Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT
Neurocomputing 56 (2004) 481–487

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Letters

A gradient BYY harmony learning rule on Gaussian mixture with automated model selection☆

Jinwen Ma[a,b,*], Taijun Wang[a], Lei Xu[a]

[a] *Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong*

[b] *Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China*

Abstract

One important feature of Bayesian Ying–Yang (BYY) harmony learning is that model selection can be made automatically during parametric learning. In this paper, BYY harmony learning with a bi-directional architecture is studied for Gaussian mixture modelling via a gradient learning rule. It has been demonstrated by simulation experiments that the number of Gaussians can be determined automatically during learning the parameters of the Gaussian mixture.

Keywords: Model selection; Gaussian mixture; Bayesian Ying–Yang learning

1. Introduction

Gaussian mixture modelling is a powerful approach for data analysis. Although there have been several statistical methods for implementing this task, e.g., maximum likelihood estimation and the EM algorithm, it is usually assumed that the number k of Gaussians in the mixture is pre-known. However, in many cases this key information is not available and the selection of an appropriate number of Gaussians must be made

☆ This work was supported by the HK RGC Earmarked grant CUHK 4336/02E and the Natural Science Foundation of China for Project 60071004.

* Corresponding author. Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong.

E-mail address: jwma@math.pku.edu.cn (J. Ma).

0925-2312/$ - see front matter © 2003 Published by Elsevier B.V.
doi:10.1016/j.neucom.2003.10.009

with the estimation of the parameters. One possible approach is to choose a best k^* by the Akaike's information criterion [1] or its extensions. But the process of evaluating a criterion incurs a large computational cost since we need to repeat the entire parameter learning process at a number of different values of k.

Proposed in 1995 [2] and systematically developed in past years [3–5], Bayesian Ying–Yang (BYY) harmony learning acts as a general statistical learning framework not only for understanding several existing major learning approaches but also for tackling the learning problem with a new learning mechanism that makes model selection automatically during parameter learning. In the following, we implement this mechanism on a bi-directional architecture (BI-architecture) of the BYY system via a gradient learning rule to solve the Gaussian mixture modelling problem.

2. Gradient learning rule

A BYY system describes each observation $x \in \mathcal{X} \subset R^n$ and its corresponding inner representation $y \in \mathcal{Y} \subset R^m$ via the two types of Bayesian decomposition of the joint density $p(x, y) = p(x)p(y|x)$ and $q(x, y) = q(x|y)q(y)$, being called Yang and Ying machine, respectively. In this paper, y is only limited to be an integer variable, i.e., $y \in \mathcal{Y} = \{1, 2, \ldots, k\} \subset R$ with $m = 1$. Given a data set $D_x = \{x_t\}_{t=1}^N$, the task of learning on a BYY system consists of specifying all the aspects of $p(y|x), p(x), q(x|y), q(y)$ with a harmony learning principle implemented by maximizing the functional

$$H(p\|q) = \int p(y|x)p(x)\ln[q(x|y)q(y)]\,\mathrm{d}x\,\mathrm{d}y - \ln z_q, \tag{1}$$

where z_q is a regularization term. The details are referred to [3].

If both $p(y|x)$ and $q(x|y)$ are parametric, i.e., from a family of probability densities with a parameter $\theta \in R^d$, the BYY system is called to have a Bi-directional Architecture (BI-Architecture). For Gaussian mixture modelling, we use the following specific BI-architecture of the BYY system. $q(j) = \alpha_j$ with $\alpha_j \geqslant 0$ and $\sum_{j=1}^k \alpha_j = 1$. Also, we ignore the regularization term z_q (i.e., set $z_q = 1$) and let $p(x)$ be the empirical density $p_0(x) = (1/N)\sum_{t=1}^N \delta(x - x_t)$, where $x \in \mathcal{X} = R^n$. Moreover, the BI-architecture is constructed with the following parametric form:

$$p(y = j|x) = \frac{\alpha_j q(x|\theta_j)}{q(x|\Theta_k)}, \quad q(x|\Theta_k) = \sum_{j=1}^k \alpha_j q(x|\theta_j), \tag{2}$$

where $q(x|\theta_j) = q(x|y = j)$ with θ_j consisting of all its parameters and $\Theta_k = \{\alpha_j, \theta_j\}_{j=1}^k$. Substituting these component densities into Eq. (1), we have

$$H(p\|q) = J(\Theta_k) = \frac{1}{N}\sum_{t=1}^N \sum_{j=1}^k \frac{\alpha_j q(x_t|\theta_j)}{\sum_{i=1}^k \alpha_i q(x_t|\theta_i)} \ln[\alpha_j q(x_t|\theta_j)]. \tag{3}$$

