Computer Science and Application 计算机科学与应用, 2021, 11(8), 2117-2125
Published Online August 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.118217

Hans汉斯

基于改进DeepFM的心脏病预测应用研究

张　笑，李　宁*

东北大学，辽宁 沈阳
Email: *lining80@163.com

收稿日期：2021年7月22日；录用日期：2021年8月17日；发布日期：2021年8月24日

摘　要

近年来，心脏病在全球已严重威胁到人类的身体和生命健康安全，通过利用人工智能等技术手段来辅助医疗诊断的科学技术日益普遍，为提高心脏病诊断的准确性，本文提出了一种在DeepFM模型的基础上改进后的较为新颖的模型——RDF模型。RDF模型由三个组件共同构成，其中因子分解机对低阶特征交互进行建模，BP神经网络对高阶特征交互进行建模，集成树则进一步提高模型的准确性和稳健性。本文在UCI数据集中的303个心脏病样本上进行实验，实验结果显示AUC值为0.8809，准确率为0.8317。

关键词

心脏病，因子分解机，前馈神经网络，集成树

Applied Research on Heart Disease Prediction Based on Improved DeepFM

Xiao Zhang, Ning Li*

Northeastern University, Shenyang Liaoning
Email: *lining80@163.com

Received: Jul. 22nd, 2021; accepted: Aug. 17th, 2021; published: Aug. 24th, 2021

Abstract

In recent years, heart disease has been a serious threat to human life and health safety, and the technology of medical diagnosis assisted by artificial intelligence is becoming more and more common. In order to improve the accuracy of heart disease diagnosis, based on DeepFM model, this paper proposes a novel model—RDF model. The RDF model is composed of three components: Factor Machine is used to model the low-order feature interaction, the BP neural network is used to

*通讯作者。

文章引用: 张笑, 李宁. 基于改进 DeepFM 的心脏病预测应用研究[J]. 计算机科学与应用, 2021, 11(8): 2117-2125.
DOI: 10.12677/csa.2021.118217

model the high-order feature interaction, and the integration tree is used to further enhance the accuracy and robustness of the model. The experiment was performed on 303 heart disease samples from the UCI datasets. Experimental results show that the AUC value is 0.8809 and the accuracy is 0.8317.

Keywords

Heart Disease, Factorization Machine, Feedforward Neural Network, Integrated Tree

1. 引言

在人口结构失衡和年轻人不良的饮食习惯和作息规律的影响下，中国患心血管疾病的人数越来越多，这不得不促使我们增加对心脏病诊断的关注和干预[1]。随着海量数据的挖掘、人工智能的迅猛崛起，数据的规模不断膨胀，分析处理数据的方式也在不断更新，在此背景下，人们仅依靠自身经验与猜想假设去探索未知领域，又或者以样本来推断总体情况是远远不够的，我们已然步入了一个新兴的时代——大数据时代[2]。人工智能与医疗健康领域的融合不断促进了医学诊断的创新和进步。在机器学习和深度学习的理论支撑下，为提高医学诊断的精准性，许多学者以数据挖掘技术来辅助心脏病的诊断。

国内外对于心脏病预测领域的研究已日渐成熟。早在 2007 年，Emre Çomak 等人利用最小二乘支持向量机对心脏瓣膜疾病进行分类[3]；王阶等人将逻辑回归算法应用到冠心病的诊断当中[4]。2008 年，陈天华等人在冠心病分类中运用了 BP 神经网络算法[5]。2017 年，王莉莉等人针对心脏病样本不平衡提出了一种改进的 AdaBoost 算法[6]。Rui Guo，逢凯，Indu Yekkala 等人先后将随机森林或优化后的随机森林模型应用于心脏病诊断中[7] [8] [9] [10] [11]。

