EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Speech emotion recognition method in educational scene based on machine learning

Yanning Zhang[1] and Gautam Srivastava[2,3,*]

[1]School of Telecommunication Engineering, Beijing Polytechnic, Beijing 100176, China
[2]Department of Mathematics and Computer Science, Brandon University, Brandon, Canada
[3]Research Centre for Interneural Computing, China Medical University, Taichung, Taiwan

Abstract

In order to effectively improve the accuracy and anti noise performance of speech emotion recognition in educational scenes, a new method based on machine learning is studied. Based on the fundamental frequency and resonance degree, the speech emotional characteristics of educational scenes are collected respectively. Using the kernel canonical correlation analysis in machine learning algorithm, the emotional feature samples are nonlinearly mapped to the high-level feature space, the correlation between different emotional features is analyzed, the nonlinear correlation between the two groups of variables is obtained, the two speech emotional features are integrated, and the feature samples are constructed. SVM is used to establish speech emotion recognition classifier, and genetic algorithm is used to determine the optimal parameters. The experimental results show that the emotion recognition rate of this method is more than 90%, and the emotion recognition rate of anger, fear, happiness and sadness is more than 95%; After adding a variety of noise, the speech emotion recognition results are completely consistent with the actual speech emotion, which shows that this method has high anti noise performance.

Keywords: Machine learning, Educational scenes, Speech emotion recognition, Kernel canonical correlation analysis, Support vector machine.

Received on 11 December 2021, accepted on 05 February 2022, published on 10 February 2022

doi: 10.4108/eai.10-2-2022.173380

*Corresponding author. Email: srivastavag@brandonU.ca

1. Introduction

Language is one of the most important tools for human communication. It has the characteristics of natural, convenient, accurate and efficient transmission of information [1]. It is a unique function of human beings, and speech is an important carrier of language [2]. As a communication tool and a sound form of language, speech is not limited by time and conditions [3]. In the process of education and teaching, nonverbal behavior is often used as an auxiliary means of speech behavior in human communication [4]. As a nonverbal behavior contained in sound, phonological emotion plays an important role in assisting knowledge teaching and transmitting teaching effect feedback [5], which has also attracted the attention of more and more researchers. Zvarevashe and olugbara studied speech emotion recognition method based on two-dimensional convolutional neural network deep learning algorithm [6], and developed a customized two-dimensional convolutional neural network, which can extract and classify speech features at the same time. The research shows that the deep learning algorithm can effectively extract the robust salient features in the data set. Hajarolasvadi and Demirel studied the three-dimensional CNN speech emotion recognition method based on K-means clustering and spectrum [7], divided each speech signal into overlapping frames of the same length, extracted 88 dimensional audio feature vectors,

and performed K-means clustering on the features of all frames of each audio signal to complete speech emotion recognition. Lee et al. established a speech emotion recognition model based on CNN's transfer learning and attention mechanism [8], and carried out speech emotion recognition with CNN model as the core and attention mechanism.

Machine learning (ML) is an interdisciplinary subject, involving probability theory, statistics, approximation theory, convex analysis, algorithm complexity theory and other disciplines [9]. Machine learning algorithms can be divided into three categories: supervised learning, unsupervised learning and reinforcement learning [10]. When a specific data set (training set) has specific attributes (labels), but other data have no labels or need prediction labels, supervised learning can be used, mainly including decision tree algorithm, support vector machine algorithm, classification algorithm, logistic regression algorithm, etc. [11]. Unsupervised learning can be used for a given unlabeled data set (data is not pre allocated) to find out the potential relationship between data. Reinforcement learning is between the two. Each prediction has a certain form of feedback, but there is no accurate label or error information, mainly including clustering algorithm, principal component analysis algorithm, independent component analysis algorithm, association analysis algorithm and singular value decomposition algorithm [12]. The machine learning algorithm is applied to the speech emotion recognition of educational scene, and the speech emotion recognition method of educational scene based on machine learning is studied to obtain accurate speech emotion recognition results of educational scene.

2. Speech emotion recognition method in educational scene

2.1 Speech emotion feature extraction in educational scenes

Feature extraction is an important step in the modelling of speech emotion recognition in educational scenes. The educational speech signal has short-term stability, so it can process the educational scene speech signal and extract the required feature parameters. Windowing and framing the speech signal of education scene can effectively use the short-term stability of the speech signal of education scene for feature extraction and analysis. Windowing is to multiply the original educational scene speech signal with a specific window function to obtain the windowed speech signal.

There are many speech features related to emotion, mainly including pitch frequency, formant and so on. These are important speech features, which have wide and important applications in the fields of speech enhancement, speech coding, speech synthesis, speech recognition, speaker recognition, emotion recognition, speech hiding, sound source location and so on, especially for Chinese.

