

Chapter 47

Audio Classification Using GA-Based Fuzzy C-Means

Myeongsu Kang and Jong-Myon Kim

Abstract The purpose of automatic audio classification is to meet the rising need for efficient multimedia content management. This paper proposes a robust audio classification approach that classifies audio streams into one of five categories (speech, music, speech with music, speech with noise, and silence). The proposed method is composed of two steps: efficient audio feature extraction and audio classification using genetic algorithm-based fuzzy c-means. Experimental result indicates that the proposed classification approach achieves higher than 96.16 % in terms of classification accuracy.

Keywords Audio classification · Fuzzy c-means · Genetic algorithm · Multimedia content management

47.1 Introduction

Typical multimedia databases often contain large numbers of audio signals that require automatic audio retrieval for efficient production and management [1], and the efficacy of audio content analysis depends on extracting appropriate audio features and using an effective classifier to precisely classify the audio stream.

In literatures, many researchers have tried to utilize classifiers with effective audio features such as mel-frequency cepstral coefficients, and zero-crossing rate for audio classification. In addition, various classifiers such as the Gaussian mixture model and support vector machine have been utilized for audio

M. Kang · J.-M. Kim (✉)

School of Electrical Engineering, University of Ulsan, Ulsan, South Korea
e-mail: jongmyon.kim@gmail.com

M. Kang

e-mail: ilmareboy@gmail.com

classification. Among many classifiers, a number of methods based on fuzzy c-means (FCM) have been proposed. Park et al. proposed different fuzzy methods in order to classify audio signals into different musical genres [2, 3]. In spite of the fact that FCM is an efficient for audio classification, the FCM-based classifiers exhibit performance degradation since FCM requires initialization [4]. To address this problem, this paper integrates FCM clustering with a genetic algorithm (GA) to globally optimize the objective function of FCM and offer better classification performance.

The rest of this paper is organized as follows. Section 47.2 presents audio features extraction based on principal component analysis and Sect. 47.3 introduces the proposed audio classification scheme. Section 47.4 analyzes experimental results and Sect. 47.5 concludes this paper.

47.2 Audio Features Extraction

The purpose of audio features extraction is to obtain as much information as possible about the audio streams. After rigorous studies of broad categories of audio features used in the previous studies, this paper extracts the following features to obtain accurate audio classification. Feature extraction is a frame-based process, and thus audio features are calculated in the processing window whose length is set to 0.02 s.

- Low Root-Mean-Square Ratio

Low root-mean-square ratio (L_R) is defined as the ratio of number of frames with root-mean-square (RMS) values below the 0.5-fold average RMS in the processing windows, as expressed in (47.1):

$$L_R = \frac{1}{2N} \sum_{m=0}^{N-1} [\text{sgn}(0.5 \times \overline{RMS} - RMS(m)) + 1], \quad (47.1)$$

where N is the total number of frames, m is the frame index, $RMS(m)$ is the RMS at the m th frame, \overline{RMS} is the average RMS in the processing window, and $\text{sgn}(\cdot)$ is 1 for positive arguments and 0 for negative arguments. L_R is suitable for discriminating between speech and music because L_R is generally high for a speech signal that consists of words mixed with silence, whereas L_R for music is low.

- Spectrum Spread

Spectrum spread (SS) is effective for determining whether the power spectrum is concentrated around the centroids or is spread out over the spectrum. Music is composed of a broad mixture of frequencies, whereas speech consists of a limited range of frequencies. Consequently, the spectrum spread efficiently determines between speech and music. Its mathematical definition is given by

