2011 IEEE International Workshop on Machine Learning for Signal Processing
September 18-21, 2011, Beijing, China

HETEROGENEOUS MIXTURE MODELS USING SPARSE REPRESENTATION FEATURES FOR APPLAUSE AND LAUGH DETECTION

Ziqiang Shi, Jiqing Han, Tieran Zheng

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
{zqshi, jqhan, zhengtieran}@hit.edu.cn

Abstract

A novel and robust approach for applause and laugh detection is proposed based on sparse representation features and heterogeneous mixture models (hetMM). The projections of the noise robust sparse representations for audio signals computed by L_1-minimization are used as feature. We consider the classifiers based on heterogeneous mixture models (hetMM) which combine multiple different kinds of distributions, since in practice the data may come from multiple sources and it is often unclear what the most suitable distribution is. Experimental results show that method with hetMM has better results than using a single distribution type and gives comparable performances with Support Vector Machines (SVMs).

Index Terms— heterogeneous mixture models, multivariate logistic distribution, sparse representation features (SRF), EM algorithm, audio event detection

1. INTRODUCTION

Audio content analysis plays an important role in video content parsing. Highlight sound effects, such as laugh and applause are usually semantically related with highlight events in general video, such as sports, entertainments, meeting, and home videos.

Most of the audio event detection algorithms resort to the two steps approach, which involves extracting discriminatory features from audio data and feeding them to pattern classifier. Feature commonly exploited for audio event detection can be roughly classified into time domain features, transformation domain features, time-transformation domain features or their combinations [1].

Recently in the statistical signal processing community, the interest on sparse representations of signals has revived [2]. The related research has been focused on solving the optimization problem. However, these work aim at representation and compression rather than inference or classification of signals. In neuroscience, the neural representation of sounds in the auditory cortex of unanesthetized animals is sparse, since the fraction of neurons that are active at a give instant is typically small [3]. It is worth mentioning that the sparse representation is naturally discriminative and robust to signal variations. Indeed, among all subsets of basis vectors, it selects the subset, which most compactly expresses the input signal and rejects all other possible but less compact representations [4]. Thus L_1-norm regularization based linear system which tends to produce sparse solutions is a more suitable framework for acoustic feature extraction [3]. In this paper, the projections of the sparse coefficients for audio signal recovered from the L_1-minimization with respect to the overcomplete dictionary of signal atoms which are learned from the training datasets are used as features.

Mixture models have provided a mathematical-based approach to the statistical modeling of training data generated from some underlying source. This statistical approach has been used successfully in a number of applications such as Gaussian Mixture models in speaker recognition [5], Beta Mixture Models in image classification [6], and mixture of Student's t-distributions in robust image segmentation [7]. These conventional mixture models are based on a single kind of distribution, but in practice the data may come from multiple sources, and it is often unclear what the most suitable distribution type for the task at hand is, it needs to combine several different possible kinds of distributions.

Various types of distributions have quite different properties and suitable application scope. For example, the probability density function (pdf) of Student t-distribution has heavier tails as compared to the exponentially decaying tails of a Gaussian and thus is not sensitive to outliers. Hence the robust model for image segmentation based on mixtures of multivariate Student's t-distribution (MSD) can account for outliers' values and thus provides smoother segmentations than the standard GMM [7]. Another example is logistic distribution which has been thoroughly studied [8] and widely used in biology [9], market [10], regression [11], and machine learning [12]. It has similar shape with the normal distribution but has heavier tails and high kurtosis. Malik and Abraham [13] suggested two families of multivariate logistic distribution (MLD) which are multivariate analogue of the single-dimensional logistic distribution.

In this paper, we propose a heterogeneous mixture models (hetMM) based on multiple different kinds of distributions, which is similar to the relation between multiple kernel learning [14] and support vector machine (SVM). In these experiments, we compare the

performance of the hetMM with mixture models based on single kind distributions.

