Postprint: Speech Emotion Recognition Based on BP Neural Network Optimized by Improved Genetic Algorithm

Authors: Chen Chuang, Ryad Chellali, Xing Yin

Date: 2018-05-20T00:00:00+00:00

Abstract

Speech emotion recognition technology is playing an increasingly important role in human life. To more effectively identify emotion types in speech signals, an improved adaptive genetic algorithm optimized BP neural network recognition algorithm (IAGA-BP) is proposed. The algorithm improves the selection operator in the adaptive genetic algorithm on one hand, and modifies the crossover and mutation probability formulas in the adaptive genetic algorithm on the other hand. Through these improvements to the adaptive genetic algorithm, the optimization performance of the genetic algorithm is enhanced, which is then utilized to optimize the initial weights and thresholds of the BP neural network. In comparisons with BP, GA-BP, and AGA-BP networks, experimental results demonstrate that the IAGA-BP network can effectively improve the speech emotion recognition rate and accelerate network convergence speed.

Full Text

Speech Emotion Recognition Based on Improved Genetic Algorithm Optimized BP Neural Network

Chen Chuang[1], Ryad Chellali[1]†, Xing Yin[2] [1]College of Electrical Engineering & Control Science, Nanjing Tech University, Nanjing 211816, China [2]College of Geomatics & Geoinformation, Guilin University of Technology, Guilin, Guangxi 541004, China

Abstract: Speech emotion recognition technology is playing an increasingly important role in human life. To more effectively identify emotion types in speech signals, this paper proposes an improved genetic algorithm optimized BP neural network (IAGA-BP). The algorithm improves the selection operator in the adaptive genetic algorithm and modifies the crossover and mutation probability formulas. Through these improvements to the adaptive genetic algorithm, the optimization performance of the genetic algorithm is enhanced, which is then used to optimize the initial weights and thresholds of the BP neural network. Experimental comparisons with BP, GA-BP, and AGA-BP networks demonstrate that the IAGA-BP network can effectively improve speech emotion recognition rates and accelerate network convergence speed.

Keywords: genetic algorithm; BP neural network; speech emotion recognition; adaptive; optimization

0 Introduction

As science and technology advance rapidly and human dependence on computers continues to grow, human-computer interaction capabilities have attracted increasing attention from researchers. Studies by Stanford University' s Reeves and Nass have shown that the key factors in human-computer interaction are consistent with those in human-to-human communication, with "emotional intelligence" being the most critical capability. Therefore, effectively recognizing emotion types in human speech signals and analyzing speakers' emotional states to make human-computer interaction more friendly represents the future direction of information processing research. Speech emotion recognition technology has been successfully applied in various fields such as assisted lie detection, remote teaching, and clinical medicine.

Numerous methods exist for speech emotion recognition, including Gaussian Mixture Models, Hidden Markov Models, Support Vector Machines, and Artificial Neural Networks. Artificial neural networks, inspired by the structure and principles of biological neural networks, simulate the human brain' s neural network to process complex input-output information and discover nonlinear mapping relationships between them. Among these, BP neural networks are the most widely used. However, BP neural networks suffer from slow convergence and a tendency to fall into local minima, with network initial weights and thresholds profoundly affecting training outcomes. Genetic algorithms, with their excellent global search capabilities, can address these issues. To further improve speech emotion recognition rates, this paper explores a recognition model combining an improved genetic algorithm with BP neural networks.

1 BP Neural Network

A BP neural network is a network with no fewer than 3 layers, comprising an input layer, hidden layer, and output layer. Its structure is shown in [Figure 1: see original paper]. It employs the gradient descent method for error function learning, with the learning process 主要包括 input signal forward propagation and error backpropagation. During forward propagation, signals are input through the input layer, processed by the hidden layer, and transmitted to the output layer. If the actual output does not match the expected output, the process switches to error backpropagation. During backpropagation, errors are transmitted layer by layer from the output layer to the input layer, continuously adjusting connection weights until the error is reduced. These two processes cycle repeatedly until the mean square error reaches a set standard or the maximum number of iterations is reached. However, in practical applications, BP neural networks have been found to converge slowly and easily fall into local minima, with the network' s initial weights and thresholds significantly influencing training results.

