基于数据智能的人车模型构建与资源分析系统

叶 飞,吴奇石
(西南交通大学 信息科学与技术学院,四川 成都 611756)

摘 要:随着生活水平的提高,越来越多的家庭开始购买不止一辆车。庞大的需求也促进了国内汽车制造商的蓬勃发展,同时世界各品牌的汽车制造厂也纷纷进入中国市场。然而无论是国内企业还是国外企业,产业的差异性越来越小,使得原来依赖于成本优势来提高竞争力的模式逐渐失效。目前企业把重心放在客户体验和需求以及如何提高汽车产品的设计上。文中以 AA 企业为原型,对其整车销售和营销模式进行分析。目前产业链协同平台已完善了汽车制造厂从生产到入库以及汽车经销商从下销售订单计划到销售结束的业务流程。然而,由于缺乏足够的数据分析支持,汽车制造厂和经销商往往不能估计到客户的真实需求,也不能准确地从海量的客户资源中挖掘出潜在客户。为解决该类问题,文中提出构建基于数据智能的人车模型以及客户资源分析系统。分析了现存的整车销售与营销模式的需求,提出一种新的混合优化算法。该算法将遗传算法在解决离散问题上的优势和群体智能算法在解决连续问题上的优势相结合,通过寻找最优特征子集和最优支持向量机参数配置(惩罚参数和核函数参数)来优化支持向量机。该算法的主要创新在于提出了三个并行的操作层。其中两个层是遗传算法操作层和群体算法操作层,第三个层是协调层,主要负责接收其他两个层的个体信息并组合成新的个体信息进行评估。随后将评估结果返回给其他两个层。因此当算法优化 SVM 时,不需要额外的映射函数来将离散变量转化为连续变量或是连续变量转化为离散变量。最后该算法被成功应用到产业链协同人车模型中。

关键词:跟踪;支持向量机;果蝇优化;遗传算法;优化算法;二值分类
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2019)12-0122-08
doi:10.3969/j.issn.1673-629X.2019.12.022

Pedestrian-vehicle Model Construction and Resource Analysis System Based on Data Intelligence

YE Fei, WU Qi-shi
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: With the improvement of living standard, more and more families begin to buy more than one car. The huge demand has also promoted the vigorous development of domestic automobile manufacturers, while the automobile manufacturers from all over the world have also entered the Chinese market. However, whether domestic or foreign enterprises, the industry difference is becoming smaller and smaller, which makes the original mode of relying on cost advantage to improve competitiveness gradually invalid. At present, enterprises are focusing on customer experience and demand and how to improve the design of automobile products. With AA enterprise as a prototype, we analyze its vehicle sales and marketing mode. At present, the industry chain collaboration platform has improved the business processes of automobile manufacturers from production to warehousing, and automobile dealers from the next sales order plan to the end of sales. However, due to the lack of sufficient data analysis support, automobile manufacturers and distributors often cannot estimate the real needs of customers. It can not accurately excavate potential customers from the vast customer resources. In order to solve these problems, we propose to build a data intelligence-based car model and customer resource analysis system. We analyze the demand of the existing vehicle sales and marketing mode, and then propose a new hybrid optimization algorithm which combines the advantages of genetic algorithm in solving discrete problems and swarm intelligence algorithm in solving continuous problems. The algorithm optimizes the support vector machine by searching for the optimal feature subset and the optimal support vector machine parameter allocation (penalty parameter and kernel function parameter). The main innovation of this algorithm is to propose three parallel operation layers. Two of them are genetic algorithm operation layer and swarm algorithm operation layer. The third layer is the

收稿日期:2019-01-04 **修回日期**:2019-05-07 **网络出版时间**:2019-09-24
基金项目:国家重点研发计划项目(2017YFB1400300)
作者简介:叶 飞(1990-),男,硕士研究生,研究方向为制造业信息化、价值链服务工程;吴奇石,教授,中组部青年千人计划引进人才,研究方向为工业互联网、传感器网络、高性能网络、网络安全、并行和分布式计算。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190924.1535.022.html

coordination layer, which is mainly responsible for receiving individual information from the other two layers and combining them into new individual information for evaluation. The evaluation results are then returned to the other two layers. Therefore, when the algorithm optimizes SVM, no additional mapping function is needed to convert discrete variables into continuous variables or continuous variables into discrete variables. Finally, the algorithm is successfully applied to the collaborative human-vehicle model of industrial chain.

