ACADEMIC
World Academic Union

ISSN 1746-7233, England, UK
World Journal of Modelling and Simulation
Vol. 1, No. 1, 2005, *pp*. 56-59

Audio classification by hybrid support vector machine / hidden Markov model *

Xin He [1 +], Xian-Zhong Zhou [2]

[1] Department of Automatic Control, Nanjing University of Science and Technology, Nanjing 210094, China
[2] School of Management and Engineering, Nanjing University, Nanjing 210093, China

(*Received March 29 2005, Accepted May 5 2005*)

Abstract. Audio is one of important information carriers in the multimedia. It contains abundant semantics and enriches information perception and acquisition. At present, it always uses vision information in the multimedia retrieval, but ignores audio information. In this paper, the problem of audio classification is discussed. The combination of Support Vector Machine and Hidden Markov Model is described and the hybrid model is used in audio retrieval experiments.

Keywords: support vector machine (SVM), hidden Markov model (HMM), audio classification, retrieval

1. Introduction

There are large amounts of multimedia data in network data and audio data is one of many computer and multimedia applications. It cannot meet the demand of retrieval by conventional techniques and demands in audio data rapidly increases. For these reasons, audio retrieval is becoming one of emphases in multimedia information retrieval.

Researches in closely related areas, such as speech recognition, speaker identification, etc. have long history. But researches on classification and retrieval of audio information are relatively new. Recently some researches are done by several research institutes. One important work is the system called “Muscle Fish”, which can classify and retrieve audio [1]. Moreover, it’s also used Neural Network (NN) and Hidden Markov Model (HMM) in classification of audio information [2], [3]. But also there exist some problems in audio retrieval, such as retrieval of apperception characters, scene analysis of audio, research of classifier and effective retrieval interface, etc. In this paper, the problem of audio classification is considered. It discusses how to classify audio clip better according to different characters.

The modeling ability in time series of Hidden Markov Model (HMM) is obvious. It’s used in speech recognition extensively and also in research of audio classification. Another classification method called support vector machine (SVM) is used, which is a statistical learning algorithm [4].

The paper is organized as follows. In Section 2, it introduces the basic theory of SVM and the structure of the hybrid model of SVM and HMM. In Section 3, it presents the hybrid model for retrieval. Section 4 describes the experiments for audio classification. Finally, the conclusion and discussions are given in Section 5.

2. Hybrid model of SVM/HMM

2.1. Basic theory of SVM

The SVMs are proposed recently by Vapnik [4], and have been used to solve some practical problems, such as face detection [9]. The SVM is a powerful new machine learning algorithm, which is rooted in statistical learning theory [10]. By constructing a decision surface hyper-plane which yields the maximal margin between position and negative examples, SVM approximately implements the Structure Risk

* This research was sponsored by the Natural Science Foundation of Jiangsu Province, P. R. China (BK2004137).

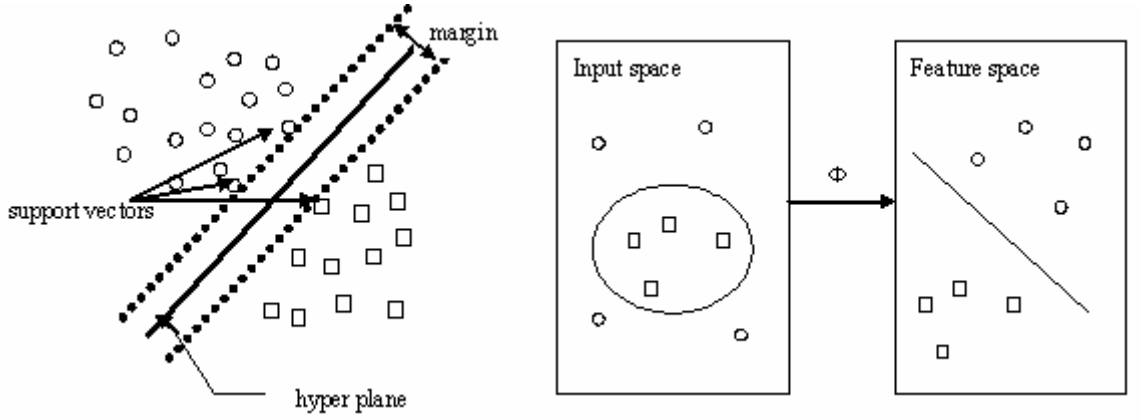
+ Corresponding author. Tel.: +86-25-84303047 ext. 8.
E-mail address: kernel_he@hotmail.com

Minimization(SRM) Principle. There are three layers in the structure of SVM: input layer, which gets data just as classification characters; hidden layer, with two functions, namely mapping the input data from low-dimension space to high-dimension by non-linear map and calculating the inner production of character vector and support vector; output layer, just showing the classification results. In the practical application, the function of hidden layer is achieved by the kernel function. The kernel function can be formulated as follows:

$$K(x, y) = (\Phi(x) \bullet \Phi(y))$$

where $K, \Phi, \bullet$ denotes respectively kernel function, non-linear mapping in high-dimension and inner production.

