最重要的部分是通过训练过程来调整网络中运算单元间连接的权重。ANN 训练前其输出是凌乱的，随着训练次数的增加，BP 神经网络的连接权重被不断调整，使目标值与 BP 神经网络输出值的误差逐渐减小直至为近似为零，此时称 BP 神经网络已收敛，训练完成。ANN 训练结束后，还需要用另一组与训练集不同的样本测试其输出是否与真实值的接近，从而验证模型的推广性。

通过已有样本的学习，将所提取的样本对应的非线性映射关系存储在权值矩阵中，在后续工作中，当我们向网络输入训练未曾见过的非样本数据时，网络也能完成输入到输出的正确映射过程。图 3 给出了神经元的模型。

图 3 的神经元模型也包含偏置 b_j，用于相应地增加或降低激活函数的网络输入。ANN 的主要结构为运算单元、层与网络三个部分。

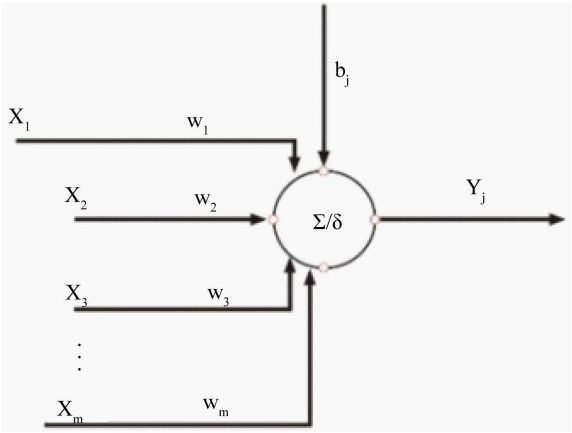

Figure 3. Neuron model of neural network
图 3. 神经网络的神经元模型

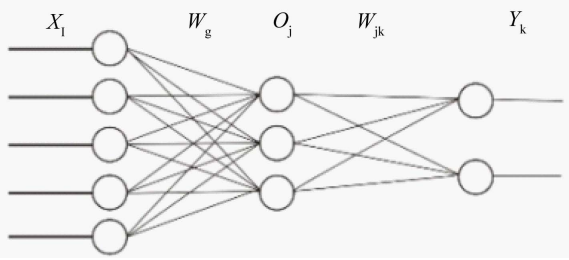

Figure 4. Neural network model [5]
图 4. 神经网络模型[5]

其中隐含层 O_i，可以是一层或多层由输入层、隐含层和输出层构成。图 4 为含一个隐藏层的 ANN 模型，其中输入层、隐含层、输出层神经元分别用 X、O、Y，神经元数分别为 5、3、2。训练时，输入由输入层神经元 X 传播至隐含层神经元，ω_{ij} 表示不同层的节点间连接权值，信息在隐含层进行加权求和并通过激活函数进行变换。从隐含层输出的信息再作为输出层的输入信息，通过 ω 加权并由输出层输出。

4) 正则化的主要思想，将一些不重要的特征的权值置为 0 或权值变小使得特征的参数矩阵变得稀疏，使每个变量都对整体有一点贡献。

对模型参数的 L1 正则项为：

$$J = J_0 + \alpha\omega_1 \tag{1}$$

设带 L1 正则化的损失函数为：

$$\Omega(\theta) = \omega_1 = \sum_j \omega_i \tag{2}$$

假设损失函数在二维上求解，则可以画出图像，见图 5。

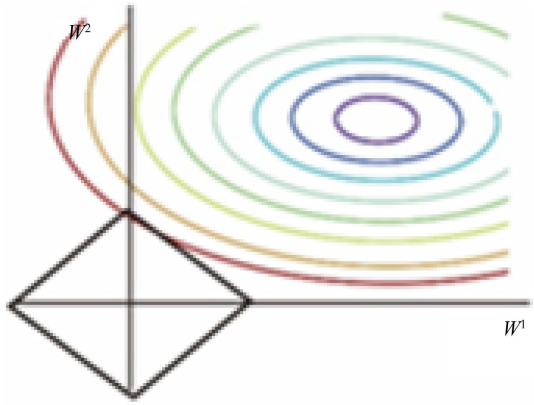

Figure 5. Image of loss function in two dimensions [6]
图 5. 损失函数在二维上的图像[6]