Characteristics of fundamental frequency

Pitch period is one of the important parameters describing excitation source in speech signal processing of educational scene. When people make voiced sound, the air flow makes the vocal cords produce relaxation oscillation vibration through the glottis, producing a quasi-periodic pulse air flow. This air flow excites the vocal tract to produce voiced sound, also known as voiced speech, which carries most of the energy in speech. The frequency of this vocal cord vibration is called the fundamental frequency [13], and the corresponding period is called the pitch period. At present, pitch detection algorithms mainly include autocorrelation function method, average amplitude difference function method, cepstrum method, and some improved algorithms based on the above algorithms.

The fundamental frequency is closely related to the size and tightness of the vocal cord. Under different emotional states, the vocal cords change accordingly. For example, when angry, the vocal cords stretch and tighten, so the fundamental frequency also changes accordingly [14]. Eight features related to the fundamental frequency are selected for speech emotion recognition in educational scenes, which are the maximum value of the fundamental frequency $F_{0\_\max}$, the minimum value of the fundamental frequency $F_{0\_\min}$, the mean value of the fundamental frequency F_{0_mean}, the standard deviation of the fundamental frequency F_{0_std}, the range of the fundamental frequency F_{0_range}, the mean of the fundamental frequency trajectory difference $F_{0\_\mathrm{diff\_mean}}$ and the standard deviation of the fundamental frequency trajectory difference $F_{0\_\mathrm{diff\_std}}$. Their calculation formulas are as follows:

$$F_{0\_mean} = \frac{\sum_{t=1}^{N} F_{0t}}{N} \tag{1}$$

$$F_{0\_\mathrm{std}} = \sqrt{\sum_{t=1}^{N} \frac{\left(F_{0t} - F_{0\_mean}\right)^2}{(N-1)}} \tag{2}$$

$$F_{0\_\mathrm{range}} = F_{0\_\max} - F_{0\_\min} \tag{3}$$

$$F_{0\_\mathrm{diff\_mean}} = \frac{\sum_{i=1}^{N-1} F_{0\_\mathrm{diff}_i}}{N-1} \tag{4}$$

$$F_{0\_diff\_std}=\sqrt{\sum_{i=1}^{N-1}\frac{\left(F_{0\_diff_i}-F_{0\_diff\_mean}\right)^2}{(N-2)}} \tag{5}$$

Where, $F_{0\_diff_i}$ represents the absolute value of fundamental frequency trajectory difference, $i=1,2,\cdots,N-1$.

Amplitude characteristics

Formant refers to some areas where energy is relatively concentrated in the spectrum of sound. Formant is not only the determinant of sound quality, but also reflects the physical characteristics of sound channel (resonant cavity) [15]. The original meaning of formant refers to the resonant frequency of sound cavity. Similar to pitch extraction, formant estimation is also plagued by many problems, including false peaks, formant merging, high pitch speech, etc., and its main methods include cepstrum method and LPC method.

Amplitude describes the intensity of speech emotional information in educational scenes, mainly in the rhythm of speech. When in the state of anger or surprise, the volume increases, while in the state of sadness, the volume is low, so the amplitude is also an indispensable voice emotional feature of educational scenes. The amplitude is measured by the short-time energy of each frame speech signal $s(n)$, and its calculation formula is as follows:

$$E_n=\sum_{m=-\infty}^{\infty}\left[s(n)w(n-m)\right]^2 \tag{6}$$

Where: $w(n)$ is the window function.

Eight features related to vibration frequency are selected for speech emotion recognition in educational scenes, which are the maximum value $E_{\text{n\_max}}$, minimum value $E_{\text{n\_min}}$, mean value $E_{\text{n\_}mean}$, standard deviation $E_{\text{n\_std}}$, value range $E_{\text{n\_range}}$, mean of trajectory difference $E_{\text{n\_diff\_mean}}$ and standard deviation of trajectory difference $E_{\text{n\_diff\_std}}$.

2.2 Feature fusion based on KCCA

According to the emotional fundamental frequency and amplitude characteristics of the collected educational scene speech, the kernel canonical correlation analysis (KCCA) algorithm in machine learning algorithm is used for feature fusion.

