CNN-Based Book Cover and Back Cover Recognition and Classification

Haochang Xia, Yali Qi, Qingtao Zeng(✉), Yeli Li, and Fucheng You

Beijing Institute of Graphic Communication, Beijing, China
zengqingtao@bigc.edu.cn

Abstract. As an important part of the national economy and an important supporting industry, printing and publishing industry is closely related to the development of national economy. In recent years, the massive publication and printing of books has made the work of storing books in databases more and more onerous. The maturity of deep learning technology has brought good news to recognition and classification of books. Convolutional neural network is a good tool. Convolutional neural network is a technology in deep learning, often used in computer vision, image recognition classification and other fields. Research results in the field of book recognition and classification are relatively lacking. There is no good book data set that can be used for neural network training. In this paper, we collected a large number of book data sets and we built a set of image classification models based on CNN to identify and classify the cover and back cover of books. Through a lot of training and testing, we have generated a set of CNN models that can effectively identify and classify the cover and back cover of books. Compared with the traditional way of manually entering books into database, the use of neural networks makes the work more efficient and saves a lot of human resources.

Keywords: CNN · Deep learning · Image recognition · Image classification

1 Introduction

In recent years, artificial intelligence has developed rapidly, and machine learning has continued to make progress in the direction of image recognition and classification. In order to improve the accuracy of image recognition, many deep learning models have achieved good results, including: CNN, RNN, DBN, GAN, etc. Deep learning has achieved good results in the fields of face recognition and handwritten digit recognition. The initial idea was to use probably the best known of ML technologies; the convolutional neural network (CNN) which is used extensively for image recognition, particularly using the massive number of images available on the internet [1]. However, research in the fields of book recognition and classification is relatively lacking. In image processing field, CNNs can be used as an efficient and high-performance classification model and have gained outstanding performance [2].

This paper builds a set of image classification models based on the research of machine learning and neural network. First, we collected a large number of images of

Y. Weng et al. (Eds.): TridentCom 2020, LNICST 380, pp. 59–70, 2021.
https://doi.org/10.1007/978-3-030-77428-8_5

the cover and back cover of the book, and processed these images as a data set. Then build a set of neural networks, use these data sets to iteratively train and predict the neural network, and finally generate a training model that can effectively identify the book cover and back cover to achieve the classification effect.

2 Overview

With the development of Internet, the printing and publishing industry is also moving in the direction of digitization and networking. Among them, book digitization is an important area. The rapid development of machine learning and the rapid maturity of image recognition and classification technologies have greatly improved the efficiency of the book digital entry system. I have read a lot of literature and found that there are already many mature algorithms and image classification models in the field of computer vision and machine learning, however, there are still relatively few studies on applying these algorithms to the recognition and classification of book covers and back covers.

The identification and classification of books is a very heavy manual labor. On major online e-commerce platforms and university libraries, the traditional way of storing books to the database is to scan the barcodes code and obtain book pictures from the publisher. The traditional way has to upload the data manually. The use of machine learning technology to realize automatic recognition and classification of book covers, however, can reduce manual workload and improve work efficiency. On the other hand, it can also effectively improve the quality of classification and solve the problem that traditional manual scanning cannot identify the incomplete books.

3 Basic Principles of Deep Learning

3.1 Neural Networks

3.1.1 Neurons

The simplest neuron structure is a model that contains data input, result output, and corresponding calculation methods. The data is passed in by the input neuron. According to the weight of each input neuron, the sum is calculated by weight, and then the result is passed to the activation function for processing. Finally, the output data is obtained by this way. In the formula, letter w represents the weight, and letter x stands for data input. As we can see, different inputs get different weight. The letter b represents a constant parameter, and y(x) represents neuron calculation results.

The calculation formula of neuron is shown as follows:

$$y(x) = f(l + b) = f\left(\sum\nolimits_{i=1}^{n} w_i x_i + b\right) \tag{1}$$

3.1.2 Multilayer Neural Network

In order to cope with a more complex environment, in practical applications, a multilayer neural network structure is generally used. This neural network structure is also called feedforward neural network. Its first layer is called the input layer, the middle layer is called the hidden layer, the hidden layer can contain multiple layers of neuron structure, and the last layer is called the output layer. The neurons in the upper layer and the neurons in the next layer are connected by means of full connections. Neurons that are not in the same layer do not communicate, and neurons in the same layer are independent of each other. Data is transferred from the input layer to the hidden layer, and the hidden layer is solved layer by layer through corresponding calculation methods, we get the output until the last layer is calculated. Figure 1 shows the general multi-layer neural network.