That is, $H(p\|q)$ becomes a harmony function $J(\Theta_k)$ on the parameters Θ_k of a finite mixture model, which was originally introduced in [2] as $J(k)$ and developed into this form in [3] using as a selection criterion of the number k. Letting $q(x|\theta_j)$ be a Gaussian density given by

$$q(x|\theta_j) = q(x|m_j, \Sigma_j) = \frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \mathrm{e}^{-(1/2)(x-m_j)^{\mathrm{T}}\Sigma_j^{-1}(x-m_j)}, \tag{4}$$

where m_j is the mean vector and Σ_j is the covariance matrix, and $\alpha_j = e^{\beta_j}/\sum_{i=1}^{k} e^{\beta_i}$ for $j = 1, 2, \ldots, k$ with $-\infty < \beta_1, \ldots, \beta_k < +\infty$. By the derivatives of $J(\Theta_k)$ with respect to β_j, m_j and Σ_j, we have the following gradient learning rule:

$$\Delta\beta_j = \eta \frac{\alpha_j}{N} \sum_{i=1}^{k} \sum_{t=1}^{N} h(i|x_t) U(i|x_t)(\delta_{ij} - \alpha_i), \tag{5}$$

$$\Delta m_j = \eta \frac{\alpha_j}{N} \sum_{t=1}^{N} h(j|x_t) U(j|x_t) \Sigma_j^{-1}(x_t - m_j), \tag{6}$$

$$\Delta \Sigma_j = \eta \frac{\alpha_j}{2N} \sum_{t=1}^{N} h(j|x_t) U(j|x_t) \Sigma_j^{-1}[(x_t - m_j)(x_t - m_j)^T - I]\Sigma_j^{-1}, \tag{7}$$

where

$$U(i|x_t) = \sum_{r=1}^{k} (\delta_{ri} - p(r|x_t)) \ln \alpha_r q(x_t|\theta_r) + 1, \tag{8}$$

$$h(i|x_t) = \frac{q(x_t|\theta_i)}{\sum_{r=1}^{k} \alpha_r q(x_t|\theta_r)}, \quad p(i|x_t) = \alpha_i h(i|x_t) \tag{9}$$

and δ_{ij} is the Kronecker function, and η is the learning rate which is usually a small positive number.

In the same way, we can construct such a gradient learning rule for the other kind of finite mixture model by considering the harmony function in Eq. (3) with $q(x|\theta_j)$ being changed into the other probability distribution instead of Gaussian one.

3. Simulation results

We conducted experiments on seven sets $\mathscr{S}_i, i = 1, 2, \ldots, 7$ of samples drawn from a mixture of four or three bivariate Gaussians densities (i.e., $n = 2$). As shown in Fig. 1, each data set of samples is generated at different degree of overlap among the clusters(Gaussians) in the mixture by controlling the mean vectors and covariance matrices of the Gaussian distributions, and with equal or unequal mixing proportions of the clusters in the mixture by controlling the number of samples from each Gaussian.