2. SPARSE REPRESENTATION FEATURES

Consider a finite training set of signals $\mathrm{A}=[\mathrm{a}_1,...,\mathrm{a}_n]\in R^{m\times n}$ which contains all audio frames. The dictionary learning problem can be defined as the optimization of the empirical cost function

$$f_n(\mathrm{D})=\frac{1}{n}\sum_{i=1}^{n} l(\mathrm{a}_i,\mathrm{D}), \tag{1}$$

where $\mathrm{D}\in R^{m\times k}$ is the dictionary, each column represents a basis vector, and $l(\mathrm{a},\mathrm{D})$ is a loss function that measures the ability of D at representing the signal a. In sparse representation, $l(\mathrm{a},\mathrm{D})$ is always defined as the optimal value of the sparse coding problem

$$l(\mathrm{a},\mathrm{D})=\min_{\mathrm{x}\in R^k}\frac{1}{2}\left\|\mathrm{a}-\mathrm{D}\mathrm{x}\right\|_2^2+\lambda\left\|\mathrm{x}\right\|_0, \tag{2}$$

where λ is a regularization parameter and $\|\cdot\|_0$ is the L_0 quasi-norm returning the number of the nonzero entries of a vector. Finding the solution to optimization problem defined in (2) is NP-hard due to the nature of the underlying combinational optimization. An approximate solution to the Eq. (2) can be obtained by replacing the L_0 norm with the L_1 norm as follows:

$$l(\mathrm{a},\mathrm{D})=\min_{\mathrm{x}\in R^k}\frac{1}{2}\left\|\mathrm{a}-\mathrm{D}\mathrm{x}\right\|_2^2+\lambda\left\|\mathrm{x}\right\|_1 \tag{3}$$

where $\|\cdot\|_1$ denotes the L_1 norm of a vector. In order to prevent the energy of the words in the dictionary to become arbitrarily large, it is always assumed that D is in the following set:

$$C=\{\mathrm{D}\in R^{m\times k}\ \text{s.t.}\ \forall j=1,...,k,\ \mathrm{d}_j^{\mathrm{T}}\mathrm{d}_j\le 1\}\ . \tag{4}$$

Then the problem (1) becomes

$$\min_{\mathrm{D}\in C,\mathrm{x}_i\in R^k}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\left\|\mathrm{a}_i-\mathrm{D}\mathrm{x}_i\right\|_2^2+\lambda\left\|\mathrm{x}_i\right\|_1\right). \tag{5}$$

Marial et al. [15] suggested an online learning algorithm based on stochastic approximations to solve (5).

After the dictionary D_T of the training set is learned, the sparse representation of a test signal can be obtained by (3). In order to build mixture models on these sparse coefficients, a random projection matrix W obtained by principal component analysis (PCA) is introduced to project the coefficient vector to a lower dimensional space.

3. HETEROGENEOUS MIXTURE MODELS AND EM ALGORITHM

Consider a set of p random variables $X_1,...,X_p$, the joint distribution function is

$$F(x_1,x_2,...,x_p)=P(X_1\le x_1,X_2\le x_2,...,X_p\le x_p). \tag{6}$$

The joint density function for this distribution is a positive real valued function p such that

$$F(x_1,x_2,...,x_p)=\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}...\int_{-\infty}^{x_p}p(u_1,u_2,...,u_p)du_1du_2...du_p \tag{7}$$

for each $x_1,x_2,...,x_p\in R$.

3.1. Heterogeneous Mixture Models

Conventional mixture models are based on a single kind of distribution, but in practice the data may be from multiple, heterogeneous sources. One can approach this problem by considering combinations of multiple different kinds of distributions, i.e.

$$f(x)=\sum_{i=1}^{N}\sum_{j=1}^{K_i}\pi_{i,j}p_i(x;\theta_i^j) \tag{8}$$

with $\pi_{i,j}\ge 0$ and $\sum_{i=1}^{N}\sum_{j=1}^{K_i}\pi_{i,j}=1$, where $p_i(x;\theta_i^j)$ for $(i=1,...,N)$ are N different kinds of distributions.

3.2. Parameter Estimation

The maximum likelihood estimation (MLE) approach is employed to find the parameters of the hetMM. In order to find the optimal value θ_i^j by the MLE approach, we introduce a set of binary indicator variables $q_{i,j}\in\{0,1\}$, where $q_{i,j}=1$ indicates that x_i belongs to the component j. Then the log joint likelihood of $X=\{x_i\}_{i=1}^n$ and $Q=\{q_{i,j}\}_{i=1..n,j=1...K}$ is

$$\log f(X,Q\mid\theta)=\sum_{i=1}^{n}\sum_{j=1}^{K}[q_{i,j}\log\pi_j+q_{i,j}\log(p(x_i;\theta_j))]. \tag{9}$$

where $K=\sum_{j=1}^{N}K_j$ is the number of all components. The corresponding auxiliary function is

$$\begin{aligned}A(\theta,\theta^{old})&=E_Q[\log f(X,Q\mid\theta)\mid X,\theta^{old}]\\&=\sum_{i=1}^{n}\sum_{j=1}^{K}E_Q[q_{i,j}\mid X,\theta^{old}][\log\pi_j+\log(p(x_i;\theta_j))].\end{aligned} \tag{10}$$

