Advances in Applied Mathematics 应用数学进展, 2022, 11(8), 5171-5182
Published Online August 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.118542

Hans汉斯

深度神经网络和Bagging分类方法的随机模拟与实证研究

李唯欣，郭磊磊

北方工业大学理学院，北京

收稿日期：2022年7月3日；录用日期：2022年7月29日；发布日期：2022年8月5日

摘　要

本文利用随机模拟实验和实证研究，考虑了深度神经网络分类和Bagging分类两种机器学习方法的建模和预测。首先，本文通过对比两种模型定义、相关的联系和区别，给出了两种方法各自的适用场景及优势和不足。之后随机生成非线性数据集进行随机模拟，结果表明，神经网络分类模型具有较高的精确度。进一步，本文选取较优的神经网络分类模型对一个实际数据集：鸢尾花分类数据集进行实证研究，并进行相关预测，结果表明，神经网络分类模型具有较高的准确度与预测精度，对于分类器的选择问题具有重要的现实意义。

关键词

深度神经网络分类，Bagging分类，随机模拟，实证研究

Stochastic Simulation and Empirical Research of Deep Neural Networks and Bagging Classification Methods

Weixin Li, Leilei Guo

School of Science, North China University of Technology, Beijing

Received: Jul. 3rd, 2022; accepted: Jul. 29th, 2022; published: Aug. 5th, 2022

Abstract

This paper uses stochastic simulation experiments and empirical research, and considers the modeling and prediction of two machine learning methods: deep neural network classification and Bagging classification. First, by comparing the definitions, related connections and differences of the two models, this paper presents the applicable scenarios and advantages and disadvantages

文章引用: 李唯欣, 郭磊磊. 深度神经网络和 Bagging 分类方法的随机模拟与实证研究[J]. 应用数学进展, 2022, 11(8): 5171-5182. DOI: 10.12677/aam.2022.118542

of the two methods. Then the nonlinear dataset was randomly generated for stochastic simulation, which showed that the neural network classification model has high accuracy. Further, this paper selects an excellent neural network classification model to conduct an empirical research on a real data set: Iris classification data set, and makes relevant predictions. The results show that the neural network classification model has high accuracy and prediction accuracy, which has important practical significance for the problem of classifier selection.

Keywords

Deep Neural Network Classification, Bagging Classification, Stochastic Simulation, Empirical Research

1. 引言

1.1. 研究背景及意义

随着大数据时代的到来，机器学习算法作为其先进工具之一也逐步得到了广泛应用，此外，人工智能的发展衍生出一种新型领域：数据挖掘，即从大量的数据中通过算法提取出内在的、隐藏的信息，且能够产生较大的价值。作为其重要分支之一，数据分类可对数据进行训练分类等一系列操作，进而提取出简洁、准确度较高、可被公众理解的知识。目前分类算法多种多样，比较常用的有贝叶斯(Bayes)分类算法、决策树分类法、随机森林(RF)、基于支持向量机(SVM)的分类器、神经网络法(MLP)、K-近邻法(KNN)等等。但分类器的优劣又会对结果的可靠性与准确度产生不同的影响，所以如何选取分类器进行数据的合理分类始终是一个至关重要的问题。

1.2. 文献综述

1.2.1. 国内文献综述

杨剑锋[1]对机器学习算法进行总结，比较多种分类算法的核心思想、优缺点以及实际应用。朱虎明[2]等人对神经网络(DNN)的发展历程与结构进行了详细的概述，并深入研究其模型数据并行性能。姚明海[3]等人对 Bagging 分类算法进行特征选择，保证了更高的预测精度。施启军[4]等人在 weka 自带鸢尾花数据进行分类对比试验，发现深度神经网络(DNN)的分类效果准确率最高，分类性能最好。马景义[5]对比随机森林与 Bagging 分类树，发现前者优于后者。任涛[6]采用 AIC 算法进行自动识别，并利用 Bagging 机器学习算法对地震事件性质进行区分。

1.2.2. 国外文献综述

Piotr F. J. Lipiński [7]引入工具对数据进行随机化，并针对其物理意义对 QSAR 模型中参数进行优化调整。Bhushan Shashi [8]在随机无响应情况下，利用仿真方法对种群方差进行最优估计。Salahuddin Zohaib [9]讨论了医学成像分析深层神经网络(DNN)可解释性的局限性，为使用可解释性方法提供了指南和未来发展。Sharafati Ahmad [10]在目前的研究中，提出了一种新开发的称为打包回归(BGR)的集成智能预测模型，并根据经典支持矢量回归(SVR)和决策树回归(DTR)模型验证了该模型。Manzanarez-Ozuna E [11]开发了一种基于遗传算法的自动搜索方法，以找到一个基于深层神经网络(DNN)并适合生物数据集的预测模型，并且还通过优化参数构建了最佳的 DNN 架构。