Key words: tracking; support vector machine; fruit fly optimization; genetic algorithm; optimization algorithm; binary classification

1　概　述

近年来,随着制造强国德国提出“工业 4.0”以及美国提出“工业互联网”的概念,中国作为制造业的后起国家,于 2015 年发布了《中国制造 2025》计划,其主要目的是实现未来制造业强国崛起战略。其中汽车产业扮演了非常重要的角色[1]。

文献[2]根据权威的汽车工业协会统计,国内汽车生产量在 2010 年已经到达 1 826.47 万辆,而在同一年大约销售了 1 806.19 万辆,同比增长均超过 32%。同时具有自主品牌的汽车市场份额也有所增加。特别是在乘用车方面,具有自主品牌汽车已经销售了 627.3 万辆,大约占全部乘用车总量的 45.6%。2016 年中国汽车销售超过 2 800 万辆车,连续八年都是全球销量最高的国家。随着汽车产业的飞速发展,传统的管理和销售模式已经不能适用于当前的汽车产业。越来越多的汽车制造企业开始利用信息化手段来促进管理水平的提高以及销售的增长。这些信息化平台通常能提供大量实用的功能给汽车制造企业,经销商和服务站。如销售订单管理,活动管理和维修相关的业务功能。然而随着汽车行业的竞争越来越激烈,普通的业务功能已经不能满足当前企业去提升市场竞争力的需要,同时当前数据分析和数据挖掘技术已经成功地应用在各行各业。

如何重构汽车模型设计以及更好的营销是汽车企业成功的核心。在传统的产业链平台上,制造企业尽管产生了大量的数据,如销售订单、生产数据以及保养数据等,然而由于缺乏数据分析的支持,汽车制造企业往往不能精准地投放广告以及推出更受消费者喜爱的新车型。数据分析将助力实现市场营销的价值,通过对海量市场营销数据(如客户档案数据和销售数据)进行分析,可以寻找目前市场营销中所存在的不足之处,对于所存在的瓶颈内容进行优化和改进,同时基于已有的分析可以对未来营销活动的开展给予相应的指导,帮助制造企业制定营销决策和改善产品设计。

正是基于这种背景,文中提出了基于数据智能的人车模型构建。该模型基于某协同销售系统以及数据空间,针对企业的客户数据和销售数据进行分析建模。建立起的模型不仅仅帮助制造企业分析用户的购买行为,并且帮助它们更好地制定营销计划和改善汽车产品设计。

基于产业链协同平台中的销售业务数据、与整车生产有关的整车出生档案数据以及客户档案数据进行研究,并通过构建人与车模型来帮助制造企业从传统的生产业务走向更加智能化的管理和营销模式。该模型采用了 B/S 架构,并使用了目前比较流行的 WEB 开发平台 ASP.NET。为了提供产品化的展示界面,文中除了使用基于标准的 ASPX 技术,还使用了 JS(Java Script)、AJAX、CSS(cascading style sheet)、JASON 等技术。此外为了能将数据进行更好的图表展示,使用了 ECHART 插件。该插件可以提供丰富的图表展示功能,能够满足大部分需求。对于模型的构建,文中主要使用了群体智能、机器学习方法以及数据挖掘等技术。

2　机器学习的研究现状

近年来随着信息技术的发展,越来越多的研究机构和企业把目光投入到人工智能领域。其中机器学习已经成为越来越热门的研究方向。文中简单回顾下机器学习的研究现状,其中着重介绍两个重要的模型,一个是支持向量机,另一个是深度学习。

支持向量机是建立在统计学习理论上的模型,最初由 Vapnik 等[3-4]提出。由于最初的支持向量机只能解决二分类问题,随后被作者扩展到可以解决多分类问题[5]。支持向量机对分类任务具有很多优势,它可以生成一个独一无二的全局超平面将不同类别的数据分开。由于支持向量机遵循结构风险最小化原则(SRM),它在训练阶段减少了风险的发生,并提高了其泛化能力。近年来,研究人员提出了一些改进的 SVM。Mangasarian 等[6]引入广义特征值近似 SVM(GEPSVM)生成两个非平行超平面。在这种方法中,每个类的模式或数据样本位于一个超平面的紧密接近处,并与其他平面保持清晰的分离。在支持向量机和 GEPSVM 的基础上,Jayadeva 等[7]提出了一种新颖的二进制分类器 twin support vector machine (TWSVM),对图案进行分类通过使用两个不平行的超平面。与传统 SVM 相比,TWSVM 解决了一对 QPP 而不是单个复杂 QPP 问题。Shao 等[8]提出了一种改进的 TWSVM,也称为 twin bounded support vector machine (TBSVM)。TBSVM 使用“连续过松弛(SOR)”技术解决了优化问题,以提高训练过程的速度,并使用正则化项来最小化 SRM 原理。就计算时间和分类精度而言,TBSVM 比 TWSVM 更有效。Kumar 等[9]提出一

