ISSN:0975-9646

Computational Auditory Scene Analysis for Acoustic Event Detection: An Improved Approach

Jincy Joseph.P, Shamla Beevi.A

Dept. of Computer Science
Mar Baselios College of Engineering and Technology, Trivandrum, India

***Abstract*— This paper presents an improved system for acoustic event classification and detection using neural networks. The adaptation of network weights using Particle Swarm Optimization (PSO) was proposed as a mechanism to improve the performance of the traditional Back Propagation algorithm used in Artificial Neural Network (ANN) for classification. This paper focuses on acoustic event classification using Neural Network. The problem concerns the identification of different environmental sounds on the basis of features extracted. Supervised method of classification is used. By using this event classification system, the unknown data can be predicted more precisely. Artificial neural networks have been successfully applied to problems of event classification. In this work, Multilayer feed- forward network is used which is trained using back propagation learning algorithm and the output is optimized by using particle swarm optimization algorithm. Experiments show that this approach achieves a good performance in classification and detection of acoustic events.**

***Key Words*— Computational Auditory Scene Analysis, Acoustic Event Detection, Artificial Neural Network, Particle Swarm Optimization**

I. INTRODUCTION

Sound is second to vision with the help of which humans sense and understand the world around him. The severity of deafness as a disability reflects the importance of sound in understanding what's happening in the physical world, including the awareness of events that occur out-of-sight and warning of danger. We thus believe that acoustic sensing and analytics of audio data could prove as significant to machine-automated monitoring of human environments as it is to the humans. Nowadays, Identifying events using acoustic data is a growing research field and getting a wide attention by the researchers due to its tremendous scope in the future.

A number of methods have been proposed to classify music, speech, and other sounds using different feature sets and algorithms[1]. In this paper, a novel automatic audio classification approach is presented. In order to discriminate different audio classes, a set of audio features is extracted to characterize audio content of different classes and a neural network approach is applied to build classifiers. To discriminate audio classes Back Propagation (BP) algorithm is used which is optimized by particle swarm optimization algorithm. During the initial stages of a global search, the particle swarm optimization algorithm was showed to converge rapidly but around global optimum, the search process will become very slow. On the contrary, the gradient descending method can achieve faster convergent speed around global optimum and at the same time, the accuracy of the convergence can be higher. So in this paper, a hybrid algorithm combining particle swarm optimization (PSO) algorithm with back-propagation (BP) algorithm is used.

II. SYSTEM DESCRIPTION

The system can be divided into four modules. The System Block Diagram is shown in Fig 1. Preprocessing is a stage which is performed prior to the feature extraction phase. The process of converting the raw pcm signal into a conditioned signal is the preprocessing stage. It is performed by converting the audio signals to a 32bit format and normalizing it, which is then fed into the feature extraction phase. Feature vectors obtained after feature extraction phase is fed as input to the artificial neural network, the output of which is the classification and detection of different acoustic events [2].

Feature extraction plays an important role in analyzing and characterizing audio content. All audio features are extracted by breaking the input signal into a succession of analysis windows or frames, each of around 10-40-ms length, and computing the feature values from each of the frames. One approach is to take the values of all features for a given analysis window to form the feature vector for the classification decision, so that class assignments can be obtained almost in real time, thus realizing a real-time classifier. In order to better discriminate different classes of audio, we consider the features which include pitch, loudness, brightness, bandwidth and band energy ratios. The statistical properties of the audio features such as average and variance are used to build the feature vector.

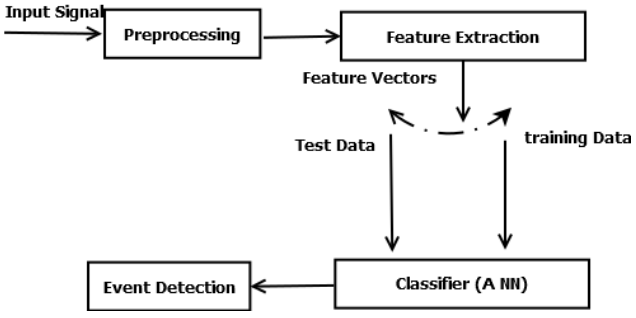

Fig 1: System Block Diagram

A. Pitch

According to the American National Standards Institute, pitch is the auditory attribute of sound according to which sounds can be ordered on a scale from low to high. Since pitch is such a close proxy for frequency, it is almost entirely determined by how quickly the sound wave is making the air vibrate and has almost nothing to do with the intensity, or amplitude, of the wave. That is, "high" pitch means very rapid oscillation, and "low" pitch corresponds to slower oscillation.

