

# 基于改进的 K-means 和 BP-Adaboost 的 寿险客户流失预测算法研究

闫 春,张馨予

(山东科技大学 数学与系统科学学院,山东 青岛 266590)

**摘 要:**针对寿险行业的客户流失问题,构建基于外在、内在、行为(EIB)属性的寿险客户指标体系。提出改进的 K-means 算法,使用改进的轮廓系数公式判断初始聚类数目,并利用欧式距离相似度与余弦相似度的测度优势提出欧式类簇空间的局部、全局离群点过滤规则。运用传统的 K-means 算法、不同离群点监测阈值下的改进 K-means 算法进行客户细分及其可视化展示,并采用 BP-Adaboost 算法对细分后的客户进行流失预测。实证表明:改进的 K-means 算法可视化噪声降低、簇内误差减小,可在后续的预测器中实现更高预测精度,为保险公司挖掘更精准的客户分类信息、挽留客户提供决策依据。

**关键词:**寿险;客户细分;客户流失;K-means 算法;BP-Adaboost 算法

**中图分类号:**F840.62

**文献标志码:**A

## Life insurance customer churn prediction algorithm based on improved K-means and BP-Adaboost

YAN Chun, ZHANG Xinyu

(College of Mathematics and Systems Science, Shandong University of Science and  
Technology, Qingdao, Shandong 266590, China)

**Abstract:** To solve the problem of customer churn in life insurance industry, a life insurance customer indicator system based on external, internal, behavioral(EIB) attributes was designed. An improved K-means algorithm was first proposed by using the improved contour coefficient formula to judge the initial number of clusters, and by adopting the advantage complement of Euclidean distance similarity and Cosine similarity to find the filtering rules of local and global outliers in Euclidean cluster space. Traditional K-means algorithm and the improved K-means algorithm under different outlier monitoring thresholds were then used to segment customers and display them visually. The churn prediction model of segmented customers based on the BP-Adaboost algorithm was established as well. The empirical results show that the improved K-means algorithm can decrease the visual noise and the intra-cluster error variance, and has a higher prediction accuracy in subsequent predictor, which provides decision basis for insurance companies in mining more accurate information of classified customers and in carrying out customer retention work.

**Key words:** life insurance; customer segmentation; customer churn; K-means algorithm; BP-Adaboost algorithm

客户流失是指某公司现有的客户,由于某些主观或客观因素,放弃消费当前公司的产品或服务,转而选择消费其他公司产品或服务的行为<sup>[1]</sup>。由于发展新客户的成本要比发展老客户高,各行各业对客户流失控制问题高度重视。中国的寿险行业虽起步较晚,但发展迅速。随着市场竞争愈发激烈,客户流失频

收稿日期:2020-2-13

基金项目:国家自然科学基金项目(61502280);全国统计科学研究重点项目(2019LZ10)

作者简介:闫 春(1978—),女,山东邹城人,教授,博士生导师,主要从事保险精算、大数据分析与管理研究。

张馨予(1998—),女,山东日照人,硕士研究生,主要从事精算学与风险管理研究,E-mail:zxxyz98@163.com

率较高,有必要深入开展客户流失预测研究,为寿险公司预防客户流失、提升盈利能力提供决策依据。

目前,学者们对寿险等行业客户流失预测的研究有一定进展,多采用单一模型进行预测,如决策树算法、BP 神经网络、二元逻辑回归模型等。Kisioglu 等<sup>[2]</sup>通过贝叶斯信念网络建模,识别出具有流失倾向的电信客户行为。周晓玮<sup>[3]</sup>将 BP 神经网络应用到寿险营销预警中,并比较其与支持向量机(support vector machine, SVM)、决策树算法的预测效果。Bi 等<sup>[4]</sup>将二元逻辑回归运用到电信客户流失预测中。Günther 等<sup>[5]</sup>将包含时间动态解释变量和相互作用的逻辑纵向回归模型拟合到非寿险数据中进行建模。梁锋<sup>[6]</sup>将寿险公司的客户数据生成库,用 IBM SPSS Modeler 工具和决策树算法建立预测模型。郑宇晨等<sup>[7]</sup>将 Logistic 模型用于证券公司客户流失预警分析。Amin 等<sup>[8]</sup>提出一种基于粗糙集理论(rough set theory, RST)的规则智能决策技术,用于提取与电信客户状态相关的重要决策规则。冯鑫等<sup>[9]</sup>以在线评论信息为基础,将情感因素引入 BP 神经网络,进行移动网络虚拟运营商的客户流失预测。张利利等<sup>[10]</sup>使用决策树方法进行航空客户流失预测,并通过 K-mans 算法进行客户价值衡量。

20 世纪 50 年代中期,基于客户实际需求不一、资源效益最大化需求,温德尔<sup>[11]</sup>最早提出客户细分的概念,指企业在特定市场和业务模式下,根据属性、行为、需求、偏好和价值等因素对客户进行分类。目前主要从市场需求、企业运营的相关条件和客户自身综合属性等几方面进行分类<sup>[12]</sup>。

综合考虑已有的客户细分标准,为了更精准、科学、立体地刻画客户行为动态,更好地进行客户细分和流失预测,本研究从外在、内在以及行为(external, intrinsic, behavior, EIB)三方面属性出发构建寿险客户指标体系。另外,考虑到传统的 K-means 算法<sup>[13]</sup>在处理大数据集时,虽有较好的可伸缩性,但也存在初始聚类数不确定以及对离群点敏感的缺陷,提出改进的 K-means 算法,将改进的轮廓系数公式作为选取初始聚类数目的依据,并综合考虑欧式距离相似度的距离测度优势与余弦相似度的方向测度优势,在聚类迭代中进行局部、全局离群点的过滤,尽可能降低可视化噪声、减小簇内误方差。使用改进后的 K-means 算法划分出不同流失风险的客户群,针对不同群体制定个性化挽留思路。吸取 BP 神经网络算法非线性拟合能力强与 Ada-boost 算法可有效提升模型泛化能力的优点,提出使用融合的 BP-Adaboost 算法构建寿险客户流失强预测器,并综合比较细分前后、K-means 算法改进前后以及单个预测器与融合预测器的效果。

1 EIB 属性与寿险客户指标体系确定

由于客户细分的标准尚未统一,本研究根据寿险行业的特点,提出客户 EIB 属性(如表 1),并以此为依据进行寿险客户指标体系设计,以便高效地进行客户细分与流失预测建模。

