Advances in Applied Mathematics 应用数学进展, 2022, 11(7), 4306-4312
Published Online July 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.117457

Hans汉斯

基于深度森林的在线课程购买行为预测研究

胡陈陈，吕卫东*，郑江怀，王一朵

兰州交通大学，数理学院，甘肃 兰州

收稿日期：2022年6月4日；录用日期：2022年6月29日；发布日期：2022年7月6日

摘　要

当前疫情环境下的生活条件与人们日益增加的教育需求，不断地推动着各种类型在线教育的发展。在传统机器学习对在线课程用户行为预测的基础上，本文采用深度学习中深度森林方法对用户的购买行为进行预测，选取准确率(Accuracy)、精度(Precision)、召回率(Recall)以及F1值作为模型评价指标，比较不同模型的预测精度。针对预测准确率最高的深度森林模型，进行多次参数调优得到最优参数模型，模型的最终准确率为98.744%。采用深度森林模型对用户购买预测，为企业进行精准营销、减少宣传投入提供重要的参考依据。

关键词

在线课程，购买行为预测，随机森林，深度森林

Research on Online Course Purchase Behavior Prediction Based on Deep Forest

Chenchen Hu, Weidong Lv*, Jianghuai Zheng, Yiduo Wang

School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou Gansu

Received: Jun. 4th, 2022; accepted: Jun. 29th, 2022; published: Jul. 6th, 2022

Abstract

The living conditions in the current epidemic environment and the increasing educational needs of people are constantly promoting the development of various types of online education. On the basis of traditional machine learning to predict online course users' behavior, this paper uses the

*通讯作者。

文章引用: 胡陈陈, 吕卫东, 郑江怀, 王一朵. 基于深度森林的在线课程购买行为预测研究[J]. 应用数学进展, 2022, 11(7): 4306-4312. DOI: 10.12677/aam.2022.117457

deep forest method in deep learning to predict users' purchasing behavior, and selects accuracy, precision, recall and F1 value as model evaluation indicators to compare the prediction accuracy of different models. Aiming at the deep forest model with the highest prediction accuracy, the optimal parameter model is obtained by multiple parameter tuning. The final accuracy of the model is 98.744%. The deep forest model is used to predict the purchase of users, which provides an important reference basis for enterprises to carry out precision marketing and reduce publicity investment.

Keywords

Online Courses, Prediction of Purchase Behavior, Random Forest, Deep Forest

1. 引言

近两年基于疫情的影响，不仅学校的授课遇到各种难题，各类线下教育机构也受到了巨大的冲击，许多机构因为生源不足面临破产的危机，相反在线教育机构在这种足不出户的环境下，拥有了很大的发展机会。基于用户过往的消费行为来预测用户未来购买的可能性，有针对的宣传，精准的营销[1]，是实现小投入高回报的目标的明智之选。

随着机器学习与深度学习的迅速发展，对于用户行为的预测分析成为国内外学者研究的热点，针对用户行为的预测，学者们已经有了许多的研究。文献[2]选择基于贝叶斯网与 Hadoop 集群相结合的系统，进行用户行为预测。文献[3]采用改进的 BP 神经网络模型对移动用户行为进行预测，在原有的 BP 神经网络基础上融合了 PSO 算法和 GA 算法，提高了预测的精度以及稳定性。文献[4]基于 CNN-LSTM 模型对用户的购买行为进行预测，相较于传统模型简化了特征选择的过程。盛钟松[5]应用 CatBoost 模型进行预测，提高了模型的精度以及鲁棒性，用户行为预测研究有了显著进展。

面对复杂多变的影响因素，传统单一的机器学习算法已经无法满足实际的需要，集成算法的发展，特别是在预测方面的应用，对于用户购买行为分析提供了一个精确的方向。深度森林作为深度学习的一种，更能够适应复杂多样的环境，对于购买行为预测有更良好的效果，文章选取深度森林作为在线课程用户购买行为预测的模型，不仅提高了预测的精度，而且相比于传统的机器学习模型来说，减少了模型预测的时间复杂度以及超参数调整的困难，是目前在用户行为预测上更为准确高效的算法。