本文将探究深度神经网络和 Bagging 分类方法对数据进行分类的优劣。就分类问题而言，实践表明，深度神经网络分类的准确度高，对噪声神经具有较强的容错能力，而 Bagging 回归能够处理不相关的特征，基于此，本文将对两种分类算法利用同一数据集进行随机模拟，深入探究其分类的准确度，在选取并基于给定的同一实际数据集：鸢尾花数据，进行实证研究，将两者进行对比分析，最终结合相关结论探究两者的优劣与其适用的具体条件。对于分类器的选择问题具有重要的现实意义。

2. 模型介绍

2.1. 深度神经网络(DNN)分类

2.1.1. 深度神经网络分类(DNN)基本结构

深度神经网络(DNN)具有很多隐藏层的神经网络，有时也叫做多层感知机(MLP)。从不同层的相关位置来看，可以将 DNN 划分为三层：输入层、隐藏层和输出层，如图 1 所示。可以看到，第一层为输入层，中间所有层为隐藏层，最后一层为输出层。

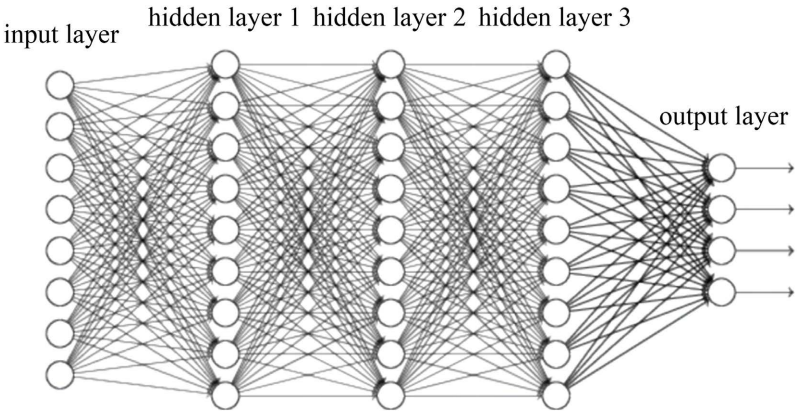


Figure 1. Basic structure diagram of DNN
图 1. DNN 基本结构图

DNN 的所有层之间都有一定联系，并且局部模型是一个线性关系加上一个激活函数 $\sigma(z)$。但是，输入层是没有这两个参数的。其中 w 为两层之间的线性相关系数，b 为偏倚，如图 2 所示，对于参数 w，上标表示层数，下标分别代表输出层(图中第三层)的索引 2 和输入层(图中第二层)的索引 4。对于偏倚 b，如图 3 所示，上标代表层数 2，下标代表所在神经元的索引 3。

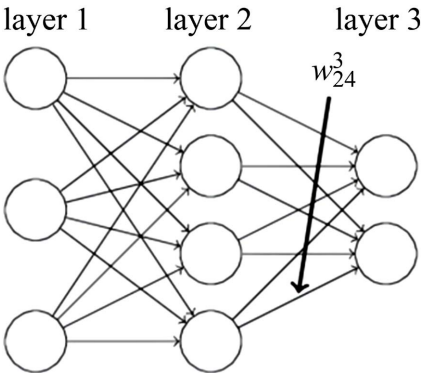


Figure 2. Parameter figure
图 2. 参数图

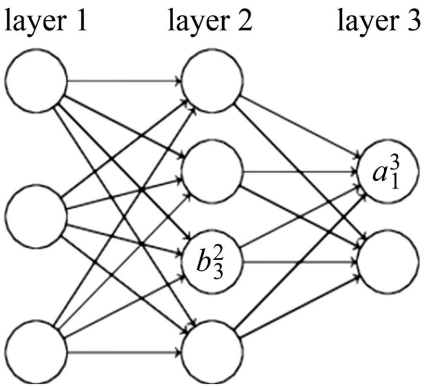


Figure 3. Plot of the bias coefficient
图 3. 偏倚系数图

2.1.2. 深度神经网络(DNN)算法原理

1) DNN 前向传播算法

首先假设选择的激活函数是 $\sigma(z)$，隐藏层和输出层的输出值为 a。可以利用上一层的输出计算下一层的输出，一直到计算到输出过程结束，具体过程展示见图 4，则用矩阵形式表示第 1 层的输出即为：

$$a^l = \sigma\left(z^l\right) = \sigma\left(W^l a^{l-1} + b^l\right) \tag{2.1}$$