KCCA nonlinearly maps the samples to the high-dimensional feature space [16], and then performs correlation analysis to obtain the nonlinear correlation between the two groups of variables. Let $X=(x_1,x_2,\cdots,x_N)$ and $Y=(y_1,y_2,\cdots,y_N)$ represent the fundamental frequency eigenvector and amplitude eigenvector of speech emotion in the educational scene respectively. KCCA acts on two sets of eigenvectors through two nonlinear mappings Φ and ψ:

$$\begin{cases}\Phi: x\to\Phi(x)\in F_x\\ X\to\Phi(X)=\left[\Phi(x_1),\Phi(x_2),\cdots,\Phi(x_N)\right]\end{cases} \tag{7}$$

$$\begin{cases}\psi: y\to\psi(y)\in F_y\\ Y\to\Phi(Y)=\left[\psi(y_1),\psi(y_2),\cdots,\psi(y_N)\right]\end{cases} \tag{8}$$

Let the kernel functions be K_X and K_Y respectively, and the kernel matrix is described as follows:

$$\begin{cases}K_X=\Phi^T(X)\Phi(X)\\ K_Y=\psi^T(Y)\psi(Y)\end{cases} \tag{9}$$

where

$$\begin{cases}(K_X)_{ij}=k_x(x_i,x_j)\\ (K_Y)_{ij}=k_y(y_i,y_j)\end{cases} \tag{10}$$

Kernel matrix centralization: zero mean of training samples is made.

$$\bar{K}=K-\frac{1}{N}1_{N\times N}K-\frac{1}{N}K1_{N\times N}+\frac{1}{N^2}1_{N\times N}K1_{N\times N} \tag{11}$$

The goal of KCCA is to find the projection directions α_Φ and β_ψ so that (12) is the largest when the following criterion function is used:

$$J(\alpha_\Phi,\beta_\psi)=\frac{\alpha_\Phi^T\Phi(X)\psi(Y)^T\beta_\psi}{\sqrt{\alpha_\Phi^T\Phi(X)\Phi(X)^T\alpha_\Phi\cdot\beta_\psi^T\psi(Y)\psi(Y)^T\beta_\psi}} \tag{12}$$

The vector α_Φ is located in the space formed by the sample $\Phi(x_1),\Phi(x_2),\cdots,\Phi(x_N)$. According to the kernel regeneration theory, there is an N-dimensional vector ξ so that $\alpha_\Phi=\Phi(X)\xi$; Similarly, there is an N-dimensional vector η and let $\beta_\psi=\psi(Y)\eta$, which can be brought into equation (13) to obtain:

$$J(\xi,\eta)=\frac{\xi^TK_XK_Y\eta}{\sqrt{\xi^TK_X^2\xi\cdot\eta^TK_Y^2\eta}} \tag{13}$$

In order to prevent meaningless canonical correlation vectors, it is necessary to introduce a regular term to constrain equation (14):

$$J(\xi,\eta)=\frac{\xi^TK_XK_Y\eta}{\sqrt{\xi^T\left((1-\tau)K_X^2+\tau K_X\right)\xi\cdot\eta^T\left((1-\tau)K_Y^2+\tau K_Y\right)\eta}} \tag{14}$$

where, $0 \le \tau \le 1$.

Therefore, KCCA is transformed into a constrained optimization problem related to sum, and the objective function is set as:

$$\max \xi^T K_X K_Y \eta \tag{15}$$

Constraints:

$$\begin{cases} \xi^T \left((1-\tau) K_X^2 + \tau K_X \right) \xi = 1 \\ \eta^T \left((1-\tau) K_Y^2 + \tau K_Y \right) \eta = 1 \end{cases} \tag{16}$$

The Lagrange multiplier method is used to solve the above constrained extreme value problem [17], and the corresponding Lagrange equation is:

$$L(\xi,\eta) = \xi^T K_X K_Y \eta - \frac{\lambda_1}{2}\left(\xi^T \left((1-\tau) K_X^2 + \tau K_X \right) \xi - 1 \right) - \frac{\lambda_2}{2}\left(\eta^T \left((1-\tau) K_Y^2 + \tau K_Y \right) \eta - 1 \right) \tag{17}$$

where λ_1 and λ_2 are Lagrange multipliers.

The partial derivatives of η with respect to ξ and $L(\xi,\eta)$ is solved respectively to make them zero, that is:

$$\begin{cases} \dfrac{\partial L}{\partial \xi} = K_X K_Y \eta - \lambda_1 \left((1-\tau) K_X^2 + \tau K_X \right) \xi = 0 \\ \dfrac{\partial L}{\partial \eta} = K_Y K_X \xi - \lambda_2 \left((1-\tau) K_Y^2 + \tau K_Y \right) \eta = 0 \end{cases} \tag{18}$$

Thus, KCCA is equivalent to solving the eigenvector problem corresponding to the generalized eigenequation, that is:

$$K_X K_Y \eta = \left((1-\tau) K_X^2 + \tau K_X \right) \xi \tag{19}$$

Solve ξ and η, and extract the nonlinear correlation features between x and y:

$$\begin{cases} u = \xi K_X \\ v = \eta K_Y \end{cases} \tag{20}$$

Where u and v are the transformed characteristic components.

