ICICS-PCM 2003
15-18 December 2003
Singapore

Applying Neural Network on the Content-Based Audio Classification

Xi Shao[#], Changsheng Xu[#], Mohan S Kankanhalli**

[#]Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
*School of Computing
National University of Singapore

Abstract

Many audio and multimedia applications would benefit if they could interpret the content of audio rather than relying on descriptions or keywords. These applications include multimedia databases and file systems, digital libraries, automatic segmentation or indexing of video (e.g., news or sports storage), and surveillance. This paper describes a novel content-based audio classification approach based on neural network and genetic algorithm. Experiments show this approach achieves a good performance of the classification.

1. Introduction

The rapid increase in speed and capability of computers has allowed the inclusion of the audio as a data type in many modern computer applications. However, audio is usually treated as an opaque collection of bytes with only the most primitive fields attached such as file name, file format, sampling rates, etc. Users accustomed to searching, scanning, and retrieving text data may be frustrated by the inability to look inside the audio objects.

Multimedia databases usually store thousands of audio recordings. These files can be music, speech and other sounds. However, it is difficult to efficiently retrieve some kinds of audio from the audio database. Moreover, compared with video, audio files cannot be browsed directly.

A number of methods have been proposed to classify music, speech, and other sounds. Saunders [1] used the average zero-crossing rate and the short time energy as features and applied a simple thresholding method to discriminate speech and music from the radio broadcast. Scheirer [2] used thirteen features in time, frequency and cepstrum domains and different classification methods to achieve a robust performance. El-Maleh [3] proposed a method to classify audio signal into speech, music and others for the purpose of parsing of news story. Kimber [4] proposed an acoustic segmentation approach that mainly applied to the segmentation of discussion recordings in meetings. Zhang [5] proposed an approach to divide the generic audio classification task into two stages. In the first stage, audio signals were segmented and classified into speech, music, song, speech with music background, environmental sound with music background, six types of environmental sound, and silence. In the second stage, further classification was conducted within each basic type. Speech was differentiated into the voice of man, woman and child. Music is classified into classics, blues, jazz, rock and roll, music with singing and the plain song, according to the instruments or types. Environmental sounds were classified into semantic classes such as applause, bell ring, footstep, wind-storm, laughter, bird's cry, and so on. Lu [6] proposed a robust two-stage audio segmentation method to segment an audio stream into speech, music, environment sound and silence.

In this paper, a novel automatic audio classification approach is presented to extend current work by using multiple audio features and efficient training algorithm of the classifier. In order to discriminate different audio classes, a set of audio features is developed to characterize audio content of different classes and a neural network approach is applied to build classifiers to discriminate audio classes. genetic algorithm and Back Propagation (BP) algorithm are used together to train the neural network instead of using BP algorithm or genetic algorithm only. This is more efficient because advantages of two algorithms are combined together.

2. Feature Selection

Feature selection is important for audio classification. The selected features should reflect the significant characteristics of different kinds of audio signals. In order to better discriminate different classes of audio, we consider the features which are related to temporal and spectral domains. The selected features include loudness, pitch, brightness, bandwidth, percentage of the low energy frames, and the statistical properties of the audio features such as derivative of loudness, pitch, brightness and bandwidth.

In our approach, we extract the perceptual features of audio content to build the feature vector.

2.1 Loudness

Loudness [9] is a commonly used perceptual feature which is approximated by the signal's root-mean-square (RMS) level in decibels. We calculate the loudness by taking a series of windowed frames of the sound and computing the square root of the sum of squares of the windowed sample values.

2.2 Pitch

Pitch is the fundamental period of a human speech waveform, and is an important parameter in the analysis and synthesis of speech signals. In an audio signal, which generally consists of pure speech as well as many other sounds, the physical meaning of pitch is lost. But we can still use pitch to characterize changes in the periodicity of waveforms in different audio signals.
The pitch is estimated by taking a series of short-time Fourier spectrum. We can calculate the pitch using the harmonic product spectrum which can be defined as:

$$P_n(e^{j\omega}) = \prod_{r=1}^{K} \left| X_n(e^{j\omega r}) \right|^2 \quad (1)$$

where $X_n(e^{j\omega r})$ is a the spectrum of a windowed frame. We store log harmonic product spectrum as a log frequency:

$$\hat{P}_n(e^{j\omega}) = 2\sum_{r=1}^{K} \log\left| X_n(e^{j\omega r}) \right| \quad (2)$$

2.3 Brightness

Brightness is computed as the centroid of the short-time Fourier transform and is stored as a log frequency. It is a measure of the higher-frequency content of the signal. For example, putting your hand over your mouth as you speak reduces the brightness of the speech sound as well as the loudness. This feature varies over the same range as the pitch, although it can not be less than the pitch estimate at any given instant.

$$\omega_C = \frac{\int_0^{\omega_0} \omega \,|F(\omega)|^2 \, d\omega}{\int_0^{\omega_0} |F(\omega)|^2 \, d\omega} \quad (3)$$

Where ω_0 is the half sampling frequency.