DeepFM 是推荐系统领域中比较成熟的用于点击率预测的模型，它是在 2017 年由 Huifeng Guo 等人提出的，DeepFM 模型的核心思想是集成因子分解机和深度学习两部分，形成一种新的神经网络架构，进行特征学习，解决实际问题[12]。陈一文将融合了 GBDT 的 DeepFM 模型应用于 CTR 的预估，具有实际的探索意义[13]。DeepFM 模型是一个继 Wide & Deep 模型后改进的更为高效的模型。2016 年由 Heng-Tze Cheng 等人提出的 Wide & Deep 模型中 Wide 代表广义线性模型，通过特征交叉实现记忆能力，Deep 代表前馈神经网络，通过生成没有出现过的高维特征提高泛化能力[14]。Wide & Deep 模型结合了 LR 和 DNN 两部分，但是由于 LR 仍需人工特征工程实现特征交叉，DeepFM 模型便以 FM 代替 LR，且和 DNN 共享同一个输入，使模型更为高效。DeepFM 模型的思想在理论上不仅可以用于点击率的预测，还能应用于具有可操作性的各种分类问题。因此，本文利用改进后的 DeepFM 模型对是否患心脏病进行分类，进一步探索数据挖掘对医疗诊断干预的重要性。

本文的工作安排如下：1) 对 DeepFM 模型结构和内容的详细展开；2) 介绍了集成学习之随机森林模型的特点和步骤；3) 在基础模型之上提出 RDF 模型，给出模型架构和输出表达；4) 实验数据的介绍和预处理；5) 将 RDF 模型应用到 UCI 心脏病数据中，并与多个模型的实验结果进行对比，根据评价指标对多组实验结果展开分析；6) 指出本文的探索意义以及有待改进的方向。

2. DeepFM 模块

DeepFM 模型是在 Wide & Deep 模型的基础上提出的用来解决点击率预估的分类模型[12]。DeepFM

模型不需要进行特征工程，能够实现端到端的训练，大大提高了训练的效率。DeepFM 结合了 FM 和 DNN 两部分，分别用来捕获低阶特征组合和高阶特征组合。先对输入的原始特征中的离散特征进行 one-hot 编码，再把编码后的稀疏特征做 Dense Embedding，转换为低维稠密向量，一起馈送到后续的神经网络架构中。在经过两部分的低阶高阶特征捕获后合并输出。

2.1. FM

FM 即 Factor Machine，因子分解机[15]。相比 Poly2 模型，FM 使用各自特征隐向量的内积替换单一的权重作为特征组合的系数这一做法不仅解决了数据稀疏性的问题，还大大减少了需要学习的参数的个数。FM 通过对特征进行建模来学习一阶特征和捕获二阶特征交互。FM 的输出由两部分组成：

$$y_{FM} = \langle w \cdot x \rangle + \sum_{j_1=1}^{d} \sum_{j_2=j_1+1}^{d} \langle V_i \cdot V_j \rangle x_{j_1} \cdot x_{j_2}$$

其中，$w \in R^d$，$V_i, V_j \in R^k$ 是隐向量，k 是给定的隐向量的维度，一般而言，$k \ll d$，$\langle V_i \cdot V_i \rangle$ 是特征 i 和特征 j 的两个维度为 k 的向量的内积。公式中的第一部分表示一阶特征的影响，第二部分强调二阶特征交互的重要性。

2.2. DNN

DNN 组件是 BP 神经网络，可以捕获高阶特征组合[12]。DNN 的输入是与 FM 共享的 enbedding 层。embedding 层的输出为

$$a^{(0)} = [e_1, e_2, \cdots, e_m]$$

其中 e_i 是第 i 个域的嵌入向量，m 是域的数量，$a^{(0)}$ 是前馈神经网络的输入，前向传播过程可以表示为：

$$a^{(l+1)} = \sigma\left(W^{(l)} a^{(l)} + b^{(l)}\right)$$

其中，σ 是激活函数，l 是隐藏层的层数，$a^{(l)}$、$W^{(l)}$、$b^{(l)}$ 分别是第 l 个隐藏层的输出、权重和偏置。假设网络中有 $|H|$ 个隐藏层，则 DNN 输出部分的结果为

$$y_{DNN} = \sigma\left(W^{|H|+1} a^H + b^{|H|+1}\right)$$

3. 集成树模块

Bagging 算法是一种并行式集成学习算法。Bagging 是基于每个划分后的样本训练出一个基学习器，每个基学习器输出一个分类结果，所有的分类结果通过投票机制得到最后归属类别。