It is linearly transformed to obtain:

$$Z = \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \xi & 0 \\ 0 & \eta \end{pmatrix} \begin{pmatrix} K_X \\ K_Y \end{pmatrix} \tag{21}$$

The projected combined features are used for the modelling and classification of speech emotion recognition in subsequent educational scenes.

2.3 Speech emotion recognition in educational scene based on machine learning

Recognition process

The implementation process of speech emotion recognition method in educational scene is as follows:

(1) The voice signal of education scene is collected and preprocessed;

(2) The fundamental frequency feature and amplitude feature are extracted respectively. Due to the different value range of each feature, in order to eliminate the impact of different range on emotion modelling, normalization must be carried out. The normalization formula is:

$$f_i' = \frac{f_i - \min(f_i)}{\max f_i - \min(f_i)} \tag{22}$$

(3) Because the fundamental frequency characteristics and amplitude characteristics describe the changing relationship between speech and emotion in educational scenes from different angles, they not only focus on each other, but also complement each other, but also have correlation, that is, information redundancy. In addition, the feature dimension is not proportional to emotion recognition, so KCCA is used to fuse the features, find the most important information in the features, and transform the original feature vector into a low dimensional vector;

(4) The emotion samples are processed by low dimension vector to reduce the data scale. The support vector machine is used to learn the training samples, establish the emotion recognition classifier, and identify the test samples to verify the effectiveness of the classifier; to sum up, the speech emotion recognition process of education scene based on machine learning (kernel canonical correlation analysis and support vector machine) is shown in Figure 1.

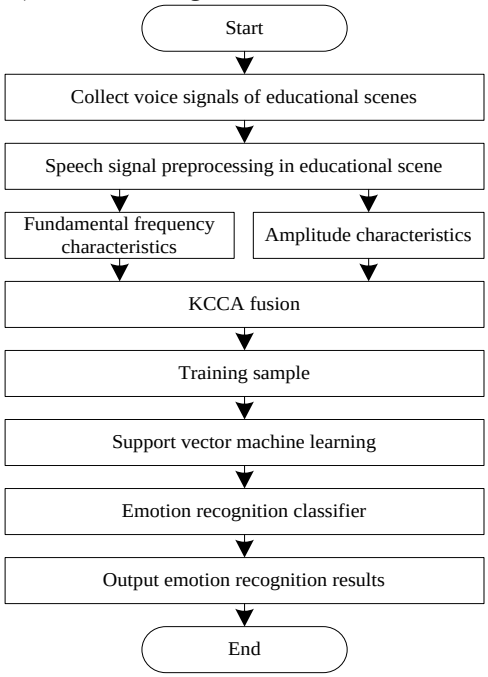

Fig 1. Implementation process of speech emotion recognition method in educational scene based on machine learning

Speech emotion recognition classifier based on SVM algorithm

Support vector machine (SVM) is a machine learning algorithm based on statistical learning theory [18]. It was first proposed by Vapnik et al. based on the principle of linear classifier. SVM can be used to solve linear and nonlinear sample classification. Its core idea is to map the linearly indivisible sample points in low-dimensional space to high-dimensional feature space through kernel function, and then construct the optimal classification hyperplane in the feature space. At this time, the data can also be segmented by hyperplane in high-dimensional space, so as to become linearly separable, and the distance between each sample and the hyperplane shall be kept to the maximum.

The basic principle of SVM classification algorithm is as follows:

The nonlinear sample sets $(x_1, y_1), \cdots, (x_n, y_n)$, $y_i \in \{-1, 1\}, i = 1, 2, \cdots, n$ are given. - 1 and + 1 in y represent two different types of samples, respectively. Firstly, the nonlinear samples in the low-dimensional space are mapped to the high-dimensional feature space through the mapping function $\varphi(x)$, and then the optimal classification hyperplane $w \cdot \varphi(x) + b = 0$ is constructed in the high-dimensional feature space, where w is the normal vector and b is the offset. Therefore, the values of w and b need to be determined.