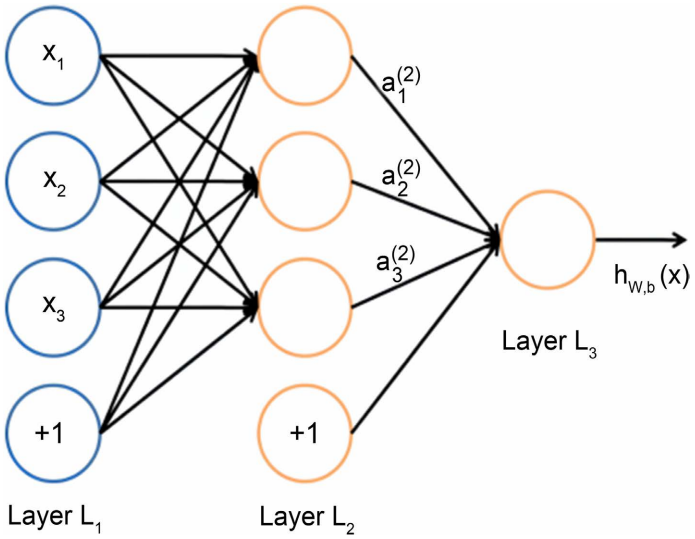


Figure 4. DNN forward propagation algorithm diagram
图 4. DNN 前向传播算法图

2) DNN 反向传播算法

通过对损失函数用梯度下降法进行迭代优化求极小值，在各层中找到最合适的参数。使用均方误差来度量损失，可以得到最小化式子：

$$J\left(W,b,x,y\right) = \frac{1}{2}a^L - y_2^2 \tag{2.2}$$

2.1.3. DNN 常用的激活函数

1) sigmoid 函数如图 5 所示。

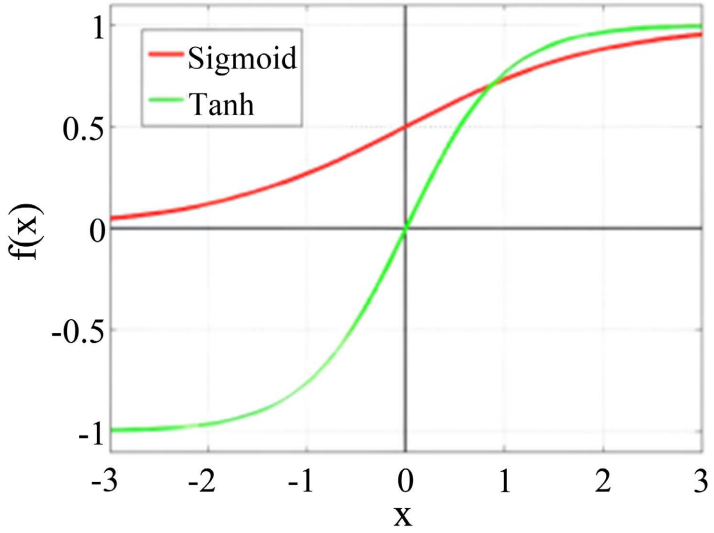


Figure 5. Sigmoid function figure
图 5. Sigmoid 函数图

2) tanh：sigmoid 的变种，表达式为：

$$\tanh(z)=\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \tag{2.3}$$

tanh 激活函数和 sigmoid 激活函数的关系为：

$$\tanh(z)=2\text{sigmoid}(2z)-1 \tag{2.4}$$

3) ReLU：$\sigma(z)=\max(0,z)$

4) PReLU：可将激活值进行一定幅度的缩小。

2.1.4. 深度神经网络分类效果

优点：深度神经网络分类可以使用统计学习方法从原始感官数据中提取高层特征，进而可以在大量的数据中获得输入空间的有效表征，并进行定量预测，准确度高，并且对噪声神经具有较强的容错能力。

缺点：若数据集太小，则深度神经网络很容易产生过拟合，进而得到比较差的结果。并且针对于没有局部相关特性的数据集，利用深度神经网络分类也无法达到较好的效果。

2.2. Bagging 分类

2.2.1. Bagging 分类主要思想

Bagging 又称自助聚集，是最早和最基本的集成技术之一，最初由 Leo Breiman 在 1996 年提出，其采用集成学习思想组合多个弱分类器的泛化性能，被认为是性能较好的分类方法。个体学习器之间不存在很强的依赖关系，一系列的个体学习器可以通过随机采样并行生成，然后使用结合策略，得到最终的集成模型，如图 6 所示，可以概括为先构建、后结合。一般常用的方法有学习法、带权投票法、投票法。

该算法的思想是使学习算法训练轮次，每轮训练集由 M 从初始训练中随机抽取一个训练样本，某个初始训练样本浓度可以在一轮训练中多次出现或没有出现(即回到抽样上)，训练序列后可以得到一个预测函数，最终的预测函数对分类问题进行表决，采用简单平均法对回归问题进行识别。