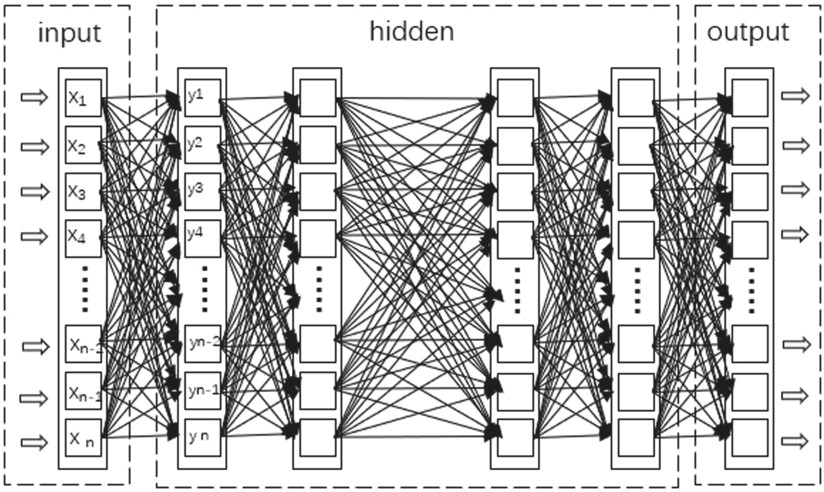

Fig. 1. Multi-layer neural network

3.2 Convolutional Neural Network

3.2.1 Convolutional Layer

Convolutional layer is the structure used for feature extraction in CNN. The convolutional layer applies several filters to the input data to perform operations. The filter is an operation in image processing. The specific operation of filtering is the sum of the product of the image pixels and the filter [3]. Here, the filters are called convolution kernels. Each part of the information is extracted through the convolution kernel, and the feature information is re-integrated through weight sharing to obtain a new feature map. The pixel value on the picture and the convolution kernel are multiplied and summed correspondingly to obtain a new feature value. Calculation process of the convolution operation is shown in the Fig. 2.

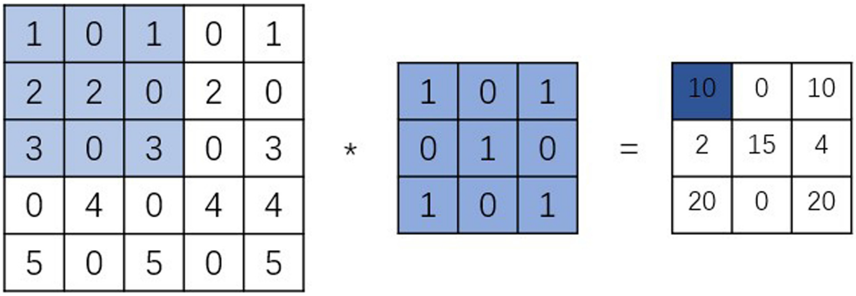

Fig. 2. The calculation process of the convolution operation

3.2.2 Pooling Layer

The pooling layer takes the down-samples for input feature map. This method can reduce the size of the feature map, thereby reducing the amount of calculation, and can also increase the translation robustness to the certain extent. Generally speaking, there are two forms of pooling layer: maximum pooling and average pooling. The maximum pooling method is to take the maximum value of the feature points in the window neighborhood, and the average pooling method is to take the average of the feature points in the window neighborhood. The two pooling methods of maximum pooling and average pooling are shown in the Fig. 3.

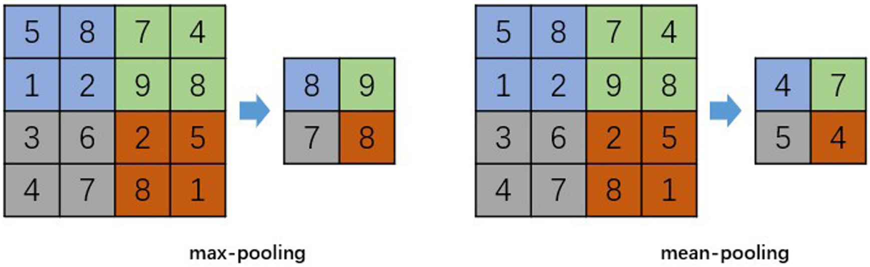

Fig. 3. Maximum pooling and average pooling

3.2.3 Activation Function

In a neural network, if there is only a convolutional layer, the relationship between input and output is just a simple linear operation. The input data x is processed by the activation function, and a new result is obtained, which improves the expressive ability of the linear model. In order to make the neural network better fit the nonlinear problem, but also help to make the deep neural network really work, it is necessary to use the nonlinear activation function. Commonly used activation functions include the following.