表 1 客户的 EIB 属性  
Tab. 1 EIB attributes of customers

属性	定义	举例	作用
E	客户身处的社会环境和社会关系	客户的地理位置、职业归属(国家行政机关、事业单位、企业、其他)	较宏观,体现直接价值
I	客户的个人基本信息	性别、年龄、家庭成员数、信用度等	较微观,体现潜在价值(忠诚与否、流失风险高低)
B	客户在购买公司的产品或服务时所产生的消费信息	缴费数量、所购险种、缴费方式、缴费次数等	较微观,体现潜在价值(忠诚与否、流失风险高低)

我国现行《保险法》第五十三条规定“投保人对本人、近亲属以及其他同意与投保人订立合同的被保险人均有保险利益”,这说明投保人与被保险人之间存在一对多的现象,且在实际情况中,这种一对多的现象,造成了客户关系的复杂性。因此需要根据“客户关系-投保人-被保险人”的对应关系,依据 EIB 属性,综合考察客户自身价值观念、生活水平以及客户和寿险公司的业务交易信息等,建立寿险客户指标体系,如表 2 所示。

表2 基于EIB属性的寿险客户指标体系

Tab. 2 Index system of life insurance customers based on EIB attribute

客户属性	客户指标	主要对象	描述
I	性别	被保险人	男=1,女=0
I	年龄	被保险人	联合国世界卫生组织提出新的年龄分段:44岁以下青年人=1,45~59岁中年人=2,60~74岁年轻老年人=3,75~89岁老年人=4,90岁以上长寿老人=5
E	职业危险级别	被保险人	纯文职人员=1,从事少量体力劳动的非纯文职人员=2,内陆渔业养殖工人、水产品加工人员等=3,农牧业人员、土木工程建筑业工人等=4,野生动物保护人员、拖拉机驾驶人员、修路工等=5,装运工人、造船业工人、高速公路工程人员等=6,高空作业人员、高压电工作人员等=7
E	家庭收入等级	投保人	低=1,较低=2,中等=3,较高=4,高=5
E	学历级别	投保人	初中及以下=1,中专或高中=2,大学专科或本科=3,研究生=4
E	婚姻状况	投保人	已婚=1,未婚=0
I	购买主导动机	投保人	碍于面子=1,获利心理=2,实际需要=3
I	信用评级	投保人	低=1,较低=2,中等=3,较高=4,高=5
B	缴费数量	投保人	所选时间段内投保人为被保险人实际缴纳的保费金额数量
B	所购险种	投保人	投保人为被保险人购买险种的种类标签
B	缴费方式	投保人	趸缴=1,年缴=2,半年缴=3,季缴=4,月缴=5
B	缴费次数	投保人	所选时间段内投保人为被保险人实际缴纳的保费次数

## 2 K-means 算法及其改进

传统 K-means 算法主要基于欧式距离测度以及最小化平方误差和准则,其步骤如下:

1) 类中心初始化。给定样本集  $D = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_m\}$ , 在  $N$  个样本中随机选择  $k$  个样本点, 作为  $k$  个初始聚类中心。 $D$  中的元素表示每个样本在欧几里得空间  $\mathbf{R}^n$  中对应的  $n$  维映射向量,  $n$  表示用于聚类的指标个数。

2) 类划分。将  $N$  个样本按照与  $k$  个聚类中心的欧式距离远近, 分别分配给距离最近的聚类中心, 形成  $k$  个簇  $C = (C_1, C_2, \dots, C_k)$ 。

3) 类中心点求解。计算  $k$  个簇中心点的平均值作为新的聚类中心。

4) 收敛判断。每次迭代中, 由  $k$  个样本点组成的质心向量  $\vec{\mu}_1, \vec{\mu}_2, \dots, \vec{\mu}_k, \vec{\mu}_i (1 \leq i \leq k)$  见式(1), 用误差平方和准则函数(式(2))判断收敛性。

$$\vec{\mu}_i = \frac{1}{|C_i|} \sum_{\vec{x} \in C_i} \vec{x}, \quad (1)$$

$$E = \sum_{i=1}^k \sum_{\vec{x} \in C_i} \|\vec{x} - \vec{\mu}_i\|_2^2. \quad (2)$$

5) 重复步骤 2) 和 3), 直到每个类的个体不再变化, 得到所有类别的最终聚类中心及其包括的个体。

本研究主要从选取初始聚类簇数和改进迭代规则两个方面, 对 K-Means 算法进行改进。

### 1) 初始聚类簇数选取的改进

传统 K-means 算法通常按照初始聚类中心来设置初始聚类簇数  $k$ , 继而进行类划分和中心点求解的迭代, 因此初始值的选取十分重要。若选取不当, 会使得聚类效果较差。Peter 在 1986 年提出轮廓系数

$$S(i) = \frac{p(i) - q(i)}{\max\{q(i), p(i)\}} \quad (3)$$

来评价聚类效果的好坏<sup>[14]</sup>。其中:  $q(i)$  表示点  $i$  到所属类中其他点的平均距离, 主要反映内聚度;  $p(i)$  表示点  $i$  到非所属类中所有点平均距离的最小值, 主要反映分离度。轮廓系数结合了内聚度、分离度两种因素, 通常数值越大, 聚类效果越好。针对传统的轮廓系数未考虑对内聚度有潜在影响的类内最小距离和对分散度有潜在影响的类间平均距离最大值的问题, 引入点  $i$  到所属类中其他点的最小距离  $s(i)$  和点  $i$  到非所属类中所有点平均距离的最大值  $r(i)$ , 提出改进后的轮廓系数公式:

$$\tilde{S}(i) = \frac{\max\{r(i) - q(i), p(i) - s(i)\}}{\max\{q(i), p(i)\}}。 \quad (4)$$

式(4)反映了各因素之间更全面的制约关系。进而得到  $N$  个样本点轮廓系数的平均值

$$\tilde{S} = \frac{1}{N} \sum_{i=1}^N \frac{\max\{r(i) - q(i), p(i) - s(i)\}}{\max\{q(i), p(i)\}}。 \quad (5)$$

## 2) 迭代规则的改进

传统的 K-means 算法在迭代过程中未考虑全局、局部离群点对平均值计算的影响。当离群点被分配到某簇中, 可能会严重影响该簇类的均值, 从而使聚类中心有较大误差, 影响最终聚类结果。以往对于 K-means 算法的离群点监测方法常常基于邻近度或密度<sup>[15]</sup>, 但这两种方法难以处理大数据集, 且对参数选择高度敏感。因此, 本研究提出一种基于相似度的离群点监测方法, 根据改进的相似度公式设置迭代中的离群点过滤规则。