Although a large number of different methods have been proposed for detecting pitch, the autocorrelation method of pitch detection is still one of the most robust and reliable methods of pitch detection.

B. Loudness

Physically, a wave involves the propagation of energy. This transfer of energy by a travelling wave is expressed in terms of the intensity. Intensity of sound waves is defined as the average energy transported per second per unit area perpendicular to the direction of propagation. The intensity of sound in air depends on the square of the frequency and the square of the amplitude. Thus, for a given frequency, the amplitude is an important factor in deciding the intensity. We can calculate the loudness by taking a series of windowed frames of the sound and computing the square root of average of the sum of squares of the windowed sample values.

C. Brightness

The term "brightness" is also used in discussions of sound timbres, in a rough analogy with visual brightness. Timbre researchers consider brightness to be one of the perceptually strongest distinctions between sounds, and formalize it acoustically as an indication of the amount of high-frequency content in a sound, using a measure such as the spectral centroid. It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights.

D. Bandwidth

Bandwidth is a measure of the width of a range of frequencies, measured in hertz. Bandwidth is computed as the magnitude-weighted average of the differences between the spectral components and the centroid.

E. ZCR and Short Time Energy

Zero-crossings occur when successive samples have different signs, and the ZCR rate is the average number of times the signal changes its sign within the short-time window. We calculate both energy and ZCR values using a window of 1024 samples with 50% overlap, at an input sampling rate of 11025 Hz

Thus for each audio data, we have 10 parameters to construct the feature vector by taking the moving average and variance of the features through the analysis frames.

III. Classification And Detection

In order to classify the events properly we need to discriminate the audio vectors in the feature vector space. We use neural network to classify different audio classes. The number of neurons in the input layer is 10 corresponding to 10 features in the feature vector. The number of the neurons in output layer corresponds to the number of audio classes we want to classify.

The neural network we used here include two hidden layers of seven neurons each. Each neuron in the network includes a nonlinear Sigmoid activation function. which is defined as follows:

$$f(s) = \frac{1}{1 + e^{(-s)}} \quad (1)$$

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

A. Training Algorithm

The hybrid PSO–BP is an optimization algorithm [9][10] combines the PSO algorithm with the BP. The reason for combing the PSO algorithm with BP lies in the fact that the BP algorithm has a strong ability to find local optimistic result, but its ability to find the global optimistic result is weak. Similar to the GA, the PSO algorithm is a global algorithm, which has a strong ability to find global optimistic result [11]. By combining the PSO with the BP, a new algorithm referred to as PSO–BP hybrid algorithm is used in this paper. The fundamental idea for this hybrid algorithm is that the PSO is employed to accelerate the training speed at the beginning stage of searching for the optimum. When the fitness function value has not changed for some generations, or value changed is smaller than a predefined number, the searching process is switched to gradient descending searching according to this heuristic knowledge.

In PSO, each particle keeps track of its coordinates in the solution space which are associated with the best solution (fitness) that has achieved so far by that particle. This value is called personal best, pbest. Another best value that is tracked by the PSO is the best value obtained so far by any particle in the neighborhood of that particle. This value is called gbest.

After finding the two best values, the particle updates its velocity and positions with following equation(2)and(3).

v[] = v[] + c1 * rand() * (pbest[] - present[]) + c2 * rand() * (gbest[] - present[]) (2)

present[] = persent[] + v[] (3)

where, v[] is the particle velocity, persent[] is the current particle (solution). pbest[] and gbest[] are defined as stated before. rand () is a random number between (0,1). c1, c2 are learning factors.Usually c1=c2=2.

The procedure for this PSO–BP algorithm can be summarized as follows:

Step 1: Initialize the positions and velocities of a group of particles randomly in the range of [0, 1].

Step 2: Evaluate each initialized particle's fitness value, and Pc is set as the positions of the current particles, while Pj is set as the best position of the initialized particles.

Step 3: When the maximum iterative generations are arrived, go to Step 8, else, go to Step 4.

Step 4: The best particle of the current particles is stored. The positions and velocities of all the particles are updated according to Eqs. (2) and (3), then a group of new particles are generated, If a new particle flies beyond the boundary [Xmin,Xmax], the new position will be set as Xmin or Xmax; if a new velocity is beyond the boundary [Vmin,Vmax], the new velocity will be set as Vmin or Vmax.

Step 5: Evaluate each new particle's fitness value, and the worst particle is replaced by the stored best particle. If the ith particle's new position is better than Pic, Pic is set as the new position of the ith particle. If the best position of all new particles is better than Pj, then Pj is updated.

Step 6: Reduce the inertia weights w according to the selection strategy.