Now the expectation-maximization (EM) algorithm can be used to estimate θ iteratively and formally. The E-step is irrespective of distribution type and it is the same for all the components. The posterior of hidden variables $q_{i,j}$ is calculated as

$$E_Q[q_{i,j}\mid X,\theta^{old}]=p(j\mid x_i,\theta^{old})=\frac{\pi_j^{old}p(x_i\mid\theta_j^{old})}{\sum_{k=1}^{K}\pi_k^{old}p(x_i\mid\theta_k^{old})}. \tag{11}$$

Then the M-step is used to find θ that maximizes $A(\theta,\theta^{old})$. The parameter updating rules are obtained by deriving $A(\theta,\theta^{old})$ with respect to θ, set them to 0, and solve them. Generally, parameters of different type components have different updating rules.

3.3. EM for MSD and MLD

In this work, we combine multivariate normal

distribution, MSD and MLD into a hetMM. Compared to other multivariate distributions, these distributions have definition domains spread over the whole Euclidean space.

A random vector $X=(X_1,...,X_p)^t$ has multivariate Student's t-distribution with $v>0$ degrees of freedom, mean $\mu\in R^p$ and scale parameter Σ, a $p\times p$ symmetric and positive definite matrix, denoted by $S(v,\mu,\Sigma)$, if its density is

$$S(x;v,\mu,\Sigma)=\frac{\Gamma[\frac{v+d}{2}]}{\Gamma[\frac{v}{2}](v\pi)^{\frac{d}{2}}}|\Sigma|^{-\frac{1}{2}}\left[1+\frac{1}{v}(x-\mu)^t\Sigma^{-1}(x-\mu)\right]^{-\frac{v+d}{2}}, \quad (12)$$

where $|\Sigma|$ represents the absolute value of the determinant of the matrix and Γ is the Gamma function. It can be shown that for $v\to\infty$, the Student's t-distribution tends to a normal distribution with covariance Σ. For convenience, we use the diagonal scale parameter Σ in this work.

Malik and Abraham [9] suggest two families of multivariate logistic distributions. The first family is used in this work. Malik and Abraham [9] define the joint distribution function of $X=(X_1,...,X_p)^t$ for MLD as

$$L(x_1,....,x_p)=[1+\sum_{k=1}^{p}\exp(-x_k)]^{-1} \quad (13)$$

and density function as

$$l(x_1,....,x_p)=p!\exp\{-\sum_{k=1}^{p}x_k\}[1+\sum_{k=1}^{p}\exp(-x_k)]^{-(p+1)} \quad (14)$$

where $-\infty<x_k<\infty, k=1,2,...,p$.

In order to build mixture models based on MLD, we introduce center μ and scale Σ into the density function and it turned into

$$l(x_1,....,x_p;\mu,\Sigma)=\left(\prod_{k=1}^{p}\sigma_k\right)^{-1} \quad (15)$$
$$p!\exp\{-\sum_{k=1}^{p}(\frac{x_k-\mu_k}{\sigma_k})\}[1+\sum_{k=1}^{p}\exp(-(\frac{x_k-\mu_k}{\sigma_k}))]^{-(p+1)}$$

where $\mu=(\mu_1,\mu_2,...,\mu_p)$ and $\Sigma=(\Sigma_1,\Sigma_2,...,\Sigma_p)$. Here, we use the fact that

$$\int f(x)=\int\Sigma^{-1}f((x-\mu)/\Sigma)dx. \quad (16)$$

When estimating the parameters of the hetMM built with these distribution types, different type components would have different updating rules. For parameters of normal components, derive $A(\theta,\theta^{old})$ with respect to means $\mu_1^g,\cdots,\mu_{K_\alpha}^g$ and covariances $\Sigma_1^g,\cdots,\Sigma_{K_\alpha}^g$ of Gaussians, and set them to 0, then obtain [5]

$$\mu_j^{g,new}=\sum_{k=1}^{n}x_k g(j\mid x_k,\mu_j^g,\Sigma_j^g)/\sum_{k=1}^{n}g(j\mid x_k,\mu_j^g,\Sigma_j^g), \quad (17)$$

$$\Sigma_j^{g,new}=\frac{\sum_{k=1}^{n}g(j\mid x_k,\mu_j^g,\Sigma_j^g)(x_k-\mu_j^{g,new})(x_k-\mu_j^{g,new})^T}{\sum_{k=1}^{n}g(j\mid x_k,\mu_j^g,\Sigma_j^g)}, \quad (18)$$

$$j=1,\cdots,K_\alpha,$$

where $g(j\mid x_k,\mu_j^g,\Sigma_j^g)$ is the posterior of x_k belonging to the jth normal component and K_α is the number of normal components.

