本文引用格式：邵延富,谢大为.基于支持向量机的车险购买意向识别方法[J].自动化与信息工程,2024,45(6):87-92.
SHAO Yanfu, XIE Dawei. A method of car insurance purchase intention recognition based on support vector machine[J]. Automation & Information Engineering, 2024,45(6):87-92.

基于支持向量机的车险购买意向识别方法*

邵延富　谢大为
（广州市景心科技股份有限公司，广东 广州 510000）

摘要：针对传统的人工方式判断车主是否有意向购买车险存在效率低、缺乏预测性等问题，提出一种基于支持向量机的车险购买意向识别方法。首先，通过标准化处理与主成分分析降维，将 30 维的通话数据映射至 10 维空间，并采用欠抽样策略解决数据样本不平衡的问题；然后，利用 SVM 模型区分有、无意向车主。实验结果表明，SNM 模型的识别召回率和误检率分别为 97.9%、4.3%。该方法可为车险公司个性化服务提供技术支持。

关键词：车险购买意向识别；主成分分析；支持向量机；通话数据
中图分类号：TP311.5　　**文献标志码：**A　　**文章编号：**1674-2605(2024)06-0013-06
DOI：10.3969/j.issn.1674-2605.2024.06.013　　开放获取

A Method of Car Insurance Purchase Intention Recognition Based on Support Vector Machine

SHAO Yanfu　XIE Dawei
(Guangzhou Joysim Technology Co., Ltd., Guangzhou 510000, China)

Abstract: Aiming at the problems of low efficiency and lack of predictability in traditional manual methods of judging whether car owners intend to purchase car insurance, a car insurance purchase intention recognition method based on support vector machine is proposed. Firstly, through standardization and principal component analysis dimensionality reduction, the 30 dimensional call data is mapped to a 10 dimensional space, and under sampling strategy is adopted to solve the problem of imbalanced data samples; Then, use SVM model to distinguish between interested and uninterested car owners. The experimental results show that the recognition recall rate and false detection rate of the SNM model are 97.9% and 4.3%. This method can provide technical support for personalized services of car insurance companies.

Keywords: car insurance purchase intention recognition; principal component analysis; support vector machine; calling data

0　引言

据统计，2022 年我国民用汽车保有量已突破 3.8 亿辆[1]。同时，车险险种的结构、费用费率不合理等问题也日渐浮现。自 2015 年第二次车险费用改革以来，我国车险行业通过市场化手段降低了车险费率，但仍存在经营粗放、竞争失序和数据失真等问题[2]，迫切需要深化改革，以满足高质量发展的要求。为此，利用车险公司积累的大量数据，通过建立准确的数学模型，对车险产品进行精细化管理和定价非常重要。其中，对不同购买意向的车主提供个性化的保险方案与服务，可提高车险的吸引力，促进车险行业向更合理、高效的方向发展。

通过车主与车险公司的通话内容，分析车主的车险购买意向，有助于车险公司为车主提供更好的服务。传统的车险购买意向主要通过人工标注和调查问卷的方式来确定，不仅效率低，而且缺乏预测性。因此，开发基于大数据和机器学习算法的车险购买意向识别技术，具有重要的理论意义和应用价值[3]。

目前，车主购买车险意向的研究主要分为 4 类：1） 根据车主过去的车险购买记录，通过数据挖掘和

* 基金项目：广东省科技计划项目（2016A050502060，2020B1010010005）；广州市科技计划项目（202206010011，2023B03J1339）。

