# A New Hybrid Audio Classification Algorithm Based on SVM Weight Factor and Euclidean Distance

YUK YING CHUNG, *ERIC H.C.CHOI, LIWEI LIU, MOHD AFIZI MOHD SHUKRAN
*DAVID YU SHI, *FANG CHEN
School of Information Technologies, University of Sydney, NSW 2006, AUSTRALIA
* ATP Research Laboratory, National ICT for Australia, NSW 1430, AUSTRALIA

*Abstract:* - The text-based classification dominates in the conventional audio classification systems, in which tedious manual work is used to notate the name, class, or sample rate. However, on most occasions, this method is not satisfying due to its opaque to real content. In order to retrieve the audio files effectively and efficiently, content-based audio classification becomes more and more necessary. In this paper, a new hybrid approach of audio classification algorithm is proposed to improve the performance of some mis-classified audio data. The proposed method has been compared with the traditional Euclidean-based K-Nearest Neighbor classifier. As to improve the accuracy for specific problems, a weight factor based on Supporting Vector will apply to the Euclidean distance and K-NN rule to achieve better accuracy. The proposed new weighted Euclidean algorithm has been proved to be more sensitive to the classification criteria. The experimental results show that it can improve the audio classification accuracy by 28% at the maximum and 7% in the overall performance. By using the new proposed algorithm, some mis-classified audio data from a conventional Euclidean distance classifier can be classified.

*Key-Words:* - audio classification, K-Nearest Neighbor, SVM, Euclidean distance

## 1 Introduction

Due to the rapidly-increasing amount of audio data in archives, it has become difficult to find out the particular audio files by using query. The audio classification becomes a very critical issue in audio retrieval. The conventional text-based audio classification is not only time-consuming, but also subjective. It seems to be helpless in many occasions. In order to classify the audio files effectively and efficiently, content-based audio classification is required. In this paper we propose the new hybrid audio classification algorithm based on SVM weight factor in order to improve the accuracy of audio classification.

As the demands and usage for audio classification and retrieval increase, the classification methods need to be improved to be more automatic and effective. The traditional text-based audio classification fails to recognize the underlying content of audio files. In this paper we propose a new hybrid audio classification algorithm based on SVM weight factor and Euclidean distance in order to improve the accuracy for audio classification. The proposed algorithm can extract the weight factor from Supporting Vector classification Model (SVM) and apply to the Euclidean measurement.

In this paper, two feature extraction algorithms namely, Original MFCCs (mel-scaled frequency cepstral coefficients) and Average MFCCs have been implemented. For Original MFCCs, the DTW (Dynamic Timing Warping) distance is used to measure the similarity of two audio files. In Average MFCCs, the Euclidean distance is directly calculated on two feature sets. Furthermore, the SVM (Supporting Vector Machine) classifier is implemented. Three classification algorithms: DTW distance, NN and K-NN have been implemented and compared with SVM classifier. The weight factor from SVM is chosen because of its outstanding ability to minimize structural risks while the other algorithms are based on enumeration. By combining the SVM weight factor and DTW distance we can consider the classification to be more aware of classification in boundary cases. The improvement of using the SVM weight factor in Euclidean distance in audio data classification has been shown in section 5.

Section 2 is the introduction to the audio classification system. Section 3 describes the feature extraction for the audio data. Section 4 explains the audio classification algorithms and the proposed new weight-KNN classification algorithm. Sections 5 and 6 present the test results and conclusion, respectively.

## 2 Introduction to Audio Classification

With the increase in the amount of audio data, audio classification technology is in high demand in order to provide effective and efficient methods. So far, audio classification has experienced two dramatic stages. At the beginning, the text-based audio classification dominated. The mechanism is similar with many popular search engines such as Google, Yahoo and Baidu. The searching result is located by using query keywords. The manual work is required to notate attributes describing audio files. For text-based technology, it can achieve a satisfactory performance, provided that the song is searched by its name. Otherwise, it will be very hard to find out the similar audio files based on specific audio features. Due to such subjective characteristics, there has been a demand for content-based audio classification technology to solve these problems.

