Improvement of neural networks learning by feature extraction methods

Sergejs Kodors

Rezekne Higher Education Institution, Brivibas street 39-9, Rezekne, LV-4601, Latvia
sk_7@inbox.lv

Abstract: *This paper discusses a comparison of the feature extraction methods, if a classifier of recognition problem is the artificial neural network with the back-propagation algorithm. The feature extraction methods can improve the classification accuracy and minimize a size of an education dataset and a signal processing time. All these improvements are satisfied by the transformation of the recognizable signal and by the minimization of the signal size. There are two problems to measure these improvements: the improvement of the recognition can be only determined in the experiment, the second problem is that a measured structure of the artificial neural network can contain the unlimited number of the layers and the unlimited number of the perceptrons in every layer. Therefore there is need to argument the chosen parameters of the experiment. The goal of this work is to organize the experiment plan to compare the feature extraction methods. The paper contains the description of the structure of the artificial neural network, the dataset and the elements which are influenced by the feature extraction methods.*

Keywords: artificial neural network, feature extraction, improvement.

Introduction

The artificial neural networks and the neurocomputers are a branch of the computer science, which is based on an imitation of the human brain behaviour. The first artificial neural networks and neurocomputers were described and proposed by the authors McCulloch, Pitts and Rosenblatt in the 1950s (Jasnickiy, 2005).
The artificial neural network is the distributed system of the computing elements – perceptrons. The perceptron imitates a structure and work of the neuron.
The neurocomputer is a type of the processor, which works using the principles of the artificial neural network.
The artificial neural networks are used in the different fields:

- The medical diagnoses;
- The lie detectors;
- The estimators of the exchange rate;
- Other fields.

The artificial neural networks have the following positive features (Haykin, 2006):

- Nonlinearity;
- Input-output mapping;
- Adaptation;
- Fault tolerance;
- Very-large-scale-integrated (VLSI) implementability;
- Uniformity of analysis and design.

The application of the artificial neural networks has some expenses to get the classification accuracy: a complexity and a size of dataset; where the complexity is the signal processing time and the complexity of the artificial neural network structure. These relations are illustrated in Fig. 1.

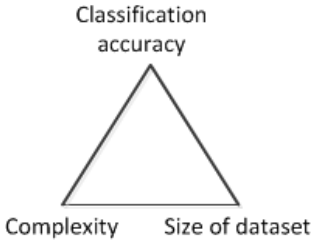

Fig. 1. **Relations of artificial neural network parameters**.

All these relations are explained by the VC-dimension.
The VC-dimension is the maximal number h of the vectors, which can be separated into two classes by 2^h possible ways. According to (Koiran and Sontag, 1996), the artificial neural networks, which use the sigmoidal activation function (1), have the VC-dimension at least as large as the square of the number of the weights.

$$\varphi(v) = \frac{1}{1+\exp(-v)} \tag{1}$$

where φ – an activation function;

v – an induced local field.

This means that the huger dataset needs the more complicated artificial neural network to get the necessary accuracy. According to the curse of dimension, the function of the higher-dimensional space is rather the more complicating than the function of the lower-dimensional space. In other words, if a signal is converted from the higher-dimensional space to the lower-dimensional space, the classification accuracy is improved, but the complexity and the size of the dataset are reduced.

The feature extraction methods improve the artificial neural network classification accuracy and provide the faster and the more cost-effective classifier reducing the number of the signal features.

Feature extraction

The supervised learning is using a training set $S^m = \{x^i, y^i\}_{i=1}^m$ to train the artificial neural network, where y^i - an expected output, x^i - an input, where $x \in \Re^N$, so N coordinates are called features (Krupka et al., 2008).

The feature extraction is a technique to select the features from a transformed space (Li et al., 2009).

The feature extraction is used:

- To improve the classification;
- To reduce the number of the dimensions;
- To get the scale, rotation and translation invariant features.