模式识别技术，预测车主未来购买车险的可能性和偏好，其优点是直接利用已有的数据资源，具有较强的实证基础，但过度依赖历史数据，忽略了车主的偏好和市场环境的变化；2） 通过分析车主的驾驶行为和风格，如驾驶速度、刹车习惯等，预测其购买车险的意向，其优点是体现了车主的个性化需求，能更精准地匹配适合其特点的车险产品，但收集和分析驾驶行为数据需要较高的技术支持和成本，且可能涉及隐私问题；3） 通过分析车主的消费习惯和偏好，如车险价格敏感度、险种选择偏好等，了解车主购买车险的行为模式，其优点是从消费者的角度为车险产品的设计和营销策略提供指导，但消费习惯受多种因素的影响，变化较大，难以准确预测；4） 车险公司的产品质量与服务态度，其优点是直接反映车险公司在市场中的竞争力，为其提升产品质量和改进服务提供指导方向，但评价标准主观性强，且受限于车主的个人体验和期望，可能存在一定的偏差[4]。

本文利用车主与车险公司的通话数据，提出一种基于支持向量机（support vector machine, SVM）的车险购买意向识别模型，分析车主购买或复购车险的意向程度。该模型利用主成分分析（principal component analysis, PCA）算法降低输入特征的维度，减少计算复杂度；通过 SVM 在 N 维空间中寻找最大化超平面，将有意向车主与无意向车主有效分离。

1 相关研究

1.1 PCA

PCA 是一种经典的数据降维技术，基本思想是将原始数据线性映射到一个新的正交空间，使映射后数据的方差最大，并舍弃贡献率较小的次要成分。具体做法是：计算数据协方差矩阵，求解其特征值和特征向量；选取前 k 个最大的特征值对应的特征向量作为新空间的基底，将原始数据投影到新空间中实现降维。

PCA 具有算法简单、无数据损失等特点，被广泛应用于图像识别、信号处理、数据压缩等领域。但它是一种无监督学习方法，数据降维后可能会丢失有用的判别信息。

1.2 SVM

SVM 是一种基于统计学习理论发展起来的有监督学习模型，主要用于模式识别、数据分析等领域。其基本工作原理是在高维特征空间中寻找一个最大边界超平面，将不同类别的样本点分开，使两类样本点到超平面的距离最大，从而达到较好的分类效果，原理如图 1 所示。

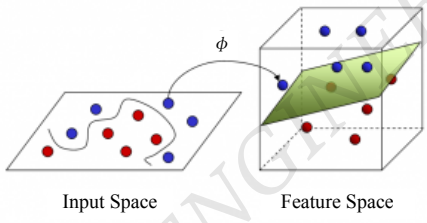

图 1 SVM 分类效果原理

SVM 具有正则化的特点，能够较好地控制模型的复杂度，避免过拟合；且泛化能力较强，在小样本、高维、非线性等情况下仍能获得较理想的性能。

2 识别方法

基于改进 SVM 的车险购买意向识别方法主要分为数据获取、数据预处理、特征降维、SVM 分类、模型预测 5 个环节。预先设定车主的车险购买意向为有意向车主和无意向车主 2 种类型。

2.1 数据获取

在预先获得授权的前提下，统计某车险公司半年内的通话数据，得出 30 个与车险购买意向相关的维度，包括是否拨打车险公司号码、通话时长、是否为主叫方等。其中，车主的手机号码经过加密脱敏处理。最终，收集到无意向车主数据 453 869 条、有意向车主数据 518 条。通话原始数据维度如表 1 所示。

表 1 通话原始数据特征维度[5]