(1) The Sigmoid function, also known as the Logistic function, is the most widely used activation function in the early days. The mathematical expression is as follows:

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

The output range of the Sigmoid function is in the (0,1) interval, and it grows continuously and monotonously, which can be well applied to classification problems.

(2) Tanh function, also known as hyperbolic tangent function, its shape is similar to sigmoid, and its mathematical expression is as follows:

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3}$$

The output value of the Tanh function is between [-1,1]. Like the Sigmoid function, there is also a saturation problem. The Tanh function can be seen as two Sigmoid functions. When the input value of the function is less than zero, the output value is less than zero, when it is equal to zero, the output value is equal to zero, and when it is greater than zero, the output value is greater than zero.

(3) The rectified linear unit (ReLU) function is the most used activation function in practical applications. Its mathematical expression is as follows:

$$f(x) = \begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases} \tag{4}$$

The expression of the ReLU function is very simple, but it works very well. Whether it is forward propagation or back propagation, the amount of calculation is very small and the training process is fast.

3.2.4 Fully Connected Layer

The processes of convolution, pooling, and activation in the neural network belong to the feature extraction stage of the image, and only the final fully connected layer accomplishes the classification task. The specific method is to map the feature space to the sample label space.

The fully connected layer performs weighted summation and offset operations on the input values. All values in the vector are weighted and summed with the weights corresponding to each convolution kernel to obtain the characteristic response. The number of characteristic responses is determined by the number of cores in the fully connected layer. The fully connected layer increases the complexity of the model, and theoretically improves the generalization ability of the model.

As shown in Fig. 1, the connection between the first input layer and the hidden layer is full connection. If the input layer has only three parameters, and the first hidden layer also has three output parameters. A simpler hierarchy like this, the calculation formula between them can be expressed as:

$$\begin{pmatrix} y1 \\ y2 \\ y3 \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{pmatrix} * \begin{pmatrix} x1 \\ x2 \\ x3 \end{pmatrix} + \begin{pmatrix} b1 \\ b2 \\ b3 \end{pmatrix} \tag{5}$$

3.2.5 Error Back Propagation

The training process of convolutional neural network is divided into two stages. The first stage is the forward propagation, which transforms the propagation of data from low-level to high-level. The neural network calculates and stores the intermediate variables of the model layer by layer in the order from the input layer to the output layer. The second stage is the back propagation. When the output result of the forward propagation does not match the expectation, the error is propagated and trained from the high-level to the bottom-level. According to the order from the output layer to the input layer, the objective function is calculated and stored layer by layer. The intermediate variables of the segment and the gradient of the parameters are modified in turn to modify the parameter values between the connections of each layer. The modification of the parameter values is determined by the error of the back-end data input.

4 Image Recognition Classification Model Based on CNN

4.1 Experimental Environment

The experiment uses Python and PyTorch development environment for programming and model training. As we know, Python is one of the most popular computer programming languages in the field of machine learning. It has a rich standard library and a strong third-party ecosystem. PyTorch is a python-based scientific computing package. It can be used as a substitute for NumPy, it uses the powerful performance of GPU for calculations, or as a highly flexible and fast deep learning platform. PyTorch is an optimized tensor library for deep learning using GPUs and CPUs. PyTorch provides the elegantly designed modules and classes torch.nn, torch.optim, Dataset, and Data Loader to help you create and train neural networks.

4.2 Data Preprocessing

4.2.1 Data Set

For each CNN model, a "multi-scale summation" module is employed to avoid overfitting that is usually caused by limited training data [4]. The recognition accuracy of the image classification model largely depends on the quality of the data set. For the design of neural network recognition system, it is very important to have a sufficient amount of labeled training data, and the data contains as many complex environmental changes as possible. The data set used in this article mainly comes from the two major e-commerce platforms of JD.com and Dangdang.com. Table 1 shows the specific categories of the data set.