The scheme of the feature extraction influence can be depicted as the following figure (Fig. 2):

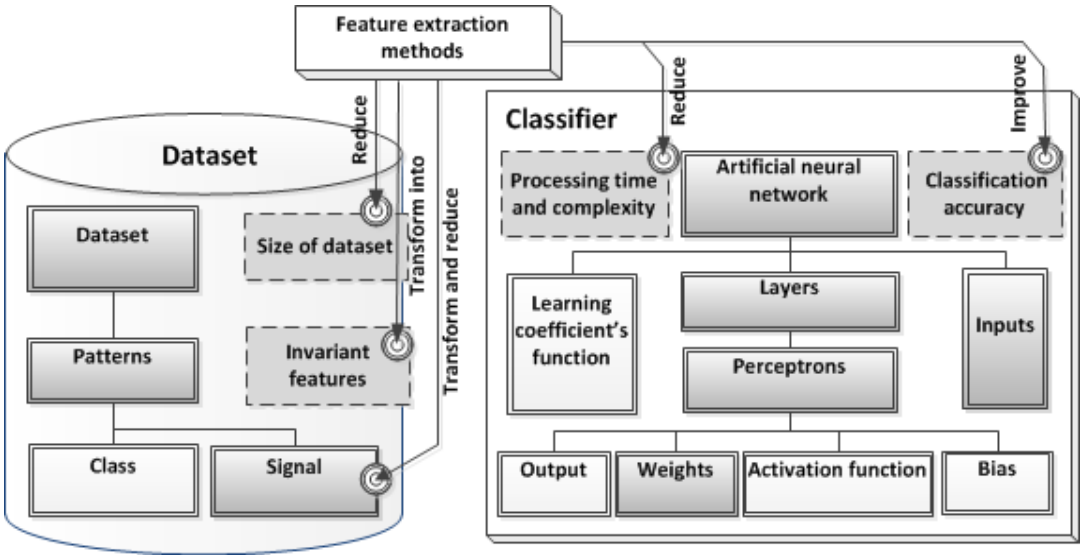

Fig. 2. **Influence of feature extraction methods**.

There are several examples of the feature extraction methods:

- Discrete Fourier transform (DFT);
- Wavelets;
- Principal component analysis (PCA);
- Independent component analysis (ICA);
- Fisher linear discriminant analysis (FLDA).

Back-propagation algorithm and feature extraction methods

The back-propagation algorithm was developed in 1986 (Haykin, 2006). It is the supervised learning algorithm to train the multilayer artificial neural network.

The back-propagation algorithm consists of two stages (Haykin, 2006; Jasnickiy, 2005; Nikolenko and Tulupjev, 2009):

- The forward pass – firstly the signal, which must be classified, is input into the artificial neural network, then it is being processed from a layer to next layer, until the output signal is generated. There are static weights and biases in this process;

- The backward pass - the output, which was generated by the forward pass, is compared with the expected output, the difference is calculated and sent back as an error signal to correct the weights.

In the training stage the feature extraction methods are used as the transformation algorithm of the dataset (Fig. 3) and as the preprocessing in the exploitation stage (Fig. 4).

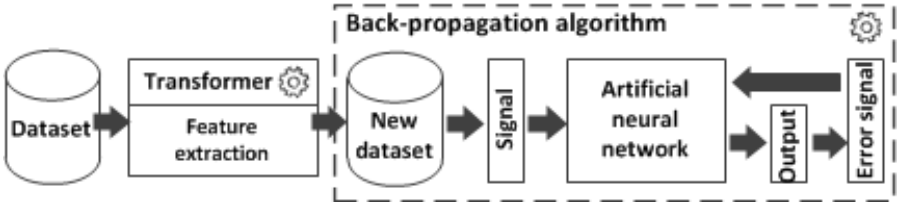

Fig. 3. **Training stage**.

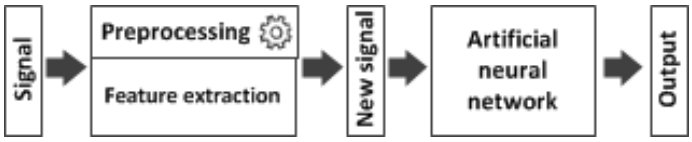

Fig. 4. **Exploitation stage**.

There is possible to see in Fig. 3 and Fig. 4, that every expense has its own field of the significance:

- The dataset is only used in the training stage, therefore the size of the dataset is not important in the exploitation stage;
- The complexity may be not so important in the training stage, but it is important in the exploitation stage.

Argumentation of experiment parameters

To design the experiment to measure the improvement by the feature extraction method, there is need to define the range of the experiment.