传统的欧氏距离相似度  $\frac{1}{1 + d(\vec{\alpha}, \vec{\beta})}$  只能从距离上测度向量之间的相似性, 而文本计算中常用的余弦相似度<sup>[16]</sup>

$$\cos(\vec{\alpha}, \vec{\beta}) = \frac{\vec{\alpha} \cdot \vec{\beta}}{\|\vec{\alpha}\| \|\vec{\beta}\|}。 \quad (6)$$

可以较好地反映变量之间的方向相关性。当向量  $\vec{\alpha}$  与向量  $\vec{\beta}$  对应维数相同且均为  $n$  维映射向量时, 可表示为  $\vec{\alpha}(a_1, a_2, \dots, a_n)$  和  $\vec{\beta}(b_1, b_2, \dots, b_n)$ 。其欧氏距离相似度与余弦相似度可分别改写为式(7)、式(8)。

$$sim_1(\vec{\alpha}, \vec{\beta}) = \frac{1}{1 + d(\vec{\alpha}, \vec{\beta})} = \frac{1}{1 + \|\vec{\alpha} - \vec{\beta}\|} = \frac{1}{1 + \sum_{i=1}^n (a_i - b_i)^2}, \quad (7)$$

$$sim_2(\vec{\alpha}, \vec{\beta}) = \cos(\vec{\alpha}, \vec{\beta}) = \frac{\vec{\alpha} \cdot \vec{\beta}}{\|\vec{\alpha}\| \|\vec{\beta}\|} = \frac{(a_1, a_2, \dots, a_n) \cdot (b_1, b_2, \dots, b_n)}{\sqrt{\sum_{i=1}^n (a_i)^2} \times \sqrt{\sum_{i=1}^n (b_i)^2}} \geq \frac{2 \sum_{i=1}^n a_i b_i}{\sum_{i=1}^n (a_i^2 + b_i^2)}。 \quad (8)$$

鉴于两种测度方法优势互补, 提出改进的相似度计算公式:

$$sim(\vec{\alpha}, \vec{\beta}) = sim_1(\vec{\alpha}, \vec{\beta}) \cdot sim_2(\vec{\alpha}, \vec{\beta}) \leq \frac{2 \sum_{i=1}^n a_i b_i}{\sum_{i=1}^n (a_i^2 + b_i^2) + \sum_{i=1}^n (a_i^2 - b_i^2) \sum_{i=1}^n (a_i^2 + b_i^2)}。 \quad (9)$$

由式(9)可见, 改进的相似度综合考虑了欧式距离相似度、余弦相似度, 且存在上限。参与聚类迭代的向量与当前簇中心向量的相似度越小, 说明其越偏离当前簇类。当低于某个阈值  $P_1$  时, 可将其对应的欧式空间样本点视为局部离群点并进行过滤; 与所有簇中心的均值向量的相似度越小, 说明其越偏离整体, 当低于某个阈值  $P_2$  时, 可将其对应的欧式空间样本点视为全局离群点并进行过滤。 $P_1$  和  $P_2$  为离群点监测的阈值参数, 在实际中, 可通过多次实验, 选取最合适的参数值。具体过滤规则如下:

在本节的 K-means 算法步骤 2) 类划分迭代中, 若向量  $\vec{x}$  与当前参与分配的簇中心向量  $\vec{\mu}_i$ 、所有簇中心

的均值向量  $\frac{1}{k} \sum_{i=1}^k \vec{\mu}_i$  分别满足相似度关系  $\text{sim}(\vec{x}, \vec{\mu}_i) < P_1$  且  $\text{sim}(\vec{x}, \frac{1}{k} \sum_{i=1}^k \vec{\mu}_i) < P_2$ , 则视为全局离群向量, 过滤其对应的样本点, 永不参与其他簇类的迭代; 满足  $\text{sim}(\vec{x}, \vec{\mu}_i) < P_1$  且  $\text{sim}(\vec{x}, \frac{1}{k} \sum_{i=1}^k \vec{\mu}_i) \geq P_2$  时, 视为局部离群向量, 将其对应样本点从当前簇类中过滤, 但继续参与其他簇类的迭代; 满足  $\text{sim}(\vec{x}, \vec{\mu}_i) \geq P_1$  且  $\text{sim}(\vec{x}, \frac{1}{k} \sum_{i=1}^k \vec{\mu}_i) \geq P_2$  时, 不进行过滤。

### 3 组合后的 BP-Adaboost 算法

BP 神经网络<sup>[17]</sup>有较强的非线性拟合能力, 理论上能够拟合任意非线性函数, 但存在收敛速度慢、泛化能力弱等缺点。而 Adaboost 算法<sup>[18]</sup>能够在迭代中降低误差, 提高模型的泛化能力。本研究将两者结合, 得到 BP-Adaboost 算法<sup>[19]</sup>来降低原始 BP 算法的预测误差, 其详细步骤如下。

1) 选择数据并进行网络初始化。随机抽取  $m$  组训练数据  $\{x_1, x_2, \dots, x_m\}$ , 初始化权重

$$w_i^t = \frac{1}{m}, i = 1, 2, \dots, m. \quad (10)$$

2) 将训练数据用 BP 神经网络弱预测器进行预测。当训练到第  $t$  个弱预测器时, 获得弱预测序列  $f_t$  的预测误差和

$$\epsilon_t = \sum_{i=1}^m w_i^t, f_t \neq y. \quad (11)$$

其中  $y$  为期望输出。

3) 计算预测序列的权重。依据  $\epsilon_t$  计算弱预测器的权重

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right). \quad (12)$$

4) 调整测试数据的权重。依据预测序列的权重  $\alpha_t$  调整新训练的样本权重

$$w_i^{t+1} = \frac{w_i^t}{Z_t} \times \exp[-\alpha_t y_i f_t(x_i)], i = 1, 2, \dots, m. \quad (13)$$

其中,  $Z_t$  称作归一化因子, 主要作用是当权重比例不变时, 使其分布之和等于 1。

5) 输出强预测器函数。迭代  $T$  次后, 得到  $T$  组弱预测器函数  $g(f_t, \alpha_t)$  合成的强预测器函数

$$h(x) = \frac{\alpha_t}{\sum_{t=1}^T \alpha_t} g(f_t, \alpha_t). \quad (14)$$

## 4 实证研究

实验数据来源于某国内保险公司网站(<http://www.chinalife.com.cn/>)2018 年 1 月 1 日—2019 年 12 月 31 日的寿险客户调查公报及其交易信息, 实验软件为 MATLAB R2014a。