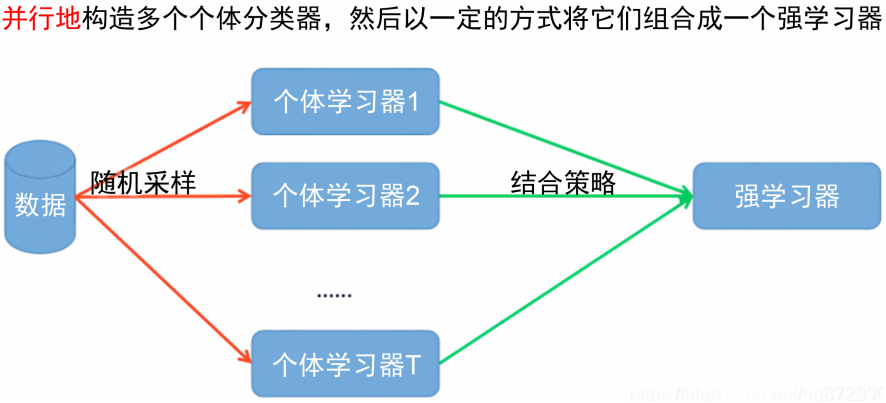


Figure 6. Bagging classification process figure
图 6. Bagging 分类过程图

2.2.2. Bagging 分类算法流程

Bagging 是直接基于自抽样方法，经过多次随机取样工作得到样本集后再进行训练，最终通过投票产生结果。即 Bagging 分类算法主要包含三步：取样、训练、分类。

1) 取样：给定一个包含 N 个样本的数据集，首先随机抽取一个样本到样本中，然后再将其放回原始数据集中，这样可以保证在下一次采样的时候仍有机率选中该样本。经过 N 次随机采样操作，可以得到一个包含 N 个样本的样本集来计算原始分布的各种统计数据。并且在理论上，自助样本 D_i 大概包含 63.2% 的原训练数据。

2) 训练：在进行完数据抽取后，将获得的随机数据集进行模拟训练，并且对于每个生成的样本训练一个特定于它的分类器。

3) 分类：利用生成的分类器分别投票产生分类结果，并且对预测结果进行表决。

2.2.3. Bagging 分类效果

优点：Bagging 分类方法由于选中每个样本的概率是相同的，所以适合用来训练多个并行的基本分类器，可以减少方差，所以一般过分拟合对其产生的影响不大。

缺点：Bagging 并不适合训练带有任何一个特定样本的数据集，并且由于 Bagging 分类算法是一种自抽样算法，所以重复放回取样改变了原有数据的正常分布，对最终结果的预测可能会产生一定的影响。

3. 随机模拟

3.1. 模拟方案

首先利用 python 模块下的 jupyter 进行非线性随机数据的生成，数据生成之后分别用两种分类方法，即深度神经网络分类与 Bagging 分类，进行随机模拟，通过调整各参数找到精度最高的参数匹配，对比两者的最终得分与分类效果，进而选择分类精度较高的分类方法进行后续的实证分析。

3.2. 模拟过程

3.2.1. 生成随机数据集

利用 python 模块下的 jupyter 通过代码实现，使用分类模型随机数生成方法进行非线性随机数据的生成，在此设置样本数为 2000，样本特征数为 10，其中多信息特征的个数为 10 个，没有样本冗余特征数和重复信息。最终将数据分为 4 类，并且每一类数据设置为由 1 个 cluster 构成。随机数据集生成，将其可视化如图 7 所示。