### 2.1 Text-based Audio Classification

Traditional text-based audio classification uses the manual effort to denote the descriptive words of the audio files in the database. The user will perform the query based on these attributes. As it is difficult for the manual classification to cover all of the required features that users need, the text-based audio classification is proved to be subjective and not user-friendly. Meanwhile, it requires enormous work to group the similar audio files which is time-consuming. Another problem is that the users usually use the query words from their own understanding. It is very difficult to match exactly the descriptive attributes in the database. Finally, the searching will fail due to the poor classification scheme.

### 2.2 Content-based Audio Classification

In the domain of content-based audio classification, the problem is converted to the computation on the feature data set. The two steps, namely feature extraction and classification method, are involved. Feature extraction refers to transforming the original audio data into the feature vector in which the specific desired characteristics are contained. It is the first step in audio processing and has proven to have a large correlation with the final accuracy. In the next step, classification methods will perform the automatic classification by analyzing the derived feature vectors. The classifiers vary from each other due to their different schemes. Generally, the sample model is needed to serve the classification.

### 2.2.1 Feature Extraction

The audio features are generally divided into two categories: time domain and frequency domain. In time domain, the audio representation is expressed as the basic audio amplitude change with time. Some features can be derived from the statistics of amplitude, such as silence rate (SR) and loudness. In the frequency domain, features are obtained by applying Fourier transform. Such features include pitch, bandwidth, brightness, harmony etc. In this paper, the MFCCs feature is adopted.

### 2.2.2 The Proposed Content-Based Audio Classification (CBAC)

The proposed Content-Based Audio Classification (CBAC) includes three steps: (1) feature extraction, (2) classification, and (3) retrieval. Feature extraction refers to transforming the original audio data into the feature vector in which the specific desired characteristics are contained. It is the first step in audio processing, and it has a large correlation with the final accuracy. The classification and retrieval function together by applying the effective formulation of a distance measure and the classification rules.

The audio features can be divided into two categories: time domain and frequency domain. In time domain, the audio representation is expressed as the basic audio amplitude change with time. Some features can be derived from the statistics of amplitude, such as silence rate (SR) and volume. In the frequency domain, features are obtained by applying the Fourier transform. Such features includes pitch, bandwidth, brightness and harmony.

Fig. 1 shows the block diagram of the proposed CBAC. In the proposed CBAC, Mel-Frequency Cepstral Coefficients (MFCCs) [1][2] were used as the feature extraction algorithm for audio signal. As the legnth of the input query audio data and the stored audio sequence data are usually different, Dynamic Time Warping (DTW) [3] algorithm was used to generate the same length of feature vectors. The Nearest Neighbor (NN) [4] rule was used for the classification and retrieval of audio data. Sections 3.1, 3.2 and 4.3 explain in more detail the MFCCs, audio classification and DTW, respectively.

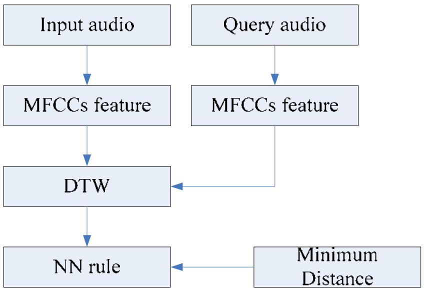

Fig. 1 The block diagram of the proposed content based audio classification system

# 3 Feature Extraction

## 3.1 Original Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs [1][2] are widely used in speech recognition systems. It adapts to human acoustic characteristics. Fig. 2 shows the block diagram of original MFCC feature extraction. The MFCCs transform was conducted in the following four steps:

1. The audio was hamming-windowed in overlapping steps;
2. The discrete Fourier transform (DFT) was computed to transform the audio data into frequency domain;
3. The spectral coefficients were perceptually weighted by a non-linear map of the frequency scale (Mel-frequency scale). Triangle filters were used here; and
4. Another DFT was used to transform the coefficients into cepstral coefficients.

