Audio Classification Framework based on DSVM

K Zin Lin
University of Computer Studies, Yangon
kzinlin78@gmail.com

Abstract

The explanation of how a decision made is important for accepting the machine learning technology, especially for applications such as multimedia. SVMs have shown strong generalization ability in a number of application areas, including audio classification. However, the poor comprehensibility hinders the success of the SVM for audio class prediction. On the other hand, a decision tree has good comprehensibility. This approach combines the SVM with decision tree. We use the SVM as a decision of binary tree to select strong instances to generate rules; it is known as decision SVM (DSVM). We considered eight audio classes: silence, non-silence, speech with music, speech with noise, music with environment sound, instrumental music, environmental sound and noise. This classification and analysis is intended to analyze the structure of the sports video.

1. Introduction

In the near future, with the arrival of the third generation of mobile networks, all this multimedia content will virtually be accessible anytime, anywhere. As such, the increasing amount of multimedia information available highlights the need to develop systems able to automatically describe this information for more efficient filtering, retrieval and, in general, management. So that the applications describing the content and the applications using the corresponding descriptions can interoperate, it is necessary to define a standard that specifies the syntax and semantics of these multimedia descriptions.

The purpose is to allow that the huge amount of multimedia content available can be filtered, searched for, managed and consumed in a thoughtful, flexible, fast and efficient way. There is an increasing need to summarize and personalize audiovisual content. Several feature sets and machine learning algorithms have been tested, providing choices of speed and performance for a target system.

In this proposed system, since the SVM usually has strong generalization ability and using the SVM as the inputs to DT, the noise, may be reduced by the process of SVMs, and some weak cases may be sieved by SVMs. This approach is to achieve high accuracy in classifying of mixed types of audio by combining two types of classifiers.

The rest of this paper is organized as follows. In Section 2, we present the related work and Section 3 describes how an audio clip is represented by low level perceptual and cepstral feature and gives an overview of linear, kernel SVM and decision tree. In Section 4, a method for multi-class classification is presented. Finally, in Section 5, we conclude for the proposed system.

2. Related work

Audio classification is to determine the category of audio file automatically according to the features under given classification system. In A. Rabauoi [1] consists of illustrating the potential of SVMs on recognizing impulsive audio signals belonging to a complex real world dataset. They presented to apply optimized one-class support vector machines (1-SVMs) to tackle both sound detection and classification tasks in the sound recognition process. They compared the sound detection and classification methods [unsupervised sound detection algorithm based on 1-SVMs] with other popular approaches [HMMs, M-SVM (1-vs-1, M-SVM (1-vs-all)].

In L. Chen [2] addressed the issue of mixed type audio data based on Support Vector Machine (SVM). In order to capture characteristics of different types of audio data, besides selecting audio features, they designed four different representation formats for audio features such as ZCR, silence ratio, harmonic ratio and sub-band energy. Their SVM-based audio classifier can classify audio data into five types: music, speech, environment sound, speech mixed with music, and music mixed with environment sound.

The work presented in L. Bai [3] used a scheme for indexing and segmentation of video by analyzing the audio track using support vector machine. This

analysis is then applied to structuring the sports video. They defined three audio classes in sports video, namely Play-audio, Advertisement-audio and Studio-audio based on the attributes of sports video. They used Support vector machine (SVM) for audio classification and then applied smoothing rules in final segmentation of an audio sequence. The results show the performance of SVM on audio classification is satisfying.

R. Shantha Selva Kumari [6] showed an improved feature vector formation technique for audio classification and categorization. This technique makes use of wavelets to extract the features of audio data. Wavelets are first applied to decompose the signal and to extract acoustical features such as sub-band power, brightness and band-width and pitch information. The additional features, such as frequency cepstral coefficients also extracted to accomplish audio classification and use a bottom-up Support Vector Machine over these acoustical features and additional features. The bottom-up Support Vector Machine categorization strategy uses an iterative procedure to match a given audio to progressively larger subsets or categories of classes.