4.2.2 Data Set Processing

The data set collected 3,600 images containing book cover and back cover from online shopping malls (JD.com and Dangdang.com). These book images have three different kinds of environments, including main-picture scene, multi-picture scene, and publishing

Table 1. Book cover and back cover data set.

Data sources	Category	Number
JD.com	Cover	900
JD.com	Back cover	900
Dangdang.com	Cover	900
Dangdang.com	Back cover	900

house scene. The main-picture scene is a single cover or back cover picture; the multi-picture scene is a picture that may contain more than one book; the publishing house scene is the cover or back cover of the book containing the content information recommended by the publisher. In view of the daily actual use, the cover and back cover images of the book are randomly selected from the three scenes at a ratio of 1:1:1 to build a training set. The number of images in the test set depends on the actual situation.

In consideration of the working mechanism of the convolutional neural network, we need to preprocess the images due to the different sizes of the collected data sets. In order to facilitate the operation of convolution operation, while keeping the basic information of the original image as much as possible, the cropping method is to first scale the image proportionally, so as to preserve the original image content to the greatest extent, and then perform center cropping to make the training data more accurate. All this operation is to make sure that the input image contains the complete central information of the original image.

4.3 Design and Construction of Neural Network Model

4.3.1 CNN (Convolutional Neural Network)

In deep learning, convolutional neural network is a special deep neural network, often used in the field of computer vision. CNN is composed of input layer, convolution layer, activation function, pooling layer, and fully connected layer. The most prominent advantage of CNN is high-speed parallel computing and processing speed not influenced by the size of image, so it is convenient for hardware implementation [5].

In the convolutional neural network, the input layer refers to the input image data, which is generally stored in the form of a matrix. The data set we collected contains pictures of various sizes. To facilitate the calculation, we processed the training set and unified the size of the training set pictures to 200 * 200 pixels.

In the convolution process, the convolution kernel extracts image features by convolution calculation operations on the input layer image. When building the neural network, we used two convolutional layers here.

The activation function is a processing function added to fit the input-output relationship of nonlinear characteristics. In CNN, each convolution operation process of the convolution kernel is to sum up the products of various positions in the template, input the accumulated value into the activation function, and then use the output value as the convolution result. This paper uses the ReLU function as the activation function of the experiment.

The pooling layer compresses the input feature map to make the feature map smaller and simplify the computational complexity of the neural network; on the other hand, the main purpose of feature compression is to extract the main features. Corresponding to the two-layer convolutional layer, we use a two-layer pooling layer, and the method used by the pooling layer is maximum pooling.

The fully connected layer contains input nodes and output nodes. The input nodes receive the output data of the previous layer, after calculation, the result is passed to the output nodes, and the output nodes transmits the data to the next layer of the neural network, thereby passing the output value to the classifier. Usually the result of CNN n-class classifier models is obtained with the selection of the maximum among the n output nodes of the final fully-connected layer [6]. This paper uses three fully connected layers, and finally two output results are obtained, corresponding to the probability values of the book cover and back cover. The Fig. 4 below shows the CNN model built in this paper.

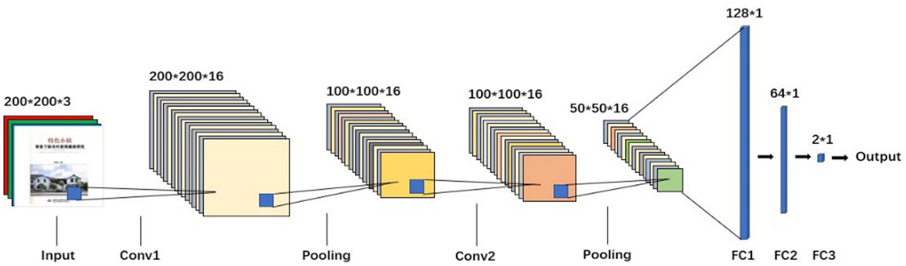

Fig. 4. CNN model

4.3.2 Building a Convolutional Neural Network

The deep neural network constructed in this paper contains a multi-layer structure, where the image size of the input layer is 200*200 pixels, and the image here has been processed in the data set preprocessing stage. In addition, the neural network constructed in this paper includes two convolutional layers, two pooling layers and three fully connected layers. The structure of the convolutional neural network is shown in Table 2.