个更快的 SVM 模型(LSTSVM),并且还表现出增强的泛化性能。LSTSVM 没有求解一对复数 QPP,而是通过求解两个线性方程来生成两个非平行的超平面。

深度学习已经成为近年来的主流,这主要得益于大量可以利用的数据以及硬件性能的极大提高。然而在早些年,由于数据的缺乏和硬件的限制,想要训练一个大型神经网络变得尤为困难。在 2006 年,Hinton[10] 等提出一种快速训练神经网络的方法,叫做深度信念神经网络(DBN)。该算法一经提出,立即引起了反响。人们随后开始使用规模越来越大以及越来越复杂的神经网络。不同于传统的神经网络,DBN 是一种无监督学习算法,其主要目的是使用无监督训练进行神经网络的权值初始化,然后对权值进行微调。在 2011 年,新的激活函数(ReLU[11])被提出。该激活函数主要解决梯度消失的问题。随后几年里,Hinton 等创建了深度神经网络架构 AlexNet[12],它采用了卷积神经网络(CNN),并且抛弃了预训练,而是直接训练在数据集上。该网络模型已经取得了 ImageNet 图像识别大赛第一名,验证了深度学习的性能。在 2014 年,Goodfellow 等[13] 提出一种新的深度学习方法,叫做对抗生成神经网络。其主要目的是训练一个生成器能产生和真实样本类似的数据。然而最初版本的对抗生成神经网络有些弊端,比如训练不稳定,并且需要恰当的参数配置。近些年来有些改进的对抗生成神经网络已经被提出。其中一个是条件对抗生成神经网络,与传统方法不同的是,生成器和判别器都增加额外信息 y 为条件,y 可以是任意信息,例如类别信息,或者其他模态的数据。刘智斌等[14] 提出启发式强化学习方法。Gulrajani 等[15] 提出一种改进的对抗生成神经网络方法,叫做 WGAN。它不仅被应用在非监督学习中来生成更加高清和相似的图像,也被应用在半监督学习中来提高神经网络的泛化能力。

3 客户资源系统架构设计

根据规划控制下二阶段设计理论,结合整车交易的瓶颈,系统开发、维护等因素,. NET 和 SQL Server 进行系统设计。整个系统可分为四个部分:数据处理层,业务逻辑处理层,用户表示层,人车模型的构建,以及最终的算法层。总体的系统架构如图 1 所示。

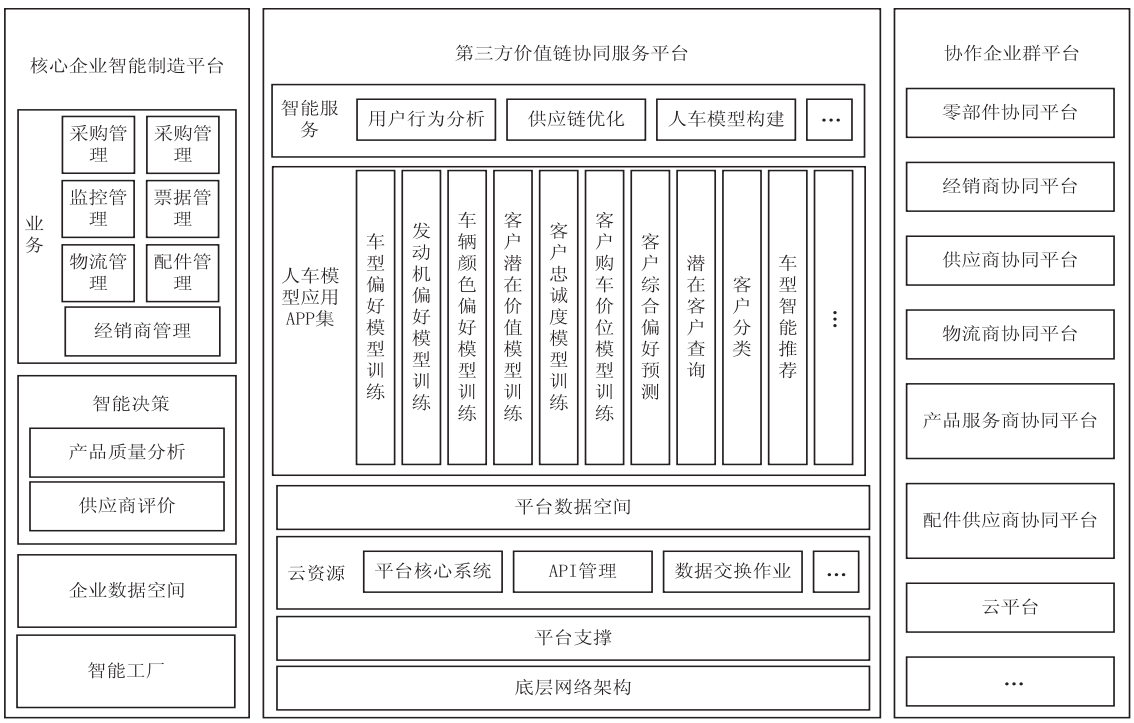

图 1 基于规划控制下多平台协同的人车模型与客户资源分析总体架构

人车模型与客户资源分析系统内部建构如图 2 所示。人车模型主要包括基础信息管理,人车模型算法配置,用户购车行为分析,购车用户分析以及销售分析等多个模块。文中侧重于人车模型的构建以及用户行为分析的功能模块。

该系统使用了整车销售数据库和服务数据库。因为仅仅通过查询整车订单和车辆基本数据是远远不够的。通过分析服务数据,文中可以提供更多因素来构建人车模型。第二层是特征抽取层,该层实时读取最新的数据并将它们转化为特定数据格式以方便建模。比如对于车辆来说,车辆的编号不适合作为特征,因为它并不是其中一个客户所关注的购买因素。第三层是数据处理层,该层的主要功能是将所有特征转化为数字矩阵,因为大部分数据如客户数据(包括职业、性别