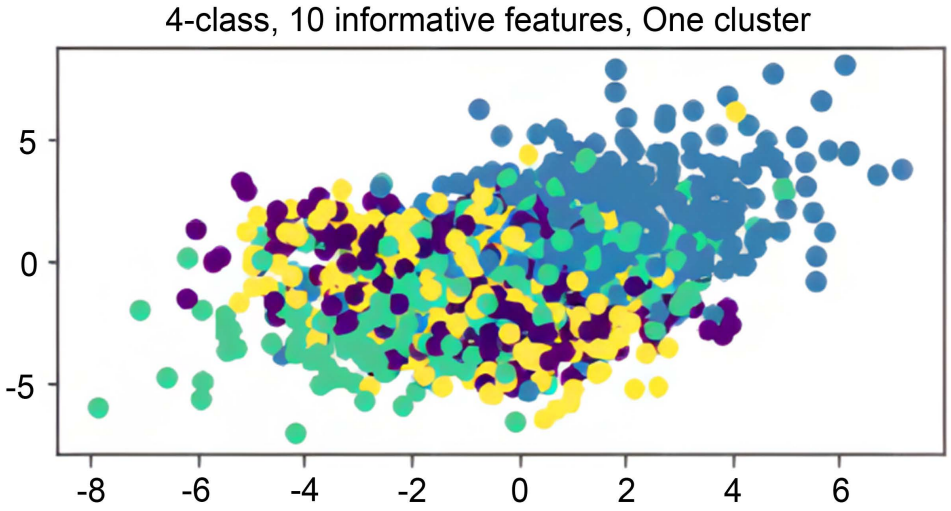


Figure 7. Random data visualization plots in figure
图 7. 随机数据可视化图

3.2.2. 深度神经网络(DNN)分类

通过代码实现将上述数据集拆分为训练集与测试集，并且设置随机种子为 888，训练的最大迭代次数设置为 500，利用训练集数据进行模拟，使模型适合数据训练集 X 和训练集 Y。利用训练后的模型对测试集数据进行得分计算，查看最终的得分，以此来判断此分类方法的准确度。

3.2.3. Bagging 分类

同样，通过代码实现将上述数据集拆分为训练集与测试集，并且设置随机种子为 888，最大弱学习器的个数设置为 2000，训练的最大迭代次数设置为 500，设置训练样本可以是任何数量的特征，利用训练集数据进行模拟，使模型适合数据训练集 X 和训练集 Y。利用训练后的模型对测试集数据进行得分计算，查看最终的得分，以此来判断此分类方法的准确度。

3.3. 模型结果

模型建立后，通过代码实现具体模拟过程，最终分别输出在 DNN 分类模型和 Bagging 分类模型下的准确率，即最终得分：深度神经网络分类模型准确率为 97%，Bagging 分类模型准确率为 90.6%，可以看出深度神经网络分类效果比较显著。

3.4. 模型可解释性

针对同一非线性随机数据集，可以由结论看出，对于此数据，深度神经网络的分类算法优于 Bagging 分类算法，即 DNN 的分类精度较高，由于深度神经网络是适用于大样本数据集，所以适用于此数据集。