2 Improved Adaptive Genetic Algorithm Optimized BP Neural Network

To address the inherent defects of BP neural networks, we can leverage the global search capability of genetic algorithms to optimize them. Genetic algorithms simulate biological evolution processes, incorporating selection, crossover, and mutation operations to obtain optimal solutions through continuous evolution. However, standard genetic algorithms also tend to fall into local optima, affecting their optimization capability. Therefore, this paper improves the genetic algorithm' s selection operator and modifies the crossover and mutation probability formulas to enhance algorithm performance and optimize the initial weights and thresholds of the BP neural network.

2.1 Encoding and Fitness Function Selection

The weights and thresholds of BP neural networks are decimals between -1 and 1 and are numerous, making binary encoding unsuitable. Here, we adopt real-number encoding rules. During the algorithm search process, the reciprocal of the mean square error serves as the standard for evaluating individual quality. The fitness function formula is as follows:

$$fitness(x) = \frac{1}{E(x)}$$

where x represents a population individual, corresponding to the arrangement of weights and thresholds in the BP network. A larger fitness value indicates a better individual.

2.2.1 Improved Selection Operator

Genetic algorithm selection operators commonly use fitness proportionate methods and elitist strategies. The fitness proportionate method suffers from random errors, where individuals with high fitness may be eliminated while those with low fitness are selected, demonstrating poor competitiveness, especially in large populations. The elitist strategy protects the best individual by directly copying it to the next generation, avoiding destruction by crossover and mutation operations, but ignores the possibility that other individuals may destroy population diversity. To address this, we propose a novel selection method that can effectively select superior individuals from the population. The algorithm steps are as follows:

a) Determine an initial population and calculate the fitness value of each individual in the population;

b) Sort individuals within the population in descending order by fitness value;

c) Directly inherit the top two individuals to the next generation, and divide the remaining individuals into four equal parts: excellent, good, medium, and poor;

d) Copy the excellent individuals twice, eliminate the poor individuals, and copy the good and medium individuals once each.

The following graphical representation illustrates this concept using a population of 14 individuals labeled with fitness values and numbers:

a) Determine an initial population with a size of 14:

[Individual fitness values and numbers shown in original]

b) Sort individuals within the population in descending order by fitness value:

[Sorted individuals shown in original]

c) Directly inherit the top two individuals to the next generation, and divide the remaining individuals into excellent, good, medium, and poor groups:

[Grouped individuals shown in original]

d) Copy excellent individuals twice, eliminate poor individuals, and copy the remaining individuals once each:

[Final selection shown in original]

This selection method improves the population' s average fitness while maintaining population diversity.

2.2.2 Crossover Operation

According to real-number encoding rules, new individuals are obtained by crossing two selected individuals from the population with a certain probability:

$$w_{kj}^l = (1-b)w_{kj}^l + bw_{lj}^l w_{lj}^l = (1-b)w_{kj}^l + bw_{lj}^l$$

where w_{kj}^l and w_{lj}^l represent the l-th gene of the k-th and j-th individuals, respectively; b is a random number in [0,1].

2.2.3 Mutation Operation

A new individual is obtained by mutating a randomly selected individual from the population with a certain probability:

$$w_{ij} = \begin{cases} w_{ij} + (w_{ij}^{\max} - w_{ij}) \cdot f(g), & r \geq 0.5 \\ w_{ij} + (w_{ij} - w_{ij}^{\min}) \cdot f(g), & r < 0.5 \end{cases}$$

where $w_{ij}^{\max}$ and $w_{ij}^{\min}$ are the upper and lower bounds of gene w_{ij}, r and g are random numbers in [0,1], g is the current iteration number, and G is the maximum evolution generation.

