ICSP2008 Proceedings

BYY HARMONY ENFORCING REGULARIZATION FOR GAUSSIAN MIXTURE LEARNING

Hongyan Wang, Jinwen Ma

Department of Information Science, School of Mathematical Sciences
Peking University, Beijing, 100871, China

ABSTRACT

In this paper, a Bayesian Ying-Yang (BYY) harmony enforcing regularization (BYY-HER) algorithm is proposed for Gaussian mixture learning with a sample dataset on both parameter estimation and model selection, i.e., selecting an appropriate number of Gaussians in the mixture, through a regularization process from the BYY harmony learning to the maximum likelihood learning. It has been demonstrated by experiments on synthetical and real sample datasets that our proposed BYY-HER algorithm can not only select the correct number of actual Gaussians in a dataset, but also obtain good parameter estimations for the parameters in the true mixture.

Keywords: Gaussian mixture; BYY Harmony learning; Automated model selection; Regularization; Maximum likelihood

1. INTRODUCTION

As a powerful tool for data clustering, Gaussian mixture model has been extensively studied in the literature for either data modeling or clustering analysis on a sample dataset. Although there have been various statistical or unsupervised competitive learning methods to do such a task, e.g. the EM algorithm [1] for Maximum Likelihood (ML), k-means algorithm [2] for the least Mean Square Error (MSE), it is usually assumed that the number of Gaussians, or clusters, in the dataset is pre-known. However, in many instances this key information is not available and then the selection of an appropriate number of Gaussians must be made before or during the estimation of the parameters in the mixture, which is a rather complicated and difficult task [3].

As the number k of Gaussians is just a scale of the Gaussian mixture model, its determination is actually referred to as model selection. Thus, the general Gaussian mixture modeling is actually a compound problem of estimation and model selection. In fact, this compound problem has been investigated by many researchers from different directions. The traditional method was to choose an optimal number of Gaussians via certain selection criterion. Among these criteria, Akaike's information criterion (AIC) [4] is well known. But the validating process is computationally consumptive because we need to repeat the entire parameter learning process at a large number of possible values of k.

Recently, the Bayesian Ying-Yang (BYY) harmony learning system [5-7] has developed a new learning mechanism that makes model selection automatically during parameter learning. In fact, the BYY harmony learning has already implemented on the Guassian mixture modeling for the parameter learning with automated model selection. In order to do so, a bidirectional architecture (BI-architecture) and a backward architecture (B-architecture) of the BYY learning system were constructed for a finite mixture such that the Gaussian mixture modeling can be transformed into a BYY harmony learning problem on them. Actually, some efficient BYY harmony learning algorithms have been already established on the BI-architecture of the BYY learning system (e.g., [8-10]). On the other hand, a direct maximization of the harmony function on the B-architecture of the BYY learning system leads to a discrete optimal problem with a hard-cut EM algorithm [5], which suffers from the difficulty of being stuck at a local maximum solution. So, the number of Gaussians cannot be determined correctly by the basic BYY harmony learning or hard-cut EM algorithm. To overcome this difficulty, Ma and Liu already proposed an annealing learning algorithm [11] for searching the global maximum of the harmony function on this architecture from the maximum likelihood learning to the BYY harmony learning with automated model selection. Furthermore, an entropy regularized likelihood learning algorithm [12] proposed to solve the model selection problem for the Gaussian mixture learning. However, it is a constant regularization of the entropy of the posterior probabilities to the likelihood function, which may lead to a deviation between the estimate and the ML solution.

In the current paper, we propose a BYY harmony enforcing regularization (BYY-HER) algorithm on the Back architecture (B-architecture) of the BYY learning system for Gaussian mixture learning. With the regularization fac-

Jinwen Ma, the corresponding author, Telephone:86-10-62760609, Email: jwma@math.pku.edu.cn.