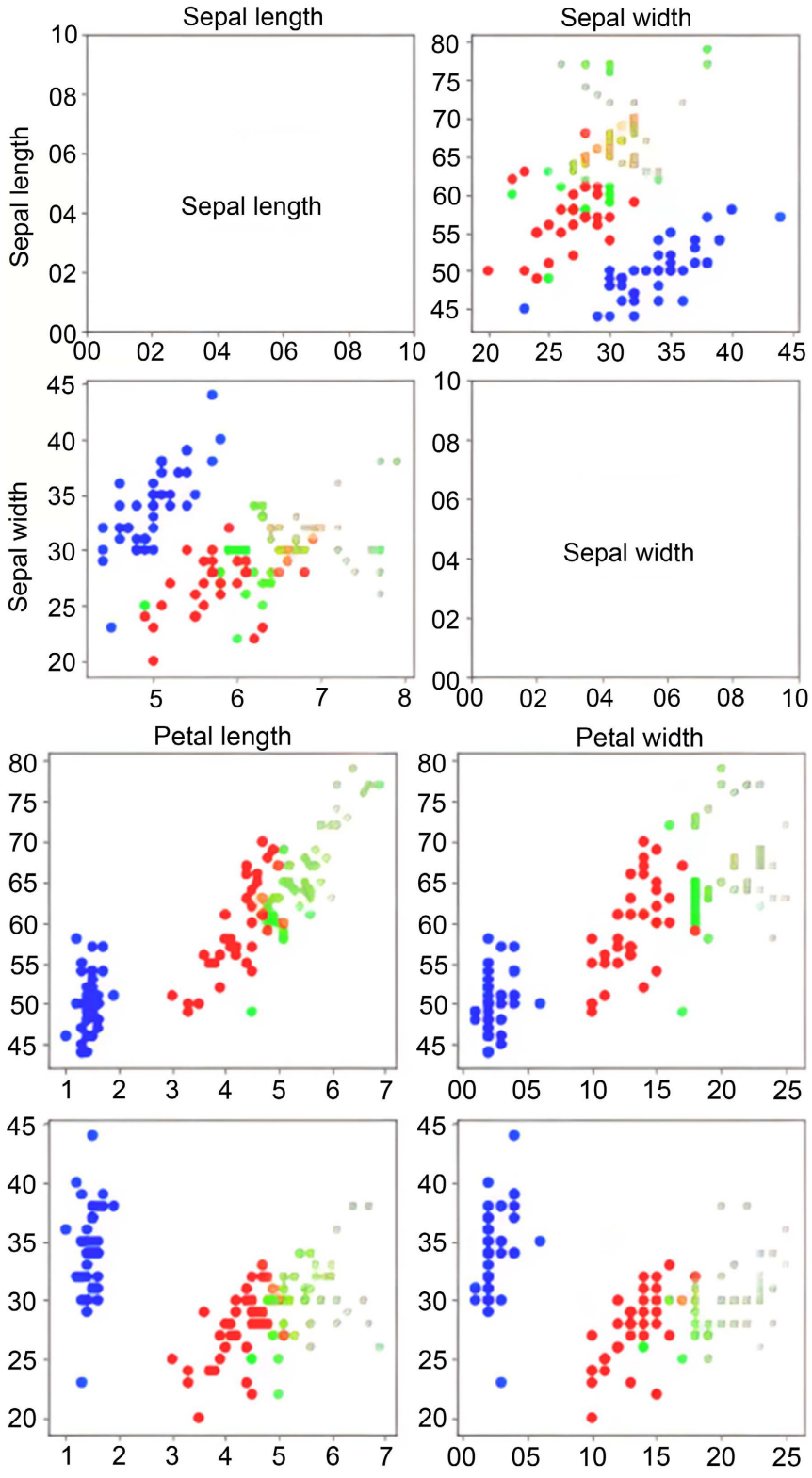
4. 实证研究

4.1. 方法概述

在进行随机模拟后，发现 DNN 分类精度较高，所以选择此方法进行鸢尾花分类数据集的实证研究。即先创建鸢尾花数据集，后搭建 DNN 模型对已经生成的鸢尾花数据集进行训练分类，最终对结果加以分析。