$$C_k = \sum_{j=1}^{24} \log(Y_j)\cos[k(j-\frac{1}{2})\frac{\pi}{24}]$$

where k = 1,2,...,P

P is the dimension of coefficient，usual choice P＝12. $\{C_k\}_{k=1,2,...,12}$ are derived MFCC coefficients. The transform formula for the Mel-frequency scale and linear frequency scale is as follows.

$$mel = \ln\left(1+\frac{f}{700}\right)\cdot\frac{1000}{\ln(1+1000/700)}$$

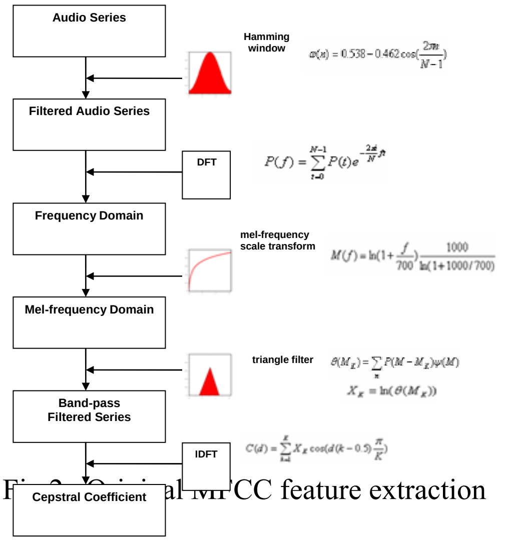
Audio Series
Hamming window
$\omega(n)=0.538-0.462\cos(\frac{2\pi n}{N-1})$
Filtered Audio Series
DFT
$F(f)=\sum_{t=0}^{N-1}P(t)e^{-\frac{2\pi i}{N}ft}$
Frequency Domain
mel-frequency scale transform
$M(f)=\ln(1+\frac{f}{700})\frac{1000}{\ln(1+1000/700)}$
Mel-frequency Domain
triangle filter
$\theta(M_k)=\sum_M P(M-M_k)\psi(M)$
$X_k=\ln(\theta(M_k))$
Band-pass Filtered Series
IDFT
$C(d)=\sum_{k=1}^{K}X_k\cos(d(k-0.5)\frac{\pi}{K})$
Cepstral Coefficient

Fig. 2 Original MFCC feature extraction

## 3.2 Average MFCCs Feature

Analyzing the MFCCs derived from hundreds of windows is too long ,while the SVM requirements for the input vectors with same length, another normalized MFCCs array is introduced in this way: to partition one music clip into three parts in time domain. Suppose that there are n hamming windows for one music clip. The one partitioning part should have n/3 windows. If there are k MFCCs coefficients in one window, we calculate the mean of corresponding coefficients of each partition so that each partition produces k MFCCs coefficients representing the n/3 windows. The vectors from three partitions are connected to form a new vector, which is used for classification. This vector should contain 3*k coefficients. Fig. 3 in the diagram shows how to extract the average MFCCs feature vectors.

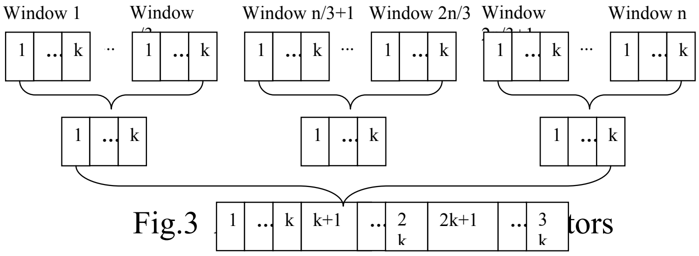

Fig.3 [illegible]tors

# 4 Audio Classification

The classification rules [1][2] are needed for the classification of the audio data. The most commonly-used algorithms are Nearest Neighbor (NN) rule, K-NN rule, HMM rule, Neural Network classifier, NFL rule and SVM rule. In this paper two distance measurements Euclidean distance and Dynamic Time Warp (DTW) distance were implemented. The Euclidean distance is used for the vectors with the same length while the DTW algorithm is used to solve the series with different lengths or different phases.