978-1-4244-2179-4/08/$25.00 ©2008 IEEE

tor shifting from 0 to 1, the BYY-HER algorithm turns the BYY harmony learning into the ML learning finally. Since the BYY harmony learning has the ability of automated model selection and the ML estimate is consistent, the BYY-HER algorithm will lead to a good estimate of the parameters with correct model selection as the regularization proceeds properly, which is actually demonstrated by the simulation and practical experiments.

In the sequel, the BYY-HER algorithm will be derived in Section 2. Some typical simulations and practical experiment are conducted in Section 3. Finally, we conclude briefly in Section 4.

2. BYY HARMONY ENFORCING REGULARIZATION ALGORITHM

According to the BYY harmony theory [7,11], we can get the following harmony function:

$$J(\Theta_k) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} p(j|x_i)\ln[\alpha_j q(x_i|m_j,\Sigma_j)], \qquad (1)$$

on the B-architecture with the parameters $\Theta_k = \{\alpha_j, m_j, \Sigma_j, p(j|x)\}$. By certain transformations, $J(\Theta_k)$ can be divided into two parts as follows:

$$J(\Theta_k) = L(\Theta_k) - O_N(p(y|x)),$$

where the first part is just the log-likelihood function:

$$L(\Theta_k) = \frac{1}{N}\sum_{i=1}^{N}\ln(\sum_{j=1}^{k}\alpha_j q(x_i|m_j,\Sigma_j)),$$

and the second part is the average Shannon entropy of the posterior probabilities $p(y|x_i)$ per sample over the sample set $\mathcal{S} = \{x_i\}_{i=1}^N$ (generated from a Gaussian mixture):

$$O_N(p(y|x)) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} p(j|x_i)\ln p(j|x_i).$$

That is, $L(\Theta_k) = J(\Theta_k) + O_N(p(y|x))$.

With this decomposition, we can consider this average Shannon entropy as the regularization term to the BYY harmony function. Moreover, a parameter $\lambda(\geq 0)$ can be introduced to dominate the intensity of the regularization. In this way, we construct the following objective function:

$$L_\lambda(\Theta_k) = J(\Theta_k) + \lambda O_N(p(y|x)), \qquad (2)$$

where λ is the regularization factor. When $\lambda = 0$, $L_\lambda(\Theta_k) = J(\Theta_k)$. This is just the BYY harmony function on the B-architecture. Maximizing $J(\Theta_k)$ with respect to $p(j|x_i)$ leads to a WTA or hard-cut version [5]. In general, the maximization of $J(\Theta_k)$ leads to the automated model selection, i.e., automatic detection of the number k^* of actual Gaussians in the sample data set $\mathcal{S}$, as long as k is initially selected to be greater than k^*. On the other hand, when $\lambda = 1$, $L_\lambda(\Theta_k)$ is just the log-likelihood function. So, maximizing $L_\lambda(\Theta_k)$ becomes the well-known maximum likelihood learning which leads to a good estimation of the parameters. If we let $\lambda \to 1$ from $\lambda_0 = 0$ appropriately, the maximization of $L_\lambda(\Theta_k)$ can make model selection automatically at the previous learning stage and converges to a good parameter estimation at the final learning stage.

Maximizing $L_\lambda(\Theta_k)$ by alternative optimization technique (refer to [11]), we can construct the BYY-HER algorithm as follows:

[The BYY-HER algorithm]

Step 1: Initialize $\lambda = \lambda_0$ (a very small constant);

Step 2: Set $t = 0$, and the initial value of $k(> k^*)$ and $\Theta_k^{(0)}$;