2.4 Bandwidth

Bandwidth is computed as the magnitude-weighted average of the differences between the spectral components and the centroid. As examples, a single sine wave has a bandwidth of zero and an ideal white noise has an infinite bandwidth.

$$B = \sqrt{\frac{\int_0^{\omega_0} (\omega - \omega_C) \,|F(\omega)|^2 \, d\omega}{\int_0^{\omega_0} |F(\omega)|^2 \, d\omega}} \quad (4)$$

Where ω_0 is the half sampling frequency and ω_C is the Brightness we got previously.

2.5 Derivative

The statistical properties of each feature are also very important for classifying the audio. We calculate the statistical properties by calculating derivatives of the four feature sequences mentioned above. The derivatives of these serials are defined as follows:

$$\mathrm{Der}(X_n(e^{jw})) = |X_n(e^{jw}) - X_{n-1}(e^{jw})| \quad (5)$$

Now we get the other four feature sequences, which are the derivatives of previous features mentioned. For these four serials, we compute the average value, the variance of the value over the trajectory in frequency domain.

2.6 Percentage of the Low Energy Frames

When we speak, there are some pauses between every tone of our speech. The energy of the frame containing pauses is lower than the other frames containing no pauses. Generally speaking, the percentage of low energy frames containing in music is lower than that containing in speech.
Now, for each audio signal, we have 17 parameters to construct the feature vector: average of the loudness, variance of the loudness, average of the pitch, variance of the pitch, average of the brightness, variance of the brightness, average of the bandwidth, variance of the bandwidth, average derivatives of the loudness, variance of the derivatives of the loudness, average derivatives of the pitch, variance of the derivatives of the pitch, average derivatives of the brightness, variance of the derivatives of the brightness, average derivatives of the brightness, variance of the derivatives of the brightness, average derivatives of the bandwidth, variance of the derivatives of the bandwidth, the ratio of low energy frames.

3. Classification

The challenge in achieving the audio classification is the proper discrimination of audio vectors in the feature vector space. We use neural network to classify different audio classes.

3.1 The Structure of the Neural Network

The neural network diagram for audio classification is illustrated in Figure 1.

The model of each neuron in the network includes a nonlinear activation function. For each neuron (perceptron), we use the Sigmoid function which is defined as follows:

$$f(s)=\frac{1}{1+\exp(-s)} \tag{6}$$

where s is the induced local field of neuron and $f(s)$ is the output of the neuron.

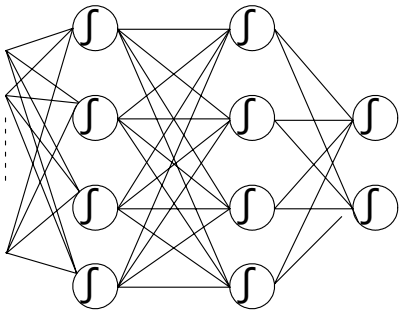

Figure 1 Audio classification diagram

The input of the network is the feature vectors we described in section 2. The dimensions of the feature vectors is 17, corresponding to the 17 parameters of features we extract from the audio content.

The network contains two layers of hidden neurons that are not part of the input or output of the network. These hidden neurons enable the network to learn complex tasks by extracting progressively more meaningful features from the input vectors. The number of neurons in each hidden layer can be specified in the experiments.

The number of the neurons in output layer is determined by the number of audio classes we want to classify. For example, if we want to classify the audio into music and speech, we have two neurons in output layer corresponding to the music and speech respectively. That means if the output value of the neurons representing music is bigger, we classify the input audio to music, and vice versa.

3.2 Training Algorithm

The basic approach in learning is to start with an untrained network, present training set to the input layer, pass the signals through the network and determine the output at the output layer. The outputs are compared with the target values and the difference corresponds to an error. The error or criterion function is some scalar function of the weights and is minimized when the network outputs match the desired outputs. Here, the training data set is some audio data which belong to a certain audio class. For each audio content, we calculate the feature vector and associate the feature vector with the desired output. Then, we can input the unlabeled audio data and classify these data using trained network.