序号	字段名
id	手机号码（加密脱敏处理）
x_1	总拨打次数
x_2	总被拨打次数
x_3	对车险公司总拨打次数
x_4	对车险公司总拨打号码数

续表

序号	字段名
x_5	对车险公司拨打总接通次数
x_6	对车险公司拨打总接通号码数
x_7	被车险公司总拨打次数
x_8	被车险公司总拨打号码数
x_9	被车险公司拨打总接通次数
x_{10}	被车险公司拨打总接通号码数
x_{11}	与车险公司平均通话时长
x_{12}	对车险公司拨打未接通次数
x_{13}	对车险公司拨打未接通号码数
x_{14}	被车险公司拨打未接通次数
x_{15}	被车险公司拨打未接通号码数
x_{16}	0~8 点时段被车险公司拨打次数
x_{17}	8~12 点时段被车险公司拨打次数
x_{18}	12~14 点时段被车险公司拨打次数
x_{19}	14~18 点时段被车险公司拨打次数
x_{20}	18~24 点时段被车险公司拨打次数
x_{21}	0~8 点时段对车险公司拨打次数
x_{22}	8~12 点时段对车险公司拨打次数
x_{23}	12~14 点时段对车险公司拨打次数
x_{24}	14~18 点时段对车险公司拨打次数
x_{25}	18~24 点时段对车险公司拨打次数
x_{26}	0~8 点时段与车险公司平均通话时长
x_{27}	8~12 点时段与车险公司平均通话时长
x_{28}	12~14 点时段与车险公司平均通话时长
x_{29}	14~18 点时段与车险公司平均通话时长
x_{30}	18~24 点时段与车险公司平均通话时长
y	是否有购买车险意向

无意向车主数据（453 869 条）与有意向车主数据（518 条）存在严重的数据不平衡。如果直接基于这种不平衡数据训练模型，将导致模型泛化性能较差。为解决这一问题，本文采用欠抽样的策略，从无意向车主数据中随机抽取 518 条，与有意向车主的 518 条数据组成平衡的混合数据集，并将该混合数据集按照 1∶1 的比例划分为训练集和验证集[6]。

2.2 数据预处理

由于通话数据特征维度的量纲不同，如是否主叫的取值范围为{0，1}，通话时长以秒为单位，因此需要对原始通话数据进行标准化处理，即将不同量纲的数据统一到同一数量级：

$$X' = \frac{X - \mu}{\sigma} \tag{1}$$

式中：X 为原始通话数据，μ 和 σ 分别为每个维度通话数据的均值和标准差，X' 为标准化后的通话数据[7]。

将训练集的通话数据标准化后，所有特征数据的均值为 0，标准差为 1。

2.3 特征降维

由于标准化后的通话数据的特征维数较高（30 个），且部分特征存在线性相关性，直接输入模型可能会导致过拟合及计算效率降低。为此，采用 PCA 算法对标准化后的通话数据进行降维处理。

设 30 维标准化后的通话数据组成的协方差矩阵为

$$cov(\boldsymbol{X}) = \begin{pmatrix} cov(x_1, x_1) & \cdots & cov(x_1, x_{30}) \\ \vdots & \ddots & \vdots \\ cov(x_{30}, x_1) & \cdots & cov(x_{30}, x_{30}) \end{pmatrix} \tag{2}$$

式中：$cov(x_i, x_j)$为第 i 个数据与第 j 个数据的协方差。

计算协方差矩阵 $cov(\boldsymbol{X})$的特征值及其对应的特征向量。将 $cov(\boldsymbol{X})$特征值按照从大到小的顺序排序后分别为$r_1, r_2, \ldots, r_{30}$，其对应的特征向量分别为$a_1, a_2, \ldots, a_{30}$。选择前 k 个最大的特征值对应的特征向量，并计算这 k 个特征值之和占全部特征值总和的百分比[8]，计算公式为

$$p_k = \frac{\sum_{i=1}^{k} r_i}{\sum_{i=1}^{30} r_i} \times 100\% \tag{3}$$

式中：p_k为前 k 个最大的特征值之和占全部特征值总和的百分比。

依次计算 $k = 1,2,3,\ldots,10$，结果如表 2 所示。

表 2　协方差矩阵 $cov(\boldsymbol{X})$特征值占比　　%

维度	特征值占全部特征值总和的比例	特征值之和占全部特征值总和的比例 p_k
$k=1$	17.0	17.0
$k=2$	15.0	32.0
$k=3$	11.0	43.0
$k=4$	10.0	53.0
$k=5$	9.0	62.0
$k=6$	8.7	70.7
$k=7$	7.4	78.1
$k=8$	7.1	85.2
$k=9$	6.5	91.7
$k=10$	6.4	98.1