2. 深度森林模型

深度森林[6] [7]算法是周志华教授于 2017 年提出的一种集成算法，针对于深度神经网络(DNN)超参数过多，同时预测结果过度依赖于参数的调整，基于 DNN 的深度森林就具有超参数少且无需过度调参的优势，同时模型的自适应性很强，训练的效率也较高，适用的范围也更加广泛。因此，本文采用深度森林模型预测在线课程用户的购买行为[8] [9] [10]，并通过具体的实例来证明了深度森林算法的运用有效性。

深度森林又称为 gcForest，其结构可以理解为：多粒度级联森林(Multi-Grained Cascade Forest)，深度森林模型主要由两部分结构所构成，一个是多粒度扫描技术，另一个便是级联森林结构。下面将对这两部分进行详细介绍。

2.1. 多粒度扫描

在应用的实际中，选取的数据的各个特征之间不能保证是完全独立的，内部可能存在着各种联系，尤其对于连续样本的一些预测问题，序列数据之间存在顺序上的关系，想要达到提高模型的预测精度的目标，就需要采用多粒度扫描技术，实现在提取数据特征的同时挖掘到特征顺序与预测精度之间的联系。这种技术就是为了实现对输入的样本数据进行特征扫描采样，多粒度扫描可以加强每一层级联森林对特征变量的学习效果。多粒度扫描过程见图 1：

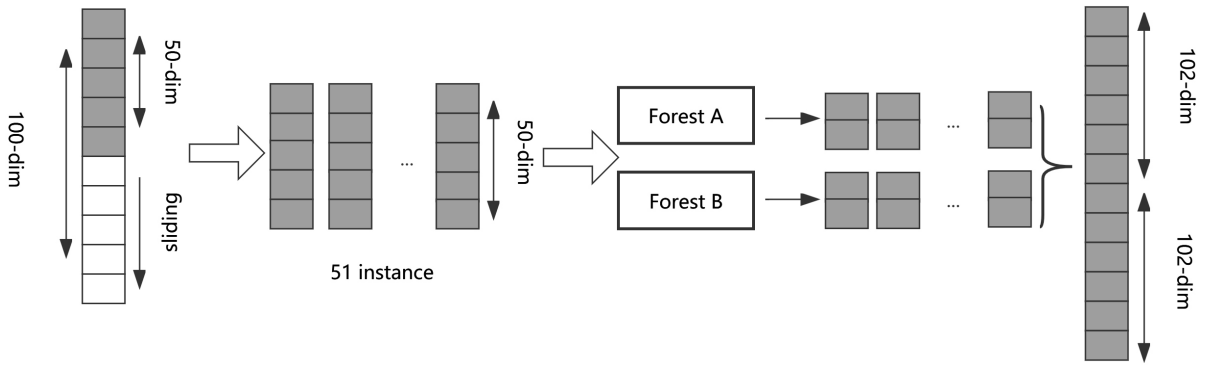

Figure 1. Multi granularity scanning process
图 1. 多粒度扫描过程

下面依图介绍多粒度扫描的过程，假定选择一个 100 维的数据作为我们的输入数据，第一步先利用滑动窗口对 100 维的数据进行滑动采样，选择 1 作为采样的步长，50 作为算法选定的滑动窗口维度。如果输入样本的特征数量为 M，滑动窗口大小为 P，采样的步长为 T，子样本数量为 Q，则有 $Q=\frac{(M-P)}{T}+1$，根据上述计算公式我们就可以得到我们所需的 51 个扫描子样本。接着分别由随机森林[11] A 和 B 对前面得到的样本来训练，因为相应的每一个子样本经过设定的森林训练就能获得一个 2 维的概率特征向量，因此对于给定的数据通过级联森林的训练就能够得到 204 维的概率特征向量，然后将所得到的 204 维的概率特征向量连接起来，作为后续级联森林的输入向量[8]。

实际在深度森林的使用过程中，滑动窗口的大小是可以根据实际需要进行调节的，而且还可以同时利用多个不同长度的窗口进行采样。也就是说，利用多粒度扫描技术可以根据所需得到多种粒度的概率特征向量，这就使得级联森林的输入包括了更多的特征，对原始数据进行了更为深层次的加工。

2.2. 级联森林

级联森林的结构很好的呈现出了深度森林在深度学习方面的思想，顾名思义，级联森林便是多个森林的联合，是由许多层的随机森林组合所得到一种层次结构，同时每一层森林有含有多个更简单的森林。这种多层多维的处理流程，对于输入数据概率特征向量的处理是非常有益的，对于所输入数据的特征表征能力具有巨大的提升效果，很大程度上会提高模型预测的准确度。