## 4.1 Euclidean Distance

The Euclidean distance is an effective way to measure the similarity between two arrays with the same dimension and phase [5]. For any dimension of i, it is defined as:

$$D=\sqrt{\sum_{i=1}^{i=n}(x_i-y_i)^2}$$

In the audio classification, we use Euclidean distance to measure the distance between two MFCCs feature vectors from two different audio files. Because each hamming window generates a feature vector with the same length, the Euclidean distance can compute the difference between two windows effectively. The Euclidean distance has been implemented in two classification methods. For the original MFCCs feature, the Euclidean distance is used to calculate the distance between feature vectors of two windows, which is a component in DTW computation. For the Average MFCCs feature, the Euclidean distance is to calculate the distance between the average feature vectors of two different audio files, and the result is directly used in NN and K-NN rules.

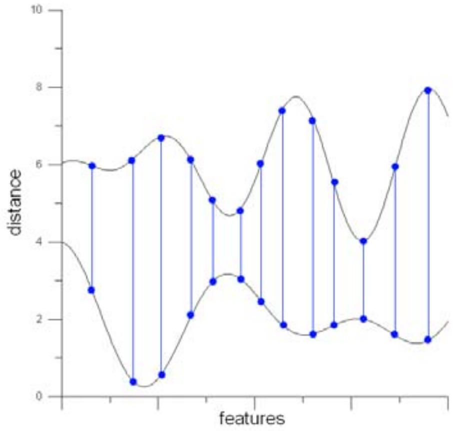

Fig.4 Euclidean distance between two vectors

## 4.2 Nearest Neighbor (NN)

The Nearest Neighbor (NN)[10] rule is used to classify the observation into the category that only depends on a collection of correctly classified samples. It can be defined in the following:

$Xk \in (X_1, X_2, \ldots X_n)$

If min $d(X_i, X) = d(X_k, X)$ (where $i=1,2\ldots n$)

The NN rule chooses to classify the X to the category $\theta_n$, where the $X_k$ is the nearest neighbor of X and $X_k$ belongs to $\theta_n$.

## 4.3 Dynamic Time Warping (DTW)

When the feature vectors are generated, one problem arises: the lengths of the input and the stored sequences are unlikely to be the same. The Dynamic Time Warping (DTW) [6] algorithm can be used to solve this problem.

DTW algorithm is defined as follows:

```
(1) Let $g(1,1) = 2d(1,1), j = 1$.
(2) Let $i_1 = \max(1, j-r), i_2 = \min(j + r, I)$,
    Compute $g(i, j), (i = i_1,\ldots, i_2)$.
(3) $j < J$?
      YES→(2)
      NO→Compute $D(T, R) = g(I, J) / (I+J)$
```

The generalized algorithm is as follows:

$$g(ck) = \min \begin{bmatrix} g(i-1, j) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i, j-1) + d(i, j) \end{bmatrix}$$

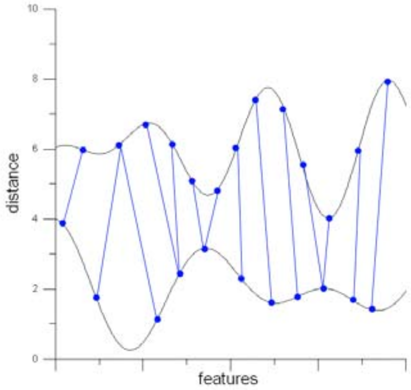

Fig.5 DTW distance between two vectors

In this paper the DTW algorithm was used to measure the distance between the MFCCs from different sound clips.

## 4.4 SVM Classifier

The SVM theory is based on binary classification. The core concept of SVM algorithm is to derive the hyperplane to separate the two classes. The one that maximizes the margin of hyperplanes between two classes is called “optimal hyperplane”. To solve the multi-dimension problem, the Kernel function serves to map the input vectors into a high-dimensional feature space in which the mapped data is linearly separable [6].The SVM method provides a way to minimize the structural risk. The multi-class classification also can be achieved by SVM by combining the binary classifications.