### 4.1 基于改进 K-means 算法的寿险客户细分

依据客户的 EIB 指标体系提取数据信息, 归一化处理后, 将客户调查公报中各指标出现的频次与全部指标出现的频次之比作为重要度权值, 对指标进行加权量化处理, 最终得到 2 000 条寿险客户样本, 部分数据如表 3 所示。

1) 轮廓系数改进前后的实验结果对比

为了获得最佳初始聚类簇数目, 选取不同的  $k$  值, 对归一化处理后的样本进行 K-means 聚类, 并统计改进前后的轮廓系数均值, 如图 1 所示。

表 3 部分处理后的寿险客户样本  
Tab. 3 Part of life insurance customer samples after processing

性别	年龄	职业危险级别	家庭收入等级	学历级别	婚姻状况	信用评级	缴费数量	所购险种	购买主导动机	缴费方式	缴费次数
0.000	0.497	0.029	0.500	0.003	0.237	0.253	0.751	0.501	0.003	0.003	0.927
0.200	0.214	0.921	0.400	0.201	0.000	0.602	0.208	0.212	0.214	0.208	0.893
0.000	0.129	0.267	0.333	0.197	0.000	0.339	0.976	0.121	0.065	0.095	0.382
0.000	0.389	0.201	0.600	0.809	0.237	0.607	0.212	0.214	0.192	0.176	0.277
0.000	0.183	0.455	0.364	0.182	0.000	0.364	0.928	0.091	0.091	0.091	0.364
0.200	0.747	0.751	0.750	0.250	0.237	0.252	0.508	0.007	0.007	0.009	0.196
0.000	0.143	0.285	0.286	0.143	0.000	0.286	0.927	0.005	0.005	0.010	0.571
0.000	0.253	0.257	0.750	0.501	0.000	0.983	0.250	0.221	0.016	0.277	0.503
0.000	0.101	0.408	0.300	0.209	0.000	0.217	0.911	0.181	0.100	0.012	0.419
0.200	0.204	0.211	0.968	0.604	0.237	0.009	0.412	0.159	0.318	0.231	0.802

由图 1 可见,在改进后的轮廓系数均值中,不同初始聚类簇数  $k$  下的系数变化幅度较改进前明显增大,表明改进后的轮廓系数均值能更全面地衡量聚类的内聚度和分离度,对于筛选合适的聚类数目更具区分度。在两种轮廓系数中,对应最大系数的  $k$  值均为 3,故选取  $k = 3$  作为初始聚类簇数。

取  $k$  值分别为 3 和 4 进行轮廓系数分布的可视化展示,如图 2 所示。

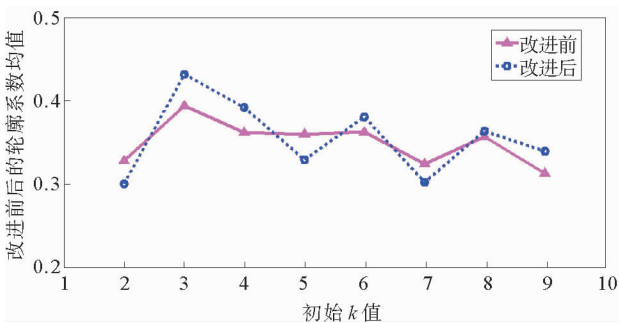


图 1 改进前后的轮廓系数均值对比

Fig. 1 Comparison of mean contour coefficients before and after improvement

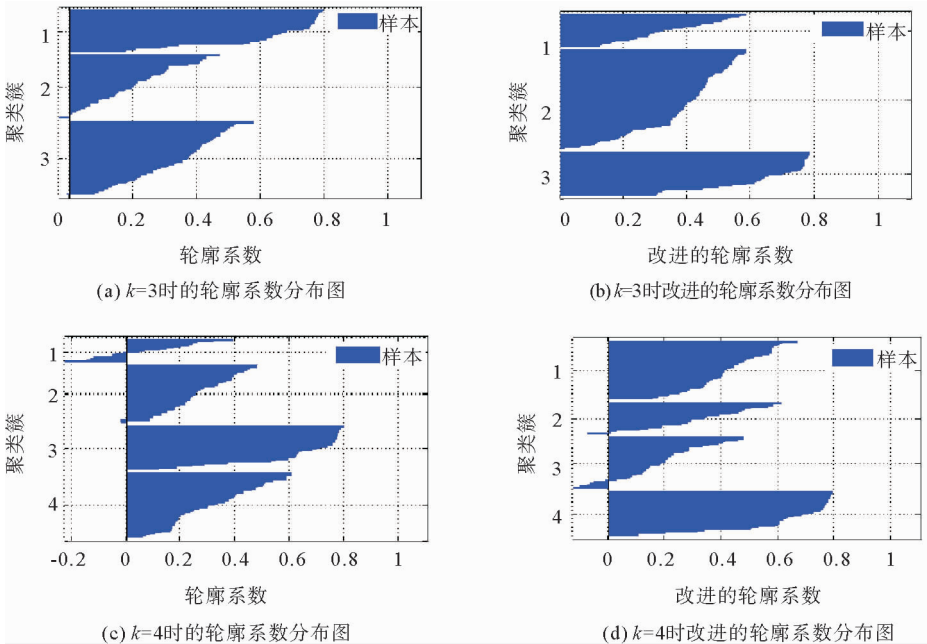


图 2 改进前后的轮廓系数分布图对比

Fig. 2 Comparison of contour coefficient distributions before and after improvement

由图2可知,  $k=3$  时轮廓系数为负的样本点更少, 且总体轮廓系数更大, 进一步这说明  $k=3$  作为初始聚类簇数的优越性; 与改进前相比, 改进后的轮廓系数为负的样本点明显减少 ( $k=3$  时几乎为0), 且总体轮廓系数明显增大。

## 2) 改进迭代规则的 K-means 算法结果分析

在尽可能减少分组的基础上, 为保证直观性, 在三维空间中进行样本点的可视化展示。可视化结果共有  $C_{12}^3$  个。由于篇幅限制, 图3仅展示4个三维空间{性别, 年龄, 职业危险级别}、{家庭收入等级, 学历级别, 婚姻状况}、{信用评级, 缴费数量, 所购险种}、{购买主导动机, 缴费方式, 缴费次数}的可视化结果, 反映3类客户群(I、II、III)的分布。

