Journal of Management Science and Statistical Decision
Vol.7 No1.(2010) p88-92

Musical Instrument Audio Identification Based on Kernel Logistic Regression

Zunxiong Liu
School of Information Engineering
East China Jiaotong University
Nanchang, China
e-mail:liuzunx@tom.com

Jinfeng Xu
School of Information Engineering
East China Jiaotong University
Nanchang, China
e-mail:xujinfeng.611@163.com

***Abstract*—Audio classification based on statistical learning has attracted widespread attention and been widely put into some commercial application, because of better theoretical foundation and simple implementation mechanisms. Based on exploration the theory of the classical logistic regression (LR) and kernel logistic regression (KLR), a novel approach for audio classifier is put forward with the help of KLR in this paper. It is used to handle music from the same type of musical instruments. Music signals are collected with violin, viola and cello, and all the signals are preprocessed to extract features. The processed samples are used in experiments, while the classification performances are compared with 3 different kernel functions. Simulation results show that KLR performs better than traditional LR on classification accuracy and has better non-linear processing ability. Furthermore, KLR model with RBF kernel function can have a better stability besides good prediction performances.**

Keywords- kernel logistic regression; audio classification; feature extraction

I. INTRODUCTION

Logistic Regression (LR) is a traditional statistical analysis method, being used to predict the probability of event occurrence by fitting data to a logistic curve. It is a generalized linear model used for binomial regression. Like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical. LR is a nonlinear model, where maximum likelihood method is usually employed to estimate model parameters. It can be proved that the maximum likelihood estimation in LR model has the property of consistency and Asymptotic Normality on the condition of random samples [1]. LR focuses on modeling a posteriori probability of the membership to each of the C classes, and need less presumption about the samples. At present, the method has been widely used in the economic, social sciences, medical and many other scientific fields.

Samples are assigned to the category directly based on their posterior probabilities with LR, however the probability estimation accuracy is limited because of the linear model used there, some scholars have applied the kernel trick used with Support Vector Machine (SVM) to extend the LR, obtaining the nonlinear counterpart, kernel logistic regression (KLR) in the Reproducing Kernel Hilbert Space (RKHS) [2-3]. KLR model overcomes the problem of estimating class conditional densities and has a clear probabilistic interpretation that allows us to quantify a confidence level for class assignments.

Automatic audio classification is one of the most important approaches to cope with audio structure and extract audio content semantic, and being also the research hotspot of content-based audio retrieval. Presently, more emphases are placed on audio feature extraction and classifier construction in this area, while there are a lot of researches associated with it. In order to improve classification accuracy, researchers have proposed different classification methods, such as the nearest neighbor (NN) rule, Support Vector Machine (SVM), Gaussian Mixture Models (GMM), Hidden Markov Model (HMM),and so on[4-8]. These approaches are usually used to discriminate the audio signals which are different evidently, for example, speech, music and environmental audio signals are samples to be handled, it’s relatively simple to classify them. With the similar audio signals, such as the music audio from the same type of musical instruments, different approaches need to be implemented.

In this paper, the multi-class KLR model is proposed to identify the same type of musical instruments [9], violin, viola and cello with audio signals from them. Meanwhile, three different kernel functions are experimented with the self-processed signals, including linear kernel, polynomial kernel and radial basis kernel. This paper is organized as follows. Features chosen with audio signal are explained in section 2, which will be used in our experiments. In section 3, logistic regression and its non-linear version with kernel function, kernel logistic regression (KLR), are explored in details. KLR is introduced to cope with multi-class classification problem in section 4. Following that, Simulation experiments with the audio signals from violin, viola and cello are carried out, and results are presented in section 5, the related analyses also given. Section 6 concludes the paper and outlines directions of future work.

