文章编号：2095-2163(2020)06-0066-03 中图分类号：T181；F724.6 文献标志码：A

机器学习在购买意图方面的应用

刘占玉，高荣芳
（西安石油大学 计算机学院，西安 710065）

摘 要：顾客是否成功购买商品，不仅与商品本身有关，而且与顾客所处区域、类型和特殊节日有关。互联网时代，各大购物网站都有海量的顾客购买信息，因此可以通过顾客对网站的使用和操作信息，使用机器学习算法来预测顾客购买此类商品的意向。本文使用随机森林算法、SVM 算法和朴素贝叶斯算法建立模型，并采用五折交叉验证的方法选出这 3 个可靠的模型，预测顾客在线购买的可能性，最终通过准确率、召回率、F1 值、AUC 对模型进行评估。实验结果表明：随机森林更适合于在线购买意图的预测。

关键词：在线购买意图；随机森林；SVM；朴素贝叶斯；五折交叉验证

Application of machine learning in purchase intention

LIU Zhanyu, GAO Rongfang
(School of Computer Science, Xi´an Shiyou University, Xi´an 710065, China)

【Abstract】 Whether a customer successfully purchases a product is not only related to the product itself, but also related to the region, type and special festival where the customer is located. In the Internet era, all major shopping sites have a large amount of customer purchase information, so you can use machine learning algorithms to predict the customer´s intention to purchase such products through the customer´s use and operation of the website. In this paper, the random forest algorithm, SVM algorithm and Naive Bayes algorithm are used to establish the model, and the three reliable models are selected by the method of 5-fold crossValidation to predict the possibility of customers buying online. Finally, the accuracy rate, recall rate, The F1 value and AUC evaluate the model. The experimental results show that random forest is more suitable for online purchase intention prediction.

【Key words】 Online purchase intention; Random forest; SVM; Naive Bayes; 5-fold cross Validation

0 引 言

随着互联网的发展，网络购物几乎成为人们最常用的消费渠道，然而不同类型的消费者、不同的购物网站、不同促销活动和特殊日期等，都会影响消费者的购买意图。社会环境不同，使得消费者的购物需求出现了个性化和多样化，如受新冠肺炎疫情影响，大部分消费者选择网上购物。

在线购买领域，国内学者也做了很多相关研究。如：袁智慧采用实证研究的方法，来探究中 UGC 不同形式的自我披露对消费者在线购买意愿的影响机理，并分析了产品熟悉度在其中的调节作用，不仅对自我披露理论的发展起到一定的补充和深化作用，也能给社会化商务平台通过 UGC 达到商家、消费者、平台三方共赢的局面提供一定的决策支持[1]。卢美丽等人考虑在线重复购买强化效应，建立顾客重复购买通用模型[2]。Verhagen 等人针对没有研究检查在线商店信念和消费者在线情感状态之间的效果等级是否因产品类型而异。研究通过检查思考层次和感觉思考层次在解释针对搜索产品与体验产品的在线购买意向以及高参与度与低参与度产品的在线购买意向中的解释能力[3]。本文使用机器学习对电商平台的顾客在线购物数据进行分析，帮助商家更好的预测并掌握消费者的购买意向。

1 机器学习

1.1 随机森林算法

随机森林是 Leo Breiman 把随机子空间算法和集成学习算法相结合，最终得到了解决决策树过拟合问题的随机森林算法。它是一种基于树的分类器，由多棵决策树构成对样本进行训练，并预测的一种分类器。对于一棵树，训练样本采用放回式，从总的训练集中随机采样出来，而训练树的结点 $\{G_1, G_2, \cdots, G_n\}$ 时，特征是从原有特征中按照一定的比例随机地无放回式抽取的，类别的输出是由各节点预测结果来决定最优的预测结果，如图 1 所示。

1.2 SVM 算法

支持向量机是由 Vapnik 等人根据统计学理论提出的一种新的机器学习方法，是通过监督学习的方式对样本数据进行二分类的广义性分类器，它主

作者简介：刘占玉(1997-)，女，硕士研究生，主要研究方向：智能计算与可视化；高荣芳(1963-)，女，硕士，副教授，硕士生导师，主要研究方向：计算机应用技术。

收稿日期：2020-03-16

要寻找一个超平面对样本数据进行分割，让训练集样本中的数据恰好分布在超平面两侧。分割原则是间隔最大化，最终转化为一个凸二次规划问题来求解[4]。给定训练数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，其中 $x_i\in X=R^n$，$y_i\in \gamma=\{+1,-1\}$，$i=1,2,\cdots,N$，x_i 为第 i 个特征向量，y_i 为 x_i 的类标记。它最基本的想法就是在训练集 D 的样本空间中找到一个划分超平面，将不同类别的样本分开，其中样本的划分存在很多个超平面，找到一个最佳的分类超平面，如图 2 所示。