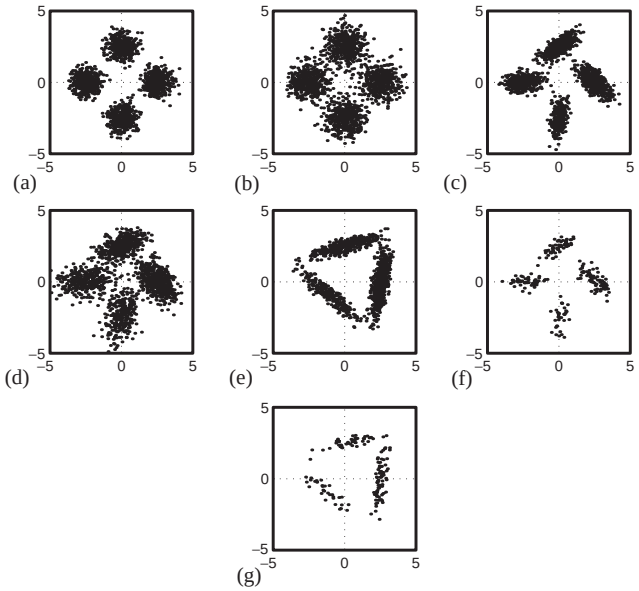

Fig. 1. Seven sets of sample data used in the experiments (a) set $\mathscr{S}_1$; (b) set $\mathscr{S}_2$; (c) set $\mathscr{S}_3$; (d) set $\mathscr{S}_4$; (e) set $\mathscr{S}_5$; (f) set $\mathscr{S}_6$; (g) set $\mathscr{S}_7$.

Using k^* to denote the number of Gaussians in the original mixture, we implemented the gradient learning rule on those seven sample data sets always with $k \geqslant k^*$. To speed up and stabilize the algorithm, we replaced the learning rate η with $\eta|\Sigma_j|/\alpha_j$ in Eqs. (5–7) and set η to be 0.1. Moreover, the other parameters were initialized randomly within certain intervals. In all the experiments, the learning was stopped when $|J(\Theta_k^{\text{new}}) - J(\Theta_k^{\text{old}})| < 10^{-7}$.

The experimental results on $\mathscr{S}_2$ and $\mathscr{S}_4$ are given in Figs. 2 and 3, respectively, with case $k = 8$ and $k^* = 4$. We observe that four Gaussians are finally located accurately, while the mixing proportions of the other four Gaussians were reduced to below 0.001, i.e, these Gaussians are extra and can be discarded. That is, the correct number of the clusters have been detected on these data sets. Moreover, the experiment has been made on $\mathscr{S}_5$ with $k = 8, k^* = 3$. As shown in Fig. 4, clusters are far from spherical shapes (actually they are very flat). Again, three Gaussians are located accurately, while the mixing proportions of the other five extra Gaussians become less than 0.001. That is, the correct number of the clusters can still be detected on such a special data set. Furthermore, the gradient learning rule was also implemented on $\mathscr{S}_6$ with $k = 8, k^* = 4$. As shown in Fig. 5, even each cluster has a small number of samples, the correct number of clusters can still be detected, with the mixing proportions of other four extra Gaussians reduced below 0.001.

The further experiments on the other sample sets had been also made successfully for the correct number detection in the similar cases. Actually, in many experiments, a failure on the correct number detection rarely happened when we initially set $k^* \leqslant k \leqslant 3k^*$. However, the gradient learning may lead to a wrong detection when

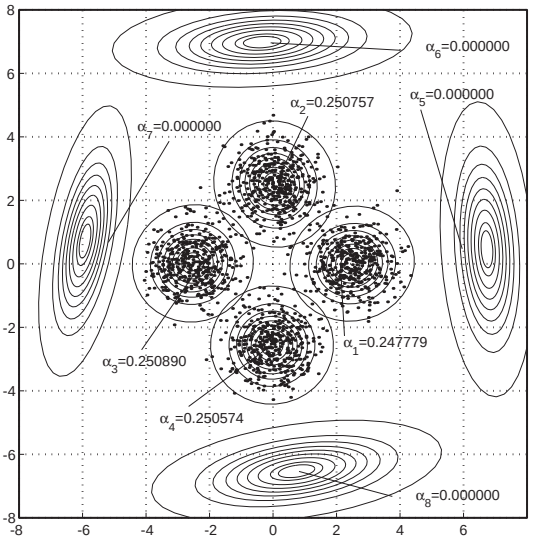

Fig. 2. The experimental result on $\mathscr{S}_2$ (stopped after 180 iterations).

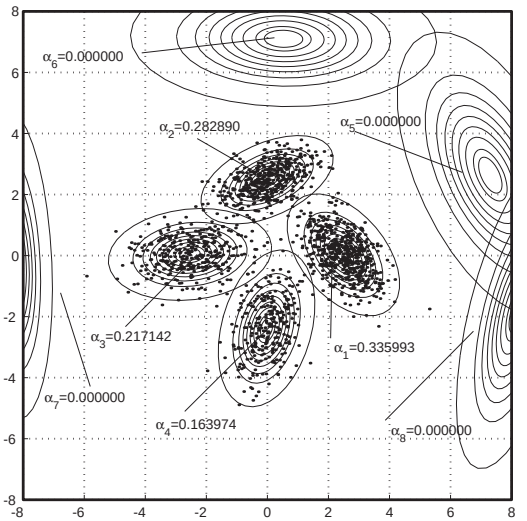

Fig. 3. The experimental result on $\mathscr{S}_4$ (stopped after 543 iterations).