随机森林(Random Forest，简称 RF)是集成学习中一种特殊的 Bagging 方法，它的两大特点是：

1) RF 由很多的决策树组成。决策树的生成算法有 ID3，C4.5 和 C5.0 等，它是通过自上而下的递归方法形成的一个树形结构，基本思想是以信息熵为判断标准构造一棵信息熵下降最快的树，其中每个内部节点表示在一个特征属性上的判断，每个分支输出一个判断结果，每个叶子节点存放一个分类结果。由多颗决策树组成的 RF 通过简单投票法得到最终分类结果。

2) RF 在训练决策树的过程中使用了 Bootstrap 方法，随机生成 m 个训练集，在节点分裂时，使用随机属性选择。Bootstrap 方法产生的这种样本扰动和属性扰动增加了 RF 中基学习器的多样性[16]。由于训练出的每棵决策树的差异性，最终集成的泛化能力也因此得到提高；因其并行式的训练方法，训练速度更有优势。随机森林算法的流程图见图 1。

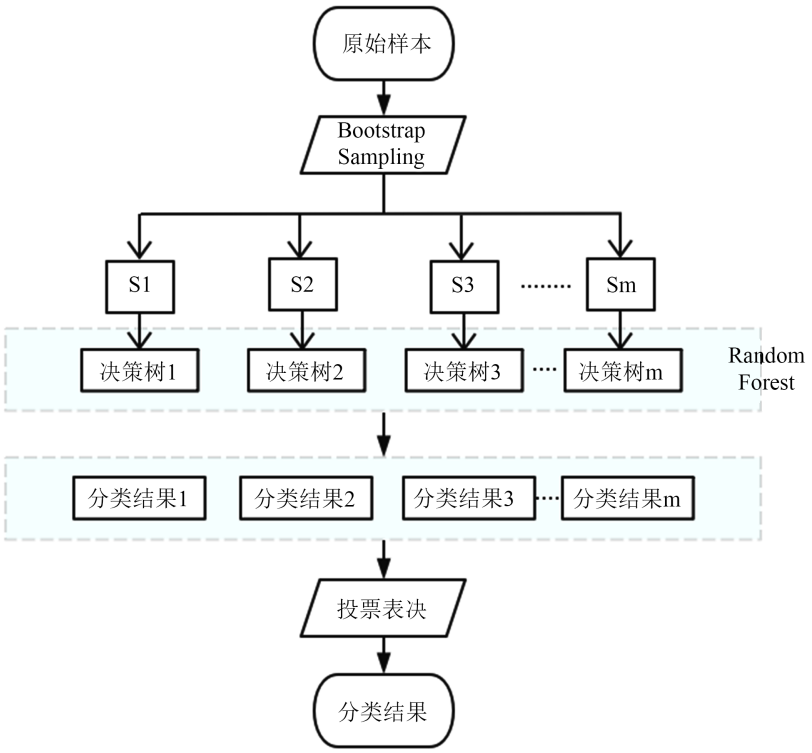

Figure 1. Flow chart of random forest algorithm
图 1. 随机森林算法流程图

4. 模型组件组合

模型整体结构如图 2 所示：