和兴趣)都是字符类型。为了能使用机器学习去训练这些数据,该系统使用统一的编码将不同的职业、兴趣等都转化为不同的数字。此外归一化主要用来将矩阵所有的元素映射到 0-1 之间以方便训练。第四层是算法层,该系统主要实现了使用 SVM 和 SVR 进行人车模型的构建。此外,SVM 和 SVR 对特征以及核函数很敏感,所以该系统使用提出的混合优化算法来得到最优的模型,提高性能和准确率。第五层主要是将训练好的模型进行应用,第六层主要是基于这些应用给汽车经销商和制造厂提供智能决策功能模块。

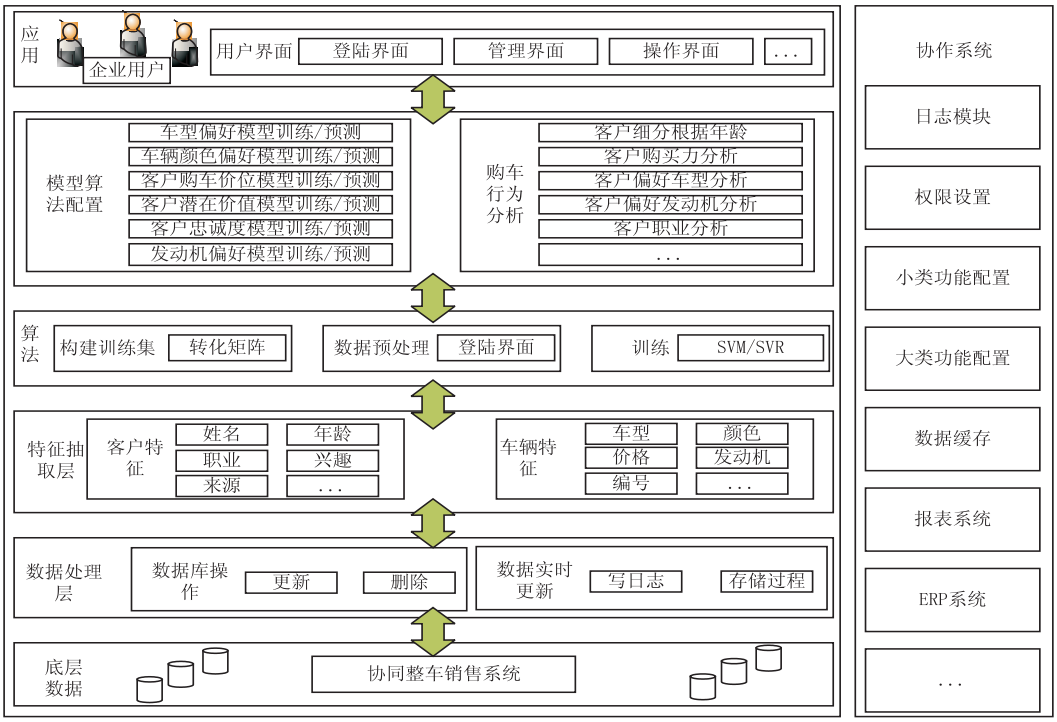

图 2　人车模型与客户资源分析系统内部架构

4　人车模型算法模型的构建与优化

文中主要针对汽车经销商与制造企业在实际的销售与营销管理中存在的不足,提出算法解决方案来构建人车模型。从企业的实际战略管理角度出发,建立以客户与车辆的关系,并以客户为中心的销售与营销模式。使用提出的混合优化算法来构建 6 个不同的子模型。

4.1　混合优化算法

本节提出一种新的进化支持向量机模型,该模型采用基于 GA 算法,结合群体智能优化技术(PSO 或 FOA)来实现 SVM 分类器的特征选择和参数优化。提出的 GAPSO-FS 和 GAFOA-FS 可以自适应地探索最佳超平面参数,并选择支持向量机模型的最优特征子集。每个算法模型的系统架构主要由三层组成。第一层是基于遗传算法的特征选择层,主要通过选择、交叉和变异的遗传操作来搜索最优特征子集,并将特征选择映射为二进制编码。第二层是参数优化层,该层主要负责通过改变每个粒子的位置和速度来动态调整 SVM 分类器的惩罚参数和超平面参数。使用该方式,并不需要一个特殊的映射函数,而支持向量机模型的连续参数可以由 PSO 或 FOA 确定。第三层是协调层,主要负责处理来自其他两层的信息,并将计算得到的结果返回给其他两个层。在第三层处理中,二进制字符串与连续参数组成一个 SVM 的参数配置。使用设计好的适应度函数可以计算出每个 SVM 参数配置的适应值,并将适应度值返回给其他两个层,方便其他两个层进行进化操作。而最优 SVM 参数配置所对应的个体将在两个算法内部指导进化操作,以便在下一代中能产生更优良的个体。

4.2　输入特征表示

在所提出的混合算法模型中,特征子集选择由 GA 实现。因为它是基于染色体的编码方式,所以可以很容易地被用来表示所选择的特征集。文中构造一组比特来表示个体染色体和特征选择之间的映射关系。输入变量 $\boldsymbol{D}$ 由 n 个样本组成 $(S_1,S_2,\cdots,S_i,\cdots S_n)^{\mathrm{T}}$,$S_i$ 代表第 i 个样本包含 d 个特征。第 i 个特征集可以用特征集 $X^i=(X_1^i,X_2^i,\cdots,X_d^i)$ 表示。因此,原始输入变量 $\boldsymbol{D}$ 可以通过以下特征矩阵表示:

$$\boldsymbol{D}^{n\times d}=(\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_n)^{\mathrm{T}}=\begin{pmatrix}X_1^1 & X_2^1 & \cdots & X_d^1\\ X_1^2 & X_2^2 & \cdots & X_d^2\\ \vdots & \vdots & \ddots & \vdots\\ X_1^n & X_2^n & \cdots & X_d^n\end{pmatrix}\quad(1)$$