For Students components, derive $A(\theta,\theta^{old})$ with respect to freedom degrees $v_1^s,\cdots,v_{K_\beta}^s$, means $\mu_1^s,\cdots,\mu_{K_\beta}^s$, and scales $\Sigma_1^s,\cdots,\Sigma_{K_\beta}^s$, set them to 0 respectively, then obtain [7]

$$\mu_j^{s,new}=\frac{\sum_{k=1}^{n}x_k s(j\mid x_k,v_j,\mu_j^s,\Sigma_j^s)/(1+v_j^{-1}\delta(x_k;\mu_j^s,\Sigma_j^s))}{\sum_{k=1}^{n}s(j\mid x_k,v_j,\mu_j^s,\Sigma_j^s)/(1+v_j^{-1}\delta(x_k;\mu_j^s,\Sigma_j^s))}, \quad (19)$$

$$\Sigma_j^{s,new}=\frac{\sum_{k=1}^{n}s(j\mid x_k,v_j,\mu_j^s,\Sigma_j^s)(1+v_j^{-1})\frac{(x_k-\mu_j^s)(x_k-\mu_j^s)^T}{(1+v_j^{-1}\delta(x_k;\mu_j^s,\Sigma_j^s))}}{\sum_{k=1}^{n}s(j\mid x_k,v_j,\mu_j^s,\Sigma_j^s)}, \quad (20)$$

$$\sum_{k=1}^{n}s(j\mid x_k,v_j,\mu_j^s,\Sigma_j^s)\{\psi(\frac{v_j+1}{2})-\psi(\frac{v_j}{2})-\frac{p}{v_j} -\log(1+\frac{1}{v_j}\delta(x_k;\mu_j^s,\Sigma_j^s))+\frac{v_j+p}{v_j}\frac{\delta(x_k;\mu_j^s,\Sigma_j^s)}{v_j+\delta(x_k;\mu_j^s,\Sigma_j^s)}\}=0, \quad (21)$$

$$j=1,\cdots,K_\beta$$

where K_β is the number of Students component, $s(j\mid x_k,v_j,\mu_j^s,\Sigma_j^s)$ is the posterior of x_k belonging to the jth Students component, $\delta(x_k;\mu_j^s,\Sigma_j^s)=(x_k-\mu_j^s)^t\Sigma_j^{s-1}(x_k-\mu_j^s)$ is the Mahalanobis squared distance, and $\psi(x)=\frac{\partial\log(\Gamma(x))}{\partial x}$ is the digamma function. The grid search method was used to find the solution to the Eq. (21), and this approximate solution is the new degree of the jth Students component.

For logistic components, derive $A(\theta,\theta^{old})$ with respect to $\mu_1^l,\cdots,\mu_{K_\gamma}^l$, $\Sigma_1^l,\cdots,\Sigma_{K_\gamma}^l$ and set them to 0 respectively, we obtain

$$\sum_{k=1}^{n}l(j\mid x_k,\mu_j^l,\Sigma_j^l)(1-(p+1)\frac{\exp(-\frac{x_{k,i}-\mu_{j,i}^l}{\Sigma_{j,i}^l})}{1+\sum_{i=1}^{p}\exp(-\frac{x_{x,i}-\mu_{j,i}^l}{\Sigma_{j,i}^l})})=0, \quad (22)$$

$$\sum_{k=1}^{n}l(j\mid x_k,\mu_j^l,\Sigma_j^l)(-\Sigma_{j,i}^l +x_{k,i}-\mu_{j,i}^l-(p+1)\frac{\exp(-\frac{x_{k,i}-\mu_{j,i}^l}{\Sigma_{j,i}^l})(x_{k,i}-\mu_{j,i}^l)}{1+\sum_{i=1}^{p}\exp(-\frac{x_{k,i}-\mu_{j,i}^l}{\Sigma_{j,i}^l})})=0, \quad (23)$$