彩色实线是 J_0 的等值线，黑色实线是 L1 正则的等值线。二维空间(权重向量只有 ω_1 和 ω_2)上，L1 正则项的等值线是方形，方形与的等值线相交时相交点为顶点的概率很大，所以 ω_1 或 ω_2 等于零的概率很大。所以使用 L1 正则项的解具有稀疏性。

相同的，也可以推广到本问题中的更大的维度空间。L1 正则项的等值线或等值面是比较尖锐的，所以这些突出的点与接触的机会更大，而在这些突出的点上，会有很多权值等于 0。由 L1 正则化导出的稀

疏性质已被广泛用于特征选择，特征选择可以从可用的特征子集中选择有意义的特征。本文提取出来的关键特征便是累计续保年。

5.3.3. 问题 1 解决

BP 神经网络搭建：

将客户特征输入神经网络中，以是否续保作为输出项训练神经网络。基于 Keras 框架的网络设置如下：

1) 模型参数设置：

在进行训练之前，需要配置模型的优化器、损失函数、指标列表。本文选择分类准确度 Accuracy 作为损失函数从而获得客户续保概率预测值与真实值之间的误差。

由于 SGD 算法收敛速度较慢，同时预测结果容易出现震荡，本文选择 Adam 作为优化器。而优化后的 Adam 算法收敛速度快，精确度也得到提高。

2) 模型训练：

设置一层全连接层 Dense。由于数据中的输出选用了是否续保，该结果的数据真实值分布约为 8:1，存在数据不均衡的情况，因此利用变权组合，给予少数据的训练结果更高的权重来更新神经网络迭代。每次训练包含的样本数 batch_size = 128，训练轮数 epoch 设置为 100。

3) 模型预测：

训练后的模型需要在测试集上进行验证，验证模型预测的精度。由于仅使用了一层 Dense 层，该层的对每一个输入都有一一对应的一个权重，使用深度学习 BP 神经网络训练投保用户已预处理过的数据信息，从网络中得到权重就及网络连接，就可直接得到该批客户的特征值排序和续保概率。

单层 Dense 的 BP 神经网络模型见图 6。

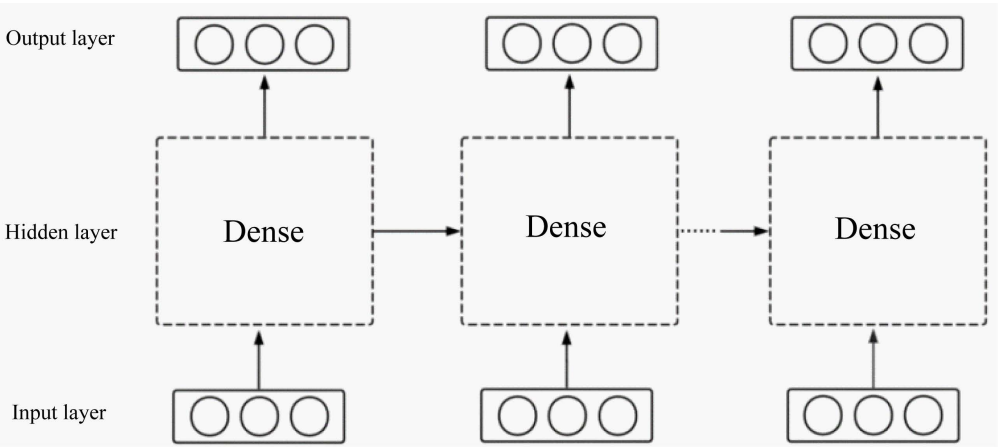

Figure 6. Single-layer Dense BP neural network model
图 6. 单层 Dense 的 BP 神经网络模型

对 BP 神经网络的所有客户特征求取平均值，选取大于平均值的权重为正值的客户特征分析，见表 1 所示。

不难发现，权重占比比较大的如续保年，签单保费等都是客户本身人的特征，而非汽车的特征。

但由于只使用了一层 Dense 层，特征种类存在交集，并非互相独立，不适合作为客户的精准画像，本文对特征种类进行了合并。结果见表 2 所示。

Table 1. Feature type with positive weight in customer information
表 1. 客户信息中权重为正值的特征种类