第一层森林利用第一部分多粒度滑动窗口采样获得的概率特征向量进行学习，将这一层森林输出训练结果与上面采样获得的特征向量结合起来作为新的向量，并将此向量传递给下一层森林进行学习。之后的每一层森林都对我们结合好的特征向量进行学习，直到整个算法达到了所制定的停止条件，算法就停止训练并输出最终的结果。若我们假定级联森林中的每一层包含两个随机森林和两个完全随机森林，则详细的级联森林结构如图 2 所示：

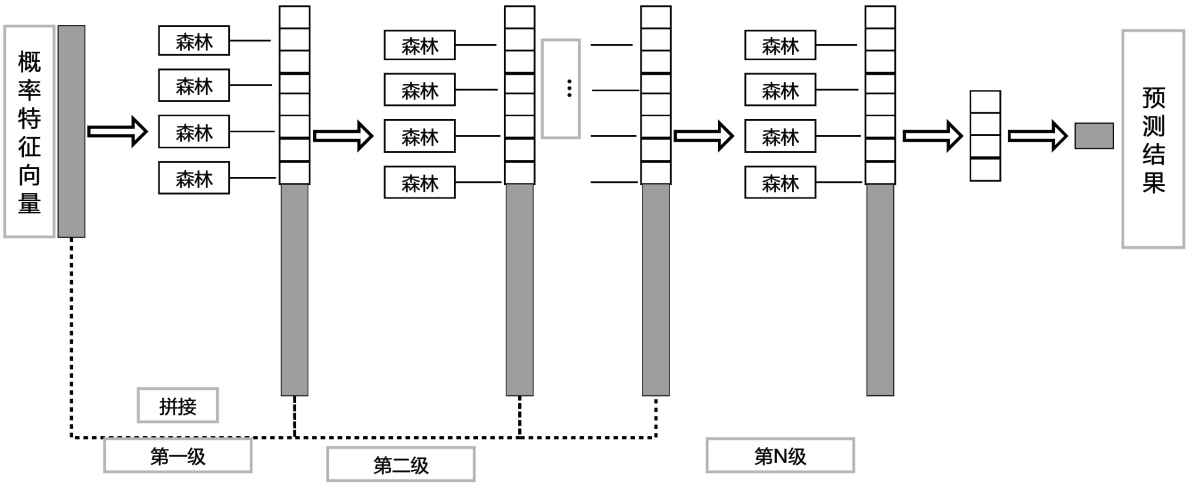

Figure 2. Cascade forest process
图 2. 级联森林过程

在上图的级联森林过程中，输入向量就是上一过程得到的最终的输出向量，在级联过程中，开始先将 204 维概率特征向量利用 4 个不一样的随机森林处理后，得到 4 个 2 维的类向量。下一步，将输出的 2 维类向量与原始概率特征向量进行组合并形成新的特征向量，用作下一级的输入特征向量[8]。反复进行上面的训练过程，一直进行到最后一级随机森林。算出最后一级随机森林输出的平均值，并将平均值中最大值对应的类别作为最终预测结果。

3. 特征分析

3.1. 数据预处理

本文使用的实验数据为 2021 年全国大学生数据统计与分析竞赛的数据集，包括用户信息表(user_info.csv)，用户登录情况表(login_day.csv)，用户访问统计表(visit_info.csv)，用户下单表(result.csv)四个部分，数据集共包含了 135,968 个用户信息。用户在信息登记过程中有信息的缺失，故对其进行预处理操作。

第一步，查看四个数据表的表结构和缺失值，具体结果见表 1，由表结果可知，用户信息表(user_info.csv)存在 28,209 个缺失值，其他数据表无缺失值。

Table 1. Data structure
表 1. 数据结构

表名	行	列	缺失值
user_info	135,968	8	28,209
visit_info	135,617	26	0
login_day	135,617	16	0
result	4639	2	0