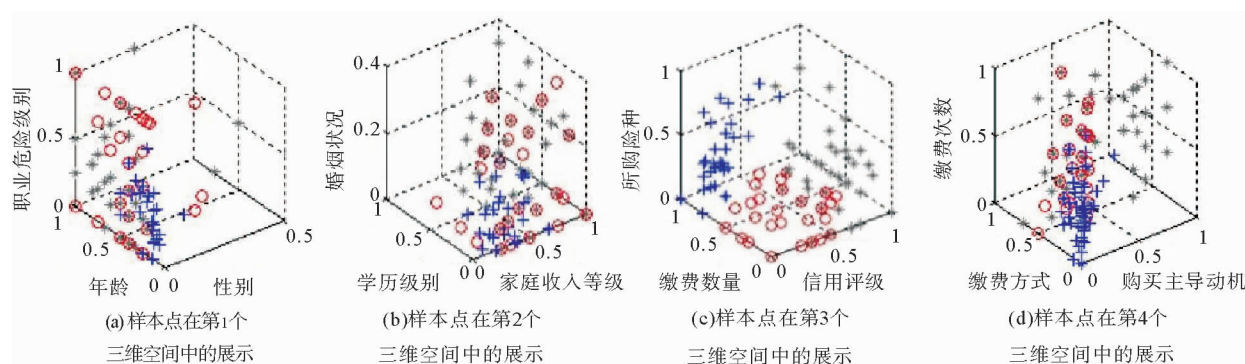


图3 K-means 算法的聚类结果可视化(+客户群 I, O 客户群 II, \* 客户群 III)

Fig. 3 Visualization of clustering results of K-means algorithm(+Customer base I, O Customer base II, \* Customer base III)

使用改进迭代规则的 K-means 算法进行聚类, 在实验过程中, 固定阈值  $P_2 = 0.005$ , 以 0.03 为起点、0.03 为步长将  $P_1$  逐步增加到 0.18。将三维指标集{信用评级, 缴费数量, 所购险种}用于改进 K-means 算法的可视化展示, 如图4所示。

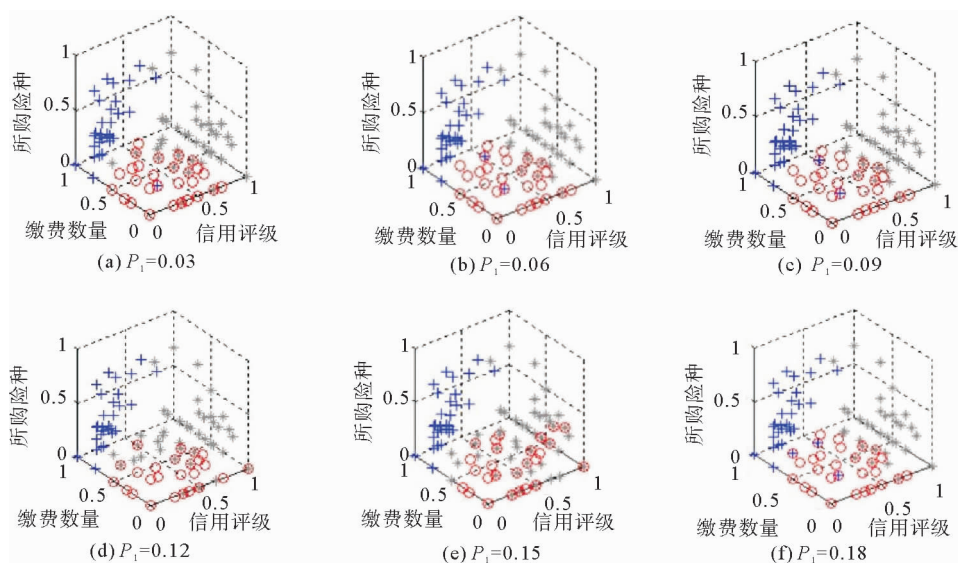


图4 不同阈值下的改进 K-Means 算法聚类结果可视化(+客户群 I, O 客户群 II, \* 客户群 III)

Fig. 4 Visualization of clustering results of improved K-means algorithm under different thresholds  
(+Customer base I, O Customer base II, \* Customer base III)



由图 4 可知,阈值  $P_1 \leq 0.09$ ,尤其是  $P_1 = 0.06$  时,可视化噪声明显较改进前的图 2 有所减小;但  $P_1 > 0.09$ ,尤其是  $P_1 = 0.15$  时,可视化噪声较大。

使用最终的簇内误方差(sum of the squared error,SSE)作为评价改进的 K-means 算法聚类结果好坏的指标,统计阈值  $P_1$  在 0.03~0.18 范围内的最终簇内误方差,结果如图 5 所示。由图 5 可见,当阈值  $P_1 = 0.06$  时获得较低的簇内误方差,而  $P_1 = 0.12$ 、0.15 时的簇内误方差较大,甚至与改进前持平。

这说明阈值  $P_1 \leq 0.09$  时,改进的 K-means 算法能合理过滤局部和全局离群点,有效提升聚类效果。分别将  $P_1$  为 0.03、0.06、0.09 时的最终聚类中心以及对应的细分客户群体进行汇总,并与改进前的结果作比较,如表 4 所示。

由表 4 可知,在不同阈值下的改进 K-means 算法中,最终聚类中心、对应客户数量在不同客户类别中的差距较改进前均有明显增大,其中最终聚类中心的变化主要表现在指标集{性别,年龄,职业危险级别,学历级别,婚姻状况,信用评级,缴费数量,所购险种,购买主导动机}中,这主要体现了局部离群点过滤的作用;改进 K-means 算法后的客户数量总和均不足 2 000,体现了全局离群点过滤的作用。

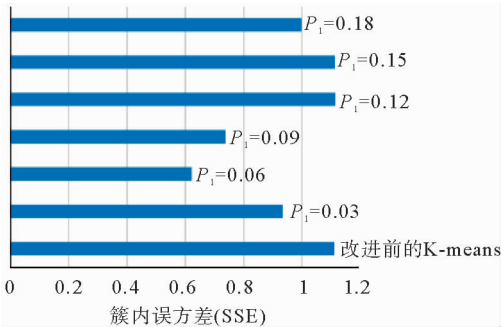


图 5 不同阈值下改进 K-means 算法的 SSE

Fig. 5 SSE of improved K-means algorithm under different thresholds

表 4 改进 K-means 算法前后的最终聚类结果对比

Tab. 4 Comparison of final clustering results before and after the improvement of K-means algorithm