The SVM is based on the optimal hyper-plane in linear separable condition. The basic idea is just as follows: transforming the input into high-dimension space by non-linear transform; then calculating the optimal linear classification plane, which is described in Figure 1.



(a) (b)

Fig. 1: (a) A linear SVM finds the maximum margin linear separating hyper-plane in the input space. (b) A non-linear SVM uses a nonlinear kernel to map the data into a high dimensional future space.

As well known, there are three main types in kernel functions:

(1) Polynomial: $K(x, y) = (x \bullet y + 1)^d$, where d is the degree of the polynomial;

(2) Gaussian Radial Basis Function: $K(x, y) = \exp(-\frac{\|x - y\|^2}{2\sigma^2})$, and the parameter σ is the width of Gaussian unction;

(3) Sigmoid Function: $K(x, y) = \tanh(k(x \bullet y) - \mu)$, k and μ are the scale and offset parameters. In this paper, the Sigmoid Function is used, just because it includes only one hidden multi-layer perception, the number of hidden-layer nodes is confirmed by the algorithm and there is no local minima problem, which exists in the method of Network Neural.

2.2. Hybrid model of SVM/HMM

Just as the HMM or SVM, the hybrid model also includes two parts: training and classification. Firstly, $\lambda = (A, B, \pi)$, parameters of model can be obtained by training, where B is the model parameter of input. Secondly, it can be calculated the probability estimate $P(\overline{X} \mid \lambda)$ by Viterbi, corresponding to the observation sequence $\overline{X} = X_1 X_2 \cdots X_N$, and compared with the threshold to obtain classification results.

But the output of SVM is numerical value. For combined with HMM easily, it should transform the value into probability form [5-8]. In the processing of calculating, it should make training sample normalization, namely $|g(x)| = 1$. Then the sample point can be described as $g(x) = \pm dw$, where d denotes the distance between sample point x and classification plane, and the sign denotes which side the point position. The probability output form is formulated as follows:

$$P(C_{\pm 1} \mid x_i) = \frac{1}{1+\exp(\mp g(x_i))}$$

2.2.1 Training of SVM

SVM is a two-class classifiers, it should be used to a multi-class classification in audio classification. There are two common schemes for this purpose: one-against-all and the one-against-one. In this paper, the one-against-all is used, namely training SVMs for each audio. The detailed processing is described as follows:

Supposing there are *m* classes audio in audio sample database, and there are *n* training sample for each class.

(1) Selecting all sample of the first audio from sample database, labeling as class I and other audio classes as class II. Using these samples as input to training a SVM and obtaining the corresponding support vector and optimal classification plane. Then labelling the SVM as ①, which can distinguish class I from other audio classes.

(2) Selecting all sample of the second audio from sample database, labeling as class I and rest audio classes as class II. Using these samples as input to training a SVM and obtaining the corresponding support vector and optimal classification plane. Then labelling the SVM as ②, which can distinguish the second class from other audio classes.

(3) Repeating steps described above, training all audio classes and achieving *m* SVMs.

2.2.2 Training of HMM

Supposing M class audio in the database, namely there should be M HMM, and each sample class has *n* sample. Audio training samples are classified by SVM training method and probabilistic outputs of SVM are calculated by the forward-backward algorithm. Then models of all audio are achieved.

2.2.3 Audio classification of the Hybrid model

Supposing there are *m* classes audio in sample database, which is for classification. And there are trained templates according to each audio class. When a new sample for classification is inputted, it can be classified by the method described as above. It's formulated as Figure 2.