$j = 1,\cdots,K_\gamma; i = 1,\cdots,p$,

where K_γ is the number of logistic components , $l(j \mid x_k, \mu_j^l, \Sigma_j^l)$ is the posterior of x_k belonging to the jth logistic component, $\mu_j^l = (\mu_{j,1}^l, \mu_{j,2}^l, \cdots, \mu_{j,p}^l)$, $\Sigma_j^l = (\Sigma_{j,1}^l, \Sigma_{j,2}^l, \cdots, \Sigma_{j,p}^l)$, $x_k = (x_{k,1}, x_{k,2}, \cdots, x_{k,p})$ and p is the dimension of the feature space. The grid search method was used to find the solution to Eq. (22) and Eq. (23), and these approximate solutions are the new center and scale parameters of the jth logistic component.

For prior probabilities, incorporating the constraints into the system, we obtain

$$J(\theta, \theta^{old}) = A(\theta, \theta^{old}) + (1 - \sum_{k=1}^{K_\alpha + K_\beta + K_\gamma} \pi_k)\lambda_j , \quad (24)$$

where λ_j are Lagrange multipliers. Derive $J(\theta, \theta^{old})$ with respect to π_j and set it to 0, we obtain

$$\pi_j^{new} = \frac{1}{n}\sum_{i=1}^{n} p(j \mid x_i, \theta^s). \quad (25)$$

That is to say

$$\alpha_j^{new} = \frac{1}{n}\sum_{k=1}^{n} g(j \mid x_k, \mu_j^g, \Sigma_j^g), (j = 1,\cdots,K_\alpha)$$

$$\beta_j^{new} = \frac{1}{n}\sum_{k=1}^{n} s(j \mid x_k, v_j, \mu_j^s, \Sigma_j^s), (j = 1,\cdots,K_\beta), \quad (26)$$

$$\gamma_j^{new} = \frac{1}{n}\sum_{k=1}^{n} l(j \mid x_k, \mu_j^l, \Sigma_j^l), (j = 1,\cdots,K_\gamma),$$

Until now, all update rules are obtained by the EM algorithm. Before applying the EM algorithm, the K-means algorithm is often used to initialize the parameters in the heterogeneous mixture models.

4. EXPERIMENTS AND RESULTS

In order to assess the effectiveness of sparse representation features and hetMM, experiments are conducted on a collected database. We downloaded about 20hours videos from Youku [16], with different programs and different language. The start and end position of all the applause and laughter of the audio-tracks are manually labeled. The database includes 800 segments of each sound effect. Each segment is about 3-8s long and totally about 1hour data for each sound effect. All the audio recordings were converted to monaural wave format at a sampling frequency of 8kHz and quantized 16bits. Furthermore, the audio signals have been normalized, so that they have zero mean amplitude with unit variance in order to remove any factors related to the recording conditions.

Audio streams were windowed into a sequence of short-term frames (20 ms long) with 10-ms overlap. SVM with Gaussian kernel and mel-frequency cepstral coefficients (MFCCs+ Δ MFCCs+ Δ^2 MFCCs) with cepstral subtraction when in noisy conditions were used as alternatives to our approaches. Following [15], we have used the regularization parameter $\lambda = 1.2/\sqrt{120}$ in all of our experiments. We evaluate our approach using 5-fold cross validation on our labeled collections: At each fold, classifiers are trained on 4/5 of the data and then tested on the remaining 1/5. For mixture model, 512 mixtures are used where the optimal component number is determined by a 5-fold cross validation on a 2hours subset. For hetMM, we chose nearly equal heterogeneous components.

In Figure 1 and 2, the recall and precision obtained by all approaches are plotted as a function of the feature space dimension. The best recall (95.56%) was achieved by hetMM based classifiers using SRF. The standard deviation of the error rates was estimated due to 5-fold cross-validation. At the best recall, its standard deviation was found to be 0.86%.

To evaluate the robustness of the sparse representation features to noisy condition, we conducted a comparison on the test dataset added with 20dB Gaussian white noise. Generally sparse representation features is more robust to audio variants than MFCC features. Sparse representation features in noisy condition have an average decrease of 0.5% and 0.72% compared to clean condition on recall and precision respectively, while the decrease is 4.12% and 6.19% for MFCCs features. However, the experimental results show that both features have comparable performance when sparse representation features have nearly same dimension as MFCCs.