The aim of training algorithm is to set the weights based on training vectors and the desired output, and minimize the error which is defined as the difference between the output of the network in a certain training data set and the desired output. The most commonly used training algorithm is BP algorithm [7]. The major problem of using BP algorithm is that it is often easy to find local minima and difficult to search the globe minima. That means by using BP algorithm, the neural network is easy to fall into local minima and cannot find the global minima.

genetic algorithm [8], however, has the advantage of finding the subspace of the potential global minima in the searching space

In our training scheme, we first use genetic algorithm to create an initial population of weight vectors **W(0)**, and we select an initial weight vector for each individual, then each individual is allowed to learn with BP for some number of trials, and the error rate at which it is performing at this time is considered to be the fitness of that individual. genetic algorithm uses these to produce a new population of weight vectors **W(0)**, and the cycle repeats itself until some stop criteria is met. Better result can be obtained by combining these two algorithms together.

The important parts of genetic algorithm are chromosomes and fitness function. A chromosome is defined as a vector that includes all the weights of the neural network. The Mean Square Error (MSE) function is selected as fitness function. The MSE function is defined as follows:

$$\varepsilon=\sum_{j=1}^{n}(\mathrm{d_j}-\mathrm{y_j})^2 \tag{7}$$

where d_j is the desired output of the neural network related to the jth sample in the training set; y_j is the real output of the neural network related to it; and n is the number of samples in the training set.

The detailed back propagation plus genetic algorithm procedure can be described as follows:

(1) Initialization: Randomly generate the *pop_size* number of chromosomes to construct the initial population.

(2) Selection: Selecting individuals for mating. In this step, we use the evaluation function to set the selection probability of the chromosome V_i. The higher the selection probability, the more likely the chromosome to be selected for mating.

The evaluation function we use here is defined as follows:

$eval(V_i)=a(1-a)^{i-1}$ $i=1,2,\cdots,pop\_size$ (8)

We use the following methods to select next generation.

(a)Assume the chromosomes in current population are V_1 , V_2 , … , V_{pop_size} ,we order these chromosomes by evaluation function. For each chromosome，we calculate the total probability q_i:

$$\begin{cases} q_0 = 0 \\ q_i = \sum_{j=1}^{i} eval(V_j) \quad i = 1,2,3,\cdots pop\_size \end{cases} \tag{9}$$

(b) Get a random number r in $[0,q_{pop\_size}]$, if $q_{i-1} < r \le q_i$, the we choose the ith chromosome $V_i(1 \le i \le pop\_size)$.

Repeat (a) and (b) *pop_size* times, then we get a new population.

(3) Mutation: There are P_m*pop_size number of chromosomes engaged in mutation in the current population, P_m is defined as mutation probability, and for each chromosome in the current population, we get a random number r in [0,1], if $r<P_m$, this chromosome will be the parent V, then we randomly choose a mutation direction d in the multidimensional space of audio features ,and replace the V with $V+M \cdot d$. M is a predefined constant.

(4) Get all the generated chromosomes as an initial population of weight vectors **W(0)**. Then, BP is used to optimize each of these, and then the MSE of each result was used for the individual's fitness. GA used these to produce a new population of weight vectors **W(0)** and go to Step (2) to begin a new cycle until the minimum MSE of the new population is less than a constant C.

4. Experimental Results

To illustrate and evaluate the proposed audio classification approach, experiments are conducted for test samples.

4.1 Dataset Collection

The audio dataset used in audio classification experiment contains hundreds of audio samples. They are collected from Internet and cover different classes such as music, speech and natural sound. All data have 22050 Hz sampling rate, stereo channels and 16 bits per sample. The audio database is shown in table1. All files have two labels. A coarse label is corresponding to the three major classes: Speech, music and sound. A fine label is corresponding to more specific classes. Each audio file is divided into frames of 256 samples with 50% overlap of the two adjacent frames.

In order to make training results statistically significant, training data should be sufficient and cover various classes of audio.

Table 1. Structure of Audio database

Class name	No. of Files	Class name	No. of Files
1.Speech	200	Violin-Pizzicato	40
Female	100	3.Sound	62
Male	100	Animal	9
2.Music	300	Bell	7
Trombone	14	Crowds	4
Cello	47	Laughter	7
Oboe	32	Machines	11
Percussion	102	Telephone	17
Tubular-bell	20	Water	7
Violin-bowed	45	Total	562

4.2 Classification Results

Two experiments have been conducted. The audio database is split into two equal parts: one for training and the other for testing.

Experiment 1: Classifying audio database into three major classes. The numbers of perceptrons in first and second hidden layer of the neural network we construct are all four.

Experiment II: Classifying audio database into 16 classes. The numbers of perceptrons in first and second hidden layer of the neural network we construct are all five.

Table 2 shows the results of two experiments using BP algorithm and BP+GA separately.

Table 2 Result of the experiment

Test	Method	Correct ratio
Experiment I	BP	99%
	BP+GA	99.5%
Experiment II	BP	89%
	BP+GA	92%

From table 2, when we classify the audio into three major classes, the classification result of BP+GA is as good as the result of BP only, and the correct ratio of two methods are all above 99%.