What is the minimal classification accuracy?

The minimal classification accuracy is the case, when there is possible to argue, that the correctly classified signals have not been guessed. Therefore there is need of the calculating the probability to guess the results and the method to verify, that the results are different enough.

So there are 2^n possible outputs, where n is a number of the output perceptrons, then there is the probability $1/2^n$ to guess the expected output.

The Chi-square test is a statistical measure, which is used to make comparison between the theoretical populations and actual data to test the goodness of fit. It can be applied to verify the difference for the artificial neural networks (Table 1).

Table 1

Chi-square test for artificial neural network

degree of freedom = 1	Correct answers	Wrong answers
Expected frequency	$p(A)\cdot N = \frac{1}{2^n}N$	$p(\bar{A})\cdot N = \left(1-\frac{1}{2^n}\right)N$
Observed frequency	m	$N-m$
Chi-square	$\chi^2 = \frac{(m-p(A)\cdot N)^2}{m} + \frac{((N-m)-p(\bar{A})\cdot N)^2}{N-m}$	
Minimal classification accuracy is satisfied	If $m > \frac{1}{2^n}N$ and $\chi^2 \geq \chi^2_\alpha$	
α - a level of significance, N – a size of the test set, n - a number of the output perceptrons, χ^2_α - the critical Chi-square, m – a number of the correct answers		

What is the minimal size of the test set?

The cross-validation method is used to measure the possibility of the artificial neural network to classify the signals, which were not used in the training. The cross-validation method divides the dataset on the training set and the test set:

- The training set is used in the learning process;
- The test set is used to measure the classification accuracy.

According to (Kothari, 2004), there are the following conditions for the application of the Chi-square test:

- The overall number of the items are at least 50;

- Every group contains at least 10 items.

These conditions can be applied to the test set. If the observed frequency is less than 10, there can be applied the following rules:

- If the number of the wrong answers is less than 10, the minimal classification accuracy is satisfied, because 40 from 50 is 80% , but the maximal expected number of the correct answers are 50% for the guess, when there are only one output perceptron;
- If the number of the correct answers is less than 10 and $N / 2^n \geq 10$, then the minimal accuracy is not satisfied.

What is the minimal size of the dataset?

According to (Haykin, 2006), the ratio of the training set to the test set is 4:1. So, if the minimal size of the test set is at least 50 items, then the training set must be at least 200 items. It means the minimal size of the dataset is 250 items.

How long must be the artificial neural network trained?

The early stopping method of training uses the cross-validation method to determine the end of the training process. The artificial neural network has been trained for *Z* iterations and then it is tested by the test set. While the classification accuracy is rising, the training process is repeated.

What is the minimal number of the hidden layers?

The universal approximation theorem states that two-layer (with one hidden layer) artificial neural network with the sigmoidal activation function can approximate the classification function of the input-output set with the unspecified size (Haykin, 2006; Jasnickiy, 2005).

What is the maximal number of the perceptrons?

For an evaluation of the number of the perceptrons in the hidden layer, it is possible to use the formula (2), which is the result of the Kolmogorov-Arnold-Hecht-Nilsen theorem (Jasnickiy, 2005).

$$\frac{N_y Q}{1+\log_2 Q} \leq N_w \leq N_y\left(\frac{Q}{N_x}+1\right)\cdot(N_x+N_y+1)+N_y \tag{2}$$

where N_y – a number of the output perceptrons;
Q – a size of a training set;
N_w – a number of the weights;
N_x – a number of the inputs.

So there is possible to calculate the number of the perceptrons in the hidden layer of two-layer artificial neural network by the formula (3).

$$N = \frac{N_w}{N_x + N_y} \tag{3}$$

where N – a number of the perceptrons in the hidden layer.

What is the minimal number of the measures?

The central limit theorem states if a sample is from the normal population, the mean of this sample is itself normally distributed; if the population is not normally distributed, the shape of the distribution depends largely on the shape of the parent population, when a size of the sample is small; but as the size is getting larger ($s > 30$), the shape is becoming more and more like a normal distribution.

What is the value of the learning coefficient and the iterations?

The training process of the artificial neural network and the adaptation process of the self-organizing map have some similarities, therefore there is possible to use the recommendations for the self-organizing maps.