II. FEATURES CHOSEN WITH AUDIO SIGNAL

As any pattern recognition problem, what features will be chosen is a very important step in audio classification. The features should be selected to fully characterize audio signals in the time domain and frequency domain, so that the samples are easily classified. The features chosen are expected to possess robustness and consistency. In general, audio features can be extracted with two different lengths of time, one is based on audio frame lasting tens of milliseconds, the other is based on the audio clip which is of few seconds. Here the original audio signals are processed into clips of 3 seconds, each clips is split into frames that are 512 sampling points in length, and an overlap of 25% in Hanning window is used to further reduce edge transients. The frequency content of each frame is determined using a 512-point Fast Fourier Transform (FFT) and the following quantities were extracted from each frame, after each frame was normalized to unit energy. Then the features are calculated on the level of the audio clip to gain the data sets of a 3s audio sample according to the frame-level features. The features chosen from audio frame are following:

A. *Frequency Centroid*

The frequency centroid is defined by the relationship:

$$\omega_c = \int_0^\pi \omega \left|F(\omega)\right|^2 d\omega \Big/ \int_0^\pi \left|F(\omega)\right|^2 d\omega$$

Thus the centroid gives the centroidal frequency contained in a frame.

B. *Bandwidth*

The bandwidth is defined by the relationship:

$$B = \sqrt{\int_0^\pi (\omega - \omega_c)^2 \left|F(\omega)\right|^2 d\omega \Big/ \int_0^\pi \left|F(\omega)\right|^2 d\omega}$$

This is a measure of the variation in frequency in the frame.

C. *Critical sub-band power ratios*

The critical sub-band power ratios are the ratios of the log power in each critical sub-band to the log-power in the entire frame. The critical sub-bands are bands of frequencies determined by subjective experiments across which there are abrupt changes in subjective response. Another crude explanation is that for each critical band the human ear has approximately the same sensitivity.

After these quantities were determined for each frame, their means and standard deviations were calculated across the entire audio sample. This thus became a crude form of temporal analysis of the frequency information. For each three second sample, the averages of the centroid, bandwidth and power ratios provided frequency information across the entire sample while the standard deviations of the same quantities provided a crude measure of the temporal information. In this way, a three second sample can be represented by a feature vector.

III. KERNEL LOGISTIC REGRESSION

Considering a binary classification problem with labels {0, 1}, the success probability of the sample X belonging to class 1 is given by $p(y=1 \mid x)$ and $p(y=0 \mid x) = 1 - p(y=1 \mid x)$ is the probability that it belongs to class 0.

In Logistic Regression (LR) the posterior probability of the class membership is modeled via the linear function $f(x) = \beta^T X$, Where β denotes the weight vector, including a bias β_0 where the sample X is augmented by a constant entry of 1. Interpreting the output of $f(x)$ as an estimate of a probability $P(X,\beta)$, we have to rearrange equation by the logit transfer function

$$\log it\{P(X,\beta)\} = \log\frac{P(X,\beta)}{1-P(X,\beta)} = \beta^T X \tag{1}$$

Then the probability is obtained as

$$P(X,\beta) = \frac{1}{1+\exp(-f(x))} \tag{2}$$

Assumed that the training data is drawn from a Bernoulli distribution conditioned on the samples X, the conditioned probability of $P(y \mid X,\beta)$ is

$$P(y \mid X,\beta) = P(X,\beta)^y (1-P(X,\beta))^{1-y} \tag{3}$$

The negative log-likelihood (NLL) of equation (3) can be written as

$$l\{\beta\} = \sum_{i=1}^{N} -y_i \beta^T X_i + \log(1+\exp(\beta^T X_i)) \tag{4}$$

To avoid over-fitting to the training data it is necessary to impose a penalty on large fluctuations of the estimated parameters β [10]

$$l(\beta) = l(\beta) + \frac{\lambda}{2}\left\|\beta\right\|_2^2 \tag{5}$$

While λ is the regularization parameter. To minimize the regularized NLL we set the derivatives $\frac{\partial l(\beta)_{ridge}}{\partial \beta}$ to zero and use the Newton-Rephson algorithm to iteratively solve equation (5).This algorithm is also referred to as iteratively re-weighted least square (IRLS) in this case,

$$\beta^{new} = (X^T W X + \lambda I)^{-1} X^T W z \tag{6}$$

$$z = X\beta^{old} + W^{-1}(y-P) \tag{7}$$

Where P is the vector of fitted probabilities with the i'th element $P(\beta^{old}, X_i)$, W is the $N \times N$ weight matrix with the entries $P(\beta^{old}, X_i)(1-P(\beta^{old}, X_i))$ on the diagonal, and I is the identity matrix.

