Appendices for:

Automated identification of steel weld defects, a convolutional neural network improved machine learning approach

Zhan SHU[a], Ao WU[a], Yuning SI[a], Hanlin DONG[b], Dejiang WANG[a,*], Yifan LI[c]

[a] School of Mechanics and Engineering Science, Shanghai University, Shanghai 200444, China

[b] School of Civil Engineering, Shanghai Normal University, Shanghai 201418, China

[c] Shanghai PinlanData Technology Co., Ltd., Shanghai 200072, China

*Corresponding author. E-mail: djwang@shu.edu.cn

Appendix A. Features extraction

A1. Mel frequency cepstral coefficients (MFCC) features

In this paper, the MFCC features add dynamic first- and second-order difference coefficients. Therefore, the MFCC coefficients are tripled relative to those of the original MFCC method [39]. The MFCC method was frequently used to extract audio features [18,40]. The MFCC features are calculated by following procedures.

1) Frame blocking and Hamming windowing

A frame gathers N sampling points into an observation unit. As expressed in Eq. (1), each frame is multiplied by the Hamming window to increase the continuity between the left and right ends of the frame. The expression of Hamming window is modeled in Eq. (2).

$$s'(n) = s(n) \times \omega(n), \quad (1)$$

$$\omega(n) = 0.54 - 0.46 \times \cos\left(\frac{2\pi n}{N-1}\right), \ 0 \le n \le N-1. \quad (2)$$

2) Fast Fourier transform (FFT)

Since it is difficult to see the effective characteristics of the acoustic signal in the time domain, the signal is converted into an energy distribution in the frequency domain. Therefore, after multiplying the Hamming window, each frame performs the FFT to obtain its power spectrum [33]. The calculation of FFT is expressed in Eq. (3).

$$x_F(k)=\sum_{n=0}^{N-1}s'(n)e^{-\frac{2\pi jk}{N}},\ 0\le k\le N\ . \quad (3)$$

3) Mel filter bank

The energy spectrum is processed by a set of Mel-scale triangular filters, defining a filter with M filters (the number of M is consistent with the number of critical bands), using a triangular filter with a center frequency of $f(m)$, and $0 \le m \le M$. The interval between each $f(m)$ widens as the value of m increases. The triangular bandpass filter, expressed in Eq. (4).

$$H_m(k)=\begin{cases}0 & ,k<f(m-1)\\ \dfrac{2\left(k-f(m-1)\right)}{\left(f(m+1)-f(m-1)\right)\left(f(m)-f(m-1)\right)}, & f(m-1)\le k\le f(m)\\ \dfrac{2\left(f(m+1)-k\right)}{\left(f(m+1)-f(m-1)\right)\left(f(m)-f(m-1)\right)}, & f(m)\le k\le f(m+1)\\ 0 & ,\ k\ge f(m+1)\end{cases}\ . \quad (4)$$

4) Logarithmic energy

The energy of a frame is important feature of audio. It is common to add the logarithmic energy of a frame making the basic audio feature of each frame one more dimension. The logarithmic energy of the filters could be computed by Eq. (5) [22].

$$s(m)=\ln\left(\sum_{k=0}^{N-1}\left|x_F(k)\right|^2H_m(k)\right),\ 0\le m\le M\ . \quad (5)$$

5) Discrete cosine transform (DCT)

The last step to obtain the MFCC coefficients is to perform the DCT as expressed in Eq. (6), and to bring the aforementioned logarithmic energy into the DCT to calculate the Mel coefficients of the L order. Here L was set 40 in the experiments, and M is the number of triangular filters.

$$C(n)=\sum_{m=0}^{N-1}s(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right),\ n=1,2,\ldots,L\ . \quad (6)$$

6) Dynamic difference coefficients

The dynamic characteristics of audio can be described by the differential spectrum of the static features $C(n)$. The dynamic first- and second-order difference coefficients are calculated by Eq. (7).

$$d_t = \begin{cases} C_{t+1} - C_t & , t < K \\ \dfrac{\sum_{k=1}^{K} k\left(C_{t+k} - C_{t-k}\right)}{2\sum_{k=1}^{K} k^2} & , K \le t < Q - K \\ C_t - C_{t-1} & , t \ge Q - K \end{cases} . \quad (7)$$

Therefore, the full composition of the MFCC features is actually composed of 3*L*-dimensional MFCC parameters (*L* MFCC parameters + *L* first order difference parameters + *L* second order difference parameters).