由表 2 可知，前 10 个最大的特征值之和占全部特征值总和的比例为98.1%，即主成分保留98%以上，故选取前 10 个特征值替代通话数据的 30 个特征值。

设标准化、降维后的通话数据 10 个特征值对应的特征向量组成的列矩阵为$\boldsymbol{A}=(a_1,a_2,\ldots,a_{10})$，标准化后的通话数据组成的矩阵为$\boldsymbol{X}_{(518\times30)}$，则最后得到的投影矩阵为$\boldsymbol{X}_{投}=\boldsymbol{XA}$。

投影矩阵$\boldsymbol{X}_{投}$为新的通话样本数据矩阵，共 10 个维度，518 个样本。

2.4　SVM 分类

经初步分析发现，有意向车主数据与无意向车主数据在一些字段上较为相似，用低维的曲线拟合容易导致欠拟合，用高维的多项式拟合容易出现过拟合。为此，采用 SVM 算法先把低维不可分的数据投射到高维空间，再用高维超平面进行是否有车险购买意向的二分类。

然而在实践中，较难通过一个超平面精准地分开两个类别。为此，引入软间隔的概念，允许某些数据点可以处于错误的分类侧。软间隔的概念包括松弛变量ξ和惩罚参数 C 两个要素。

1） 松弛变量ξ：对于每个样本 i，引入一个非负的松弛变量ξ_i来表示该样本距离正确分类边界的距离。如果该样本被正确分类，则$\xi_i=0$；如果该样本被正确分类，但位于边界内部，没有完全满足间隔要求，则$0<\xi_i<1$；如果该样本恰好位于分割超平面上，则$\xi_i=1$；如果该样本被错误分类，则$\xi_i>1$。

2） 惩罚参数 C：引入一个正的调整参数 C，其决定了分类的严格程度。较大的惩罚参数 C 意味着模型会对错误分类的样本给予更大的惩罚，从而可能导致模型过拟合；较小的惩罚参数 C 使模型对错误分类的样本惩罚减小，从而可能导致模型欠拟合。

设 $\boldsymbol{w}$ 为所求超平面的法向量，本文目标是找到这个法向量 $\boldsymbol{w}$ 以及参数ξ_i，以最大化两个类别之间的边界或间隔，同时确保所有样本能被正确分类。这可以表达为以下问题：

$$\min_{w,b,\xi}:\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{N}\xi_i \tag{4}$$

为了确保所有的样本尽可能地被正确分类（允许一些错误分类存在），对每个样本 i 施加以下约束：

$$y_i(\boldsymbol{w}^{\mathrm{T}}n_i+b)\geq 1-\xi_i \tag{5}$$

$$\xi_i>0,i=1,2,\ldots,518 \tag{6}$$

式中：y_i为样本 i 的真实类别（有意向车主或无意向车主），b 为超平面的截距，n_i为由样本 i 各维度（10 个）数值组成的向量。

利用拉格朗日乘子法求解上述优化问题，对应的拉格朗日函数为

$$L(w,b,\xi,\alpha,\mu)\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{N}\xi_i-$$

$$\sum_{i=1}^{N}\alpha_i\left(y_i\left(\boldsymbol{w}\cdot n_i+b\right)-1+\xi_i\right)-\sum_{i=1}^{N}\mu_i\xi_i \tag{7}$$

式中：μ_i为拉格朗日算子。

这样，把 $\boldsymbol{w}$ 和 b 的约束最小化 L 问题转换为α的约束最大化 L 问题，从而得到对偶问题：

$$\max_a:\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}y_iy_ja_ia_jn_i^{\mathrm{T}}n_j \tag{8}$$