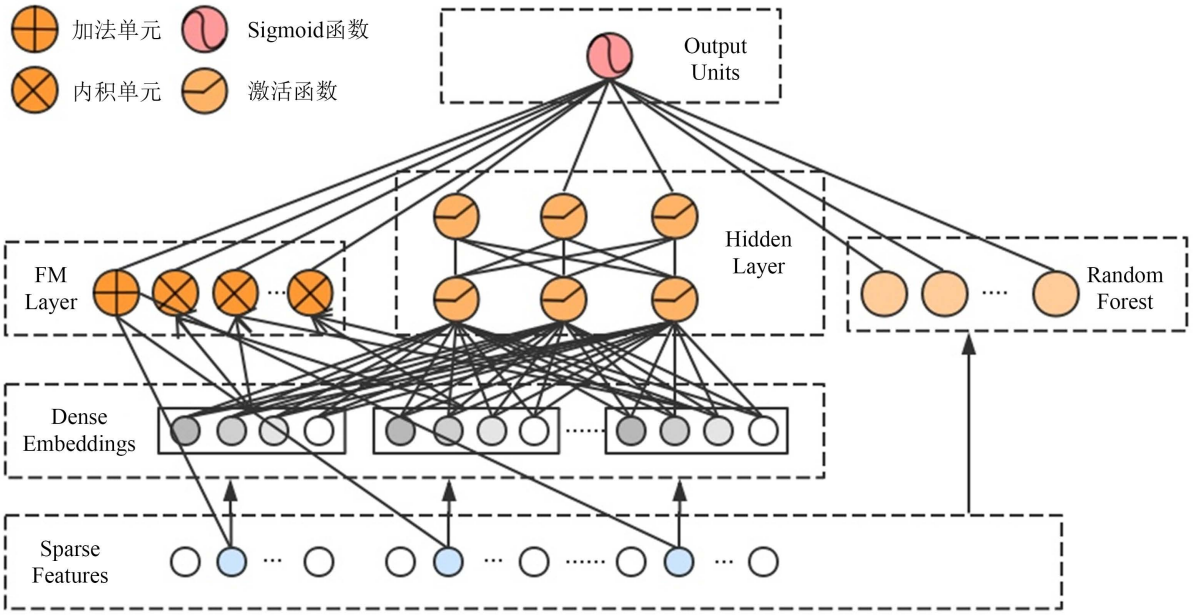

Figure 2. Modular block diagram of model components
图 2. 模型组件组合框架图

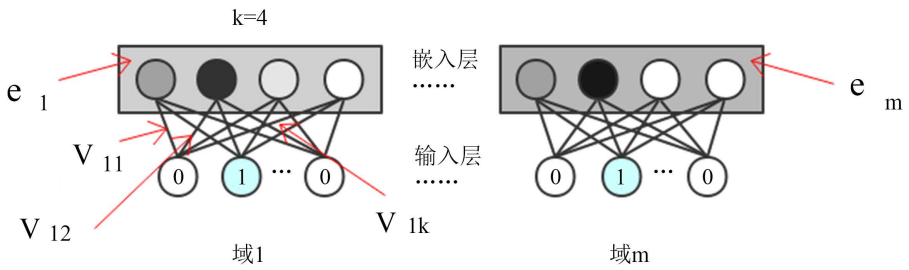

Figure3. Embedded layer structure diagram
图 3. 嵌入层结构图

图 3 详细展示了模型组合中从输入层到嵌入层的转化形式。我们把进行 one-hot 之前的每个特征都看作一个域，one-hot 后的各个特征维度是不同的，这个嵌入过程的独特之处就在于通过 embedding vector 把 one-hot 之后维度不同的域转换成了维度相同且都为 k 的 embedding 向量。这样不仅能够解决 one-hot 带来的特征稀疏化问题，还能减少模型学习的参数数量。这里的 embedding vector 不需要由 FM 进行预训练用做初始化，而是以端到端的方式联合训练整个网络。

DeepFM 组件和随机森林组件通过绘制学习曲线自适应地学习组合的权重，以达到模型组合下最优的预测准确率，这种自适应调节权重的方法不仅能达到最优目标，还能增强模型的稳定性，提高泛化能力。