The source (Haykin, 2006) advices the following parameters:

- The learning coefficient $\eta \in [0.01;0.1]$;
- The number of the iterations to organize the map is equal to 1000;
- The formula to calculate the learning coefficient:

$$\eta = \eta_0 \cdot \exp(-\frac{n}{\tau}) \tag{4}$$

where n – an iteration;
η_0 – an initial learning coefficient;
τ – some coefficient.

- The recommended parameters for the formula (4): $\eta_0 = 0.1$ and $\tau = 1000$.

Results and discussion

All the argumentations and the ranges of the experiment, which were discussed in the previous section, can be formed as the following experiment design:

1. Prepare at least 250 items for the dataset;
2. Convert the dataset by the feature extraction method;
3. Randomly take at least 50 items for the test set from the dataset;
4. Calculate the maximal number of the weights using the formula (2);
5. Calculate the maximal number of the hidden perceptrons N_h using the formula (3);
6. Prepare the artificial neural network with one hidden layer and one hidden perceptron, the activation function is the sigmoidal function (1);
7. Train the artificial neural network using the early stopping method of training. The training time is 1000 iterations before the testing, the formula of the learning coefficient is (4), the initial learning coefficient is 0.1 and the coefficient $\tau = 1000$;
8. Repeat the 7th step 99 times and calculate the mean observed frequency of the correct answers;
9. If the mean observed frequency is greater than $1/2^n$, the difference is verified by the Chi-square test, otherwise go to the 11th step;
10. If the Chi-square test has showed that the observed frequency deviates from the expected frequency, output the classification accuracy and the related number of the hidden perceptrons, otherwise go to the 11th step;
11. If the number of the hidden perceptrons is less than N_h, increase the number of the hidden perceptrons by 1 and back to the 7th step, otherwise output the best result of the classification accuracy and the related number of the hidden perceptrons.

Conclusion

This work was prepared with a goal to compare different feature extraction methods for RGB image recognition. Using the described experiment design the scientist can define the minimal number of the hidden perceptrons to classify the dataset. The group of the parameters: the minimal number of the hidden perceptrons, the number of the features, the number of the output classes and the dataset; can form the baseline to compare the feature extraction methods. If two experiments have the same datasets, comparing the minimal number of weights to classify the dataset, one can compare the improvements of two feature extraction methods. If the number of the features are the same, one can compare the number of the perceptrons, otherwise there is need to calculate the weights by the formula (5).

$$W = (i+1) \cdot p_h + (p_h + 1) \cdot p_{out} \tag{5}$$

where W – a number of the weights;
i – a number of the features;
p_h – a number of the hidden perceptrons;
p_{out} – a number of the classes (output perceptrons).

This paper has not description of the complication of the dataset, but this parameter is important to compare the results in the case, if the different datasets were used in the experiments. There is need to add, that the described experiment is only useful for the huge populations, for example, the characters recognition with the image size equal to 3x5, then the population will be 2^{15} items, if there are only the black and white pixels.

References

Haykin, S., 2006. Neironye seti: polnii kurs, izdanie vtoroe (Neural networks: a comprehensive foundation, second edition), Viljams, Moskva, 1104 p. (In Russian)

Jasnickiy, L., 2005. Vvedenie v iskusstvennyi intellekt (Introduction to artificial intelligence), Akademija, Moskva, 176 p. (In Russian)

Koiran, P., and Sontag, E.D., 1996. Neural networks with quadratic VC dimension. Advances in Neural Information Processing Systems, 8, pp.197-203.

Kothari, C.R., 2004. Research methodology: methods and techniques, 2nd ed., New Age International Publishers, New Delhi, 418 p.

Krupka, E., Navot, A., Tishby, N., 2008. Learning to select features using their properties. Journal of Machine Learning Research, 9, pp.2349-2376.

Li, Y., Lu, B.-L., Zhang, T.-F., 2009. Combining feature selection with extraction: unsupervised feature selection based on principal component analysis. International Journal on Artificial Intelligence Tools, 18, pp.883-904.

Nikolenko, S.I. and Tulupjev, A.L., 2009. Samoobuchajushiesja sistemy (Self-instruction systems). Moskovskii center nepreryvnogo matematicheskogo obrazovanija, Moskva, 288 p. (In Russian)