2.3 Adaptive Crossover and Mutation Probability

Standard genetic algorithms use fixed crossover and mutation probabilities, which are difficult to tune optimally in practice. When the crossover rate is too small, the algorithm tends toward premature convergence; when too large, it easily destroys excellent individuals. When the mutation rate is too small, the algorithm struggles to generate new pattern structures; when too large, it becomes a purely random algorithm. To address this issue, Srinivas et al. proposed the Adaptive Genetic Algorithm (AGA), where individuals in the population can adaptively adjust crossover and mutation probabilities based on their environment:

$$P_c = \begin{cases} k_1 \frac{f_{\max} - f'}{f_{\max} - f_{\text{avg}}}, & f' \geq f_{\text{avg}} \\ k_2, & f' < f_{\text{avg}} \end{cases}$$

$$P_m = \begin{cases} k_3 \frac{f_{\max} - f}{f_{\max} - f_{\text{avg}}}, & f \geq f_{\text{avg}} \\ k_4, & f < f_{\text{avg}} \end{cases}$$

where f' is the larger fitness value of the two individuals to be crossed; $f_{\max}$ is the population' s maximum fitness; f_{avg} represents the population' s average fitness; f represents the fitness of the mutated individual; and k_1, k_2, k_3, k_4 are adaptive control parameters. Srinivas et al. believed that when an individual' s fitness is below the population' s average fitness, higher crossover and mutation probabilities should be used if the individual is selected to improve its performance; when an individual is close to the population' s maximum fitness, lower crossover and mutation probabilities should be used to preserve its excellent patterns. However, AGA is not obvious in the early evolution stage because the better individuals in the population hardly change during this period, and these excellent individuals are not necessarily the global optimal solution, which increases the possibility of evolution falling into local convergence.

2.4 Improved Adaptive Crossover and Mutation Probability

In genetic algorithms, $E(X)$ represents the average level of population fitness values, and $D(X)$ represents the discrete degree of population fitness values. The similarity coefficient formula is defined as:

$$\varphi = \frac{1}{1 + E(X) + D(X)}$$

where:

$$E(X) = \frac{f_1 + f_2 + \cdots + f_M}{M} = \frac{\sum_{i=1}^{M} f_i}{M}$$

$$D(X) = \frac{(f_1 - E(X))^2 + (f_2 - E(X))^2 + \cdots + (f_M - E(X))^2}{M} = \frac{\sum_{i=1}^{M} (f_i - E(X))^2}{M}$$

where φ represents the similarity coefficient of the current population, M represents the population size, and f_i represents the fitness value of the i-th individual in the population. As the genetic algorithm' s evolution generations increase, the population' s average fitness level continuously improves, and duplicate individuals gradually increase, reducing the discrete degree of individual fitness values in the population and making individuals increasingly similar. Based on the AGA algorithm' s design principles and the similarity coefficient definition, the improved adaptive genetic algorithm' s crossover and mutation probability formulas are:

$$P_c = \begin{cases} k_1 \cdot \frac{f_{\max} - f'}{f_{\max} - f_{\text{avg}}} \cdot e^{\varphi} + k_2, & f' \geq f_{\text{avg}} \\ k_3 \cdot \frac{f_{\max} - f'}{f_{\max} - f_{\text{avg}}} + k_4, & f' < f_{\text{avg}} \end{cases}$$

$$P_m = \begin{cases} k_5 \cdot \frac{f_{\max} - f}{f_{\max} - f_{\text{avg}}} \cdot e^{\varphi} + k_6, & f \geq f_{\text{avg}} \\ k_7 \cdot \frac{f_{\max} - f}{f_{\max} - f_{\text{avg}}} + k_8, & f < f_{\text{avg}} \end{cases}$$

where P_c is the crossover probability; P_m is the mutation probability; k_1 to k_8 are adaptive control parameters. From the above analysis, the similarity coefficient φ continuously increases. Since e^{φ} is an increasing function, P_c increases accordingly. Similarly, the mutation probability increases. The adaptive crossover probability P_c is set between 0.6 and 0.9, and the mutation probability is set between 0.001 and 0.1. k_1 and k_2 are constants, so the control parameters are set as: $k_1 = 0.6$, $k_2 = 0.1$, $k_3 = 0.6$, $k_4 = 0.9$, $k_5 = 0.198$, $k_6 = 1$, $k_7 = 0.001$, $k_8 = 0.1$.