特征种类	权重	特征种类	权重
累计续保 1 年	3.14789	投保车上人员	0.52791
累计续保 2 年	2.80372	新车价格	0.432727
累计续保 3 年	2.44013	党政机关、事业团体用车	0.415859
累计续保 4 年	2.24108	带拖挂的载货汽车	0.380481
累计续保年 5 年	1.86614	NCD	0.361888
保单性质	1.85213	特种车一	0.350993
累计续保 7 年	1.20996	交商全保	0.313121
累计续保 6 年	0.955628	轻微型载货汽车	0.303959
10 吨及 10 吨以上挂车	0.942373	出租租赁	0.295914
客户类别：机构	0.804504	渠道：个人代理	0.292227
签单保费	0.768047	低速载货汽车	0.271022
累计续保 8 年	0.755821	企业非营业用车	0.268568
带拖挂汽车	0.710418	低速货车和三轮汽车	0.254239
特种车三	0.710355	中型载货汽车	0.251581
临时车牌	0.656487	保单性质：交强险	0.249178
2 吨及 5 吨以下货车	0.597618	10 座及 20 座以下客车	0.213471
特种车	0.567696	已决赔款	0.210001

Table 2. Feature types without intersection after merging
表 2. 合并后无交集的特征种类

精准画像	权重
续保年份	15.420369
货车	4.0858
挂车	3.3054
保单性质	1.92754612
客户类别：机构	1.85213
签单保费	0.768047
临时车牌	0.656487
投保车上人员	0.52791
新车价格	0.432727
NCD	0.361888
保单性质：交强险	0.249178
已决赔款	0.210001

4) 为对比神经网络对于传统做法——PCA + SVM 的优越性。本文还运用 SVM 模型进行对比考察。

首先，将客户信息数据导入 MATLAB，进行 PCA 降维。通过对生成的降维特征矩阵和原有数据进行矩阵运算即可得到权重转移矩阵 w，对 w 进行求和排序便得到了客户特征值的排序。利用 PAC 降维提取的 12 个特征我们建立一个以是否续保作为因变量的 SVM 预测模型。SVM 多元线性回归模型的一般形式：

$$Y_i = \beta_0 + \beta_{1X_{1i}} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \left(i = 1,2,\cdots,n\right) \tag{3}$$

其中 k 为解释目前主成分分析得到的特征变量的数目，$\beta_j \left(j = 1,2,\cdots,k\right)$ 称为回归系数(regression coefficient)。上式也被称为总体回归函数的随机表达式。它的非随机表达式为：

$$E\left(Y \middle| X_{1i}, X_{2i}, \cdots, X_{ki}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{k_i} \tag{4}$$

β_j 也被称为偏回归系数(partial regression coefficient)线性模型使用最小二乘法求解参数，本文我们借助 python 运用机器学习库来实现最小二乘法。

支持向量机要求在超几何空间中找到一个超平面，使得两类样本距离最大，也就是续保和不续保。

通过对比由 PCA + SVM 和 BP 神经网络得到的客户特征值排序，可以相互验证两种结论的正确性和准确度，也同时提高了由 BP 神经网络所得出的客户续保概率的可信效度。但是通过对比这两种方法，我们发现神经网络的准确度与速度都高于 SVM，见下表 3。

Table 3. Comparison of BP neural network and SVM accuracy
表 3. BP 神经网络与 SVM 准确度对比