第二步，去除完全重复的行数据，并对所有数值型数据做描述统计，通过描述统计分析，我们对数据有了初步的认识。

3.2. 特征分析

3.2.1. 用户城市分布不均

很多在线课程公司因为宣传的侧重，导致用户所在城市地区分布非常不均匀。就本文数据对用户所在城市进行统计，发现分布在重庆市的用户最多，且远远超过其他城市，用户数量排名紧随其后的四位城市分别是成都、运城、广州、北京。由此看出，用户所在的城市对用户后续是否会购买有着比较大的影响，在特征的选择上用户所在城市是重要的特征量。

3.2.2. 用户转化率较低

从用户浏览课程到付款，每个环节都会存在用户流失。部分有明确需求的用户在浏览完课程后会下单购买，但是有的用户可能在浏览之后就离开，或者试听完免费和体验课后就离开，这取决于用户的感兴趣程度以及课程设置的吸引力大小。本实验用户从注册到登录，转化率为 99.74%。用户从登陆到购买课程，转化率为 3.41%，有较多用户在过程中就流失掉了，所以在对用户最后的购买行为影响方面，整个过程中的浏览以及体验感受都十分重要。

3.2.3. 用户增量有季节变化趋势

在线课程用户下单量的多少和时间是有关系的。如初高中阶段课程就可能在寒暑假或期中考试后购买率最高，针对于成人用户的可能就会在节假日的下单率比较高一些。本文数据的用户就在 12 月份到次年的 1 月份增加量最高，也可以大体看出此课程更加倾向于成年人。因此在预测的阶段，用户登录等行为的时间特征对最后是否会下单的影响也是比较大的。

4. 模型预测

4.1. 算法流程

1) 本文先对所用数据集进行清理、去重以及合并，提高数据的质量，达到后续实验所需数据要求。

2) 将处理过后的数据按照 7:3 的比例划分为训练集与测试集，利用训练集样本对深度森林模型进行训练，得到对线上课程购买预测的模型。

3) 对上述得到深度森林模型，选用测试集样本进行预测，对得到的预测结果运用相应的评价指标进行衡量，评估模型的预测精度。

4.2. 评价指标

首先根据模型输出结果，定义如下混淆矩阵，我们定义 TP 表示购买了课程也预测为购买的用户人数，FP 为未购买但预测为购买的用户人数，FN 为实际购买但预测为未购买的用户人数，TN 表示实际未购买也预测为未购买的用户人数。根据上述定义的混淆矩阵，采用经典的准确率(Accuracy)、精度(Precision)、召回率(Recall)、F1 值以及 AUC 曲线对模型的预测精度进行评估，其中：

准确率表示整个样本预测正确的样本数量与总体数量的比值，即模型预测的准确率，计算公式如下：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{TN}} \tag{1}$$

精度表示的是预测为正例的样本中预测结果正确的样本所占的比例，计算如下：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

召回率表示的为预测正确的样本中预测结果为正例的样本所占的比例，计算公式为：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

F1 值可以理解为精度与召回率的加权平均值，计算公式为：

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

4.3. 结果分析

深度森林模型的关键是模型的建立，而在模型建立的过程中，森林的建立又是最关键的，想要得到一个精确度更高的模型，只有通过不断调整森林的各个参数来实现。在深度森林中，对模型精度影响较大的参数就是 n_estimators (每个级联层中估计器数量)，以及 n_trees (每个估计器中的决策树个数)。本实验通过调整相关参数，得到用时最短精确度最高的参数组合，其精确度评估指标选用了 Accuracy 指标。如图所示，在实验中 n_trees 选择了 50, 75, 100, 125, 150 四个数值，相应的 n_estimators 选取了 2 和 3 两个。由图 3 中变化趋势可以看出，当估计器数量为 2 时，AUC 值随着估计器中树的数目增加是增大的趋势，但当估计器数目为 3 时，随着树的数量增加，AUC 的值却不增反降，综合模型运行时间与模型的准确度，最终选择参数 n_estimators = 3，n_trees = 50，模型的准确度为 98.744%。