Table 2. The CNN structure of the book cover and back cover classification.

Layer	Depth	Number	Size
Input layer	3	–	200 × 200
Convolution layer 1	3	16	3 × 3
Pooling layer 1	–	–	2 × 2
Convolution layer 2	16	16	3 × 3
Pooling layer 2	–	–	2 × 2
Fully connected layer 1	–	–	40000(in) 128(out)
Fully connected layer 2	–	–	128(in) 64(out)
Fully connected layer 3	–	–	64(in) 2(out)

4.4 Model Training

Model training is divided into several stages. First, the input layer obtains the processed 24-bit jpg format RGB three-channel image. After the first convolution operation, the output image is a 16-channel 200 * 200 pixels image. Then we perform the first pooling operation. The pooling layer uses a 2 * 2 Max Pooling and outputs a 100 * 100 * 16 matrix. Next it is the second convolution operation, after that we can obtain a new matrix of 100*100*16.

After all this, it is the second pooling operation, the image is reduced to 50 * 50 at this time. At last, we have three full connections options, and the input nodes here are separately 40,000, 128, and 64. The number of output nodes of the last one is 2. These two nodes correspond to two output values, including the probability of the book cover and back cover. Figure 5 shows the recognition result.

This paper provides a convolutional neural network model built on the pytorch framework. Pytorch offers a rich API interface for the construction of deep neural networks. At the same time, it takes full advantage of the good computing speed of the GPU. Selecting the pytorch framework to use the GPU for calculations saves a lot of time for model training and testing. The experimental operating system is Windows 10, and the experimental environment is Intel Core i7-9750H CPU and NVIDIA GeForce GTX 1660Ti GPU.

5 Model Assessment

In the experiment, different training times and learning rates were set for the constructed model to observe the experimental data. For each CNN model, a "multi-scale summation" module is employed to avoid overfitting that is usually caused by limited training data [4]. In order to improve the classification effect of the convolutional neural network model and achieve good predictions on new data, the data set is divided into three subsets (training set, validation set and test set). Because the data set is relatively scarce, it can't cope with all situations. But we can simulate more training data by increasing the number of training. A too high learning rate will increase the risk of overfitting, and a too low

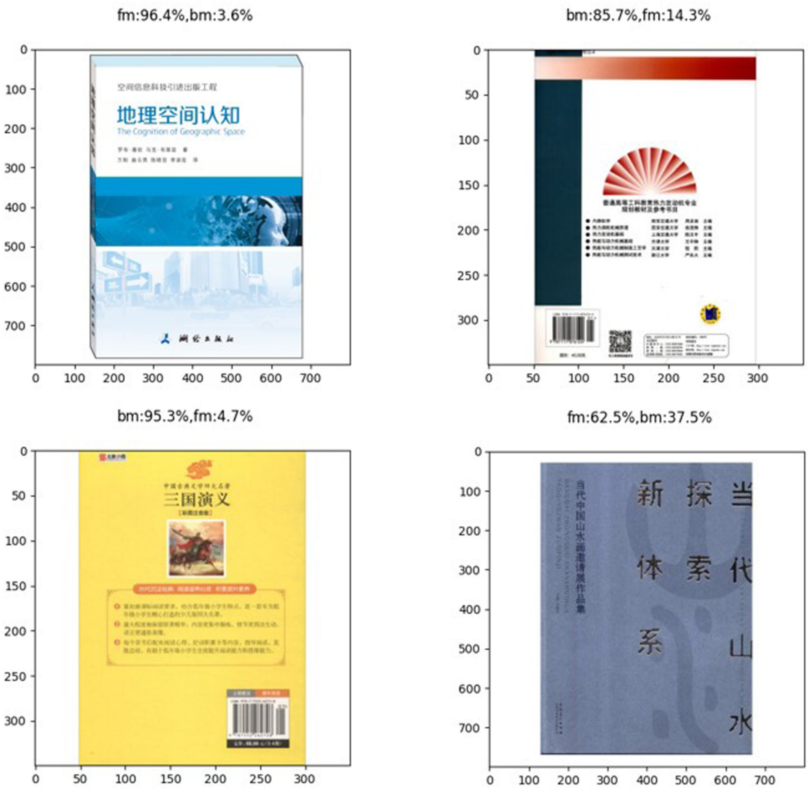

Fig. 5. Recognition result of book cover and back cover

learning rate combined with relatively scarce training data may cause the learning model to fail to learn specific feature values. Part of the training data is shown in Table 3.