A2. continuous wavelet transform (CWT) features

The mathematical representation of CWT is shown in Eq. (8),

$$C(a,b) = \int_R s(t)\psi_{a,b}(t)dt \ , \quad (8)$$

where *C* is the 2-dimensional matrix of wavelet transform coefficients, *s(t)* is the audio signal, and $\psi_{a,b}(t)$ is a doubly-indexed family of wavelets generated by scaling and translating, represented in Eq. (9),

$$\psi_{a,b}(t) = |a|^{-1/2}\psi\left(\frac{t-b}{a}\right), \quad (9)$$

where $a,\ b \in R,\ a \neq 0$, *a* is the scaling factor, and *b* is the translating factor [41,42].

A3. Short-time Fourier transform (STFT) features

The signals are converted to spectrogram with information of frequency domain and time domain by using STFT. The mathematical representation of STFT [18] is shown in Eq. (10),

$$X(t,\omega) = \int s(t)w(\tau - t)e^{-j\omega\tau}dt \ , \quad (10)$$

where $s(t)$ is the audio signal as well, and $\omega(t)$ is the Hamming window.

Appendix B. machine learning (ML) algorithms

A brief introduction to the concept of four traditional ML algorithms is shown in Fig. 16. random forest (RF) is a typical ensemble learning [32]. The basic classifier of RF is decision tree. In addition, the decision trees of RF are independent of each other and can be generated at the same time.

The mechanism of k-nearest neighbors could be described as pre-determining a value for *k*, and the classifier will find *k* nearest samples of the test sample and count their labels, characterizing different classes. The nearest sample depends on the calculated distance between the samples. There are several different distance formulas. Then choose the label with the most samples as the prediction result of the test sample. Hence, *k* is a crucial parameter. The predicted results vary significantly with different *k*. In addition, there are many ways to calculate the distance. Euclidean distance is used in the experiment.

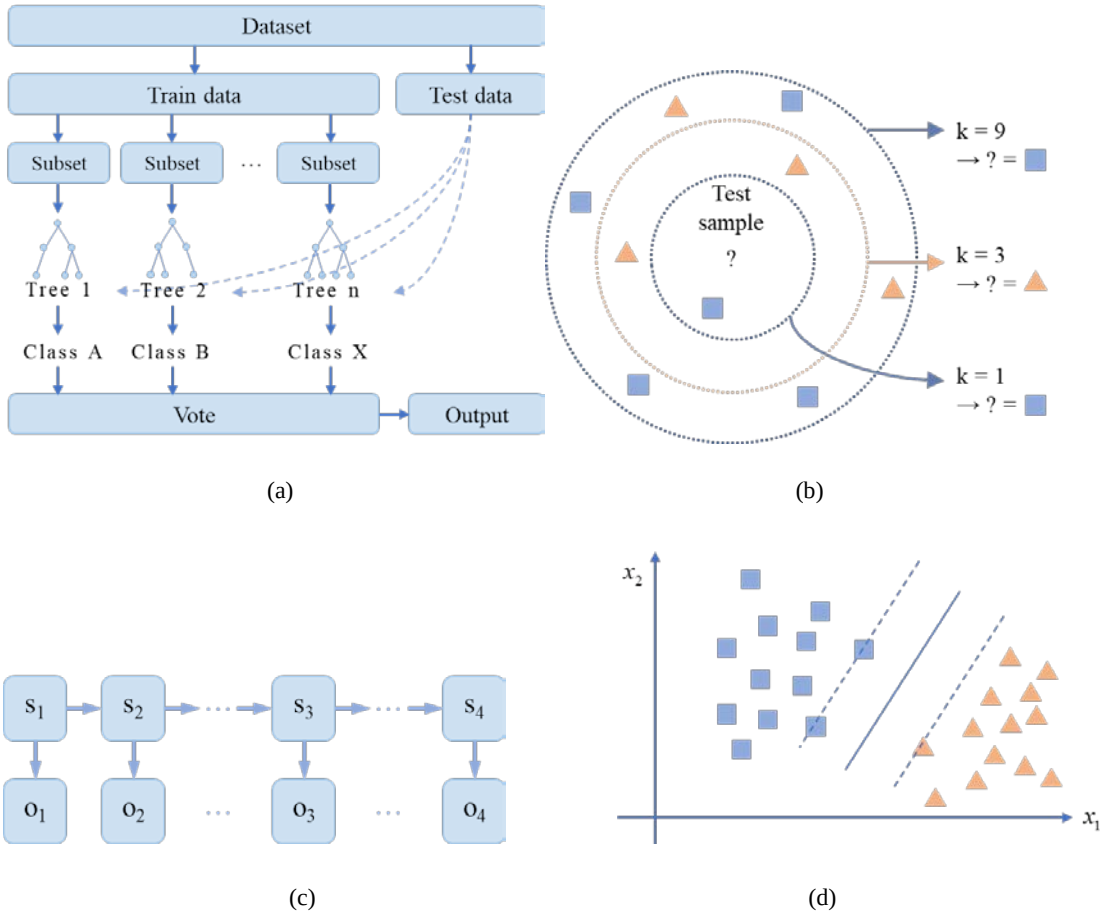

Fig. 16 Three ML classifiers schematic map.