SX^i 代表被选择的特征集。如果 $f_j^p = 1$,则选择了第 j 个特征;如果 $f_j^p = 0$,则第 j 个特征没有被选择。因此,原来的特征 $D^{n\times d}$ 通过第 P 个个体所有代表的特征掩模转化为新的矩阵 $D_p^{n\times fd}$,并且 fd 是所选择的特征的数目。最后将特征矩阵分为训练集和测试集。并结合标记向量 $Y = (Y_1, Y_2, \cdots, Y_i, \cdots, Y_n)^{\mathrm{T}}, Y_i \in \{-1, 1\}$ 作为输入变量和 SVM 参数来计算第 p 个个体的分类精度和其他性能标准。

$$SX^i = (X_1^i \cdot f_1^p, X_2^i \cdot f_2^p, \cdots, X_j^i \cdot f_j^p, \cdots, X_d^i \cdot f_d^p) \quad (2)$$

4.3 混合算法的适应度函数设计

个体适应度主要由三个评价标准决定。这些评价标准是分类精度、所选择的特征子集的大小和特征成本。因此,为了在进化过程中找到最佳个体,具有低特征成本和高分类精度的小特征子集往往是最优解。所提出的混合模型采用单一的混合适应度函数,将这三个评价标准组合为一个函数。混合算法的适应度函数主要先接受来自 GA 优化层的二进制字符编码和来自群体智能层的 SVM 参数配置(特征子集和 SVM 参数),然后对 SVM 进行训练和测试,获得分类准确度。最后根据适应度函数计算每个个体的适应度值,并将更新的适应度信息转发给两个优化层。GA 层主要用于提供所选择的特征子集和特征成本,而(PSO 或 FOA)层主要用于计算惩罚参数和超平面参数。通过式 3 获得个体的适应度。

$$\text{fitness\_all} = W_{\mathrm{a}} \times \text{accuracy} + W_{\mathrm{f}} \times \left(1 - \frac{1}{n}\sum_{i=1}^{n} F_i \times C_i\right) \quad (3)$$

其中,fitness_all 表示个体的适应度值;W_a 表示分类准确度的权重;accuracy 表示当前个体所获得的分类精度;W_{f} 表示所选择特征的特征成本数量大小的权重;C_i 表示第 i 个特征的成本。一般来说,用户可以改变一个特征的成本,以满足不同任务的要求。F_i 表示第 i 个特征的掩码值。$F_i = 1$ 表示第 i 个特征被选择为 SVM 分类器的输入特征,$F_i = 0$ 表示第 i 个特征被忽略。

该适应度函数设计的主要优势在于,在模型的训练过程中,不光考虑 SVM 模型的分类精度,同时也希望有更好的泛化能力。在该适应度函数设计中,希望有更少的特征成本与特征数量。因为当用更少的特征数量来构建 SVM 的输入空间时,SVM 模型本身的参数会更少。

4.4 混合算法流程

本节描述了所提出的 GAPSO-FS 和 GAFOA-FS 算法的细节。GAPSO-FS 和 GAFOA-FS 方法使用不同的智能群体算法优化 SVM 参数,即 GAPSO-FS 使用 PSO,GAFOA-FS 使用 FOA 来优化 SVM 参数和核函数参数。两个方法的基本过程可以归纳为九个独立步骤:

步骤 1:数据预处理。

面对来自数据库的数据,这些数据通常不能直接用于 SVM 训练,将每个数据集转换为特征矩阵和标签向量。

步骤 2:参数初始化。

在参数初始化阶段,GA 和 PSO 的种群大小被设置为相同的值,并且必须适当地设置迭代的最大次数。此外,当设置 GA 的参数时,突变概率和交叉概率一般分别设置为 0.15 和 0.75。交叉点和选择操作采用单点交叉和轮盘选择方法。在 GAPSO-FS 方法中,PSO 参数一般设置如下:C1i,C1f,C2i,C2f 加速度系数分别设置为 2.5、0.5、0.5 和 2.5。$W_{\max}, W_{\min}$ 惯性权重分别设置为 0.4 和 0.9。在 GAFOA-FS 方法中,参数分别设置为 20,10,20 和 10。

步骤 3:人口初始化。

(1)在人口初始化阶段,生成一组随机二进制串 $G = \{G_1, G_2, \cdots, G_i, \cdots, G_p\}$,其中每个二进制串 G_i 由一组比特串组成,其数目等于特征数。每个二进制字符代表一个特征掩码。