The interval between the two types of samples is $\frac{2}{\|w\|}$. According to the principle of SVM, $\frac{2}{\|w\|}$ should be maximized. Based on this principle, the optimization problem is transformed into a convex quadratic programming problem for variables w and b, as shown in equation (23):

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}\|w\|^2 \\ subject\ to\ y_i\left(w \cdot \varphi(x) + b\right) \geq 1, i = 1, 2, \cdots, n \end{cases} \tag{23}$$

Equation (23) belongs to a nonlinear programming problem with constraints, which can be solved by introducing Lagrange function, that is, introducing a Lagrange multiplier α for each constraint, as shown in (24):

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i\left(w^T \cdot \varphi(x_i) + b\right) - 1\right) \tag{24}$$

Problem (23) is transformed into finding the minimum value of equation (24). Firstly, L is used to find the partial derivatives of w and b respectively, so that $W(\alpha) = L(w, b, \alpha)$, $W(\alpha)$ and constraints are as follows:

$$\begin{cases} W(\alpha) = \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum\limits_{j=1}^{n} \alpha_j \\ subject\ to \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C, i = 1, 2, \cdots, n \end{cases} \tag{25}$$

Thus, the minimization problem of equation (24) is transformed into the dual problem of maximizing function $W(\alpha)$, where C is a parameter to control the weight between finding the maximum hyperplane and ensuring the minimum deviation of data points. $K(x_i, x_j)$ is the kernel function that maps linearly non-separable samples into high-dimensional space, and $K(x_1, x_2) = (\varphi(x_1) \cdot \varphi(x_2))$ is satisfied for any samples x_1 and x_2. The optimal solutions α^*, w^* and b^* corresponding to α, w and b in the above formula can be solved respectively, and then the optimal classification function can be obtained, as shown in formula (26):

$$f(x) = sgn((w^* x) + b^*) = sgn\left(\sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x_j) + b^*\right) \tag{26}$$

How to select the penalty factor C, the type of kernel function $K(x_i, x_j)$ and its parameter φ has become the primary problem in training an SVM classifier. In the research of speech emotion recognition using SVM, the penalty factor C and kernel function parameter φ are generally determined by experiment or experience. And in classification training, fixed parameter values are usually selected. The size of the sample set and the number of experiments has a certain impact on the selection of parameters, and the SVM parameters of different training sets should also be different.

Genetic algorithm to determine the optimal parameters

Genetic algorithm is an efficient global optimization search algorithm that combines the survival of the fittest in the process of biological evolution with the random information exchange system of chromosomes in the population [19]. Genetic algorithm parameter optimization is to encode the parameters to be optimized to form chromosomes and randomly generate the initial population. In genetic evolution, the selection strategy based on fitness function is used to simulate the survival law "survival of the fittest" to select individuals, and crossover and mutation are used to produce the next generation population. The population is continuously

optimized until the expected termination conditions are met. The last generation of chromosomes is regarded as the global optimal solution, and the optimal parameters are obtained by decoding.

In this paper, genetic algorithm is used to optimize the parameters of different training sets to find the optimal parameters belonging to the training set, and then SVM model is trained and identified. Figure 2 is the experimental flow chart of speech emotion recognition in educational scene with genetic algorithm optimizing SVM parameters.

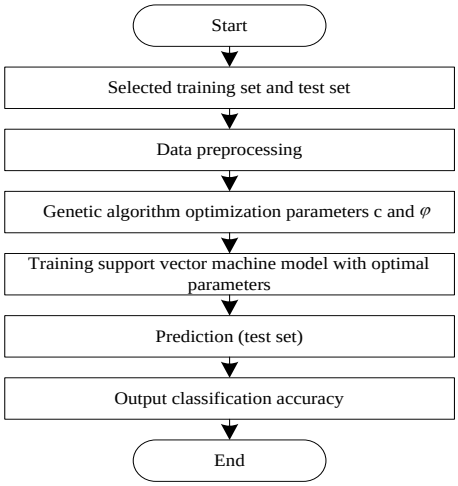

Fig 2. Optimization of support vector machine parameters by genetic algorithm

In this paper, genetic algorithm is used to optimize SVM parameters. The specific steps are as follows:

Step 1: initialize the parameters, and make binary coding on the parameters C and φ of SVM classification model. Each variable is represented by 20 binary bits, and then randomly generate the initial population.

Step 2: decode C and φ, substitute them into the SVM algorithm function, and take the trained classification recognition rate as the fitness value. The higher the fitness id, the greater the probability of inheriting to the next generation is, and the lower the fitness is, the less the probability of inheriting to the next generation is.

Step 3: selection operation, simulate the "survival of the fittest" in each generation of evolution through individual fitness [20], select excellent individuals from the group as the parent generation, and then generate a new group.

Step 4: crossover operation, select the individuals after the selection operation, and generate new individuals according to the crossover probability.

Step 5: mutation operation: in the population individual string, change the gene of a locus according to the mutation probability to generate a new individual.

Step 6: decode and calculate the fitness value, compare the classification recognition rate between the offspring and the parent, and update the optimal individual.

Step 7: judge whether the number of iterations or fitness value reaches the set termination value. If not, repeat steps 3 to 6; if the requirements are met, proceed to step 8.

Step 8: when the end condition is reached, the optimal solutions C and φ are output.