It can be seen from the experimental data that when the number of training rounds is 3 and the learning rate is 0.001, the accuracy of the test reaches the highest. When the number of training rounds is 5, the accuracy rate reaches the highest and the learning rate is 0.001 as well. With the same number of training rounds, it seems that the accuracy of the model decreases as the learning rate increases. However, when the number of training rounds is adjusted to 10 rounds, the highest point of the accuracy of the model falls in the case of the lowest learning rate at 0.0001. After constant adjustment and testing of the parameters, we finally set the number of training rounds to 20 rounds and the learning rate of 0.01 as the final parameters of our model, and the test results are good, with an accuracy of 97.6%. In the future work, we will investigate many supervised learning algorithms such as: SVM [7], k-nearest neighbors [8] and Boosted Regression Trees [9, 10] to classify an image. If I combine multiple machine learning methods, the accuracy will be further improved.

Table 3. Training data

Train number	Epoch	Learning rate	Test accuracy
1	3	0.0001	0.5799476
2	3	0.001	0.7421927
3	3	0.01	0.5049149
4	5	0.0001	0.7311714
5	5	0.001	0.9067465
6	5	0.01	0.5064889
7	10	0.0001	0.93220395
8	10	0.001	0.8237559
9	10	0.01	0.5196011
10	20	0.0001	0.9116753
11	20	0.001	0.9758636
12	20	0.01	0.50095326

6 Conclusion

Machine learning (ML) is an algorithm set especially suited to prediction. These ML methods are easier to implement and perform better than the classical statistical approaches [11]. This paper builds a set of convolutional neural networks through the research of deep learning to identify and classify the cover and back cover of books. By revising the different parameters in the construction model, a relatively good accuracy rate has been achieved, which provides a certain reference for book identification and classification. Due to the limited ability of the author, I failed to provide a more perfect solution for book classification. Machine learning is often inseparable from big data. My next step is to improve the book classification data set and classify the data set according to book categories. Through these works, we can provide various special diagnosis attributes for the training of neural networks to adapt to various complex situations. At the same time, we can optimize algorithms and modify models to improve recognition accuracy.

Acknowlegments. 1. Beijing science and technology innovation service capability construction project (PXM2016_014223_000025).

2. Major special project of science and technology of Guangdong Province, No: 190826175545233.

3. BIGC Project (Ec202007).

References

1. Padfield, N.: Exploring the classification of acoustic transients with machine learning. In: Proceedings of ACOUSTICS 2019, vol. 10, no. 13 (2019)

2. Leng, J., Li, T., Bai, G., et al.: Cube-CNN-SVM: a novel hyperspectral image classification method. In: 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1027–1034. IEEE (2016)
3. Yin, Q., Zhang, R., Shao, X.L.: CNN and RNN mixed model for image classification. In: MATEC Web of Conferences. EDP Sciences, vol. 277, p. 02001 (2019)
4. Zhang, M., Li, W., Du, Q.: Diverse region-based CNN for hyperspectral image classification. IEEE Trans. Image Process. **27**(6), 2623–2634 (2018)
5. Qingling, J.: Edge detection for color image based on CNN. Int. J. Adv. Inf. Sci. Serv. Sci. **3**(10), 61–69 (2011)
6. Jhang, K.: Gender prediction based on voting of CNN models. In: 2019 International Conference on Green and Human Information Technology (ICGHIT), pp. 89–92. IEEE (2019)
7. Piyush, R.: Hyperplane based classification: perceptron and (Intro to) support vector machines. In: CS5350/6350: Machine Learning (2011)
8. Domeniconi, C., Gunopulos, D., Peng, J.: Large margin nearest neighbor classifiers. IEEE Trans. Neural Netw. **16**(4), 899–909 (2005)
9. Rokach, L., Maimon, O.: Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. (2008). ISBN: 978-9812771711
10. Friedman, J.H.: Greedy function approximation: a gradient boosting machine (1999)
11. Breiman, L.: Statistical modeling: the two cultures. Stat. Sci. **16**, 199–215 (2001)