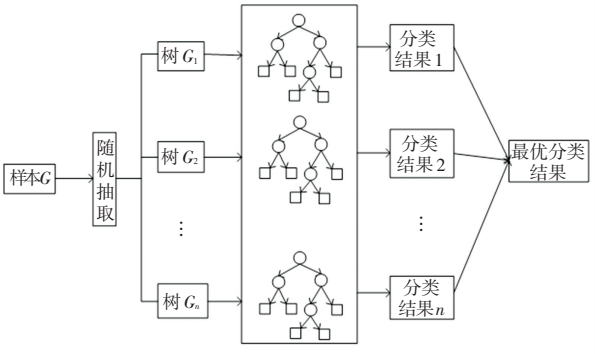

图 1　随机森林

Fig. 1　Random forest

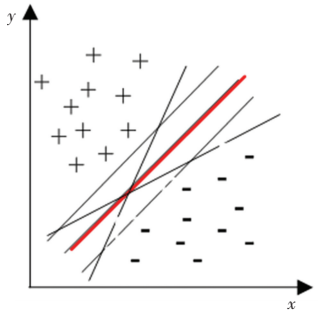

图 2　支持向量机

Fig. 2　Support vector machines

对线性不可分情况的 SVM，选择恰当的核函数 $K(x_i,x_j)$ 和恰当的参数 C，构造并求解最优问题，如公式(1)：

$$\max W(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}y_iy_ja_ia_jK(x_i,x_j). \quad (1)$$

其中：a_i 为拉格朗日乘子，$K(x_i,x_j)$ 为核函数，C 为惩罚系数。

支持向量机最终的判别函数，如公式(2)：

$$f(x)=sign\left(\sum_{i=1}^{n}a_i^*y_iK(x_i,x_j)+b^*\right). \quad (2)$$

1.3　朴素贝叶斯算法

朴素贝叶斯算法是结合贝叶斯原理和特征条件假设的分类方法。有 n 维特征向量 $X=\{x_1,x_2,\cdots,x_n\}$，类变量 $Y=\{y_1,y_2,\cdots,y_m\}$。根据朴素贝叶斯基本理论，其后验概率，如公式(3)：

$$P(Y\mid X)=\frac{P(Y)\prod_{i=1}^{n}P(x_i\mid Y)}{P(X)}. \quad (3)$$

朴素贝叶斯算法根据样本的特征 X，计算所有类别的概率，最终概率最大的类别即为该样本所属的类。$P(X)$ 是不变的，在比较后验概率时，只比较上式分子部分，得到一个样本数据属于类别 y_i 的朴素贝叶斯计算方法，如公式(4)：

$$P(y_i\mid x_1,x_2,\cdots,x_n)=\frac{P(y_i)\prod_{j=1}^{n}P(x_j\mid y_i)}{\prod_{j=1}^{n}P(x_j)}. \quad (4)$$

2　在线购买意向预测

本文实验使用 Anaconda 3 5.0.1 环境，UCI 网站公开的 Online Shoppers Purchasing Intention Dataset Data Set 数据集，该数据集包含 12330 个实例和 18 个字段，字段包括 BounceRates(跳出率)、ExitRates(退出率)、SpecialDay(特殊日期)、Region(区域)、PageValues(页面值)、VisitorType(访客类型)等，其中 Revenue 是类标签。

实验使用大部分样本数据进行模型训练，小部分数据进行模型预测。使用清洗过的数据集建立随机森林、SVM、朴素贝叶斯模型。为了选出可靠的模型，每个模型都进行 k 折交叉验证，参数 cv 设置为 3、5、10，通过实验验证，得到效果最好是 $cv=5$，即 3 个模型采用五折交叉验证，结果如表 1 所示。支持向量机模型的参数 c 表示惩罚系数，通过多次实验取得 $c=10$ 的模型训练效果最好。

表 1　模型五折交叉验证评估结果

Tab. 1　Model 5-fold cross-validation evaluation results

	五折交叉验证				
随机森林	[0.891 590 68	0.895 083 63	0.894 069 94	0.901 115 62	0.886 916 84]
SVM	[0.842 451 87	0.841 358 34	0.841 865 18	0.842 799 19	0.842 799 19]
朴素贝叶斯	[0.801 418 44	0.795 742 52	0.796 756 21	0.788 032 45	0.812 880 32]

(下转第 73 页)