The implicit parallelism and powerful global search ability of genetic algorithm can search the global best in a very short time; the optimization process is completed automatically without manual intervention, which avoids the errors caused by manual operation and improves the efficiency of optimization.

3. Experimental analysis

3.1 Data corpus

The phonetic databases used in this experiment are Berlin affective phonetic database and Chinese affective corpus of Chinese Academy of Sciences. The Berlin affective corpus was recorded by the Technical University of Berlin. There are 10 non professional actors, 5 men and 5 women. They have anger, boredom, disgust, fear, happiness, neutrality, sadness and 10 recorded scripts. A total of 800 emotional sentences were recorded, and then 20 volunteers listened and recognized them. Among the 800 emotional sentences, some sentences are short and difficult to recognize; There are also some sentences with serious colloquialism. Therefore, the samples of emotional sentences were screened. Finally, 535 sentences were retained. The Chinese emotion corpus was recorded and provided by the human-computer speech interaction research group of the State Key Laboratory of pattern recognition, Institute of automation, Chinese Academy of Sciences. There are two male and two female professional speakers. They use six emotional states: anger, fear, happiness, sadness, surprise and neutrality. They have 50 recording scripts and finally get 1200 sound emotions. Both data sets are stored in 16000 sample rate, 16 bit quantization and wav format.

3.2 Kernel function selection in feature fusion

The selection of kernel function of KCCA is very important to the recognition results. At present, there are mainly polynomial kernel function, fractional power polynomial kernel function and Gaussian kernel function. In the case of different training samples, the average recognition accuracy of different kernel functions is shown in Figure 3.

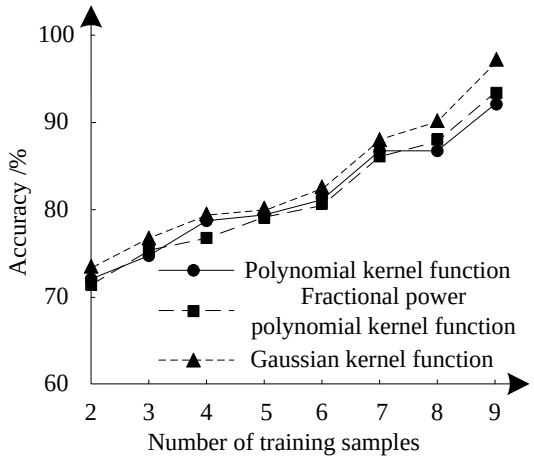

Fig 3. Comparison of average recognition accuracy of different kernel functions

It can be seen from Figure 3 that under the conditions of different training samples, Gaussian kernel function has the highest emotion recognition rate. Therefore, the method in this paper selects Gaussian kernel function as the kernel function of KCCA in the feature fusion process for feature fusion.

3.3 Parameter optimization experiment of support vector machine

Through the experiment, the SVM kernel function is selected. It is known that the polynomial kernel function (i.e. t = 1) is better for the seven emotion experiments of Berlin speech data set; The experiment of five emotions in Berlin speech set and Chinese emotion data set is better by using linear kernel function (i.e. t = 0). Empirical cross validation technology is used in speech emotion recognition experiment. For each group of experiments, the parameters C and φ are optimized by genetic algorithm, and then the SVM model is trained and identified by the optimal parameters. The experimental parameters are set as follows: the crossover rate and variation rate are 0.7 and 0.035 respectively. The parameters are binary coded. The population size is 20 and the number of iterations is 100. Figure 2 is a set of experimental results obtained by using genetic algorithm to optimize SVM parameters for seven emotions on the Berlin data set.

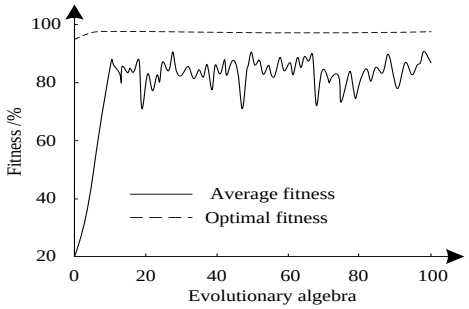

Fig 4. Optimization results of support vector machine parameters by genetic algorithm

By analysing Figure 4, the optimal parameters of this group of support vector machine models are C value of 12.3836 and φ value of 0.0896. This parameter is used for training and classification prediction of support vector machine model. The results are shown in Figure 5. At the same time, in order to further illustrate the performance advantages of the proposed method in emotion recognition, taking the methods in reference [6] and reference [7] as the comparison method, two comparison methods are used for emotion recognition on the test data set, and the emotion recognition results of the proposed method and the two comparison methods are compared. The results are shown in Figure 5.