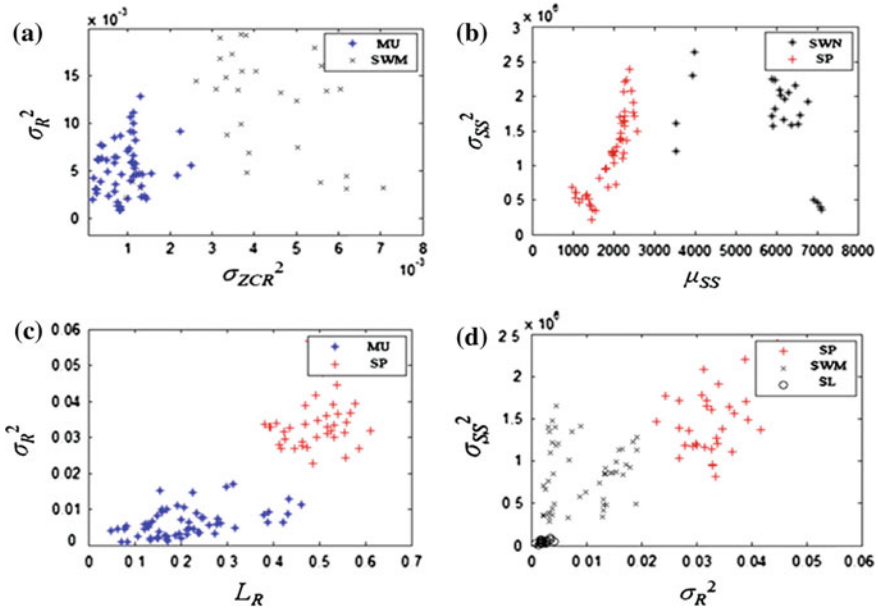


Fig. 47.1 Feature distribution based on $[L_R, \sigma_R^2, \mu_{SS}, \sigma_{SS}^2, \sigma_{ZCR}^2]$. Classification, **a** between music and speech with music, **b** speech and speech with noise, **c** music and speech, and **d** among speech, speech with music, and silence

$$SS(n) = \sqrt{\frac{\sum_{m=0}^{K-1} [(k - SC(n))^2 \times |A(n, m)|^2]}{\sum_{m=0}^{K-1} |A(n, m)|^2}}, \tag{47.2}$$

where K is the order of the discrete Fourier transform (DFT), m is the frequency bin for the n th frame, $SC(n)$ is spectral centroid at the n th frame, and $A(n, m)$ is the DFT of the n th frame of the given signal. $SC(n)$ and $A(n, m)$ are computed as

$$SC(n) = \frac{\sum_{m=0}^{K-1} m \times |A(n, m)|^2}{\sum_{m=0}^{K-1} |A(n, m)|^2}, \quad A(n, m) = \sum_{k=0}^{N_{samples}-1} x(k) e^{-j \cdot (\frac{2\pi}{W}) \cdot k \cdot m},$$

where $N_{samples}$ is the total number of samples in the audio stream.

- Zero-Crossing Rate

Zero-crossing rate (ZCR) value is defined as the number of zero-crossings within a processing window, as shown in (47.3):

$$ZCR(n) = \frac{1}{N_{samples} - 1} \sum_{m=0}^{N_{samples}-1} |sgn[x(m)] - sgn[x(m - 1)]|, \tag{47.3}$$

where $x(m)$ is the value of m th sample in the processing window, and $sgn(\cdot)$ is a sign function as mentioned in (47.1). Voiced and unvoiced speech sounds have

low and high zero-crossing rates, respectively. This results in high ZCR variation, whereas music typically has low ZCR variation.

47.2.1 Feature Vector Configuration

Audio streams are classified into the following five categories: *silence (SL)*, *speech (SP)*, *music (MU)*, *speech with music (SWM)*, and *speech with noise (SWN)*. According to our experiments, statistical values of RMS, SS, and ZCR represent the characteristics of target audio signals for classification well. Consequently, this paper finally selects five audio features $[L_R, \sigma_R^2, \mu_{SS}, \sigma_{SS}^2, \sigma_{ZCR}^2]$ for more accurate audio classification, as shown in Fig. 47.1.

47.3 GA-Based FCM for Audio Classification

Let an unlabeled data set $X = \{x_1, x_2, \dots, x_n\}$ represent n number of features. The FCM algorithm sorts the data set X into c clusters. The standard FCM objective function with the Euclidian distance metric is defined as follows:

$$J_m(U, V) = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m d^2(v_i, x_k), \quad (47.4)$$

where $d^2(v_i, x_k)$ represents the Euclidian distance between the centroid v_i of the i th cluster and the data point x_k , and u_{ik} is the degree of membership of the data x_k to the k th cluster, along with the constraint $\sum_{i=1}^c u_{ik} = 1$. The parameter m controls the fuzziness of the resulting partition, with $m \geq 1$, and c is the total number of clusters. Local minimization of the objective function $J_m(U, V)$ is achieved by repeatedly adjusting the values of u_{ik} and v_i according to the following equations:

$$u_{ik} = \left[\sum_{j=1}^c \left(\frac{d^2(v_i, x_k)}{d^2(v_j, x_k)} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad v_i = \frac{\sum_{k=1}^n u_{ik}^m x_k}{\sum_{k=1}^n u_{ik}^m} \quad \text{where } 1 \leq i \leq c. \quad (47.5)$$