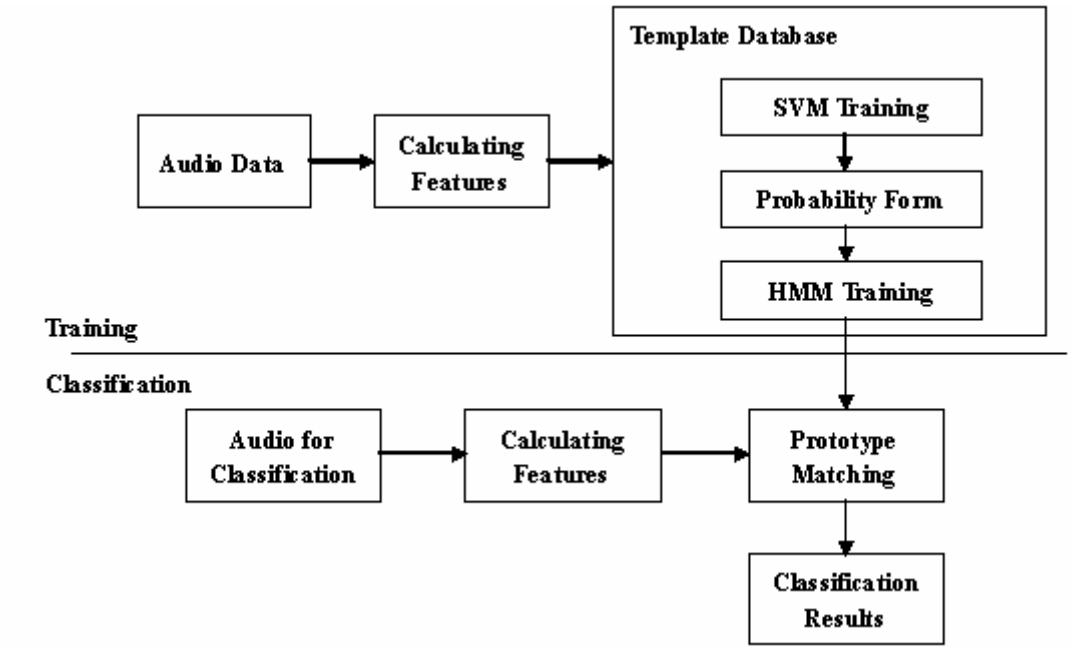


Fig. 2: Hybrid model of SVM/HMM

3. Experiments

The database used in experiments is classified into three classes, namely pure speech, background music and

claps, collected from TV programs, the internet and music CDs with each clip labeled in terms of pre-defined classes. All audio clips are 8-bit, mono-channel, down-sampled into 16 KHz and saved as ".wav". It's partitioned into a training set of 60 minutes and a test set of about 20 minutes. For comparison, the commonly method HMM is also used in experiments, just shown as Table 1.

Table 1. Experiment results of different classifying type

Perc	Background music	Claps	Pure speech
SVM/HMM	83%	92.4%	96.2%
HMM	76%	90%	95%

4. Conclusions

In this paper, it presents in detail the approach that uses SVM/HMM for classification of an audio clip. Experiments show that the hybrid SVM/HMM achieves high classification accuracy. For future research, we will consider how to apply SVM to solve multi-classes problems and improve classification scheme to discriminate more audio classes.

5. Acknowledge

The authors are very grateful to Dr. Ying-chun Shi and Dr. Bing Huang to give us many valuable comments and advices. We also thank the anonymous referees for their helpful comments and suggestions.

6. References

[1] Wood Erling, et al, "*Content based classification, search and retrieval of audio*". IEEE Multimedia, 27-36, 1996.

[2] Fei Wu, Yueting Zhuang, Yin Zhang, et al, "*Hidden Markov Model based audio semantic retrieval*". Pattern Recognition and Artificial Intelligence, 14(2001), 1, pp. 104-108.

[3] Jian Lu, Yi-song Chen, Zheng-xing Sun, et al, "*Automatic audio classification by using Hidden Markov Model*". Journal of Software, 13(2002), 8, pp. 1593-1597.

[4] N. Vapnik Vladimir, "*The Nature of Statistical Learning Theory*", Springer, New York, 1995.

[5] Dong Xin, Yingchun Yang, Zhaohui Wu, "*Speaker verification with the hybrid use of Support Vector Machine and Hidden Markov Model*", Journal of computer-aided Design & Computer graphics, 14(2002), 11, pp. 1080-1082.

[6] A. Ganapathiraju, J. Hamaker, J. Picone, "*Hybrid SVM/HMM architectures for speech recognition*", http://www.isip.msstate.edu/publications/conferences/icslp/2000/hybridasr/paper v0.pdf, 2002.

[7] J. T. Kwok, "*Moderating the outputs for support vector machine classifiers*", Proceedings of International Joint Conference of Neural Networks, Washington DC, 1999, pp. 943-948.

[8] J. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods", In Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, USA, 2000.

[9] E. Osuna, R. Freund, and F. Girosi, "*Training support vector machines: An application to face detection*", In Proc. CVPR, 1997.

[10] V. N. Vapnik, "*Statistical learning theory*", John Wiley & Sons, New York, 1998.