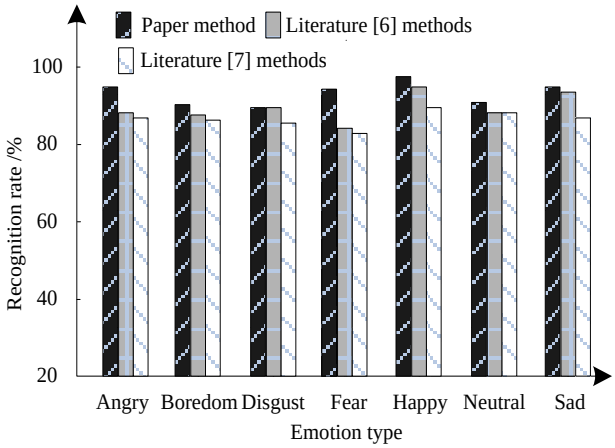

Fig 5. Comparison results of emotion recognition

By analysing Figure 5, the average recognition rates of the method in this paper, the recognition method in reference [6] and the recognition method in reference [7] are 93.6%, 87.7% and 83.7% respectively. Moreover, the recognition rate of emotion in the method in this paper's

generally higher than that in the two comparison methods, and the recognition rate is more than 90%. Among them, the recognition rate of emotion in anger, fear, happiness and sadness is more than 95%, showing that this method has a high speech emotion recognition rate.

3.4 Anti-noise performance

In order to test the anti-noise performance of the evaluation method in this paper, the speech emotion results recognized by this method under different Gaussian noise, salt and pepper noise and Gaussian filter operator standard deviation are shown in Table 1.

Table 1. Identification results under different interference States

External interference		Voice emotion	Identification results	Is it consistent
Superposition density of salt and pepper noise /%	0.003	Angry	Angry	Agreement
	0.007	Fear	Fear	Agreement
	0.028	Angry	Angry	Agreement
	0.039	Sad	Sad	Agreement
Gaussian noise variance	0.480	Boredom	Boredom	Agreement
	0.006	Neutral	Neutral	Agreement
	0.012	Fear	Fear	Agreement
	0.029	Angry	Angry	Agreement
	0.065	Happy	Happy	Agreement
Gaussian filter operator standard deviation	0.490	Happy	Happy	Agreement
	0.541	Happy	Happy	Agreement
	0.942	Boredom	Boredom	Agreement
	1.329	Disgust	Disgust	Agreement
	2.300	Neutral	Neutral	Agreement
	3.320	Angry	Angry	Agreement

By analysing table 1, with the increase of Gaussian noise variance, salt and pepper noise superposition density and Gaussian fuzzy filter operator standard deviation, the speech emotion recognition results obtained by the method in this paper are completely consistent with the actual speech emotion, which shows that this method has high noise resistance.

The above experimental results show that compared with the recognition methods in reference [6] and reference [7], the emotion recognition rate of this method is more than 90%, and the emotion recognition rate of anger, fear, happiness and sadness is more than 95%, indicating that the recognition rate of this method is high; The anti noise performance is tested under different Gaussian noise, salt pepper noise and Gaussian filter operator standard deviation. The results show that the speech emotion recognition results are completely consistent with the actual speech emotion, indicating that the proposed method has high anti noise performance.

4. Discussion

This paper studies the speech emotion recognition method of educational scene based on machine learning, and the experimental results show that this method has good emotion recognition results. This is mainly because the kernel canonical correlation analysis (KCCA) algorithm of machine learning algorithm is used for feature fusion. Kernel canonical correlation analysis algorithm introduces the idea of kernel function into correlation analysis algorithm. The idea is to map low-dimensional data to high-dimensional feature space (kernel function space), and carry out correlation analysis in kernel function space conveniently through kernel function. High precision feature fusion results are obtained by kernel canonical correlation analysis algorithm. According to the feature fusion results, the support vector machine algorithm in machine learning algorithm is used for speech emotion recognition. Support vector machine (SVM) was first proposed by Corinna Cortes and Vapnik in 1995. It is based on the VC dimension theory of statistical learning theory and the principle of structural risk minimization. According to the limited sample information, the complexity of the model (i.e., the learning accuracy of specific training samples) and learning ability (i.e., the ability to identify any sample without error), the best compromise between them is found in order to obtain the best promotion ability. Support vector machine method shows many unique advantages in solving small sample, nonlinear and high-dimensional pattern recognition, and can be extended to other machine learning problems such as function fitting. In machine learning, support vector machine (SVM, also support vector network) is a supervised learning model related to related learning algorithms. It can analyse data, identify patterns, and be used for classification and regression analysis.