Hidden Markov model (HMM) is a kind of dynamic Bayesian network, which is a famous directed graph model. There are two assumptions of HMM. One of these is, at an instant of time, the observed vector o_t is only decided by the state at the same instant of time s_t. Another is, at an instant of time, s_t is only decided by the preceding state s_{t-1}, which is called Markov chain [32]. HMM performs well in time series modeling (e.g., speech recognition [43,44]). Based on the HMM method, the Gaussian-HMM uses a Gaussian distribution to fit the probability density function from the hidden state to the observed vector, thereby the Gaussian-HMM is suitable for continuous observed vectors.

Support vector machine (SVM) was originally designed for binary classification task [45,46]. The main idea of SVM is to find a hyperplane to divide the samples with high accuracy. Compared with the binary classification tasks, there are two decision function shapes to use SVM to implement the multi-class tasks. One is “one versus rest,” which needs the same number of SVMs as the number of classification types. Every SVM only decides a sample is one kind or the rest. Another decision function shape is one versus one, which means every SVM can only obtain some samples of two types as training test and divide the rest into these two labels corresponding to two types. In other words, the principle of this method is to decide whether each sample is first class or second class.

Appendix C. Convolutional neural network (CNN)-enhanced approach

CNN is usually used to process problems about images [47]. Moreover, it has been proved that the CNN, as a typical deep learning method, can provide higher accuracy than traditional ML algorithms in classification tasks [48–51]. A typical CNN network, as shown in Fig. 17, has some traditional kinds of layers including convolutional layer, BN layer, activation layer, pooling layer, and fully connected (FC) layer. In addition, pooling layer can be divided into MP layer and average pooling layer.

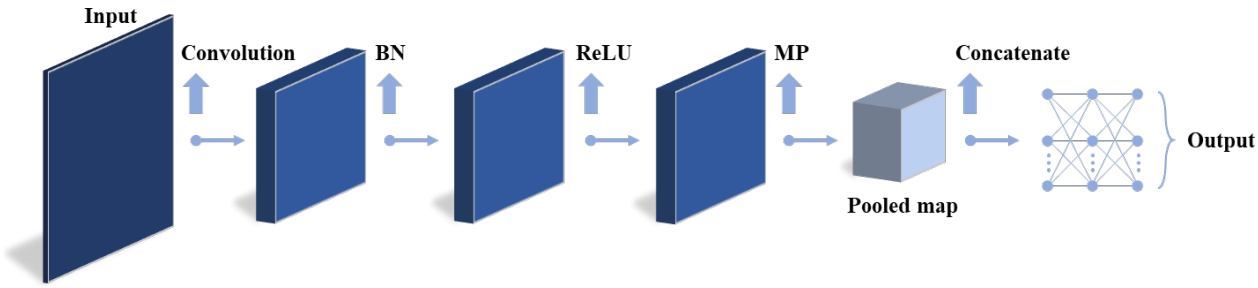

Fig. 17 The architecture of a typical CNN.

1) Convolutional layer

Convolutional layer is the most important layer of a CNN. The core of convolutional layer is the convolutional kernel, which acts as a weight matrix. The size of convolutional kernel is specified manually and multiplied with a part of the feature map. Furthermore, the convolutional kernel is initialized randomly and can be trained. The CNN can extract the weighted combination of features by the convolutional operation.

2) Batch normalization (BN) layer and activation layer

BN layer can effectively prevent gradient vanishing and accelerate network training. The function of the activation layer is to nonlinearly map the convoluted results. The common activation function includes sigmoid function, tanh function, and Rectified Linear Unit (ReLU). In addition, the ReLU has been widely used since its development.

3) Max pooling (MP) layer

The feature maps from the convolutional layer are sent into the MP layer, and then the pooling matrices are spit out as the output. The MP layer is designed to lessen the number of the coefficients and decrease the effect of the overfit. Through the MP layer, the maximum of every kernel size of the feature maps could be chosen as an element of the pooling matrices.

4) Adaptive average pooling (AAP) layer

In this paper, one of the fully-connected layers is replaced by an AAP layer [50,51], as the AAP layer can greatly reduce the number of the coefficients. The output size is set in the definition of the network. The characteristic of “adaptive” is reflected on the size of kernel and stride. In this layer, all the 2-D matrices are calculated into a number.

5) Fully connected layer

There are usually two fully-connected layers in a CNN. “Fully-connected” means that every element in a former layer is connected to every element in a latter layer. Therefore, the number of coefficients in this layer is large beyond imagination. The function of FC layer is to combine the features of the former layer.