As J_m is iteratively minimized, v_i becomes more stable. The iteration of the FCM algorithm is terminated when the terminal condition $\max_{1 \leq i \leq c} \{abs(v_i^t - v_i^{t-1})\} < \varepsilon$ is satisfied, where v_i^{t-1} are the centroids of the previous iteration, $abs()$ denotes the absolute value, and ε is the predefined termination threshold. Finally, all data points are distributed into clusters according to the maximum membership u_{ik} . As noted in Sect. 47.1, FCM starts with randomly initialized centroids, which has strong effects on its performance. To deal with this drawback, this paper employs a GA for obtaining more accurate classification performance. To do this, centroids are initially selected by GA, and these centroids

are used for calculating membership values of FCM. According to [5], suitable coding (representation of chromosomes) for the problem must be devised before GA is performed. Likewise, the fitness function, which assigns a figure of merit to each coded solution, is required. During the process, parents must be selected for reproduction, and combined to generate offspring.

47.3.1 Initialization

The initial population, which consists of randomly produced initial individuals, is generated, whose number is M_{cla} . We set the population size as $M_{cla} = 100$. Data are normalized within the range $[0:1]$, and then each individual is encoded in the population such as $Chrom^i = \{v_{11}^i, v_{12}^i, \dots, v_{15}^i, v_{21}^i, v_{22}^i, \dots, v_{25}^i, \dots, v_{51}^i, v_{52}^i, \dots, v_{55}^i\}$, where $Chrom^i$ is i th individual in population, $i \in [0:(M_{cla} - 1)]$.

47.3.2 Generating GA Operators

In this process, genetic operators such as selection, crossover, and mutation are set.

- *Selection*: The stochastic universal sampling method is utilized to select potentially useful chromosomes for recombination.
- *Crossover*: This paper selects an intermediate recombination technique which is suitable for use with real-valued code. In this method, different variable (or dimension) values of offspring are chosen somewhere near the values of parents. If there is a population of N_{var} dimension data in each individual, offspring are produced according to the following rule:

$$Var_i^0 = Var_i^{P_1} \times a + Var_i^{P_2} \times (1 - a), \quad (47.6)$$

where Var_i^0 indicates the value of the i th dimension of offspring, and $Var_i^{P_1}$ and $Var_i^{P_2}$ are values of the i th dimension of the first and second parents, respectively, and $i \in (1, 2, \dots, N_{var})$, $a \in [-d, 1 + d]$. Here, a is a scaling factor that is chosen by random over an interval $[-d, 1 + d]$ for each variable. The value of the parameter d defines the size of the area for possible offspring. In this paper, we set $d = 5$ because the number of features is five, and $N_{var} = 25$ because we classify audio data into five clusters with five features.

- *Mutation*: Individuals are randomly altered by mutation. These variations (mutation steps) are mostly small. The probability of mutating a variable is inversely proportional to the number of variables (dimensions). If one individual has more dimensions, the possibility of mutating a variable becomes smaller. In general, a mutation rate of $1/N_{var}$ has produced good results for a wide variety of objective functions. For audio classification, the mutation rate is set to 0.04 since

N_{var} is set to 25 for this paper. Furthermore, the objective function using (47.4) is computed in this process, and the fitness assignment is based on the objective function as follows:

$$f_i = \frac{1}{\sum_{l=1}^c \sum_{k=1}^n u_{ik}^m d^2(v_l^i, x_k)}. \quad (47.7)$$

47.3.3 Checking Termination Criterion

The optimization criterion is checked if $abs(f_{best}^i - f_{best}^{i-1}) \leq T_{clafinal}$ is satisfied or not, where f_{best}^i and f_{best}^{i-1} are the best fitness values for chromosomes in i th (current) and $(i - 1)$ th (previous) generations, respectively. Moreover, $T_{clafinal}$ is the pre-defined termination threshold. Therefore, if the terminal criterion is satisfied, we move on to the classification process. Otherwise, we turn back to *Generating GA operators* step.