When we classify audio database into 16 classes, the classification result of BP+GA is more reliable than the result of BP only. This can be explained by comparing the performance of the BP+GA combination to the performance of BP and GA used independently. As figure 2 and figure 3 shows.

In figure 2, the generation average mean squared error (MSE) of the BP+GA hybrid is plotted on a logarithmic scale, as a function of generation and labeled **GA+BP(Avg)**. The minimum of these MSE in each generation is drawn as **GA+BP (Min)**. The generation average of population in which GA was used to select initial weights and the MSE of these individuals is taken immediately (i.e., without any BP learning) is labeled **GA(Avg)**. For comparison with BP used in isolation, we plot the performance of BP algorithm alone in figure 3.

From figure 2 and figure 3, we find that after about 50 generations, use of the GA+BP hybrid is able to find strictly better individuals than could be found by 5000 independent BP runs, and ultimately finds a much better one.

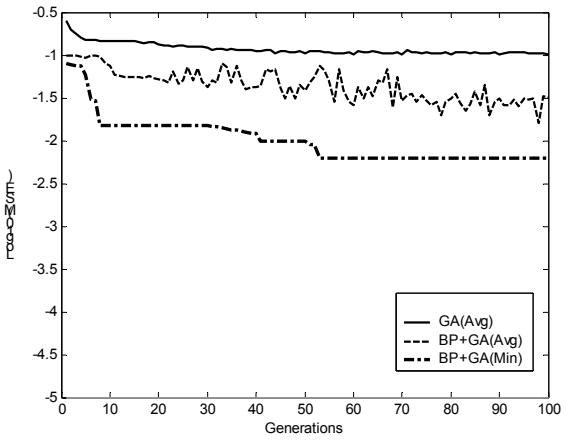

Figure 2 Performance of GA+BP algorithm

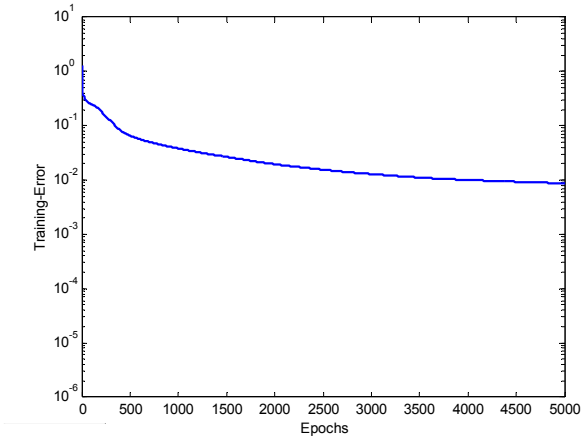

Figure 3 Performance of BP algorithm used in isolation

5. Conclusions and Future Work

The classifier we have built has provided excellent and robust discrimination among speech, music and other sounds. We first extracted the features from the audio content and built the feature vectors, then we applied the neural network to classify the audio, and we used the genetic algorithm and BP algorithm together to train the network instead of using BP lgorithm or genetic algorithm only. This is more efficient because the advantages of two algorithms are combined together.

There are many interesting directions that can be explored in the future. The first direction is to make the classification result more accurate. To achieve this goal, we need to explore more audio features that can be used to characterize the audio content. The second direction is to improve the computational efficiency for neural network.

References

[1] J. Saunders, "Real-time Discrimination of Broadcast Speech/Music", In *Proc. ICASSP-96*, pp.993-996, 1996.

[2] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Music/Speech Discriminator", In *Proc. ICASSP97*, Vol.2, pp.1331-1334, 1997.

[3] K. El-Maleh, M. Klein, G. Petrucci and P. Kabal, "Speech/Music Discrimination for Multimedia Application", In *Proc. ICASSP00*, 2000.

[4] D. Kimber and L. Wilcox, "Acoustic Segmentation for Audio Browsers", In *Proc. Interface Conference,* Sydney, Australia, 1996.

[5] T. Zhang and C.-C. Kuo, "Video Content Parsing Based on Combined Audio and Visual Information", In *Proc. SPIE 1999,* San Jose, USA, Vol.4, pp.78-89, 1999.

[6] L. Lu, H. Jiang and H. J. Zhang, "A Robust Audio Classification and Segmentation Method", In *Proc. ACM Multimedia 2001,* Ottawa, Canada, 2001.

[7] S. Haykin.-2nd ed, "Neural Networks: a Comprehensive Foundation", Prentice Hall,1999.

[8] D. Dasgupta, Z.Michalewicz, "Evolutionary algorithms in engineering applications", Springer-Verlag (March 1997).

[9] E.Wold, T.Blum, D.Keislar and J.Wheaton(1996),Content-based classification, search and retrieval of audio, *IEEE multimedia Mag.3*,pp.27-36