4.2. 生成鸢尾花数据集

直接通过代码导入实际数据集鸢尾花数据集(Iris)，并且将此真实数据进行可视化如图 8 所示。



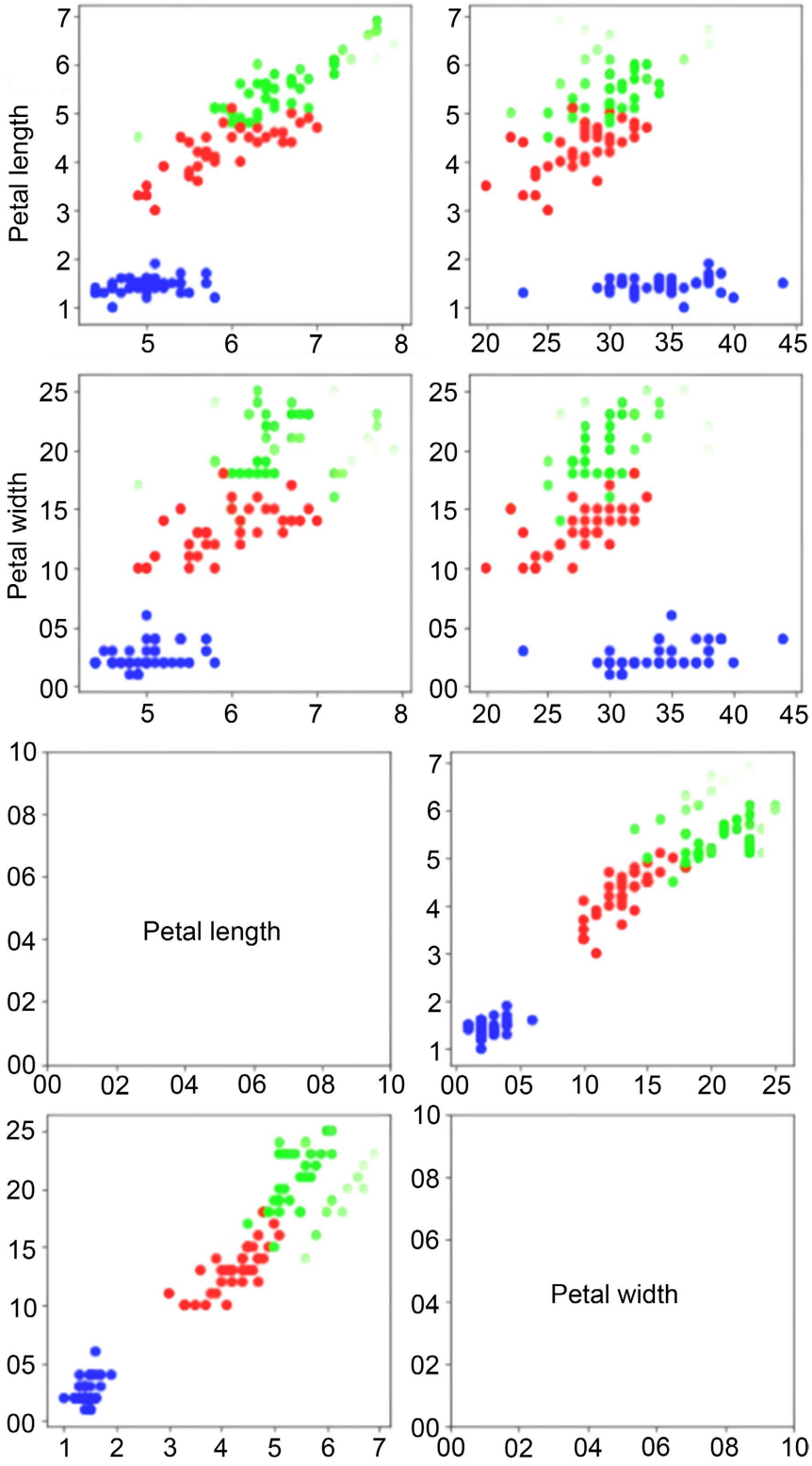


Figure 8. Visual plot of Iris dataset
图 8. 鸢尾花数据集可视化图

该数据集一共有 150 个样本，里面的每一个样本都包含着四个特征，这四个特征是：花萼长度(sepal

length)、花萼宽度(sepal width)、花瓣长度(petal length)、花瓣宽度(petal width)。并将样本一共分为三类，分类序号如表 1 所示。

Table 1. Classification table of the Iris dataset
表 1. 鸢尾花数据集分类表

样本编号	类别
1-50	Se-tosa
51-100	Versicolour
101-150	Virginica

4.3. DNN 的搭建与实验

搭建深度神经网络来解决鸢尾花数据集的分类问题，具体搭建结构如图 9 所示。

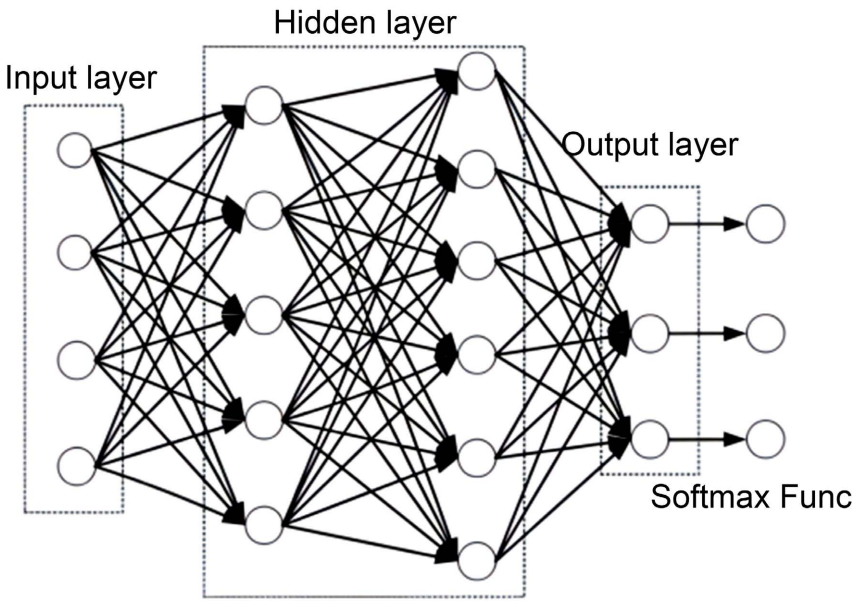


Figure 9. DNN construction structure diagram
图 9. DNN 搭建结构图

由图 9 可以看出，本文搭建的 DNN 实验架构由四个部分组成，分别是输入层、隐藏层、输出层、损失函数四个部分组成。其中，输入层由四个神经元组成，也对应为鸢尾花数据集得四个特征，中间设置了两个隐藏层，分别有 5 个和 6 个神经元，最后为三个输出层，也对应为鸢尾花的三个类别，层与层之间的权重与偏倚系数都是随机生成的。最后是选择 softmax 函数，损失函数选择交叉商损失函数。构建完 DNN 模型后，利用此模型对表 2 中的实验数据集进行训练。

Table 2. Experimental dataset table
表 2. 实验数据集表

数据集名称	特征数目	类别数目	样本数目
Iris	4	3	150

4.4. 模型结论

在本次实验中，训练的最大迭代次数设置为 500，利用训练集数据进行模拟，使模型适合数据训练集 X 和训练集 Y，最终输出在该 DNN 模型下的准确率，发现在迭代 500 次后准确率达到了 97.37%，接近 100%，可以看出分类效果非常显著。

4.5. 模型预测

首先假设一朵鸢尾花的 4 个特征分别为 6.3，3.3，5.3，1.3，分别将这四类代入到上述所创建的 DNN 模型中，进行预测，最终发现将具有此特征的鸢尾花划分为 versicolor 类别中，联系现实，发现此预测结果符合现实性，即证明此 DNN 分类模型预测效果良好。

5. 结论

5.1. 基本结论

1) 深度神经网络分类模型适用于大量数据的特征提取和定量预测，准确度高，并且对噪声神经具有较强的容错能力，并且适用于具有局部相关特性的数据集。但对于小样本则容易产生过拟合，进而得到较差的结果。