组合后模型的整体输出表达式为：

$$\hat{y} = sigmoid\left(y_{FM} + y_{DNN} + y_{RF}\right)$$

将组合后的模型称为 RDF (Random Forest & DeepFM)模型，此模型无需进行复杂的特征工程，训练效率高，泛化能力强，且模型的解释能力强。

5. 实验数据

5.1. 数据来源

该实验数据选自 UCI 数据集中的 Heart Disease Data Set，来源于克利夫兰基金会。数据集包括 303 例样本，每例样本包括 13 个生理监测指标，一个心脏病分类指标。数据集属性见表 1。

Table 1. Data set attribute
表 1. 数据集属性

特征名称	特征描述	特征类型
age	年龄	连续型
sex	性别(1 = 男性；0 = 女性)	离散型
cp	胸痛类型(0：典型的心绞痛；1：非典型心绞痛；2：非心绞痛；3：无症状)	离散型
trestbps	静息血压，以毫升 Hg 计	连续型
chol	血清胆固醇水平(mg/dl)	连续型
fbs	空腹血糖(>120 mg/dl，1 = 真；0 = 假)	离散型
restecg	静息心电图结果(0：正常；1：有 ST-T 波异常；2：左心室肥大)	离散型
thalach	达到的最大心率	连续型
exang	运动引起的心绞痛(1 = 是；0 = 否)	离散型

Continued

oldpeak	运动相对于休息引起的 ST 压低	连续型
slope	最高运动 ST 段的斜率(1：上坡；2：平坦；3：下坡)	离散型
ca	被荧光显色的主要血管数目(0~3)	连续型
thal	一种称为地中海贫血的血液疾病(1 = 固定缺陷；2 = 正常；3 = 可逆缺陷)	离散型
target	心脏病(0 = 否，1 = 是)	离散型

5.2. 数据预处理

首先处理缺失值，由数据信息可知该数据集属于任意缺失类型，对缺失值的处理分为删除和填充两种，填充方法大致有 K 近邻法、均值插补、多重填补和特殊值填充等。由于缺失值数量少且考虑数据的完整性，本文将样本中缺失的属性归为同一类，用新的数字标记。

再处理离散型数据，离散型数据中的定类数据是按照事物属性进行分类，类所表示的数据大小并没有实际意义，不能进行计算和大小比较，在将此类数据输入到模型之前需将不同数据大小所代表的属性形式转换成同等地位的形式。所以对此进行 one-hot 编码，one-hot 编码是将分类变量转换为机器学习算法易于读取的形式的过程，例如，我们将性别特征[“男”，“女”]转换成向量形式，[1, 0]代表“男”，[0, 1]代表“女”。one-hot 编码后特征维度由原来的 13 扩增到 26。

由于各生理监测指标的属性不同，通常具有不同的量纲。当不同特征之间的量纲差异很大时，若直接使用原始数据建立模型，则会突出数值较高的特征在问题分析中的作用，相对减弱数值较低特征的作用。因此，为了使特征之间具有可比性、确保结果的真实性，本文对数据作标准化处理。本文选用标准差标准化，z-score 标准化过程如下：

$$x' = \frac{x - \bar{x}}{s}$$

其中 $\bar{x}$ 是均值，s 是标准差。

6. 实验结果

ROC 曲线，又称接受者操作特征曲线，它是根据不同阈值组成的混淆矩阵(见表 2)画出的。ROC 曲线是评估一个模型好坏的综合指标，反应模型的预测能力。一个二分类模型的阈值可以在 0 到 1 之间任意取值，每个阈值对应一组假正率(FPR)和真正率(TPR)。其中，TPR 表示分类器预测为正例的正样本占实际正例样本数量的比例，FPR 表示分类器预测为正例的负样本占实际负例样本数量的比例。ROC 曲线就是以 FPR 为横坐标，以 TPR 为纵坐标，由不同阈值所对应的点(FPR, TPR)连成的曲线。ROC 曲线下的面积被称作 AUC，它表示预测的正例排在负例前面的概率。AUC 值越大，分类的正确率和稳健性越高。Accuracy 表示被正确分类的样本占所有样本的比例。