K (x, y) is an arbitrary kernel function which satisfies the Mercer condition, where the nonlinear mapping Φ: $X \to \Phi(X)$ works in the Reproducing Kernel Hilbert Space. Every β lies in the span of all $\Phi(X_i)$:

$$\beta = \sum_{i=1}^{N} \alpha_i \Phi(X_i) \tag{8}$$

Introducing the kernel matrix K with $K_{ij} = \left(\Phi(X_i)\Phi(X_j)\right) = K(X_i X_j)$, we can write equation (2) as

$$P(y \mid X, \alpha) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{N} \alpha_i K(X_i, X)\right)} \tag{9}$$

then,

$$\alpha^{new} = (K + \lambda W^{-1})^{-1} K W \tilde{z} \tag{10}$$

$$\tilde{z} = (K\alpha^{old} + W^{-1}(y - P)) \tag{11}$$

In this paper, the used three kinds of kernel functions in the KLR model are respectively:

Linear kernel function : $k(x,y) = xy$

Polynomial kernel function : $k(x,y) = (x \cdot y + 1)^d$

Gaussian radial basis kernel function:
$k(x,y) = \exp(\frac{-\|x - y\|}{2\sigma^2})$

IV. Multi-class problems

The multi-class classification problem refers to assigning each of the observations into one of C classes. Kernel logistic regression could be extended to multiclass problems. The common one-versus-one approach is employed, where a classifier learns to discriminate one class from one other class. This leads to $C(C-1)/2$ pairwise classification rules [11].

Considering the set of events $\{y_i\}_{i=1}^{c}$, let the probabilities $r_{ij} = \mathrm{P}(y_i \mid y_i \text{ or } y_j)$ of a class y_i given a sample vector X belong to either y_i or y_j. From the i'th and j'th classes of a training set, the KLR model is constructed. For any new X, r_{ij} can be calculated. The goal is to couple the r_{ij} into a set of probabilities $p_i = (y_i \mid X)$. A new set of auxiliary variables $\mu_{ij} = \frac{p_i}{p_i + p_j}$ which are in some sense "close" to the observed r_{ij} , assume $r_{ij} + r_{ji} = 1$. A suitable closeness measure is the Kullback-Leibler divergence between r_{ij} and μ_{ij} :

$$\begin{aligned} l(p) &= \sum_{i \neq j} n_{ij} r_{ij} \log \frac{r_{ij}}{\mu_{ij}} \\ &= \sum_{i<j} n_{ij} (r_{ij} \log \frac{r_{ij}}{\mu_{ij}} + (1 - r_{ij}) \log \frac{1 - r_{ij}}{1 - \mu_{ij}}) \end{aligned} \tag{15}$$

n_{ij} is the number of training data in the i'th and j'th classes. To minimize (15), letting $\partial l(p) / \partial p_i = 0, i = 1, 2..c$, we should find a point that satisfies

$$\sum_{j:j \neq i} n_{ij} \mu_{ij} = \sum_{j:j \neq i} n_{ij} r_{ij},$$

$\sum_{i=1}^{k} p_i = 1$,and $p_i > 0$, $i = 1, \ldots C$.

Such a point is obtained by the following algorithm:

- Step1.Starting with an initial guess for the p_i and corresponding μ_{ij}
- Step2.Repeat (i=1,…C,1,…)

  $\alpha = (\sum_{j:j \neq i} n_{ij} \mu_{ij}) / (\sum_{j:j \neq i} n_{ij} r_{ij})$

  $\mu_{ij} \leftarrow \alpha \mu_{ij} / (\alpha \mu_{ij} + \mu_{ji}), \mu_{ji} \leftarrow 1 - \mu_{ij}, \text{for all } j \neq i$

  $p_i \leftarrow \alpha p_i$

  Normalize P, until α close to ones.
- Step3. $p \leftarrow p / \sum_{i=1}^{k} p_i$

We finally obtain the posterior probabilities for class membership of sample X.