需要满足条件：

$$0\leq a_i\leq C,\ \sum_{i=1}^{n}a_iy_i=0 \tag{9}$$

原始问题的解若要与对偶问题的解一致，需要满足 KKT 条件，即

$$\alpha_i\left(y_i\left(\boldsymbol{w}\cdot n_i+b\right)-1+\xi_i\right)=0 \tag{10}$$

$$\mu_i\xi_i=0 \tag{11}$$

使用序列最小优化算法求解上述对偶问题（详细过程不是本文重点，不再赘述），可以得到 SVM 模型的超平面法向量$\boldsymbol{w}$以及超平面的截距b。

对于需要预测的样本点n_{i+1}，SVM 模型的输出结果为

$$u_{i+1}=\boldsymbol{w}\cdot n_{i+1}+b \tag{12}$$

$u_{i+1}>0$ 为无意向车主，$u_{i+1}<0$ 为有意向车主[9]。

综上所述，基于 SVM 模型的车险购买意向识别方法的流程如图 2 所示。

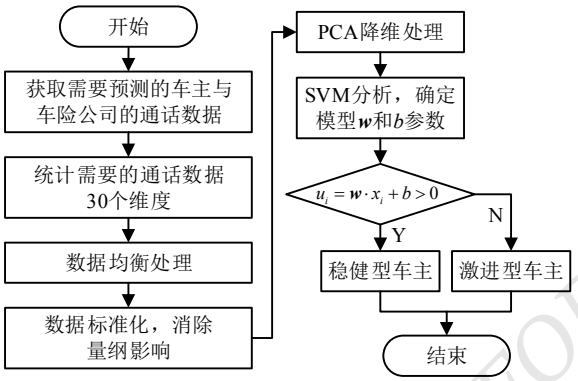

图 2　基于 SVM 模型的车险购买意向识别方法的流程

2.5　模型预测

利用验证集的 518 条数据进行 SVM 模型预测验证，其中 288 条数据为有意向车主数据，占比为 55.6%；230 条为无意向车主数据，占比为 44.4%。

对验证集数据进行数据预处理、特征降维（10 维）后，输入 SVM 模型（公式 12），得出的预测结果如表 3 混淆矩阵所示[10]。

表 3　验证集对模型验证的结果混淆矩阵　　单位：个

		预测	
		有意向车主	无意向车主
实际	有意向车主	282	6
	无意向车主	10	220

由表 3 可知，SVM 模型预测的召回率为 97.9%（282/(282+6)），误检率为 4.3%（10/(220+10)），表明该模型的预测性能良好[11]。

3　结论

本文针对车主的车险购买意向识别问题，提出了一种基于支持向量机的车险购买意向识别方法。1）根据车主与车险公司的通话数据，统计需要分析的 30 个维度；2）利用欠抽样策略随机抽取部分车主数据，构建平衡数据集，缓解模型训练过程中的数据偏差；3） 对 30 维的原始通话数据进行标准化预处理，消除不同特征量纲的影响；4） 采用 PCA 算法将 30 维数据映射到 10 维空间，并保留了 98%以上的方差贡献，有效降低了数据冗余；5） 采用 SVM 算法建立分类模型，将是否有意向购买车险的数据进行分类。经验证集测试，SVM 模型的召回率为 97.9%，误检率为 4.3%，具有良好的分类性能。本文 SVM 模型还存在需要改进的地方，如模型进一步优化、模型参数的自动调优等。

参考文献

[1] 杨怡.关于推动我国交通运输高质量发展的几点思考[J].人民公交,2024(11):69-72.

[2] 房弢,陈晨,华圆.中小财产车险公司差异化发展路径与服务实践研究[J].浙江工商职业技术学院学报,2024,23(1):13-17.

[3] 唐金成,肖思文.新能源车险发展困境与创新策略[J].中国保险,2024(5):29-33.

[4] 谭征.基于 K-Means 和 SEM 的消费者互联网保险购买意愿研究——以 TPB 和 TAM 为分析框架[J].重庆理工大学学报(自然科学),2019,33(2):198-207.

[5] 邓育辉,李鹏,邵延富,等.基于神经网络与粒子群算法的骚扰电话识别研究[J].数据通信,2022(5):28-30.

[6] HAN Jiawei, KAMBER Micheline, Pei Jian. Data mining: Concepts and techniques[M]. Waltham: Morgan Kaufmann Publishers, 2012:265-270.