Step 3: At t time with $\lambda(t)$, iterate the following λ-EM algorithm until convergence:

$$\begin{aligned}
p^+(j|x_i) &= \frac{[\alpha_j q(x_i|m_j,\Sigma_j)]^{1/\lambda(t)}}{\sum_{j=1}^{k}[\alpha_j q(x_i|m_j,\Sigma_j)]^{1/\lambda(t)}};\\
\alpha_j^+ &= \frac{1}{N}\sum_{i=1}^{N} p^+(j|x_i);\\
m_j^+ &= \frac{1}{\sum_{i=1}^{N} p^+(j|x_i)}\sum_{i=1}^{N} p^+(j|x_i)x_i;\\
\Sigma_j^+ &= \frac{1}{\sum_{i=1}^{N} p^+(j|x_i)} \times \\
&\quad \sum_{i=1}^{N} p^+(j|x_i)(x_i - m_j^+)(x_i - m_j^+)^T.
\end{aligned}$$

Step 4: Let $t = t + 1$, and increase λ according to certain rule (refer to the experiments for detail);

Step 5: If $\lambda < 1 - \epsilon$ (ϵ is a very small positive constant), go to step 3; otherwise terminate.

Unlike the BYY annealing learning algorithm [11] in which λ attenuates along time, the regularization is finally enhanced in our BYY-HER algorithm. With $\lambda \to 1$ from 0, we can get both the correct model selection and the maximum likelihood estimate of the parameters in the Gaussian mixture from the BYY-HER learning. Ueda and Nakano already proposed a deterministic annealing EM (DAEM) algorithm [13] for the maximum likelihood estimation problem. Actually, the annealing parameter β in the DAEM algorithm serves as $1/\lambda$ in our BYY-HER algorithm. However, unlike our λ, the DAEM algorithm makes $1/\beta$ gradually tend to 1 from $+\infty$ (i.e., β from 0 to 1) so that it can search for the global maximum of the likelihood function to overcome the local maxima problem associated with the

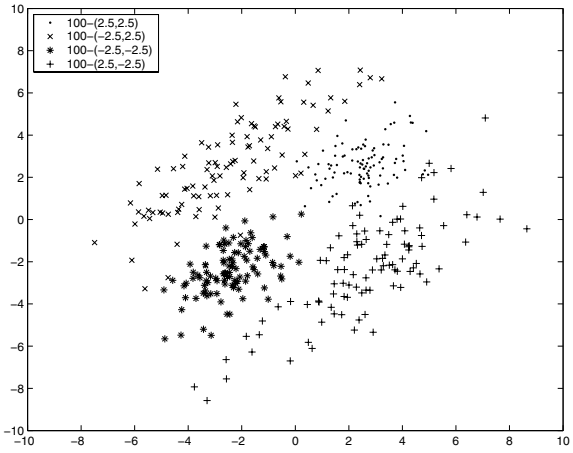

Fig. 1. The first sample dataset $\mathbf{S}_1$.

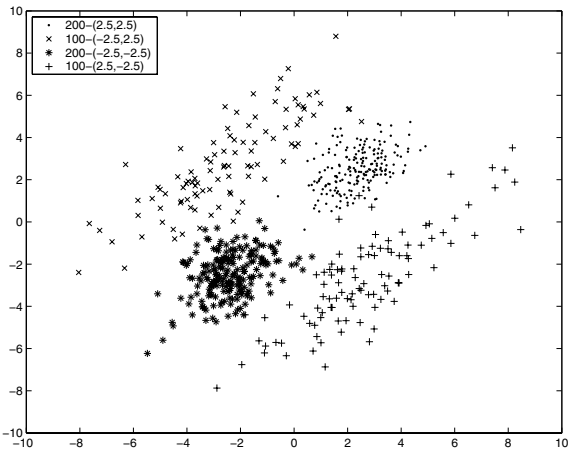

Fig. 2. The second sample dataset $\mathbf{S}_2$.

conventional EM algorithm at the setting $k = k^*$. Therefore, the DAEM algorithm leads to a good maximum likelihood estimate, but it has no ability to make model selection for the Gaussian mixture.