Step 7: If the current Pj is unchanged for ten generations, then go to Step 8; else, go to Step 3.

Step 8: Use the BP algorithm to search around Pj for some epochs, if the search result is better than Pj, output the current search result; or else, output Pj.

IV. Experiment And Analysis

To illustrate and evaluate the proposed audio event detection approach, experiments are conducted. The audio dataset used in audio classification experiment contains hundreds of audio samples. They are collected from Internet and cover different classes natural sounds as we are focusing on the classification and detection of natural sounds. All data have 11025 Hz sampling rate, mono channels and 16 bits per sample. The audio database is shown in table1. Each audio file is divided into frames of 1024 samples with 50% overlap of the two adjacent frames.

TABLE I: DATASET USED FOR EXPERIMENT

Class Name	Number Of Files	Class Name	Number Of Files
Animals	40	Firecrackers	30
Bell	25	Households	30
City	35	Machines	35
Classroom	35	Nature	35
Crowd	35	Water	30
Disasters	30	Weapons	35
Electronic Equipments	35	Weather	30
Emergency	25	Vehicles	35

Two experiments have been conducted. In both the experiments, the numbers of perceptrons in first and second hidden layer of the neural network we construct are all seven.

Experiment 1: Event classification and detection are performed using Back Propagation algorithm only.

Experiment II: Event classification and detection are performed using hybrid PSO-BP algorithm.

The performance of the classification used can be evaluated using the confusion matrix. A confusion matrix contains information about actual and predicted classifications done by a classification system. It shows the error in classification of a particular class if that class had been wrongly classified as another one. This in turn helps in understanding and analyzing the performance of any classifier. In order to reduce the size of the confusion matrix, the whole data set is divided into 4 classes. From each class 32 audio files are used for testing, 8 files each from 4 sub classes.

Class A: City, Crowd, Emergency and Vehicles.

Class B: Animals, Nature, Water and Weather.

Class C: Households, Machines, Electronic Equipments and Weapons.

Class D: Bell, Classroom, Firecrackers and Disasters.

TABLE II: CONFUSION MATRIX FOR EXPERIMENT 1

		Predicted Class			
		Class A	Class B	Class C	Class D
Actual Class	Class A	26	1	3	2
	Class B	2	25	2	3
	Class C	4	-	26	2
	Class D	3	2	2	25

TABLE III: CONFUSION MATRIX FOR EXPERIMENT 2

		Predicted Class			
		Class A	Class B	Class C	Class D
Actual Class	Class A	29	1	3	-
	Class B	1	30	1	2
	Class C	2	-	30	-
	Class D	2	1	-	28

TABLE IV: SHOWS THE RESULTS OF TWO EXPERIMENTS USING BP ALGORITHM AND BP+PSO SEPARATELY

Test	Method	Correct ratio
Experiment I	BP	80%
Experiment II	BP+PSO	91%

V. Conclusion

The classifier we have built has provided a robust discrimination among different audio events. At first, the features are extracted from the audio samples and built the feature vectors, then we applied the neural network to classify the audio, and we used the hybrid PSO-BP algorithm to train the network. The experimental result clearly shows that the improved method of training the neural network with the hybrid algorithm enhances the accuracy of the acoustic events detected by the system. In many directions the work can be extended in the future to obtain more accurate results. To achieve this goal, we need to explore more audio features that can be used to uniquely identify the audio content.

REFERENCES

[1]. Annamaria Mesaros, Toni Heittola, Antti Eronen, Tuomas Virtanen, "Acoustic Event Detection In Real Life Recordings", 1Department of Signal Processing Tampere University of Technology Korkeakoulunkatu Finland 2005.

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

[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] Vikramjit Mitra and Chia J. Wang,"A Neural Network based Audio Content Classification", Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007.

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

[8] 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

[9] R.C. Eberhart, J. Kennedy, A new optimizer using particles swarm theory, in: Proc. of Sixth Int. Symp. on Micro Machine and Human Science, Nagoya, Japan (1995) 39–43.

[10] Jing-Ru Zhang, Jun Zhang, Tat-Ming Lok, Michael R. Lyu, 2006," A hybrid particle swarm optimization–backpropagation algorithm for feedforward neural network training"

[11] Marco Gori, Alberto Tesi, On the problem of local minima in back-propagation, IEEE Trans. Pattern Anal. Mach. Intell. 14 (1) (1992) 76–86

[12] X. Yao, A review of evolutionary artificial neural networks, Int. J. Intell. Syst., 8(4) (1993) 539–567.

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

[14] Albert S. Bregman. "Auditory Scene Analysis". MIT Press, Cambridge, 1990