算法及相关 参数设置	最终聚类中心												群体 编号	客户 数量	客户 总数
	性别	年龄	职业危 险级别	家庭收 入等级	学历 级别	婚姻 状况	信用 评级	缴费 数量	所购 险种	购买主 导动机	缴费 方式	缴费 次数			
K-means 算法	0.004	0.241	0.307	0.586	0.198	0.053	0.334	0.913	0.255	0.077	0.069	0.559	I	431	2 000
	0.008	0.410	0.315	0.523	0.292	0.085	0.439	0.278	0.161	0.092	0.115	0.901	II	556	
	0.009	0.432	0.332	0.684	0.413	0.141	0.618	0.384	0.193	0.193	0.178	0.944	III	1 013	
改进 K-means 算法	0.004	0.242	0.302	0.506	0.198	0.053	0.334	0.980	0.274	0.077	0.071	0.558	I	380	1 993
	$P_1=0.03$	0.008	0.420	0.318	0.550	0.277	0.068	0.453	0.188	0.102	0.088	0.087	0.944	II	507
		0.009	0.431	0.345	0.597	0.473	0.178	0.763	0.421	0.233	0.236	0.226	0.925	III	1 106
	0.004	0.242	0.309	0.523	0.198	0.053	0.334	0.980	0.274	0.077	0.071	0.558	I	347	1 992
	$P_1=0.06$	0.008	0.434	0.323	0.580	0.305	0.087	0.448	0.203	0.114	0.112	0.105	0.920	II	502
		0.014	0.447	0.361	0.603	0.461	0.167	0.803	0.417	0.231	0.223	0.219	0.948	III	1 143
	0.004	0.218	0.312	0.576	0.178	0.037	0.305	0.958	0.289	0.063	0.053	0.527	I	365	1 994
	$P_1=0.09$	0.008	0.422	0.329	0.529	0.342	0.119	0.470	0.217	0.133	0.131	0.135	0.977	II	511
		0.025	0.457	0.352	0.605	0.567	0.177	0.697	0.494	0.191	0.337	0.293	0.937	III	1 118

考虑到“信用评级”指标在聚类可视化结果中展示出良好的区分度且与客户消费行为密切关联,故将其用于客户相对流失风险识别。“信用评级”的高低与流失风险水平呈负相关,因此得到不同风险客户细分{I=“高流失风险客户群”,II=“中流失风险客户群”,III=“低流失风险客户群”},其所含客户数量按高、中、低流失风险客户群依次减少。

低流失风险客户群所含客户数量最多,对应聚类中心的年龄最大、职业危险级别最高、学历最高、婚姻状



况倾向于“已婚”、购买主导动机倾向于“实际需要”、缴费数量适中,反映了该群体对保险的需求心理、理性思维方式和一定的经济实力。这类客户在购买寿险产品时,多考虑自身或家庭成员的需要,因此不易流失,能给公司带来长期的稳定利润。

高流失风险客户群所含客户数量最少,对应聚类中心的年龄最小、职业危险级别最低、学历最低、婚姻状况倾向于“未婚”、购买主导动机倾向于“碍于面子”、缴费数量相对较高。该类客户在购买保险产品时,缺乏理性购买动机,容易跟风购买一些价格相对较高的寿险产品。虽然该群体有一定购买力,但存在较大的流失风险。

中流失风险客户群所含客户数量适中,对应聚类中心的年龄适中、职业危险级别适中、学历适中、缴费数量相对较低。这一类客户对寿险产品有一定的需求,购买主导动机多样化,能给公司带来为数不多但较为稳定的利润,流失的风险性介于上述两种群体之间。

#### 4.2 基于客户细分和 BP-Adaboost 算法的寿险客户流失预测

为进一步证明改进 K-means 算法的优越性并展开寿险客户流失预测研究,统计改进前后 K-means 算法的客户细分结果,分别运用 BP 弱预测器、BP-Adaboost 强预测器对不同客户群体进行流失预测建模,并综合比较其预测误差。

##### 1) 基于二分类的寿险客户状态观测

寿险客户在观测期内的状态有两种,用二分类集合{流失,未流失}来表示。本研究从反映客户与公司业务往来的属性 B 中选取合适的规则,作为判断客户流失与否的标志。如表 2 所示,属性 B 对应的 4 个指标中,缴费数量、缴费方式和缴费次数存在数值关系。令二分类变量为  $Y$ ,设置客户状态的观测方法如下:

对于一次性缴清所有保费的趸缴客户,其状态容易观测。将含有“退保”和“犹豫期退保”字样信息的客户识别为流失客户( $Y=1$ ),其余识别为未流失客户( $Y=0$ )。

对于期缴客户,投保人为每位被保险人缴费的数量即为所购险种的每期保费之和。虽然目前大部分险种均采用均衡保费方法(即每期缴费数目相同),但部分险种的保费随时间变化而增加,不便于观测客户状态,因此根据“缴费次数”、“缴费方式”来综合识别期缴客户的状态。令观测期为  $H$  年,缴费方式对应的每次缴费间隔时间为  $G$  年(通常  $G \leq 1$ ),客户实缴保费次数用  $f$  表示,当满足实缴保费次数不足应缴次数,即  $f < \frac{H}{G}$  时,说明该客户在此期间已流失,将其识别为流失客户( $Y=1$ );反之,将其识别为未流失客户( $Y=0$ )。

##### 2) BP 算法与 BP-Adaboost 算法实验结果对比

将 BP 神经网络设置为 3 层:输入层为{性别,年龄,职业危险级别,家庭收入等级,学历级别,婚姻状况,购买主导动机,信用评级,缴费数量,所购险种};输出层为客户状态集  $Y=\{0,1\}$ ;隐藏层神经元数量的设置采用试凑法,即首先选取较少隐含层神经元训练 BP 网络,观测预测精度或误差,随后增加隐含层神经元数量,直到预测精度不再增加为止,最终确定网络各层神经元数量依次为 10、5、1。

根据预测结果调整样本权重,把预测误差大于 0.1 的测试样本作为应该加强学习的样本训练 BP 神经网络弱预测器,最终获得由 10 组弱预测器生成的 BP-Adaboost 强预测器。以  $P_1=0.06$  时改进 K-means 算法细分的低流失风险客户群为例,在 1 143 条样本中,随机选择 943 条作为训练样本、200 条作为测试样本进行实验,实验数据的误差均方曲线见图 6。