S. O. Sadjadi An unsupervised clustering method is proposed [7], based on one-class support vector machines (OCSVM) and inspired by the classical K-means algorithm, which effectively classifies speech/music signals. First, relevant features are extracted from audio files. Then in an iterative K-means like algorithm, after initializing centers, each cluster is refined using a one-class support vector machine. The experimental results show that the clustering method, which can be easily implemented, performs better than other methods implemented on the same database.

3. Background

3.1 Audio feature extraction

In order to obtain high accuracy for audio classification, it is critical to select good features that can capture the temporal and spectral characteristics of audio signal and are robust for circumstance changing. In this work, the audio stream is segmented into clips that are 1 second long with 0.5 second overlapping with the previous ones. Each clip is then divided into frames that are 440 samples long for sampling frequency of 22 kHz. Features are analyzed and extracted in two levels: frame-level and clip-level. The features extracted from one clip are combined as one feature vector after normalization. We extracted the frame level features that include: zero-crossing rate (ZCR), short time energy (STE) and spectral flux (SF).

In our work, audio clip-level features are computed based on the frame-level features and used a clip as the classification unit. For ZCR, STE and SF, we compute its mean of all frames in a given clip respectively as base clip-level features which is proved to be effective for distinguishing speech, music and speech with background sound [5,6].

Zero-crossing rate is proved to be useful in characterizing different audio signals. In general, speech signals are composed of alternating voiced sounds and unvoiced sounds in the syllable rate, while music signals do not have this kind of structure. Hence, for speech signal, its variation of ZCR will be in general greater than that of music.

$$ZCR = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} |sgn[x(m+1)] - sgn[x(m)]| \quad (1)$$

where sgn[.] is a sign function and x(m) is the discrete audio signal, m=1…..N.

STE is the audio feature that is widely used and the easiest. It is also called volume. STE is a reliable indicator for silence detection. Normally STE is approximated by the rms (root mean square) of the signal magnitude within each frame.

$$E(m) = \sum_m (x(n)W(n-m))\text{^}2 \quad (2)$$

where m is the time index of the short-term energy, x(n) is the discrete time audio signal, W(n) is the window (audio frame) of length N where n=0,1,2,…..,N-1.

Spectrum flux (SF) is defined as the average variation value of spectrum between the adjacent two frames in a given clip.

$$SF_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])\text{^}2 \quad (3)$$

where $N_t[n]$ and $N_{t-1}[n]$ are the normalized magnitude of the Fourier transform at the current frame t, and the previous frame t-1, respectively. The spectral flux is a measure of the amount of local spectral change.

For speech signal, the spectrum flux (SF) of environment sounds is among the highest and change more dramatically than those of speech and music. Based on our previous work [4], this feature is especially useful for discriminating some strong periodicity environment sounds such as tone signal, from music signals. SF is a good feature to discriminate among speech, environment sound and music.

3.2 Support vector machine

SVM algorithm is a classification algorithm that provides state-of-the-art performance in a wide variety of application domains. There are two main reasons for using the SVM in audio classification.

First, many audio classification problems involve high dimensional, noisy data. The SVM is known to behave well with these data compared to other statistical or machine learning methods.

Second, the feature distribution of audio data is so complicated that different classes may have overlapping or interwoven areas. However, a kernel based SVM is well suited to handle such as linearly non-separable different audio classes.

3.2.1 Linear support vector machines

SVM transforms the input space to a higher dimension feature space through a nonlinear mapping function. Construct the separating hyper plane with maximum distance from the closest points of the training set.

Consider the problem of separating a set of training vectors belonging to two separate classes, $(x_1; y_1), \ldots, (x_l; y_l)$, where $\boldsymbol{x}_i \in R_n$ is a feature vector and $y_i \in \{-1, +1\}$ is a class label, with a separating hyper-plane of equation $\boldsymbol{w}\cdot\boldsymbol{x}+b = 0$; of all the boundaries determined by $\boldsymbol{w}$ and b. On the basis of this rule, the final optimal hyper-plane classifier can be represented by the following equation:

$$f(x) = sgn\left(\sum_{i=1}^{l} \bar{\alpha} i y_i x_i x + \bar{b}\right) \quad (4)$$

where α and b are parameters for the classifier; the solution vector $\boldsymbol{x}_i$ is called as Support Vector with α_i being non-zero.