(2)在 GAPSO-FS 方法中,PSO 过程需要初始化一组粒子位置和一组粒子速度。每个粒子 X 的位置和相应的速度 V 形成二维阵列:X_i 粒子的位置由参数范围内的随机值给出,粒子的速度 V_i 在范围 $[V_{\min}, V_{\max}]$ 中随机初始化。在 GAFOA-FS 方法中,根据果蝇群的位置 $(X_{\text{axis}}, Y_{\text{axis}})$ 对每个果蝇的位置 (X_i, Y_i) 进行初始化。

步骤 4:获得个体表示的解。

提出的混合算法模型中的个体 (individual_i) 包含两个不同的子个体;一个是在 GA 中使用的个体 $(\text{individual}_i^{\text{GA}})$,用于表示所选择的特征的原型;另一个是 PSO 个体 $(\text{individual}_i^{\text{PSO}})$ 或 FOA 个体 $(\text{individual}_i^{\text{FOA}})$,用于表示 SVM 参数和超平面参数。个体 $(\text{individual}_i^{\text{FOA}})$ 的详细设计如图 3 所示,$\text{individual}_{(i,j)}^{\text{GA}}$ 表示 GA 的第 i 个个体的第 j 个维度,也代表第 i 个特征掩码的第 j 个维度。$\text{individual}_{(i,j)}^{\text{PSO}}$ 和 $\text{individual}_{(i,j)}^{\text{FOA}}$ 分别表示 PSO 或 FOA 的第 i 个个体的第 j 个维度。例如,第一个个体的参数可以用 $(\text{individual}_{(0,j)}^{\text{GA}}$ 和 $\text{individual}_{(0,j)}^{\text{PSO}})$ 或 $(\text{individual}_{(i,j)}^{\text{PSO}}$ 和 $\text{individual}_{(i,j)}^{\text{FOA}})$ 表示。因此,第 i 个个体 (individual_i) 可以方便地表示其对应的原型(特征掩码,惩罚参数和核函数参数)。

步骤 5:计算适应度值。

使用在步骤 4 中获得的个体所表示的原型(所选

择的特征子集、惩罚参数和超平面参数),计算适应度值如下:

(1)将 SVM 模型与每个($individual_i$)所代表的原型进行训练,然后使用 SVM 模型进行预测,以确定其准确性、灵敏度和特异度。

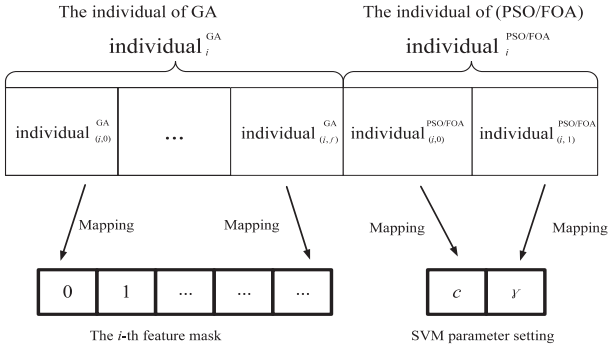

图 3　个体详细设计

(2)计算每个个体的适应度值,根据准确度、灵敏度、特异性和所选择的特征的数量等几个指标。

步骤 6:将适应度值分配给 GA 和群体智能优化算法(PSO/FOA)。

该步骤将适应度信息分配给 GA 和 PSO 或 FOA 操作层,以确定两个层的最优个体。GA 中的最佳个体是通过选择种群间的最大适应值来确定的。对于 GAPSO-FS 方法,当所有粒子中新的粒子的最佳适应度大于所有粒子的最佳适应度,gbest 会进行更新。对于 GAFOA-FS 方法,通过选择果蝇群中最大气味浓度值的果蝇来确定果蝇群的位置 (X_{axis},Y_{axis})。

步骤 7:群体智能技术(PSO/FOA)层的过程。

在 GAPSO-FS 方法中,通过更新迭代 t 中的第 i 粒子的 d 维的当前速度,并更新每个粒子的位置。

步骤 8:GA 优化层的过程。

该步骤执行 GA 的操作,包括复制、选择、交叉和变异操作,以搜索更合适的特征子集。

步骤 9:检查终止条件。

如果满足任何终止条件(最大迭代次数),则算法的迭代停止,并且获得最优 SVM 模型;否则,重复步骤 4 ~步骤 8。

4.5　人车模型建模

本节主要介绍如何利用提出的混合优化算法来进行人车模型的构建。文中主要构建了 6 个不同的模型,分别是车型偏好模型、发动机偏好模型、车辆颜色偏好模型、客户潜在价值模型、客户忠诚度模型和客户购车价位模型。

4.5.1　人车模型指标体系建立

为了构建 6 个模型,客户指标体系的建立尤为重要。因为不同的客户有着不同的购买偏好以及价值。这些都需要依赖于客户本身的属性。当然并不是所有客户的属性都能成为指标体系。表 1 为人车模型指标体系。

表 1　人车模型指标体系

因素	分析指标	符号
客户指标体系	年龄	C_1
	兴趣	C_2
	职业	C_3
	来源	C_4
	性别	C_5
	学历	C_6
购买信息	主要用途	C_7
	维修次数	C_8
	购车时间	C_9
	二次购买	C_{10}
	重大故障问题次数	C_{11}
车型和发动机	车辆价格	C_{12}
	车型型号	C_{13}
	发动机型号	C_{14}
	车辆颜色	C_{15}