由图 6 可见,误差均方曲线逐渐收敛,在第 17 步达到最好的测试效果 0.065 281,之后逐渐趋向于平缓,误差值几乎不变化,效果较好。

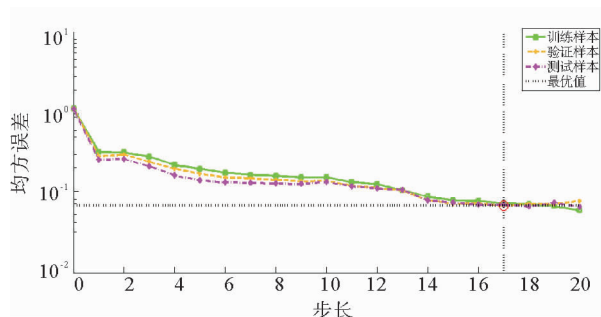


图 6 误差均方曲线

Fig. 6 Curve of mean squared errors

图 7 为 10 组 BP 神经网络弱预测器的平均误差绝对值和对应 BP-Adaboost 强预测器的误差绝对值。可以看出,在细分客户样本的预测误差值中,除极个别样本的强预测器预测误差高于弱预测器以外,总体上,用 Adaboost 调整后得到的强预测器预测的误差绝对值要普遍小于弱预测器。在 200 个预测样本中,传统 BP 网络算法的测试误差绝对值区间为  $[0, 0.2]$ ,样本点的误差绝对值有不少超出 0.1; BP-Adaboost 算法的测试误差的绝对值区间绝大多数都在  $[0, 0.1]$  之间,样本点的误差绝对值几乎都接近 0。模型的拟合效果显示,强预测器预测的训练集  $R=0.952\ 97$ 、验证集  $R=0.940\ 35$ 、测试集  $R=0.961\ 06$ 、总体  $R=0.952\ 51$ ,说明模型的拟合结果较好。

### 3) 全部实验结果对比

对于每次实验,将预测误差绝对值超过 0.2 的样本点剔除,计算剩余样本点的预测误差平均值。汇总全部实验结果如图 8 所示。

由图 8 可见:BP-Adaboost 算法的预测误差较传统的 BP 算法小,说明 Adaboost 在迭代中对 BP 算法进行了有效提升;细分前客户样本的预测误差要明显大于细分后,说明客户细分对于提高客户流失预测的精度有一定作用;改进的 K-means 算法细分的客户群与传统的 K-means 算法相比,在后续的流失预测中,预测误差几乎全部变小,进一步证明改进的 K-means 算法实现的客户细分结果更为精准,且对后续客户流失预测的精度提升有明显作用。

### 4.3 对寿险公司的建议

寿险公司在实际的营销过程中,客户细分对于客户流失预测有重要意义。客户的挽留管理有助于公司经济效益的提升。公司要充分利用已有客户信息,挖掘并掌握不同客户群体的特征,采取不同的措施对不同的客户群体制定个性化服务。以本文的实验结果为例,对不同流失风险的客户群提出建议如下。

1) 低流失风险客户群。这类客户的年龄相对较大、职业危险性相对较高,在购买保险产品时倾向于理性和满足实际需求,且有充足的资金支持续保。这类客户是当今寿险市场的主流客户,且客户数量庞大,能为公司带来长期稳定的利润。公司应当对这类群体给予高度重视,并根据每一位客户在时间、空间上的需求变化,尽可能地为其量身定制更适合的寿险服务,使这类客户更加忠诚地续保。

2) 高流失风险客户群。这类客户的年龄相对较小、职业危险性相对较低,在购买保险产品时缺乏理性考虑,容易受保险推销员或周围朋友的影响购买一些用处不大却价格昂贵的寿险产品,给公司带来的利润虽多,但较不稳定。这类客户有一定购买力,但对寿险产品的热衷程度还不够。公司可以举办一些形式丰富的活动,来提高其对寿险产品的购买欲望,培养其与公司的感情。例如:定期对客户进行回访询问,节假日举办一些促销活动,以抽签方式赠送小礼品,等等。通过公司服务水平的提升,客户的忠诚度、满意度也会随之上升,流失风险随之降低。

3) 中流失风险客户群。这类客户的年龄、职业危险级别、学历处于中等水平,对寿险产品有一定的需求,缴费数量较低但相对稳定。作为寿险公司的营销对象,有一定的发展潜力。因此,可以综合高、低流失客户

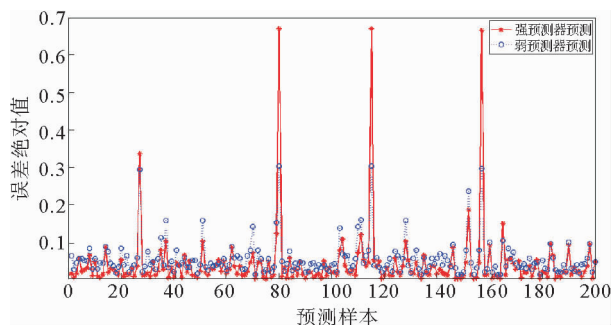


图 7 预测误差的绝对值

Fig. 7 Absolute values of prediction error

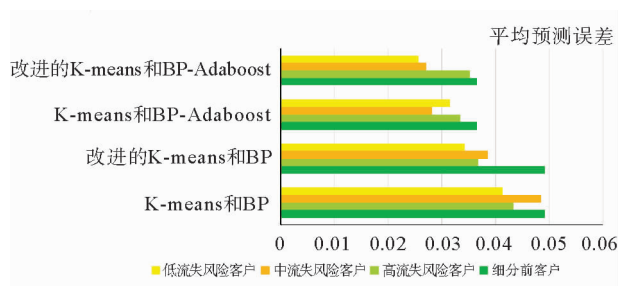


图 8 改进 K-means 算法前后的平均预测误差对比

Fig. 8 Comparison of average prediction errors before and after the improvement of K-means algorithm

群体的措施进行客户挽留管理。公司在为其进行节假日促销活动的同时,还可以挑选一些幸运客户,同低风险群体共同参与量身定制产品活动,或者开展价格相对高的寿险产品的首单优惠活动,激发此类客户对该类产品的购买欲望,提高公司的盈利水平。