[7] 包研科.数据分析教程[M].北京:清华大学出版社,2011:68-76.

[8] 王学民.应用多元分析[M].4 版.上海:上海财经大学出版社,2014:188-214.

[9] TAN Pangning, STEINBACH Michael, Vipin Kumar. Mining

W I D. Introduction to data mining[M]. New Jersey: Pearson Education, Inc, 2006:127-189.
[10] 李文楷,刘原池,刘子越,等.基于正样本-背景数据的校正混淆矩阵[J].海南大学学报(自然科学版),2023,41(3):293-302.
[11] 高瑞.混淆矩阵在商业决策中的应用研究[J].当代经济,2021(4):110-113.

作者简介：

邵延富，男，1972 年生，硕士研究生，高级工程师，主要研究方向：软件工程和大数据工程。E-mail: 13632102858@139.com
谢大为，男，1991 年生，本科，高级工程师，主要研究方向：大数据工程。E-mail: 13798088446@139.com

（上接第 79 页）

[4] KUNDU S, GAUTAM M, HERATH A, et al. People detection in an elevator car using computer vision[C]. Proceedings of the 7th Baltic Mechatronics Symposium, Aalto-yliopisto, 2022.
[5] WU D, WU S, ZHAO Q, et al. Computer vision-based intelligent elevator information system for efficient demand-based operation and optimization[J]. Journal of Building Engineering, 2024, 81:108126.
[6] KOLB A, BARTH E, KOCH R. ToF-sensors: New dimensions for realism and interactivity[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2008: 1-6.
[7] HE Y, LIANG B, ZOU Y, et al. Depth errors analysis and correction for time-of-flight (ToF) cameras[J]. Sensors, 2017, 17(1):92.
[8] TANNER R, STUDER M, ZANOLI A, et al. People detection and tracking with TOF sensor[C]. 2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance, IEEE, 2008: 356-361.
[9] STEC M, HERRMANN V, STABERNACK B. Using time-of-flight sensors for people counting applications[C]. 2019 Conference on Design and Architectures for Signal and Image Processing (DASIP), IEEE, 2019: 59-64.
[10] JANG Jun-Woo, LEE Sehwan, KIM Dongyoung, et al. Sparsity-aware and re-configurable NPU architecture for Samsung flagship mobile SoC[C]. 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), IEEE, 2021:15-28.
[11] KIM S, PARK G, YI Y. Performance evaluation of INT8 quantized inference on mobile GPUs[J]. IEEE Access, 2021, 9:164245-164255.
[12] TERVEN J, CÓRDOVA-ESPARZA D, ROMERO-GONZÁLEZ J A. A comprehensive review of Yolo architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS[J]. Machine Learning and Knowledge Extraction, 2023,5(4): 1680-1716.
[13] IOFFE Sergey, SZEGEDY Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]. Proceedings of the 32nd International Conference on Machine Learning, 2015,37:448-456.
[14] QIU M, HUANG L, TANG B-H. ASFF-YOLOv5: Multielement detection method for road traffic in UAV images based on multiscale feature fusion[J]. Remote Sensing, 2022,14(14): 3498.
[15] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015,37(9):1904-1916.
[16] ELFWING S, UCHIBE E, DOYA K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning[J]. Neural Networks, 2018,107:3-11.
[17] GLOROT Xavier, BORDES Antoine, BENGIO Yoshua. Deep sparse rectifier neural networks[C]. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011,15:315-323.

作者简介：

唐其伟，男，1979 年生，硕士研究生，主要研究方向：智能传感与智能控制。E-mail: tangqiwei@hitachi-helc.com
杜卉然，男，1993 年生，硕士研究生，主要研究方向：机器学习与机器视觉。E-mail: duhuiran@hitachi-helc.com
程伟，男，1984 年，硕士研究生，主要研究方向：智能控制。E-mail: chengwei@hitachi-helc.com
陈刚，男，1989 年生，硕士研究生，主要研究方向：机械电子工程。E-mail: chengang39802@hitachi-helc.com