4.5.2　数据预处理与量化

(1)数据整理与清洗。

数据清洗是指发现并纠正数据文件中可识别的错误的最后一道程序,包括检查数据一致性,处理无效值和缺失值等。

(a)无效数据剔除。

客户的服务生命周期内的服务数据来源于销售系统与售后服务系统,在实际业务协同中存在客户在某企业群购买车辆却不在该企业群进行售后服务的情况。因此部分客户的服务及消费记录严重缺失。信息缺失的客户信息需要被首先剔除。

(b)缺失数据处理。

在销售及售后服务过程中,原有的业务单据存在有些非业务核心数据项没有进行验证的情况,业务处理过程中操作员出于工作便利未填写实际值,导致数据缺失。文中采取使用均值的办法进行缺失数据处理。

(c)错误数据纠正。

在销售与售后系统中,工作人员可能因为个人便利而没有认真输入正确的信息。这些信息在数据库中也没有进行验证,造成数据的不真实。大多数错误信息比较明显,所以根据常识或是均值进行修改。

(2)数据离散化。

由于大量的指标并不是数字类型,比如性别、职业等。而如果要使用机器学习来构建模型,所有的数据集必须转化为数字矩阵。为了解决这个问题,文中采

用给一些非数字指标进行编码,以及数字指标直接赋值的方法。表 2 为人车模型的数字化结果。

表 2 人车模型指标数字化结果

指标名称	指标取值
年龄	实际取值
兴趣	将所有不同兴趣按照 0 到 N 编码
职业	将所有不同职业按照 0 到 N 编码
来源	将所有不同兴趣按照 0 到 N 编码
性别	男:1 女:2
主要用途	将所有不同的使用通途按照 0 到 N 编码
学历	1:小学及以下 2:中学 3:高中及职高 4:大专及本科以上
维修次数	实际取值
购车时间	当前时间减去购车时间(精确到月)
重复购买	实际取值
重大故障问题次数	实际取值
车辆价格	实际取值
车型型号	根据不同的车型按照 0 到 N 编码
发动机型号	根据不同的发动机型号进行 0 到 N 编码
车辆颜色	根据每种颜色进行 0 到 N 编码

(3)数据规范化。

当人车模型指标数字化后,可以得到一个数字矩阵。然而为了防止某些特征值远大于其他所有特征,接下来要对数字矩阵进行归一化处理。此外由于核函数使用内积运行,如果特征数据相差太大可能会给计算带来困难。从效率和性能两方面考虑,在训练前需要进行归一化处理。

假设 $X_{\max}$ 表示每维数据的最大值, $X_{\min}$ 表示每维数据的最小值。[a,b]表示归一化数据的目标区间。一般常用[-1,1]和[0,1]。则对于矩阵的某个 X_i ,归一化计算如下:

$$\left[\frac{X_i - X_{\min}}{X_{\max} - X_{\min}}\right] \times (b - a) + a \tag{4}$$

4.5.3 模型训练、分析与应用

(1)车型偏好模型的训练与预测。

文中实现了在产业链协同平台的模型实时训练。每次训练前,通过读取数据库可以获得最新的数据。并将实时数据转化并归一化为一个数字矩阵方便训练。该模型的训练特征为(年龄,兴趣,职业,来源,性别,主要用途,学历,购车价格等)。训练集合的标签为车型型号。因为是实时训练,大约有 3 000 多个客户数据,文中采用 70% 作为训练集,剩下 30% 作为测试集。表 3 展示了部分客户数据以及转化后的数字矩阵。最终的结果是 80.1% 的准确度。

表 3 部分客户数据转化后的数字矩阵

Num	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	类别
1	30	6	1	6	3	2	1	53 000	2
2	34	2	2	11	2	1	2	42 000	1
3	35	5	1	6	6	3	2	45 500	3
4	37	3	2	5	4	10	1	24 930	4
5	29	11	1	6	1	2	3	30 800	5
6	51	1	1	3	4	2	3	35 600	6
…	…	…	…	…	…	…	…	…	…

(2)发动机偏好模型的训练与预测。

同车型偏好模型一样,发动机偏好模型也是一个分类模型。该模型主要目的是预测客户的发动机型号偏好。主要的训练特征为(年龄,兴趣,职业,来源,性别,主要用途,学历,购车价格)。而每个客户样本的 Y 值为 0-9,分别表示 10 个不同的发动机型号。最终的结果是 76.5% 的准确度。

(3)车辆颜色偏好模型的训练与预测。

车辆颜色是客户选择车辆的重要因素之一。该模型也是一个分类模型,其主要目的是预测客户对车辆颜色的喜好。主要的训练特征为(年龄,兴趣,职业,来源,性别,主要用途,学历,购车价格)。而每个客户样本的 Y 值为 0-9,分别表示 10 个不同的车辆颜色。最终的结果是 50.4% 的准确度。