	输出	
模型	BP 神经网络	SVM
权重提取方式	Dense 层	PCA
Accuracy	0.99	0.93
F1 score	0.98	0.85

5.3.4. 问题二解决

通过运用神经网络 L1 正则化，对投保客户的所有特征合集进行稀疏处理，可以得到客户的关键特征的权重排序。以稀疏得到的结果：累计续保年作为对客户群体的重要考察指标。本文依据不同续保年在 BP 神经网络中的权重排序，对不同续保年的客户进行区间合并统计，相对准确地将客户划分为五种等级：高收益客户(累计续保 1 年)，较高收益客户(累计续保 2 年)，中收益客户(累计续保 3 年)，低收益客户(累计续保 4 年)，极低收益客户(其他年份数)，见下图 7。

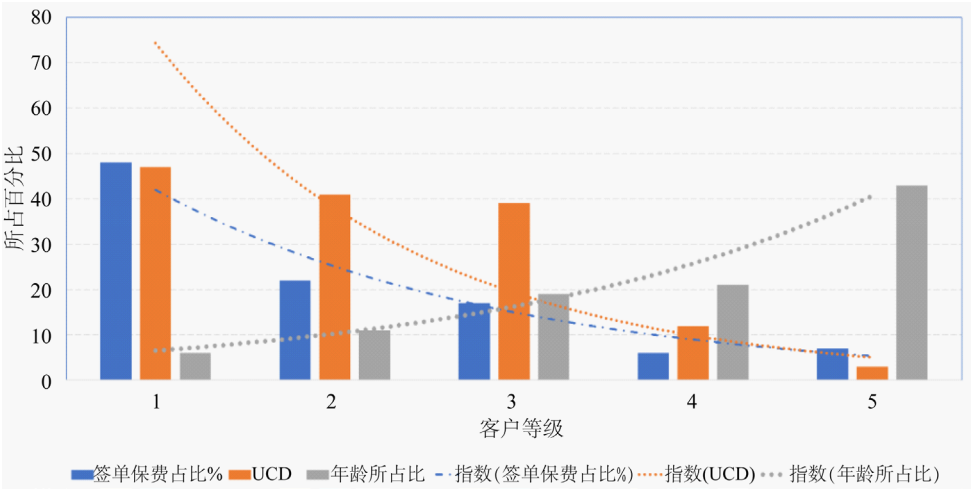

Figure 7. Comparison of weight attributes between different levels of customers
图 7. 不同等级客户之间权重属性的比较

分析可发现：从高价值到低价值，客户的签单保费与 UCD 成指数下降，平均年龄从整体上看也呈现反比趋势，在中收益与低收益客户间，年龄的差距不明晰。

可见稀疏化后的特征：累计续保年能比较好的划分开五类客户。进一步分析：高续保率的客户往往更愿意支付更多的签单保费。而且，签单保费金额在种类 1 与种类 2、3 之间的区别并不是特别大。中等的客户与较高等的客户在优惠的让利刺激下，稍微多支付些许签单保费则种类 2 和种类 3 的客户可大量成为种类 1——高续保率客户。

对于客户是否续保这一二分类问题，对于样本集合，可以用数学形式表示：

$$(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m),x\in R_n,y\in(-1,1) \tag{5}$$

支持向量机寻找一个最优超平面，使得分类间隔最大，在图 2 中，H1、H2 为 2 类分类样本离分类线最近且平行于分类线的直线，这 2 条线之间的距离称之为分类间隔。

设分类方程为：

$$X\cdot W+b=0,y\in(-1,1),y_i\left[-(Wx_i+b)\right]>O,i=1,2,3,\cdots,m \tag{6}$$

则分类间隔是 $2/\|w\|$，使分类间隔最大等价于$\|w\|$最小，满足上述约 m 束条件，并且使得 $\|w\|$ 最小的分类线就是最优分类线。这样原本就可以完成二分类问题，实现对客户是否续保做出评估。

5.4. 方案优化

基于第一问所得的影响保户续保的主要影响因子，再根据五个不同客户等级的特征分析，可以初步得到基于不同等级的客户优惠方案。

具体的实行细则如下：取优惠下限为五折，考虑当前客户与高续保客户保费的比值和与上一级保户保费的比值，对初步优惠方案进行探讨，得到初步的优惠方案

在运用此车险优惠方案之后，对不同等级的客户的收费变见表 4。

Table 4. Changes in customer auto insurance charges for different levels after the initial offer
表 4. 运用初步优惠后不同等级的客户车险收费变化