3. EXPERIMENTAL RESULTS

In this section, two experiments are carried out to demonstrate the performance of the BYY-HER algorithm for automated model selection as well as parameter estimation on the set of sample data randomly generated from a Gaussian mixture with certain degree of overlap. Moreover, the BYY-HER algorithm is applied to classification of the Iris data.

3.1. On synthetic datasets

As shown in Figs 1 & 2, two synthetic datasets of Gaussian mixture are used in our experiments. They consist of four ellipse-shaped Gaussians or clusters with different numbers of samples. The initial value of λ is near 0. Experientially, it is $1e-100$. As for the increasing procedure of λ, we set $\lambda = 1/[1 + exp(\frac{-(t-b)}{a})]$, where a can be selected in $[1.8, 2.2]$, while b can be chosen according the complexity of the problem. In our experiment, it is reasonable to set it by $100 \sim 200$. t is the number of updates for λ, being initialized by 0 and increased by 1 at each time. Moreover, as the BYY-HER algorithm shifting to the ML learning, we can eliminate the Gaussian whose mixing proportion α_j is less than 0.08 so that the model selection can be made automatically during the regularization process.

Before running the BYY-HER algorithm, we can implement the RPCL algorithm [14] on the sample datset to get a set of reasonable initial values for the mean vectors. In this way, the BYY-HER algorithm rarely converges to a lo-

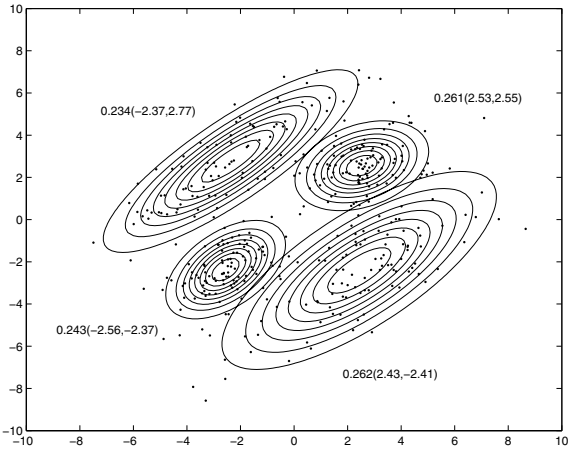

Fig. 3. The experimental result on $\mathbf{S}_1$.

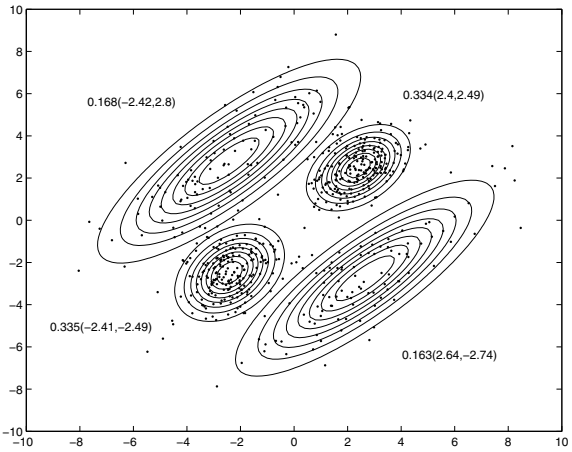

Fig. 4. The experimental result on $\mathbf{S}_2$.

cal maximum solution. The experiment results on the two synthetic datasets in the case of $k = 8$ are shown in Figs 3 & 4, respectively. It can be observed that four estimated Gaussians can match the actual ones in each dataset accurately, with the extra Gaussians being canceled automatically. Moreover, it can be further found that the estimations of the parameters are as good as the ML estimators.

3.2. On the Iris data

We further apply the BYY-HER algorithm to the classification of the Iris data (from http://archive.ics.uci.edu/ml/). This dataset consists of 150 samples of three classes which are Versicolor, Iris Virginica and Iris Setosa. Each class contains 50 samples and each sample or datum is 4-dimensional with measures of the plants morphology. Because the BYY-HER algorithm is in an unsupervised learning mode, we ignore the indexes of these samples. As the learning process has been accomplished, each sample is classified according to its maximum posterior probability $p(j|x_t)$.