(4)客户潜在价值模型的训练与预测。

与以上三个模型不同的是,客户潜在价值模型是一个回归模型。文中的主要目的是给某个客户打分。分数越高他的潜在价值越高。所以使用 SVR 来进行训练。该模型的训练特征为(年龄,兴趣,职业,来源,性别,主要用途,学历,购车价格)。为了计算每个客户的潜在价值,文中考虑到 2 个重要指标(维修次数,购车时间)。因为维修次数更多以及购车时间更早的客户更倾向于购买第二辆车。因此每个客户样本的 Y 值计算为维修次数加上购车时间。这样文中可以建立客户与潜在价值的模型,SVR 可以学习每个客户属性对客户潜在价值的影响。

然而在模型的实际训练与预测过程当中,客户的潜在价值通常并不是 0 到 100 之间。为了更好地展示客户潜在价值,使用式 5 将预测后的结果映射到 0 和 100 之间。

$$p_i = \frac{V_i}{V_{\max} - V_{\min}} \times 100 \tag{5}$$

其中, V_i 为第 i 个客户的潜在价值; $V_{\max}$ 和 $V_{\min}$ 分别为最大和最小的客户潜在价值; P_i 为最终的客户潜在价值。

此外为了在客户分类中帮助汽车经销商精准地区

分不同潜在价值的客户,定义了 4 个不同类别的客户,分别是普通客户、中等客户、一级客户和高级客户。其中普通客户的潜在价值区间为 65-75,中等客户的潜在价值区间为 75-85,一级客户的潜在价值区间为 85-90,高级客户的潜在价值区间为 90 以上。

(5)客户忠诚度模型的训练与预测。

同客户潜在价值类似,客户忠诚度模型同样也是一个回归模型。该模块功能主要给客户进行忠诚度打分。该模型的训练特征为(年龄,兴趣,职业,来源,性别,主要用途,学历,购车价格)。为了计算每个客户的忠诚度,文中考虑到 1 个重要指标(重复购买次数)。因为客户如果选择进行二次购买并且购买的车辆属于同类车型说明该客户的忠诚度较高。因此每个客户样本的 Y 值计算为重复购买次数。这样文中可以建立客户与忠诚度的模型,SVR 可以学习每个客户属性对客户忠诚度的影响。原始的客户忠诚度并没有在 0 到 100 之间,所以最终预测的客户忠诚度将会被转化为 0 到 100 之间。

(6)客户购车价位模型的训练与预测。

客户购车价位模型也是一个回归模型,它根据客户的属性来预测客户的购车价位。使用 SVR 去学习客户属性以及购车记录,主要特征为(年龄,兴趣,职业,来源,性别,主要用途,学历)。而每个客户样本的 Y 值为购车价格。

5　结束语

文中在产业链协同平台的基础上,分析了整车销售与营销的业务现状,并以 AA 企业为原型,对其销售和营销模式进行了研究。以汽车经销商和制造厂为核心,站在企业管理者的角度,分析具体的需求。由于现阶段大量的汽车制造企业和经销商想要利用信息技术和数据分析来扩展他们的销售渠道和改善他们的营销方式,所以数据分析和模型构建是文中的重点。

参考文献:

[1] 孙林夫. 网络协同制造与智能工厂[C]//网络协同制造与智能工厂学术研讨会. 成都:西南交通大学,2018:5-19.

[2] 刘小苗. 论汽车销售模式[J]. 大众科技,2005(12):224-226.

[3] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[4] DEVROYE L, GYÖRFI L, LUGOSIG. Combinatorial aspects of Vapnik-Chervonenkis theory[M]//A probabilistic theory of pattern recognition. [s. l.]: Springer, 1996: 215-232.

[5] LIN C J. A comparison of methods for multiclass support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(4): 1026-1027.

[6] MANGASARIAN O L, WILD E W. Multisurface proximal support vector machine classification via generalized eigenvalues[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2006, 28(1): 69-74.

[7] JAYADEVA, KHEMCHANDANI R, CHANDRA S. Twin support vector machine for pattern classification[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(5): 905-910.

[8] SHAO Y H, CHEN W J, HUANG W B, et al. The best separating decision tree twin support vector machine for multi-class classification[J]. Procedia Computer Science, 2013, 17: 1032-1038.

[9] KUMAR M A, GOPAL M. Least squares twin support vector machines for pattern classification[J]. Expert Systems with Applications, 2009, 36(4): 7535-7543.

[10] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2014, 18(7): 1527-1554.

[11] KANG Y, CHOI S. Restricted deep belief networks for multi-view learning[C]//European conference on machine learning and knowledge discovery in databases. Athens, Greece: Springer, 2011: 130-145.

[12] BALLESTER P, ARAUJO R M. On the performance of GoogLeNet and AlexNet applied to sketches[C]//Thirtieth AAAI conference on artificial intelligence. Phoenix, Arizona: AAAI Press, 2016: 1124-1128.

[13] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th international conference on neural information processing systems. Montreal, Canada: MIT Press, 2014: 2672-2680.

[14] 刘智斌,曾晓勤,刘惠义,等. 基于 BP 神经网络的双层启发式强化学习方法[J]. 计算机研究与发展,2015,52(3):579-587.

[15] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of Wasserstein GANS[C]//Advances in neural information processing systems. [s. l.]: [s. n.], 2017: 5767-5777.