等级	1	2	3	4	5
实施优惠前的保费占比	48.00%	22.00%	17.00%	6.00%	7.00%
该等级客户与等级 1 客的户保费的比值	100.00%	46.00%	35.00%	12.50%	14.60%
与上一等级保客户保费的比值		46.00%	77.27%	35.30%	111.60%
费用优惠方案	10 折	7 折	6 折	6 折	5.5 折
使用方案后新的保费占比	54.00%	24.00%	14.70%	6.20%	1.10%

将初步优惠方案实行后新的客户样本分布输入 BP 神经网络概率模型进行仿真实验，可以发现均值明显升高，故而初步优惠方案研究发现可行度很高。对优惠福利方案进行深入设计，给出三个系列的方案。

实行不同的优惠福利方案如下：

1) 系列 1：

1. 等级一——不打折，

2. 等级二——7 折，

3. 等级三、等级四——6 折，
4. 等级五——5.5 折。

2) 系列 2：

1. 等级一——9 折；
2. 等级二——7 折，
3. 等级三——6 折，
4. 等级四——5.5 折，
5. 等级五——5 折。

3) 系列 3：无论等级，全体客户一致享受 7 折。

其中系列一为不做优惠方案系列，系列 4 是人为设置的对比组，用来对比其他方案的可行性。若方案三的得出的模拟效果与其他系列效果相当、或比其他方案更优，则说明其他方案的可行性极差；反之，若其他系列的续保率增量明显大于次组，即从另一方面验证了其他系列的可行性。

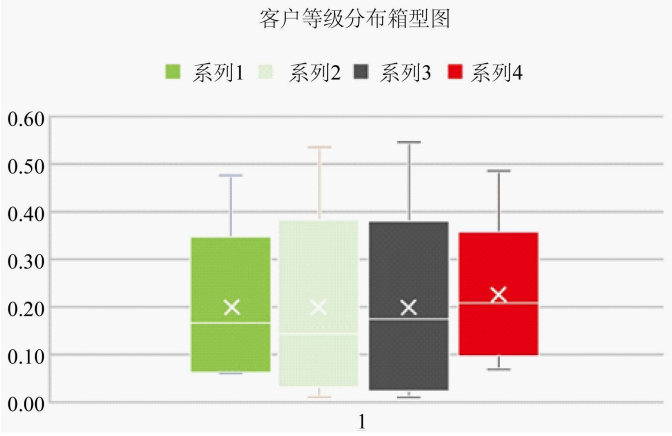

Figure 8. Four series of graded distribution box diagrams
图 8. 四种系列的等级分布箱型图

根据 BP 神经网络续保概率模型，得出的客户等级分布箱型图，见图 8。

可以看出，系列 4 的效果最差，对客户群体的区分度很小，也从侧面印证了系列 2 和系列 3 对客户续保有一定积极的效果。

当新的保费占比形成后，此时，不同种类的客户类型所占客户总人数也同样发生相应的变化，见表 5。

Table 5. Comparison of total renewal customer ratio of the four schemes
表 5. 四种方案的总续保客户占比对比

客户等级	原始	方案 1	方案 2	方案 3
1	48.00%	54.00%	55.00%	49.00%
2	22.00%	24.00%	22.00%	21.00%
3	17.00%	15.00%	18.00%	23.00%
4	6.00%	6.00%	4.00%	14.00%
5	7.00%	1.00%	1.00%	7.00%
总续保客户占比	18.70%	22.80%	23.10%	12.10%

从而我们可以得出的结论：采用系列 2，即系列 2：等级一——9 折；等级二——7 折，等级三——6 折，等级四——5.5 折，等级五——5 折这一方案，能进一步提高客户的续保概率。