IAGA-BP Algorithm Implementation Steps

a) Provide experimental samples, divide them into training and test samples, and perform normalization;

b) Set operational parameters such as population size N, maximum evolution generation G, and initial hidden layer node count;

c) Randomly generate N individuals as the initial population and perform real-number encoding;

d) Decode each individual to obtain the initial connection weights and thresholds of the hidden layer; train the BP neural network and calculate individual fitness values according to equation (1);

e) Sort individual fitness values in descending order and select individuals according to the improved selection operator in section 2.2.1;

f) Perform crossover and mutation operations from sections 2.2.2 and 2.2.3 on selected individuals using the adaptive crossover and mutation probabilities from equations (10) and (11);

g) Check if the maximum evolution generation is reached; if satisfied, output the optimal individual; otherwise, return to step d);

h) Decode the optimal individual to obtain the BP network' s initial connection weights and thresholds, and establish the IAGA-BP model.

3 Speech Emotion Recognition Based on IAGA-BP Algorithm

The speech emotion database used in this paper comes from the Berlin Emotional Speech Database. The Berlin Emotional Speech Database was created by the Technical University of Berlin, with 10 non-professional actors (5 male and 5 female) performing 7 emotions (anger, happiness, calmness, sadness, fear, disgust, and boredom) on 10 recording scripts, totaling 800 utterances. After 20 volunteers listened and identified the emotions, 535 utterances were retained. This experiment selected the first 5 speech emotions, retaining 400 utterances, with training and test samples allocated in a 1:1 ratio.

To effectively represent emotional information in speech signals, this paper extracted 4 categories of features and their derivative coefficients: short-term energy, pitch frequency, formants, and Mel-Frequency Cepstral Coefficients (MFCC), forming a total of 140-dimensional speech emotion features. Specifically: short-term energy, pitch frequency, voiced frame differential pitch, first/second/third formants, 0-12 order MFCC and their first-order differential maximum, minimum, mean, and variance; jitter of short-term energy, linear regression coefficients, and mean square error of linear regression coefficients; percentage of energy in the 0-250Hz band relative to total energy; first and second-order jitter of pitch frequency; first-order jitter of the first/second/third formants; and maximum, minimum, and mean of the second formant ratio. Specific calculation methods are detailed in reference [15].

After multiple experiments, a three-layer BP network structure of 140-19-5 was selected, with specific parameters: maximum iterations 3000, target accuracy 0.001, and learning rate 0.1. GA parameters were set as: initial population size 50, maximum iterations 200, crossover rate 0.8, and mutation rate 0.05. Based on the Berlin emotional dataset, BP, GA-BP, AGA-BP, and IAGA-BP networks were used for 5-emotion recognition, with results shown in .

As shown in , the IAGA-BP network achieved the highest recognition rate of 94.50% for the 5 emotions, followed by AGA-BP at 92.50%, GA-BP at 91.00%, and BP at 90.00%. The optimized BP networks showed significant improvement in sadness recognition, increasing from 67.74% with BP to 87.10% with GA-BP and 83.87% with AGA-BP, with IAGA-BP further improving sadness recognition to 96.77%. This indicates that IAGA-BP is more sensitive to negative emotions. Generally, negative emotions have the greatest impact on people, and effectively identifying these negative emotions can help improve individual cognition and work efficiency while reducing factors that affect cognitive and work abilities. In the 5-emotion recognition task, IAGA-BP achieved recognition rates of 100%, 82.35%, 94.87%, 96.77%, and 93.94% for each emotion, respectively. Except for a slightly lower recognition rate for happiness, all other emotions maintained high recognition rates, demonstrating that IAGA-BP provides balanced recognition across various speech emotions with excellent comprehensive performance.

shows the recognition results for different networks.

The fitness value curves of the optimal individuals for GA, AGA, and IAGA are shown in [Figure 2: see original paper]. As seen in [Figure 2: see original paper], IAGA demonstrates significantly stronger optimization capability than AGA and GA. Although GA and AGA converge quickly, they do not find the optimal individual. During evolution, GA and AGA experience multiple fitness drops, indicating that the optimal individual may be destroyed during each evolution. IAGA protects the optimal individual through elitist preservation and improves overall performance through group selection and elimination. When GA evolves to generation 24 and AGA to generation 29, both converge and cannot escape old patterns, whereas IAGA shows better adaptive performance by adjusting crossover and mutation rates to escape old patterns and drive the algorithm forward with new patterns. IAGA exhibits a stepwise upward trend, escaping local convergence and demonstrating strong adaptive performance.