V. Simulation Experiments

A. Data Set Description and Setting

In the following experiments, a comprehensive database of real instrument recordings is used, available for research purpose at [12]. Music audio signals from 3 instruments, violin, viola and cello in the data set are chosen, all of them recorded at 44.1kHz and 16 bit/sample, the formats are all AIFF, which transformed into WAV format. Then the original audio samples are also split into 3s clip. Each three-second audio sample is split into frames that are 512 sampling points in length. An overlap of 25% was used and a Hanning window was used to further reduce edge transients. 300 samples are collected, and there are 100 audio clips for each musical instrument. The following figures present the related important features used.

Fig.1 shows the waveforms and spectrums with 3 different original audio signals, the first column is the waveforms with the three instrument audio signals respectively, the second column presents the spectrums of the corresponding signals. Here the spectrums are limited in the frequency scope of 0~400Hz, the spectrums are drawn again, given in last column. It can be seen that their pitch frequencies are different from each other with 3 musical instrument signals, the pitch frequencies with violin, viola and cello are 200~400Hz, 150~300Hz and 100~250Hz respectively.

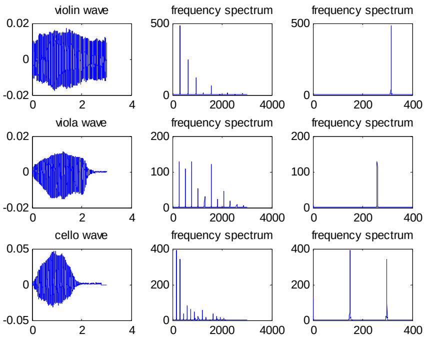

Figure 1. Waveforms and spectrums with 3 different original audio signals

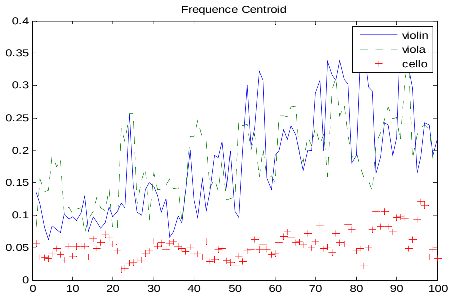

Figure 2. Frequency mean values for each sample of 3 musical instruments

The frequency mean values for each sample of 3 musical instruments are given in Fig.2, where the horizontal axis refers to 100 samples, the vertical coordinates denotes the frequency mean value. The difference of mean and standard deviation value between three musical instruments is obvious, among them the frequency mean values with violin audio are generally bigger because the violin plays more strident sound. So frequency mean can be taken as a feature for classification problem.

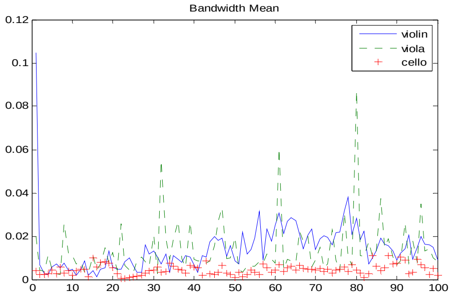

Figure 3. Bandwidth mean values for each sample of 3 musical instruments

Fig.3 throws out the bandwidth mean values for each sample of 3 musical instruments. Twenty-two critical bands were used in this paper, yielding 22 ratios per frame of 512 sampling points. After these quantities were determined for each frame, their means and standard deviations are calculated across the entire audio sample. All these become a crude form of temporal analysis of the frequency information. For each three second sample, the averages of the frequency centroid, bandwidth and power ratios provide frequency information across the entire sample while the standard deviations of the same quantities provided a crude measure of the temporal information. Thus in all, a three second sample was represented by a 48-dimensional feature vector (22 frequency means, 22 standard deviations, mean and standard deviations of centroid and bandwidth).

B. *Result*

The accuracy and the stability are evaluated with the proposed classifier in the musical instrument audio data set. In 100 samples for each category, 60 samples are randomly selected to training the classifier model, the left 40 samples make testing samples. The training dataset with 180 samples and the testing dataset with 120 samples and the testing are obtained. The datasets are experimented with the KLR classifier for six times. Here 3 different kernel functions are employed, and LR approach is also carried out as a comparison. The error rate mean and the error rate standard deviation for six experiments are shown in table 1.The two quantities can describe the performance and stability of each algorithm.