以上定义如下：

Table 2. Confusion matrix
表 2. 混淆矩阵

Predict Actual	0	1
0	TN	FN
1	FP	TP

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$$

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}}$$

将本文的模型应用到心脏病数据集中，在混淆矩阵中的表现如图 4 所示，由图可知，模型将测试集中的 61 个样本正确分类个数为 51 个，将 4 例健康的样本误判为患心脏病，将 6 例患心脏病的样本误判为健康，模型的总体正确率大致为 83%。ROC 曲线如图 5 所示，得到的 AUC 值大致为 0.881。

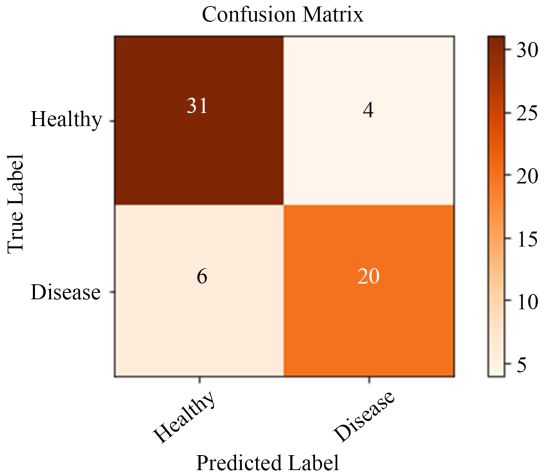

Figure 4. The confusion matrix of RDF model
图 4. RDF 模型的混淆矩阵

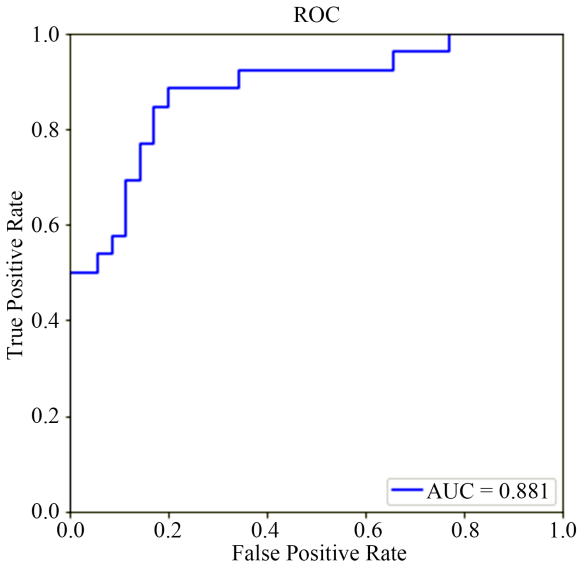

Figure 5. ROC curve
图 5. ROC 曲线图

不同算法在该测试集上的实验结果如表 3 所示，本文提出的 RDF 模型的预测准确率均高于 RF 模型、

FM 模型、DNN 模型以及衍生的 DeepFM 模型，分类效果较好。RDF 模型的 AUC 值也高于 DeepFM 模型，由此可见，融合集成树算法能提高模型的稳健性。

Table 3. The realization result graph of different algorithms
表 3. 不同算法的实验结果图

	FM	DNN	RF	DeepFM	RDF
AUC	0.8505	0.8838	0.9077	0.8696	0.8809
Accuracy	0.7771	0.8049	0.8033	0.8038	0.8317

7. 结束语

为了进一步提高心脏病预估的准确性，本文设计了一种融合了 DeepFM 和随机森林的模型，其中因子分解机部分学习稀疏化特征后的低阶特征交互，前馈神经网络部分能够学习高阶特征交互，再融合随机森林以提升模型的准确性和稳定性。通过实验表明，该模型的确在心脏病数据集上表现更优，具备解决心脏病诊断问题的实际能力。由此可见，不论是从提高模型预估准确性上还是模型的可解释性角度，本文的探索都是具有实际意义的。