The training error curves for GA-BP, AGA-BP, and IAGA-BP networks are shown in [Figure 3: see original paper]. The GA-BP network reaches target accuracy at iteration 1608; AGA-BP at iteration 1392; while IAGA-BP reaches target accuracy at iteration 1276, indicating faster convergence speed.

4 Conclusion

This paper proposes a speech emotion recognition model based on an improved adaptive genetic algorithm optimized BP neural network. The improved adaptive genetic algorithm selects the top two optimal individuals for direct inheritance to the next generation based on fitness value ranking, while performing group selection and elimination on remaining individuals. Additionally, it introduces the concept of similarity coefficient and modifies the adaptive crossover and mutation probabilities. The improved adaptive genetic algorithm is applied to optimize the initial weights and thresholds of the BP neural network. Based on 140-dimensional emotional feature parameters extracted from the Berlin dataset, GA-BP, AGA-BP, and IAGA-BP networks were used to recognize 5 speech emotions. Experimental results demonstrate that the IAGA-BP model effectively improves speech emotion recognition rates and accelerates network convergence speed.

References

[1] Cheng Y M, O' Shaughnessy D, Mermelstein P. Statistical recovery of wideband speech from narrowband speech [J]. IEEE Trans on Speech and Audio Processing, 1994, 2(4): 544-548.

[2] Han Wenjing, Li Haifeng, Ruan Huabin, et al. Survey on speech emotion recognition research progress [J]. Journal of Software, 2014, 25(1): 37-50.

[3] Fragopanagos N, Taylor J G. Emotion recognition in human-computer interaction [J]. Journal of the International Neural Network Society, 2001, 18(1): 32-80.

[4] Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction [J]. Neural Networks, 2005, 18(4): 389-405.

[5] France D J, Shiavi R G, Silverman S, et al. Acoustical properties of speech as indicators of depression and suicidal risk [J]. IEEE Trans on Bio-medical Engineering, 2000, 47(7): 829-837.

[6] Ayadi M E, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases [J]. Pattern Recognition, 2011, 44(3): 572-587.

[7] Xing H H, Lin H Y. An intelligent method optimizing BP neural network model [J]. Advanced Materials Research, 2013, 607: 2470-2474.

[8] Li X W, Ouyang H B. Evaluating artistic voice of singing objectively using BPNN [J]. Advanced Materials Research, 2012, 546-547: 1240-1244.

[9] Hou W, Liu X, Chen D, et al. Finance-based scheduling for construction project in China with improved genetic algorithms [J]. Journal of Convergence Information Technology, 2012, 7(18): 29-38.

[10] Liu Haoran, Zhao Cuixiang, Li Xuan, et al. Research on a neural network optimization algorithm based on improved genetic algorithm [J]. Chinese Journal of Scientific Instrument, 2016, 37(7): 1573-1580.

[11] Srinivas M, Patnaik L M. Adaptive probabilities of crossover and mutation in genetic algorithms [J]. IEEE Trans on Systems Man and Cybernetics, 2002, 24(4): 656-667.

[12] Wang Xiaoping, Cao Liming. Genetic Algorithm: Theory, Application and Software Implementation [M]. Xi' an: Xi' an Jiaotong University Press, 2002: 73-74.

[13] Liu Hao, Qiu Jianxin. Improved adaptive genetic algorithm in optimal layout of leather rectangular parts [J]. Advances in Natural Science, 2015, 8(3): 221-224.

[14] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech [C]// Proc of INTERSPEECH. 2005: 1517-1520.

[15] Liang Ruiyu, Zhao Li, Wei Xin. Speech Signal Processing Experimental Tutorial [M]. Beijing: Mechanical Industry Press, 2016: 231-234.

[16] Zhang Xiaodan, Hu Feng, Zhao Li. Practical speech emotion recognition based on improved shuffled frog leaping algorithm and support vector machine [J]. Signal Processing, 2011, 27(5): 678-689.

Note: Figure translations are in progress. See original paper for figures.