## 5 结束语

针对寿险行业的客户流失问题,构建了基于 EIB 属性的寿险客户指标体系。在 K-means 算法的改进中,使用改进后的轮廓系数确定初始聚类中心,并综合欧式距离相似度和余弦相似度的测度优势,在类划分中进行局部、全局离群点的过滤。使用改进前后的 K-means 算法分别进行客户细分,利用 BP 算法、BP-Adaboost 算法对细分后的客户建立流失预测模型。算例实证结果表明改进后 K-means 算法的簇内误差方差变小,最终聚类中心和客户数量在不同类别中的差距增大、可视化噪声降低,且基于改进 K-means 算法客户细分的流失预测误差较改进前有明显降低。本算法不仅为寿险公司的客户流失风险预警及挽留管理提供参考,也为寿险及相关行业的客户流失预测研究给供借鉴。本研究从“客户流失风险”角度出发,在特定的时间、空间范围内开展客户细分和流失预测建模,可以视为对客户画像的局部研究。未来可综合考虑客户各项指标在时间、空间上的变化,开展更全面的研究。

### 参考文献:

- [1]任红娟,夏国恩.客户流失研究综述[J].中国商论,2018(32):166-167.  
REN Hongjuan,XIA Guoen.A summary of customer churn studies [J].China Journal of Commerce,2018(32):166-167.
- [2]KISIOGLU P, TOPCU Y I. Applying Bayesian belief network approach to customer churn analysis: A case study on the telecom industry of Turkey[J]. Expert Systems with Applications, 2011, 38(6): 7151-7157.
- [3]周晓玮. BP 神经网络技术在寿险营销系统中的应用研究[D]. 上海: 上海交通大学, 2014: 23-31.  
ZHOU Xiaowei. Research and application of BP neural network technology in life insurance marketing system[D]. Shanghai: Shanghai Jiao Tong University, 2014: 23-31.
- [4]BI T, LIU Y, LI P, et al. Telecom customer churn prediction method based on cluster stratified sampling logistic regression [C]// International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things. Hsinchu, China, Dec. 4-6, 2014: 282-287.
- [5]GÜNTHER C C, AAS K, Tvete I F, et al. Modelling and predicting customer churn from an insurance company [J]. Scandinavian Actuarial Journal, 2014, 1-14(1): 58-71.
- [6]梁锋. 数据挖掘技术在寿险客户流失中的应用[J]. 电子科学技术, 2015, 2(1): 104-107.  
LIANG Feng. DM techniques application in the loss of life insurance customers[J]. Electronic Science & Technology, 2015, 2(1): 104-107.
- [7]郑宇晨, 吕王勇. 基于 logistic 模型的证券公司客户流失预警分析[J]. 郑州航空工业管理学院学报, 2016, 34(5): 80-88.  
ZHENG Yuchen, LÜ Wangyong. Customer churn warning analysis on securities companies based on logistic model[J]. Journal of Zhengzhou University of Aeronautics, 2016, 34(5): 80-88.
- [8]AMIN A, ANWAR S, Adnan A. Customer churn prediction telecommunication sector using a rough set approach[J]. Neurocomputing, 2017, 237(10): 242-254.
- [9]冯鑫, 王晨, 刘苑, 等. 基于评论情感倾向和神经网络的客户流失预测研究[J]. 中国电子科学研究院学报, 2018, 13(3): 340-345.  
FENG Xin, WANG Chen, LIU Yuan, et al. The customer churn prediction based on emotional polarity and BPNN[J]. Journal of China Academy of Electronics and Information Technology, 2018, 13(3): 340-345.
- [10]张利利, 马艳琴. 基于数据挖掘技术的航空客户流失与细分研究及 R 语言程序实现[J]. 数学的实践与认识, 2019, 49(6): 134-142.  
ZHANG Lili, MA Yanqin. Analysis of the airline customer churn and customer segmentation based on data mining algorithm using R[J]. Mathematics in Practice and Theory, 2019, 49(6): 134-142.
- [11]束晓君. 基于数据挖掘的保险公司精准营销研究[D]. 西安: 西安工业大学, 2014: 13-14.  
SHU Xiaojun. Research on insurance precision marketing of insurance company based on data mining[D]. Xi'an: Xi'an Technological University, 2014: 13-14.

- [12] 刘洋. 基于聚类分析的电子商务企业客户细分研究[D]. 重庆: 重庆大学, 2017: 3-4.  
LIU Yang. Research on customer segmentation of ecommerce enterprises based on cluster analysis [D]. Chongqing: Chongqing University, 2017: 3-4.
- [13] 马廷博, 刘太安, 徐建国, 等. 基于改进的 K-means 聚类算法的汽车市场竞争情报分析[J]. 山东科技大学学报(自然科学版), 2019, 38(1): 74-84.  
MA Tingbo, LIU Taian, XU Jianguo, et al. Information analysis of auto market competition based on improved K-means clustering algorithm[J]. Journal of Shandong University of Science and Technology (Natural Science), 2019, 38(1): 74-84.
- [14] 李亚, 刘丽平, 李柏青, 等. 基于改进 K-means 聚类 and BP 神经网络的台区线损率计算方法[J]. 中国电机工程学报, 2016, 36(17): 4543-4552.  
LI Ya, LIU Liping, LI Baiqing, et al. Calculation of line loss rate in transformer district based on improved K-means clustering algorithm and BP neural network[J]. Proceedings of the CSEE, 2016, 36(17): 4543-4552.
- [15] 张良均. Python 数据分析与挖掘实战[M]. 北京: 机械工业出版社, 2019: 160-161.
- [16] 夏修臣, 王秀英. 基于余弦相似度的改进 C4.5 决策树算法[J]. 计算机工程与设计, 2018, 39(1): 120-125.  
XIA Xiuchen, WANG Xiuying. Improved C4.5 decision tree algorithm based on cosine similarity[J]. Computer Engineering and Design, 2018, 39(1): 120-125.
- [17] 闫春, 厉美璇, 周潇. 基于改进的遗传算法优化 BP 神经网络的车险欺诈识别模型[J]. 山东科技大学学报(自然科学版), 2019, 38(5): 72-80.  
YAN Chun, LI Meixuan, ZHOU Xiao. Improved genetic algorithm for vehicle insurance fraud identification model based on BP neural network[J]. Journal of Shandong University of Science and Technology (Natural Science), 2019, 38(5): 72-80.
- [18] 武小军, 孟苏芳. 基于客户细分和 AdaBoost 的电子商务客户流失预测研究[J]. 工业工程, 2017, 20(2): 99-107.  
WU Xiaojun, MENG Sufang. E-commerce customer churn prediction based on customer segmentation and Adaboost[J]. Industrial Engineering Journal, 2017, 20(2): 99-107.
- [19] 王小川. MATLAB 神经网络 43 个案例分析[M]. 北京: 北京航空航天大学出版社, 2013: 42-49.

(责任编辑: 齐敏华)