TABLE I. AVERAGE ERROR RATES AND ITS STANDARD DEVIATIONS FOR SIX EXPERIMENTS IN 3 MUSICAL INSTRUMENT AUDIO SIGNALS

	ER(mean)	ER(standard deviation)
LR	0.272	0.0110
KLR(Kernel-liner)	0.0101	0.0082
KLR(Kernel-poly)	0.1158	0.2699
KLR(Kernel-RBF)	0.0130	0.0038

In table 1, the results tells that introducing the kernel trick into the LR algorithm, the original features are mapped into RHKS, producing KLR model, the classification performance of KLR model is superior to LR model. What's more, KLR with linear kernel function outperforms all other Algorithms, and KLR models with the RBF kernel function get the best stability through the standard deviation.

TABLE II. THE CLASSIFICATION ACCURACIES IN 3 MUSICAL INSTRUMENT AUDIO SIGNALS

	violin	viola	cello
LR	100%	91.84%	100%
KLR	100%	96.11%	100%

The classification accuracies with LR and KLR 3 musical instrument audio signals, are calculated, and given in Table 2.

It can be seen that the discrimination on viola audio signal is pretty good, not successful as on the two others, while the violin and cello audio signals are discriminated perfectly. According to the experiment results, It is demonstrated that the features with the 3 musical instrument audio signals are effective and the KLR model is viable and robust to solve multi-class problems.

VI. CONCLUSIONS

In this paper, the features with audio signal and LR theory are explored, and KLR model is introduced based on LR and kernel trick. The KLR model is proposed to classify the audio signals from the same type of musical instruments, violin, viola and cello. Concerning multi-class problems, we use a pairwise coupling procedure. The audio signals are first preprocessed and their features are calculated, then are used in the experiments with KLR and LR models. Experimental results effectively demonstrate that the KLR performs better than traditional LR on classification problem. Furthermore, KLR model with RBF kernel function attains a level of accuracy comparable to the linear kernel function, while additionally providing a better stability.

As kernel methods are popular, and their sparse problems attract more research interests, and thus the system computation performance can be improved. Motivated with the related technologies, the sparse KLR model will become our future research subject.

REFERENCES

[1] Hastie T, Tibshirani R, Friedman J ,The Elements of Statistical Learning: Data Mining, Inference and Prediction, Berlin, Germany: Springer, 2001.

[2] Jaakkola T S, Haussler D, “Probabilistic kernel regression models,” Proceedings of the Conference on AI and Statistics,San Francisco, USA,Morgan Kaufmann,1999, pp.99-108.

[3] Roth V, “Probabilistic discriminative kernel classifiers for multi-class problems,” Lecture Notes in Computer Science, London, UK, Springer-Verlag.2001,pp.246-253.

[4] Wold E, Blum T, Keislar D, et al, “Content-Based classification, search and retrieval of audio,” IEEE Multimedia Magazine,1996, pp.27-36.

[5] J.Lu, Y.Chen, Z.Sun, “Automatic audio classification by using hidden Markov Model,” Journal of Software.2002, vol.13 (8), pp.1593-1597.

[6] Mubarak O M, Ambikairajah E, Epps J, “Novel features for effective speech and music discrimination,” Proc of the IEEE Int’l Conf on Engineering of Intelligent Systems, 2006, pp.22-23.

[7] L.Bai, S.Lao, and J.Chen, “Audio classification and segmentation based on support machines,” Computer science.2005, vol.32(4), pp.87-91.

[8] Janet Marques, Pedro J Moreno, “A study of musical instrument classification using gaussian mixture models and support vector machines,” COMPAQ: Cambridge Research Laboratory, 1999.

[9] Jean Julien Aucouturier, Francois Pachet, Mark Sandler, “The way it sounds: timbre models for analysis and retrieval of music signals ,”IEEE Transactions on Multimedia, 2005, vol.7(6), pp.1-8.

[10] Ian T. Nabney, “Efficient training of rbf network for classification,” Artificial Neural Networks-ICANN 1999.1999,vol 1,pp.210-215

[11] Trevor Hastie, Robert Tibshirani, “Classification by pairwise coupling,” Advances in Neural Information Processing System 10, Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, Eds. MIT Press, Cambridge, MA, USA,Jun 1998

[12] L. Fritts,”Musical instrument samples,” on http://theremin.music.uiowa.edu/, The University of Iowa.