3.2.2 Kernel support vector machines

In the linearly non-separable but non-linearly separable case, the SVM replaces the inner product by a kernel function K(x,y), and then constructs an optimal separating hyper-plane in the mapped space. According to the Mercer theorem, the kernel function implicitly maps the input vectors into a high dimensional feature space in which the mapped data is linearly separable. In our method, we use the Gaussian Radial Basis kernel.

3.3 Decision tree

Learned trees can also be re-represented as sets of if-then rules to improve human readability. In a set of records, each record has the same structure, consisting of a number of attribute/value pairs. One of these attributes represents the category of the record. The problem is to determine a decision tree that, on the basis of answers to question about the non-category attributes, predicts correctly the value of the category attribute. In the decision tree, each node corresponds to a non-categorical attribute and each arc to a possible value of that attribute. A leaf of the tree specifies the expected value of the categorical attribute for the records described by the path from the root to that leaf.

4. Design of proposed system

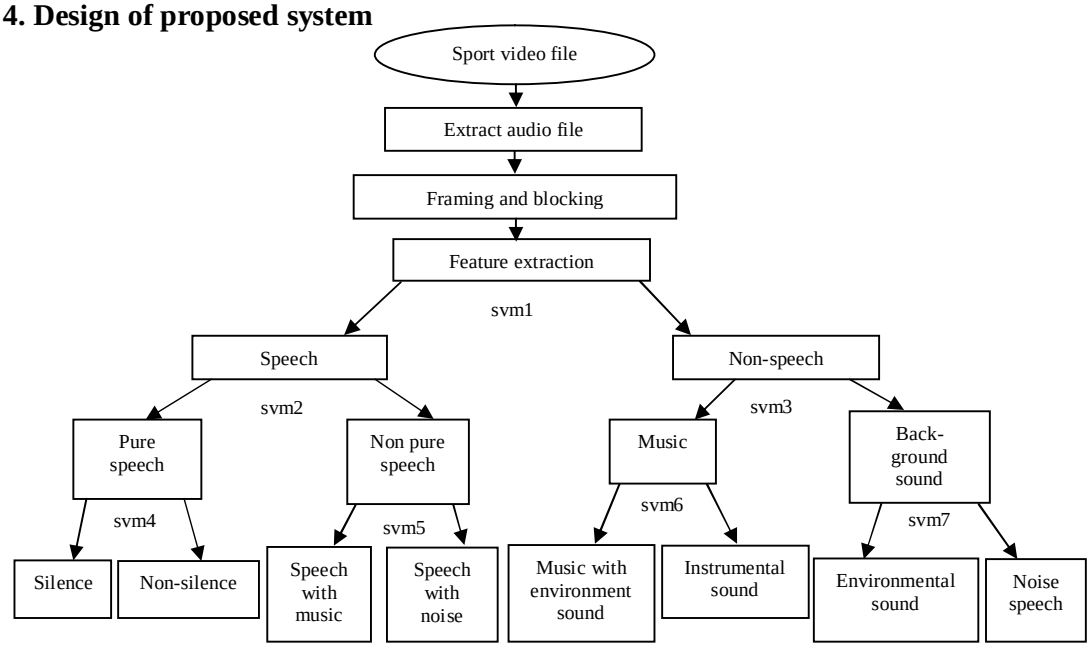

Figure 1. Mixed type audio classification by using DSVM

Many efforts in this area focus on audio data that contains some built-in semantic information structure such as in broadcast news, or focus on classification of audio that contains a single type of sound such as clear speech or clear music only. We should describe goodness of decision tree based SVM for mixed type audio.

In this work, an audio clip is classified into eight classes. We should figure out seven SVMs for classification as SVM is a two-class classifier. SVMs are learnt with the training data.

The motivation of combining the SVM and decision tree is to combine the strong generalization ability of the SVM and the strong comprehensibility of rule induction. In this research, we discuss about combining Support

Vector Machine and decision trees for multi class audio classification as shown in Figure 1.

Firstly, audio file is extracted from video file for feature calculation. To extract this audio file, we put the audio quality with the sampling frequency of 22 kHz, bit rate of 128 kbps and mono channel. Then this audio stream is split 15sec long audio wav file.