Generally, hetMM has better performance than SDT mixture models and comparable performance with SVM. It should be noted that although the overall performanc of hetMM is better than that of SDT mixture models, part of hetMM results are slightly worse than SDT mixture models.

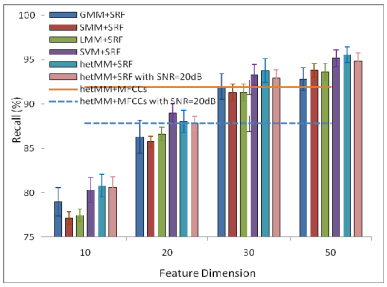

Figure 1: *Recall for various features and classifiers.*

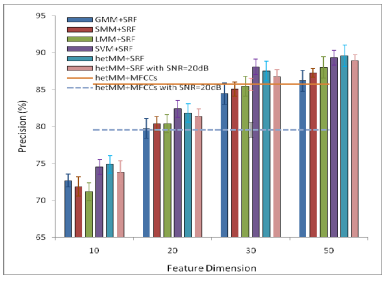

Figure 2: *Precision for various features and classifiers.*

5. CONCLUSIONS

We have presented an effective and robust audio event detection framework based on sparse representation features and heterogeneous mixture model. Experimental evaluations showed that the sparse representation features is more effective than popular audio features such as MFCCs, and classifier based on hetMM has a better performance than the that based on SDT mixture models such as GMM, SMM, LMM and gives comparable performance with SVM.

6. ACKNOWLEDGEMENTS

This work was supported by the grant from the National Basic Research Program of China (973 Program), No. 2007CB311100.

7. REFERENCES

[1] Umapathy, K., Krishnan, S., and Rao, R.K, "Audio signal feature extraction and classification using local discriminant bases," IEEE Transactions on Audio, Speech, and Language Processing, 15(4), pp. 1236-1246, 2007.

[2] Baraniuk, R. G., Candes, E., Nowak, R. and Vetterli, M. "Special issue on Sensing, Sampling, and Compression," IEEE Signal Processing Magazine, vol. 25, no. 2, Mar. 2008.

[3] Koerding, K.P., Koenig, P. and Klein, D.J., "Learning of sparse auditory receptive fields," in Proc. of the International Joint Conference on Neural Networks, pp. 1103-1108, 2002.

[4] Wright, J., Yang, A., Ganesh, A., Sastry, S., and Yi Ma "Robust Face recognition via sparse representation," IEEE Trans. Pattern Analysis and Machine Intelligence, 31(2), pp. 210–227, Feb. 2009.

[5] Reynolds, D., Quatieri, T., and Dunn, R., "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing 10, pp. 19-41, 2000.

[6] Ma, Z. and Leijon, A., "Beta mixture models and the application to image classification," in Proc. IEEE International Conference on Image Processing, pp. 2045-2048, 2010.

[7] Sfikas, G., Nikou, C., and Galatsanos, N., "Robust image segmentation with mixtures of Student's t-distributions," in Proc. IEEE International Conference on Image Processing, IEEE. pp. 273-276, 2007.

[8] Balakrishnan, N., "Handbook of the logistic distribution," CRC, 1992.

[9] Verhulst, P., "Recherches mathematiques sur la loi daccroissement de la population," Mem. de lAcademie Royale des Sci. et Belles-Lettres de Bruxelles 18, pp. 1-41, 1845.

[10] Modis, T., "Conquering uncertainty," McGraw-Hill, 1998.

[11] Hilbe, J., "Logistic regression models," CRC, 2009.

[12] Auer, P., Burgsteiner, H., and Maass, W., "A learning rule for very simple universal approximators consisting of a single layer of perceptrons," Neural networks 21, pp. 786-795, 2008.

[13] Malik, H. and Abraham, B., "Multivariate logistic distributions," The Annals of Statistics 1, pp. 588-90, 1973.

[14] Bach, F.R., Lanckriet, G.R.G., and Jordan, M.I., "Multiple kernel learning, conic duality, and the SMO algorithm," in Proc. of the twenty-first international conference on Machine learning, pp. 1-8 2004.

[15] Mairal, J., Bach, F., Ponce, J. and Sapiro, G., "Online dictionary learning for sparse coding," in Proc. of the 26th Annual International Conference on Machine Learning, pp. 689-696, 2009.

[16] "Youku", Available: http://www.youku.com