By setting $k = 6$, $a = 2.0$ and $b = 200$, $\lambda_0 = 1e-200$, we implement the BYY-HER algorithm on the Iris data set with $0.01 \leq \lambda \leq 0.99$, and t being increased by 0.1 at each time. It has been found by the experiments that the BYY-HER algorithm can detect the three actual categories and the average accuracy is 96.7% (five samples in the second class were misclassified).

4. CONCLUSIONS

We have proposed the BYY harmony enforcing regularization (BYY-HER) algorithm on Gaussian mixture for both model selection and parameter estimation. The BYY-HER algorithm implements strongly the BYY harmony learning for automated model selection at the previous learning stage and gradually transforms to the maximum likelihood (ML) learning for good estimation of the parameters at the final learning stage. It is demonstrated by the simulation and practical experiments that the BYY-HER algorithm can detect the number of actual Gaussians in the sample dataset and obtain accurate estimation of the parameters in the Gaussian mixture.

ACKNOWLEDGEMENTS

This work was supported by the Natural Science Foundation of China for grant 60771061.

5. REFERENCES

[1] R. A. Render, H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.* vol. 26, no.2, pp. 195-293, 1984.

[2] P. A. Devijver, J.kitter, "A Statistical Approach," *Englewood Cliffs, Prentice-Hall, NJ*, 1982.

[3] J. A. Hartigan, "Distribution problems in clustering," *Classification and clustering*, J. Van Ryzin Eds., pp. 45-72, New York: Academic Press, 1977.

[4] H. Akaike. "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control,* vol. AC-19, pp. 716-723, 1974.

[5] L. Xu, "Bayesian Ying-Yang machine:, clusteing and number of cluters," *Pattern recgnition Lett.,* vol.18, pp. 1167-1178, 1997.

[6] L. Xu, "Best harmony, unified RPCL and automated model selection for unsupervised and supervised learning on Gaussian mixtures, three-layer nets and ME-RBF-SVM models," *Int. J. Neural Syst.,* vol.11, no.1, pp. 43-69, 2001.

[7] L. Xu, "BYY harmony learning, structural RPCL, and topological self-organizing on mixture modes," *Neural Networks,* vol.15, pp. 1231-1237, 2002.

[8] J. Ma, T. Wang, L. Xu, "A gradient BYY harmony learning rule on Gaussian mixture with automated model selection," *Neurocomputing,* vol.56, pp. 481-487, 2004.

[9] J. Ma, L. Wang, "BYY harmony learning on finite mixture: adaptive gradient implementation and a floating RPCL mechanism," *Neural Processing Lett,* vol.24, no.1, pp. 19-40, 2006.

[10] J. Ma, X. He, "A fast fixed-point BYY harmony learning algorithm on Gaussian mixture with automated model selection," *Pattern Recognition Letters*, vol.29, no.6, pp. 701-711, 2008.

[11] J. Ma and J. Liu, "The BYY annealing learning algorithm for Gaussian mixture with automated model selection," *Pattern Recognition,* vol.40, pp. 2029-2037, 2007.

[12] Z. Lu and J. Ma, "A Gradient Entropy Regularized Likelihood Learning Algorithm on Gaussian Mixture with Automatic Model Selection," *Lecture Notes in Computer Science*, vol.3971, pp: 464-469, 2006.

[13] N. Ueda and R. Nakano, "Deterministic annealing EM algorithm," *Neural Networks,* vol.11, pp. 271C282, 1998.

[14] L. Xu, A. Krzyak, E. Oja, "Rival penalized competitive learning for clustering analysis, RBF net, and curve detection," *IEEE Trans. Neural Networks,* vol.4, no.4, pp. 636C649, 1993.