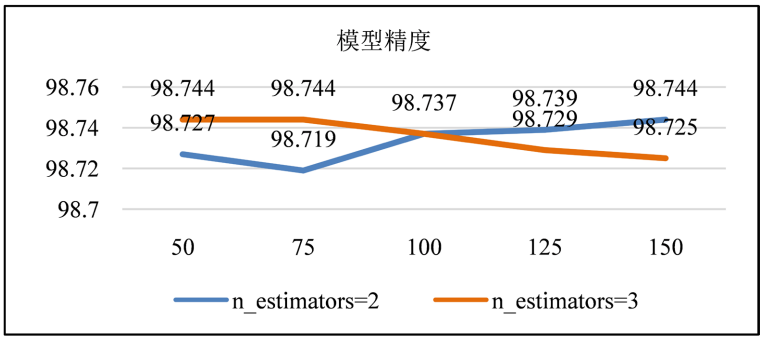

Figure 3. Model tuning
图 3. 模型调优

4.4. 模型对比

为了体现深度森林的模型预测的优越性，对上述数据又用传统的机器学习方法进行了预测，将各个模型的评价指标进行对比。传统的机器学习模型虽然运行时间较短一些，但是模型的各个评价指标都不如深度森林的结果优秀。表 2 详细数据说明深度森林整体效果更佳。

Table 2. Model comparison
表 2. 模型对比

模型	Accuracy	Precision	Recall	F1
逻辑回归	97.25%	75.38%	28.36%	41.24%
决策树	97.94%	69.86%	69.15%	69.50%
随机森林	98.52%	91.50%	64.74%	75.83%
深度森林	98.74%	85.71%	75.60%	80.34%

5. 结论

本文利用深度森林对线上课程的购买情况进行预测，经过数据的预处理、特征选择、模型建立、模型调优、模型预测等流程，利用现有数据集，对用户的行为进行分析。基于本文的预测分析结果，对在线课程公司后续对于宣传力度与营销投入的用户选择上提供了参考依据，针对预测为购买的用户可加大宣传力度，提高用户粘性，购买可能性较小的用户便可相应的减小宣传投入，节省开支，更好的达到小投入大回报的经营目标。

致　谢

本文章的顺利完成，感谢国家自然科学基金的项目支持，感谢导师吕卫东副教授的耐心指导与帮助，有了老师的教导才使得此论文能够如此顺利地完成。感谢合作伙伴郑江怀与王一朵的陪伴与支持，因为她们提供的积极活泼的氛围，让我在愉快的状态下完成了此篇论文。

基金项目

国家自然科学基金资助项目(11961040)。

参考文献

[1] 吴丽文, 蔡少霖. 基于数据挖掘的农产品精准营销路径研究——以广东省汕尾市为例[J]. 农业与技术, 2021, 41(22): 143-148. https://doi.org/10.19754/j.nyyjs.20211130040

[2] 冯辉, 邓明, 陈宝国. 基于大数据平台的用户行为预测系统设计[J]. 赤峰学院学报(自然科学版), 2020, 36(12): 17-21. https://doi.org/10.13398/j.cnki.issn1673-260x.2020.12.005

[3] 陈春玲, 陈红, 余瀚. 改进的 BP 算法对移动用户行为预测的研究[J]. 计算机技术与发展, 2018, 28(7): 178-181+186.

[4] 胡晓丽, 张会兵, 董俊超, 吴冬强. 基于 CNN-LSTM 的用户购买行为预测模型[J]. 计算机应用与软件, 2020, 37(6): 59-64.

[5] 盛钟松. 基于 CatBoost 集成算法的用户购买预测研究[J]. 现代计算机, 2021(9): 15-18.

[6] Zhou, Z.-H. and Feng, J. (2019) Deep Forest. *National Science Review*, **6**, 74-86.

[7] 夏恒, 汤健, 乔俊飞. 深度森林研究综述[J]. 北京工业大学学报, 2022, 48(2): 182-196.

[8] 张宾, 付玥, 周晶, 王帅, 李晓明. 基于深度森林的电商平台用户行为预测方法[J]. 信息技术, 2021(6): 96-101. https://doi.org/10.13274/j.cnki.hdzj.2021.06.018

[9] 葛绍林, 叶剑, 何明祥. 基于深度森林的用户购买行为预测模型[J]. 计算机科学, 2019, 46(9): 190-194.

[10] 范瑛. 基于 Python 数据挖掘实现客户预测的技术分析[J]. 信息记录材料, 2021, 22(9): 173-174. https://doi.org/10.16009/j.cnki.cn13-1295/tq.2021.09.081

[11] 董师师, 黄哲学. 随机森林理论浅析[J]. 集成技术, 2013, 2(1): 1-7.