本文提出的 RDF 模型在虽然此心脏病数据集上取得较好的实验结果，但仍存在很多需要补充和改进的地方：1) RDF 模型没有在更多数据集上得到验证；2) 后续可以在特征约简等工作上进一步展开研究；3) 下一步可以尝试将更多深度学习的思想引入到模型中，例如注意力机制，以进一步提高模型的分类准确率。

参考文献

[1] 马丽媛, 吴亚哲, 陈伟伟. 《中国心血管病报告 2018》要点介绍[J]. 中华高血压杂志, 2019, 27(8): 712-716.

[2] 秦文哲, 陈进, 董力. 大数据背景下医学数据挖掘的研究进展及应用[J]. 中国胸心血管外科临床杂志, 2016, 23(1): 55-60.

[3] Çomak, E., Arslan, A. and Turkoglu, İ. (2007) A Decision Support System Based on Support Vector Machines for Diagnosis of the Heart Valve Diseases. *Computers in Biology and Medicine*, **37**, 21-27. https://doi.org/10.1016/j.compbiomed.2005.11.002

[4] 王阶, 李军, 姚魁武, 衷敬柏. 冠心病心绞痛证候要素和冠脉病变的 Logistic 回归分析[J]. 辽宁中医杂志, 2007(9): 1209-1211.

[5] 陈天华, 郑彧, 韩力群, 唐海滔. 基于神经网络的冠心病无创诊断方法研究[J]. 航天医学与医学工程, 2008, 21(6): 513-517.

[6] 王莉莉, 付忠良, 陶攀, 胡鑫. 基于主动学习不平衡多分类 AdaBoost 算法的心脏病分类[J]. 计算机应用, 2017, 37(7): 1994-1998.

[7] Guo, R., Wang, Y.Q., Yan, H.X., *et al.* (2015) Analysis and Recognition of Traditional Chinese Medicine Pulse Based on the Hilbert-Huang Transform and Random Forest in Patients with Coronary Heart Disease. *Evidence-Based Complementary and Alternative Medicine*, **2015**, Article ID: 895749. https://doi.org/10.1155/2015/895749

[8] 逄凯. 三种机器学习方法在冠心病筛查中的比较研究[D]: [硕士学位论文]. 长春: 吉林大学, 2016.

[9] Yekkala, I. and Dixit, S. (2018) Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection. *International Journal of Big Data and Analytics in Healthcare*, **3**, 1-12. https://doi.org/10.4018/IJBDAH.2018010101

[10] 赵金超, 李仪, 王冬, 张俊虎. 基于优化的随机森林心脏病预测算法[J]. 青岛科技大学学报(自然科学版), 2021, 42(2): 112-118.

[11] Madhumita, P. and Smita, P. (2021) Prediction of Heart Diseases Using Random Forest. *Journal of Physics*: *Conference Series*, **1817**, Article ID: 012009. https://doi.org/10.1088/1742-6596/1817/1/012009

[12] Guo, H.F., Tang, R.M., Ye, Y.M., *et al.* (2017) DeepFM: A Factorization-Machine Based Neural Network for CTR

Prediction.

[13] 陈一文. 一种改进的基于 DeepFM 算法的高效 CTR 预估方法[D]: [硕士学位论文]. 长春: 吉林大学, 2020.

[14] Cheng, H.-T., Koc, L., Harmsen, J., *et al*. (2016) Wide & Deep Learning for Recommender Systems. *Proceedings of the* 1*st Workshop on Deep Learning for Recommender Systems*, Boston, 15 September 2016, 7-10. https://doi.org/10.1145/2988450.2988454

[15] Rendle, S. (2010) Factorization Machines. *IEEE International Conference on Data Mining*, Sydney, 13-17 December 2010, 995-1000. https://doi.org/10.1109/ICDM.2010.127

[16] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 180.