5. Conclusion

This paper studies the speech emotion recognition method of educational scene based on machine learning, uses machine learning algorithm to analyze the correlation between different emotional features, constructs feature samples, uses SVM to establish speech emotion recognition classifier, and uses genetic algorithm to determine the optimal parameters. Kernel canonical

correlation analysis algorithm and support vector machine algorithm are applied to the field of speech emotion recognition in educational scenes. The experimental results show that the emotion recognition rate of the speech emotion recognition method based on machine learning is more than 90%, which has a high recognition rate; The speech emotion recognition results obtained in various noise environments are completely consistent with the actual speech emotion, which shows that this method has high anti noise performance. It can provide a better scientific basis for speech emotion recognition in educational scenes.

References

[1] Andy, C. & Kumar, S. (2020). An appraisal on speech and emotion recognition technologies based on machine learning. International Journal of Automotive Technology, 8(5), 2266-2276.

[2] Akalya, C., Karthika, D. & Soundarya, S. (2019). An eeg based emotion recognition and classification using machine learning techniques. International Journal of Emerging Trends & Technology in Computer Science, 5(4), 744.

[3] Dissanayake, T, Rajapaksha, Y, Ragel, R. & Nawinne, I. (2019). An ensemble learning approach for electrocardiogram sensor based human emotion recognition. Sensors, 19(20), 4495-.

[4] Alex S B, Mary L, Babu B P (2020). Attention and Feature Selection for Automatic Speech Emotion Recognition Using Utterance and Syllable-Level Prosodic Features. Circuits Systems and Signal Processing, 39(3):1-29.

[5] Liu S, Liu D, Muhammad K, et al (2021). Effective Template Update Mechanism in Visual Tracking with Background Clutter. Neurocomputing, 458, 615-625.

[6] Zvarevashe, K. & Olugbara, O. O. (2020). Recognition of speech emotion using custom 2d-convolution neural network deep learning algorithm. Intelligent Data Analysis, 24(5), 1065-1086.

[7] Hajarolasvadi, N. & Demirel, H. (2019). 3d cnn-based speech emotion recognition using k-means clustering and spectrograms. Entropy, 21(5), 479.

[8] Lee, J. H., Yoon, U. N. & Jo, G. S. (2020). Cnn-based speech emotion recognition model applying transfer learning and attention mechanism. Journal of KIISE, 47(7), 665-673.

[9] Dnvsls, I, Lakshmi, B., Prasanna, H., Pavani, C. & Vandana, G. (2020). Assessment of patient health condition based on speech emotion recognition (ser) using deep learning algorithms. European Journal of Translational and Clinical Medicine, 7(4), 2020.

[10] Liu X, He J, Song L, et al (2021). Medical Image Classification based on Adaptive Size Deep Learning Model. ACM Transactions on Multimedia Computing Communications and Applications, 17(3S): 102.

[11] Kacur J, Puterka B, Pavlovicova J (2021), et al. On the Speech Properties and Feature Extraction Methods in Speech Emotion Recognition. Sensors, 21(5):1888.

[12] Aishwarya, R. (2020). Feature extraction for emotion recognition in speech with machine learning algorithm. International Journal of Advanced Trends in Computer Science and Engineering, 9(4), 4998-5002.

[13] Manisha, S., Nafisa, H. S., Gopal, N. & Anand, R. P. (2021). Bimodal emotion recognition using machine learning. International Journal of Engineering and Advanced Technology, 10(4), 189-194.

[14] Anandharasu, A., Bharathi, R. S., Sakthivel, A. & Gopikrishnan, R. (2019). Human emotion recognition system using machine learning for videos. International Journal on Recent and Innovation Trends in Computing and Communication, 7(3), 29-36.

[15] Shuai L, Chunli G, Fadi A, et al (2020). Reliability of Response Region: A Novel Mechanism in Visual Tracking by Edge Computing for IIoT Environments, Mechanical Systems and Signal Processing, 138, 106537

[16] Gao P, Li J, Liu S (2021). An Introduction to Key Technology in Artificial Intelligence and big Data Driven e-Learning and e-Education. Mobile Networks & Applications, 26(5): 2123-2126

[17] Alhalaseh, R. & Alasasfeh, S. (2020). Machine-learning-based emotion recognition system using eeg signals. Computers, 9(4), 95.

[18] E Lieskovská, Jakubec, M, Jarina, R. & M Chmulík. (2021). A review on speech emotion recognition using deep learning and attention mechanism. Electronics, 10(10), 1163.

[19] Deepika C (2020). Speech Emotion recognition feature Extraction and Classification. International Journal of Advanced Trends in Computer Science and Engineering, 9(2):1257-1261.

[20] Zhang, H. Huang, He. Li, W. Huang, Z. (2021). AA-LSTM Network-Based Speech Emotion Recognition. Computer Simulation, 38(3), 183.