## 4.5 The Proposed Weight-KNN Classification

In this paper we propose a new hybrid music classification algorithm based on the SVM and Euclidean distance. In the Euclidean distance, each dimension is calculated equally, while for some classification criteria, some dimensions may not be as important as other dimensions. The purpose of this method is to extract the weight factor from the training SVM model. Due to the outstanding property of minimized structural risks in the Supporting Vector Classifier, the model can imply the bias for each dimension under the classification criteria. To implement the weight factor in Euclidean distance calculation, the Euclidean distance will be more sensible to the classification criteria. It can become more flexible to solve different classification by offering different results. In order to analyze the problem in the same feature space, we choose to use linear decision function to derive the weight [8].

The explanation of the Weight-KNN is as following:

The weight vector is

$w^* = \sum_{i=1}^{l} y_i \alpha_i^* x_i = (w^*{}_1, \ldots, w^*{}_n)$ in the decision function $f(x) = \operatorname{sgn}(w \cdot x_i + b)$ of the SVM [8] .It can indicate the correlation level of each dimension regarding to the specific classification criteria.

The improved Euclidean distance can be derived as

$$d_{weight} = \sqrt{\sum_{l=1}^{n} (w^l)^2 (x_i{}^l - x_j{}^l)^2} \quad .$$

The effect of Weight-KNN can be proved as following:

In the decision function,

$g(x) = w \bullet x_i + b$ , $w = (w^1, w^2, \ldots, w^n)$, $n$ is the dimension of feature vector $x$ .To unfold the decision function, we derive

$f(x) = \operatorname{sgn}\{(\ldots + w^l x_i^l + \ldots + w^n x_i^n) + b\}$.For any $i$ , $x_i$ is constant which refers to the input vector.We suppose that for $w = (w^1, w^2, \ldots, w^n)$, $w^n$ is constant when $n \neq k$ .We change $|w^k|$ to observe the effects.

Considering the sign of $x_i^k$ and $w^k$ , the $w^k x_i^k = |w^k||x_i^k|$ or $w^k x_i^k = -|w^k||x_i^k|$.Suppose constant

$$\varphi_0 = w^1 x_i^1 + \ldots + w^{k-1} x_i^{k-1} + w^{k+1} x_i^{k+1} + \ldots + w^n x_i^n + b$$

The original decision function is converted to $f(x) = \operatorname{sgn}\{w^k x_i^k + \varphi_0\}$.

It is clear that when $x_i^k$ is unchanged, the larger the $|w^k|$ is, the more the deviation is from $\varphi_0$ , which indicates the influence on the final objective value.

Likewise, for any given $k \leq n$ , the larger the $w^k$ is, the more it will enlarge the influence of $x_i$ .

This classification method is used in our tests to classify the two classes that are difficult to discriminate each other with the conventional Euclidean-based KNN method. By applying the SVM-based weight, the Euclidean distance can be more sensitive to the classification boundary, which, in the experiments, is the boundary of different instruments.

The function of the weight factor as applied to the Euclidean distance. The proposed new algorithm takes the advantages of Euclidean distance, NN rule, and the SVM. As the SVM is outstanding for its minimized structural risks, the hyperplane can indicate the distribution of each dimension. For Euclidean distance and NN rule, they are methods based on enumeration which serve well in retrieval. By combining these two algorithms together, the enumeration method can be aware of the boundary and criteria of the practical classification. Therefore it can achieve better results in both classification and retrieval.

## 5 Testing Results

We have tested and validated the positive effect brought by SVM-based weight factor. In this test, the saxophone and guitar are mis-classified in the classification of some sets. These two classes of audio data are taken out from the sample set and perform the test. We use the Average MFCCs (mel-scaled frequency cepstral coefficients feature) feature vectors, and Euclidean distance with K-NN is carried out to do binary classification. The effects of adding the SVM-based weight will be observed.

In this test, 28 saxophone clips and 28 guitar clips are used. The audio data files are divided into 4 parts, with each part of 7 clips for guitar and 7 clips for saxophone. For each time, one part is taken as the testing set while the others are used to train the model. To validate the weight effects for the general guitar and saxophone discrimination, we apply the same weight in the four sets of tests, which is from the training model of set 1. The test has chosen MFCC12 as the basis to derive the Average MFCCs feature. The hamming window shifts every 256 samples, and there are 512 samples in each hamming window. The SVM model is trained by a linear classifier, with C=2. Fig.9 shows how we test the effect of SVM weight factor added to the Euclidean distance measurement.