2) Bagging 分类对于选取到每个样本的概率是相同的，所以适合多个并行的基本分类器，可在一定程度上减小方差和减小过拟合。但不适合于带有任何一个特定样本的数据集，由于重复放回抽样，其结果准确度可能欠佳。

3) 利用两种分类模型对生成的同一非线性数据集进行随机模拟，输出结果：深度神经网络分类模型准确率为 97%，Bagging 分类模型准确率为 90.6%，由对比可以看出深度神经网络分类效果比较显著。

4) 利用深度神经网络分类模型对实际数据集：鸢尾花数据集进行训练，最终输出在该 DNN 模型下的准确率，发现在迭代 500 次后准确率达到了 97.37%，接近 100%，可以看出分类效果非常显著。随后利用此模型进行相关预测，假设一朵鸢尾花的 4 个特征分别为 6.3，3.3，5.3，1.3，分别将这四类代入到上述所创建的 DNN 模型中，进行预测，最终发现将具有此特征的鸢尾花划分为 versicolor 类别中，联系现实，发现此预测结果符合现实性，即证明此 DNN 分类模型预测效果良好。

5.2. 总论

本文利用随机模拟实验和实证研究，考虑了深度神经网络分类和 Bagging 分类两种机器学习方法的建模和预测，由随机模拟实验测试出神经网络分类模型具有较高的精确度，并且对于噪声神经具有较强的容错能力，进而利用神经网络分类模型进行实际数据集(鸢尾花数据)的模拟，进一步论证其具有较高的准确度与预测精度，对分类器的选择具有重要的现实意义，在未来也可以对其他多种分类器进行比对模拟，运用更广泛的数据集来检验算法的运行效率，使得数据挖掘领域的发展更进一步深入。

基金项目

北京市教委科研计划项目资助(KM201910009001)。

参考文献

[1] 杨剑锋, 乔佩蕊, 李永梅, 王宁. 机器学习分类问题及算法研究综述[J]. 统计与决策, 2019, 35(6): 36-40. https://doi.org/10.13546/j.cnki.tjyjc.2019.06.008

[2] 朱虎明, 李佩, 焦李成, 杨淑媛, 侯彪. 深度神经网络并行化研究综述[J]. 计算机学报, 2018, 41(8): 1861-1881.

[3] 姚明海, 赵连朋, 刘维学. 基于特征选择的 Bagging 分类算法研究[J]. 计算机技术与发展, 2014, 24(4): 103-106.

[4] 施启军, 龙福海, 苟辉鹏, 苏浩锜, 谢雨寒. 基于深度神经网络的多分类方法研究[J]. 网络安全技术与应用, 2021(9): 41-43.

[5] 马景义, 谢邦昌. 用于分类的随机森林和 Bagging 分类树比较[J]. 统计与信息论坛, 2010, 25(10): 18-22.

[6] 任涛, 林梦楠, 陈宏峰, 王冉冉, 李松威, 刘晓雨, 刘杰. 基于 Bagging 集成学习算法的地震事件性质识别分类[J]. 地球物理学报, 2019, 62(1): 383-392.

[7] Lipiński, P.F.J. and Szurmak, P. (2017) SCRAMBLE'N'GAMBLE: A Tool for Fast and Facile Generation of Random Data for Statistical Evaluation of QSAR Models. *Chemical Papers*, **71**, 2217-2232. https://doi.org/10.1007/s11696-017-0215-7

[8] Bhushan, S. and Pandey, A.P. (2021) Optimal Estimation of Population Variance in the Presence of Random Non-Response Using Simulation Approach. *Journal of Statistical Computation and Simulation*, **91**, 3814-3827. https://doi.org/10.1080/00949655.2021.1948547

[9] Salahuddin, Z., *et al.* (2022) Transparency of Deep Neural Networks for Medical Image Analysis: A Review of Interpretability Methods. *Computers in Biology and Medicine*, **140**, Article ID: 105111. https://doi.org/10.1016/j.compbiomed.2021.105111

[10] Sharafati, A., Asadollah, S.B.H.S. and Al-Ansari, N. (2021) Application of Bagging Ensemble Model for Predicting Compressive Strength of Hollow Concrete Masonry Prism. *Ain Shams Engineering Journal*, **12**, 3521-3530. https://doi.org/10.1016/j.asej.2021.03.028

[11] Manzanarez-Ozuna, E., Flores, D.-L., Gutiérrez-López, E., *et al.* (2018) Model Based on GA and DNN for Prediction of mRNA-Smad7 Expression Regulated by miRNAs in Breast Cancer. *Theoretical Biology and Medical Modelling*, **15**, Article No. 24. https://doi.org/10.1186/s12976-018-0095-8