6. 结论与讨论

在通过 BP 神经网络和传统 PCA + SVM 做法比较之后，我们采用 BP 神经网络这一更具准确度和效率的建模工具。将投保客户的各项保单数据通过 BP 神经网络训练，以续保概率为输出，得到投保客户的精准画像，从中我们可以得影响续保概率的主要影响因子是续保年份，说明客户以往的续保历史将极大程度上影响未来的续保概率。

通过运用神经网络 L1 正则化、合并有效区间之后，我们可以将投保客户区分成 5 大类：高收益客户(累计续保 1 年)，较高收益客户(累计续保 2 年)，中收益客户(累计续保 3 年)，低收益客户(累计续保 4 年)，极低收益客户(其他年份数)。并通过给中受益客户和较高收益客户提供有力度的优惠方案，即可将这两个类别的客户大部分转换成高收益客户，从而提高客户整体的续保概率和企业的盈利水平。根据文献，对于低价值用户，尤其是续保等级也较低、续保难度较高的低价值用户，应适当地舍弃[7]。

由于客户是否续保这一分类标准下的两种样本数量不均衡，在 BP 神经网络训练的过程中会导致对数量较多的样本过拟合了[8]，而对另外类别的样本欠拟合。本文采取变权组合的方法，在训练过程中，数量少的样本更新权重的占比更大，从而规避上述问题，但并不能完全解决这一现象，需要获取更多的客户投保信息特征的数据样本。

同时，模型的优缺点主要如下：

优点：1) 与传统的统计方法相比，人工神经网络具有许多的优点，常规的影响因素分析方法如线性回归模型、logistic 回归模型等往往要求样本服从正态分布，且自变量、因变量之间的关系为线性关系，但在保户的信息数据中可能不满足上述条件。

2) 避免了在建模过程中根据假设所设定的模型函数表达式，扩充了回归建模研究中的函数类型，能揭示保户续保事件中内在的关系。

缺点：本文不能把所有的因素都考虑到模型中去，只能把一些次要的因素忽略，但现实生活中有一些小因素可能对是否续保产生影响。

其次，考虑模型的推广：

1) 模型的横向推广：模型依托大数据背景，当后期数据不断增加时，模型精度能进一步提高[9]。

2) 模型的纵向推广：模型基于多维度的客户信息特征，不论输入多大的特征维度，都能进行稀疏化提取关键影响参数。

参考文献

[1] 陈立杰, 王丹丹. 基于 UBI 的车险费率厘定模式与方法研究[J]. 现代经济信息, 2016(22): 376.

[2] 唐俊虎, 王文志, 胡冰倩. 数据挖掘驱动下的车险续保流程再造[J]. 中国保险, 2016(3): 51-56.

[3] 倪琪, 刘骅飞, 田雪颖. 车险续保率影响因素模型[J]. 企业研究(理论版), 2011(5): 110-111.

[4] 黄沛, 李剑. 基于粗糙集理论的续保规则挖掘模型[J]. 上海交通大学学报, 2004, 38(4): 641-645.

[5] 赵东方. 数学模型与计算[M]. 北京: 科学出版社, 2007.

[6] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2010.

[7] 司守奎. 保险与数学[M]. 北京: 国防工业出版社, 2011.

[8] 吴锴镔. 贪婪算法在加工流水线的应用[J]. 数字技术与应用, 2019, 37(1): 131+136.

[9] 刘欣. 我国车险经济市场的博弈分析[J]. 中国商论, 2018(22): 143-144.

知网检索的两种方式：

1. 打开知网首页：http://cnki.net/，点击页面中“外文资源总库 CNKI SCHOLAR”，跳转至：http://scholar.cnki.net/new，搜索框内直接输入文章标题，即可查询；
   或点击“高级检索”，下拉列表框选择：[ISSN]，输入期刊 ISSN：2163-145X，即可查询。
2. 通过知网首页 http://cnki.net/顶部“旧版入口”进入知网旧版：http://www.cnki.net/old/，左侧选择“国际文献总库”进入，搜索框直接输入文章标题，即可查询。

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：hjdm@hanspub.org