47.3.3.1 Training and Classification

Based on the training process, we determine the training centroid of audio types $C = \{c_1, c_2, \dots, c_5\}$ and define clusters based on minimum distance from the training centroid. We then utilize the highest membership of each data point in order to classify audio streams into proper clusters: *SP*, *MU*, *SWM*, *SWN*, and *SL*. For the training process, this paper includes 15 pieces of silence, 15 pieces of pure speech, 15 pieces of music which involve various musical instruments, 15 pieces of speech with music, and 15 pieces of outside interviews that are composed of different background noise levels.

47.4 Experimental Results

For audio classification simulation, we utilize two datasets composed of Korean news broadcasts obtained from Ulsan Broadcasting Corporation (www.ubc.co.kr). The first dataset was used for testing and the other dataset was used for training. We employ GA-based FCM to classify audio streams. In this experiment, the degree of the fuzziness and the termination condition for GA-based FCM were set to 2 and 0.001, respectively, because Bezdek et al. experimentally determined the optimal intervals for the degree of fuzziness and termination threshold, which range from 1.1 to 5 and 0.01 to 0.0001, respectively [6]. To evaluate classifications, we utilize the standard of correctness that has been widely accepted in recent studies, which is as follows:

Table 47.1 Classification result of the proposed method

	Classified					Correctness (%)	
	<i>SP</i>	<i>MU</i>	<i>SWM</i>	<i>SWN</i>	<i>SL</i>		
Input audio stream	<i>SP</i>	54	0	0	0	0	100.0
	<i>MU</i>	0	6	0	0	0	100.0
	<i>SWM</i>	0	0	5	0	0	100.0
	<i>SWN</i>	2	0	0	21	0	91.3
	<i>SL</i>	6	0	0	0	51	89.5

$$Correctness = \frac{\# \text{ of correctly classified audio-segments}}{\# \text{ of all audio-segments}} \times 100 (\%). \quad (47.8)$$

Table 47.1 presents classification results of the proposed approach and several misclassified results are obtained when attempting to distinguish between speech and silence, because some silence signals include speech components at the beginning and end. Likewise, misclassifications of speech with noise signals occurred mostly when the amplitudes of noise components were small and unclear.

47.5 Conclusion

This paper proposed a robust audio classification approach to address the rising demand for efficient multimedia content management. To classify audio streams into one of five categories (speech, music, speech with music, speech with noise, and silence), this paper explored 19 audio features. Among these audio features, this paper selected the five most suitable features such as $[L_R, \sigma_R^2, \mu_{SS}, \sigma_{SS}^2, \sigma_{ZCR}^2]$, and utilized these extracted features as inputs of GA-based FCM for audio classification. Our experimental results showed that the proposed classification method achieved very accurate classification performance.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2013R1A2A2A05004566 and NRF-2012R1A1A2043644).

References

1. Foote J (1999) An overview of audio information retrieval. *Multimedia Syst* 7(1):2–10
2. Park D-C, Tran CN, Min B-J, Park S (2006) Modeling and classification of audio signals using gradient-based fuzzy C-means algorithm with a Mercer Kernel. *Lect Notes Comput Sci* 4009:1104–1108
3. Park D-C (2009) Classification of audio signals using fuzzy C-Means with divergence-based Kernel. *Patt Recogn Lett* 30(9):794–798

4. Cheng WS, Ji CM, Liu D (2009) Genetic algorithm-based fuzzy cluster analysis for flood hydrographs. In: International workshop on intelligent systems and applications, Wuhan, pp 1–4
5. Beasley D, Bull DR, Martin RR (1993) An overview of genetic algorithms: Part 1. *Fundam Univ Comput* 15:58–69
6. Bezdek JC, Keller J, Krisnapuram R, Pal N (2005) *Fuzzy models and algorithms for pattern recognition and image processing*. Springer, Berlin