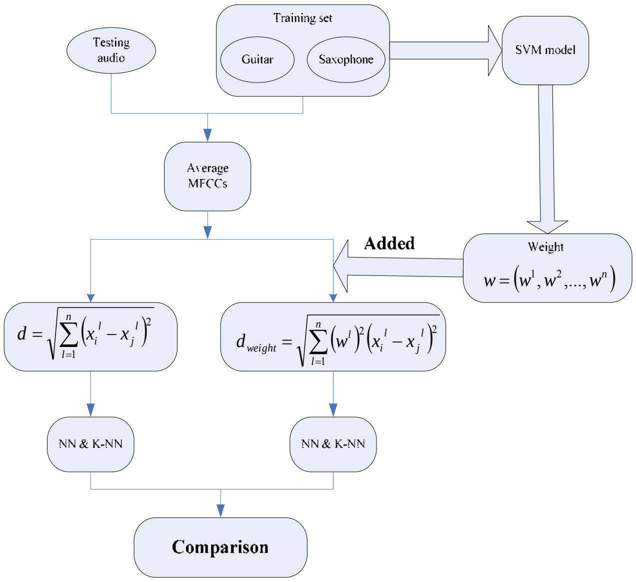

Fig.9 Testing on the effect of weight factor

According to the testing results, we find that that the Weight-KNN can improve the classification result in discriminating the Guitar music from the Saxophone music. In each set of testing, those classes with bad performance when no weight added have shown the increased accuracy in the Weight-KNN testing. The trends are working for all sets of testing, by increasing the accuracy 28% in Saxophone recognition. The classification performance is improving when we apply the SVM Weight factor in Euclidean distance measurement.

## 6 Conclusion

In this paper a new hybrid audio classification algorithm has been proposed. Due to the poor classification accuracy of the traditional Euclidean-based algorithm, we have proposed a new algorithm by extracting the weight factor from Supporting Vector Model (SVM) and applying it to the Euclidean distance measurement. From the test results shown in section 5, it has shown that the proposed weight scheme in Euclidean distance is successful to improve the recognition accuracy by 28% in an individual experiment and 7% for the overall. Due to the minimized structural risks for Supporting Vector Classification, the weight can indicate the relevance for each dimension regarding a certain classification criteria. By using the weight factor of Supporting Vector, the Euclidean-based distance measurement becomes more sensitive to the classification boundary. The superior results in audio data classification have presented in this paper. It can give a more promising direction to the use of the SVM-based weight in the Euclidean distance and K-NN rule.

## *References:*

[1] Wold, E., Blum, T., Keislar, D., et al. Content-Based classification, search and retrieval of audio. *IEEE Multimedia Magazine*, 1996,3(3):27~36.

[2] Liu, Zhu, Wang, Y., Chen, T. Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, 1998,20(1/2):61~79.

[3] J. Foote, “Content-based retrieval of music and audio”,in Multimedia Storage and Archiving Systems II, Proc.of SPIE, C. C. J. Kuo et al., Eds., 1997, vol. 3229, pp.138–147.

[4] Philippe Simard,Marcel Mitran.The Nearest Neighbor Rule: A Short Tutorial http://cgm.cs.mcgill.ca/~soss/cs644/projects/simard/

[5] Keogh, E. and Ratanamahatana, C.A. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7. 358~386.

[6] Li, S.Z.a.G., G. Content-based Audio Classification and Retrieval using SVM Learning. *Microsoft Research China*, 2000.

[7] Chang, C.-C. and Lin, C.-J., {LIBSVM}: a library for support vector machines.2001 Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm

[8] zhou, C.Z., Lei, L. and Zheng-an, Y. Feature-Weighted K-Nearest Neighbor Algorithm with SVM. *ACTA SCIENTIARUM NATURAIdUM UNIVERW ATIS SUNYATSENI*, 44 (1).