In the second stage, the audio stream is analyzed into 1sec audio clips with 0.5 sec overlap and each clip is divided into frames. The frame size is 20ms for sampling frequency of 22 kHz and it is analyzed by non-overlapping.

In the next stage, features are analyzed and extracted in two levels: frame-level and clip-level by using audio features ZCR, STE and SF which is proved to be effective for distinguishing speech, music, speech with background and environment sound.

But the characteristics of the feature components are so different that it is not appropriate to just put these features into a feature vector. Each feature component should be normalized to make their scale similar. The normalization is processed as:

$$x_i' = \frac{(x_i - \mu_i)}{\sigma_i} \quad (5)$$

where x_i is ith feature component, the corresponding mean μ_i and standard derivation σ_i can be calculated from the ensemble of the training data. The normalized feature vector is considered as the final representation of an audio clip.

After that SVM are trained on each class at each level of the tree and the SVM which is more successful in predicting a class at that level is selected as the decision in that node. Thus a tree is constructed with different SVM in each node. And the tree constructed is used for classifying the multi class audio.

SVM1 discriminate between speech and non-speech in the first level. After that those speech clips are classified into pure speech and non-pure speech by using SVM2 in the second level. At the same level, the non-speech clips are classified into music and background sound by using SVM3. Then, in the third level, pure speech, non-pure speech, music and background sound are classified into silence, non-silence, speech with music, speech with noise, music with environment sound, instrumental sound, environmental sound and noise speech by using SVM4, SVM5, SVM6 and SVM7, respectively.

Thus classification of an audio stream is achieved by classifying each clip into an audio class in sports video. The performance of the result is measured by classification accuracy defined as the number of correctly classified clips over total number of clips.

5. Conclusion

This proposed system focus on developing an effective scheme to apply audio content analysis to improve video structure parsing and indexing process. In this paper, combining SVM and decision tree is presented for effective multi-label audio classification. Audio file will be extracted in two levels from video file. The three features set ZCR, STE and SF will be used. ZCR is used to discern non-speech. Usually non-speech has a low short-term energy but a high zero crossing rate. Combining ZCR and STE to prevent low energy unvoiced speech frames from being classified as silent. STE is a reliable indicator for silence detection. SF is a good feature to discriminate among speech, environment sound and music. Moreover binary-class approach for multi-label classification is used. Since the SVM as a decision of binary tree to select strong instances to generate rules is used, we don't need to train the whole training set when we discriminate each audio clip and better accuracy and time saving can be expected upon the whole architecture of the system.

6. References

[1] A. Rabaoui, H. Kadri, Z. Lachiri, and N. Ellouze, "One-Class SVMs Challenges in Audio Detection and Classification Applications", Hindawi Publishing Corporation EURASIP Journal on Advances in Signal Processing, Volume 2008, Article ID 834973, 14 pages.

[2] L. Chen, S. G¨und¨uz, M. Tamer O¨ zsu, "Mixed Type Audio Classification With Support Vector Machine", Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, ICME 2006.

[3] L. Bai, S. Lao, H. Lia0, J. Chen, "Audio Classification And Segmentation For Sports Video Structure Extraction Using Support Vector Machine", IEEE, Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006.

[4] L. Lu, H. Jiang, H. J. Zhang, A Robust Audio Classification and Segmentation Method. Proc. of the 9th ACM international conference on Multimedia, pp. 203–211, 2001

[5] L Lu, HJ Zhang, SZ Li, Content-based Audio Classification and Segmentation by using SVM, Multimedia Systems, 2003-Springer Digital Object Identifier (DOI) 10.1007/s00530-002-0065-0, Journal Article (2003)

[6] R. Shantha Selva Kumari, D. Sugumar, V. Sadasivam, "Audio Signal Classification Based on Optimal Wavelet and Support Vector Machine", IEEE, International Conference on Computational Intelligence and Multimedia Applications (ICCIMA) 2007.

[7] S. Omid Sadjadi, S.M. Ahadi, O. Hazrati, "Unsupervised Speech/Music Classification Using One-Class Support Vector Machines", IEEE, Information, Communications & Signal Processing, 2007 6th International Conference on Volume , Issue.