$k > 3k^*$. Also, it is observed that the gradient learning rule enforces a mechanism of rewarding and penalizing competitive learning among the Gaussians through their mean vectors, which is very similar to that of rival penalized competitive learning (RPCL) [6]. Therefore, the theory may provide BYY harmony learning a new approach to the theoretical analysis of RPCL.

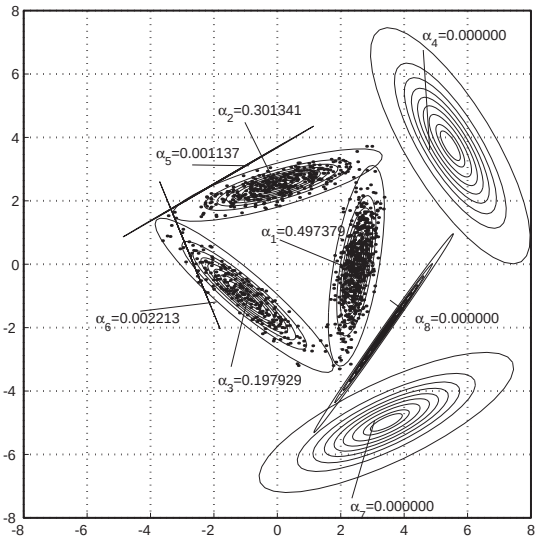

Fig. 4. The experimental result on $\mathscr{S}_5$ (stopped after 33 000 iterations).

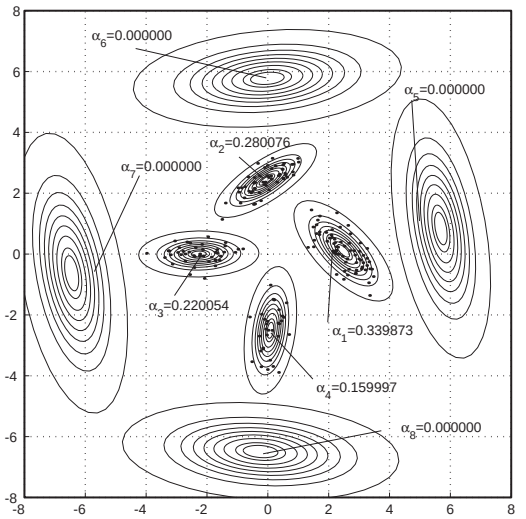

Fig. 5. The experimental result on $\mathscr{S}_6$ (stopped after 1216 iterations).

In addition to the correct number detection, we further compared the converged values of parameters (discarding the extra Gaussians) with those parameters in the mixture from which the samples come from. We checked the results in all the above empirical experiments and found that the gradient learning converges with a lower

average error between the estimated parameters and the true parameters being less than 0.1.

Furthermore, we tested the gradient learning rule for clustering on some sample data sets in which each cluster is not subject to a Gaussian. The experiment results have shown that the correct number of clusters can be still detected when those clusters can be separated in the similar degree as above. Also, under the principle of the maximum posteriori probability $p(j|x_t)$ of the converged parameters Θ_k, the clustering result is generally as good as the k-means algorithm with $k = k^*$. However, when two or more clusters are joined together like iris data, the gradient learning rule can only find out the separated clusters in the sample data set.

4. Conclusions

The automatic model selection feature of BYY harmony learning has been demonstrated on Guassian mixture modelling with a BI-architecture of the BYY system. In help of the gradient learning rule derived, a number of experiments have demonstrated that as long as the overlap among the Gaussians or clusters in a data set is not too serious, the number of Gaussians can be correctly detected automatically during learning with a good estimation on parameters of each Gaussian component density, even on a data set of a small sample size.

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Control AC-19 (6) (1974) 716–723.

[2] L. Xu, Ying–Yang machine: a Bayesian–Kullback scheme for unified learnings and new results on vector quantization, Proceedings of the 1995 International Conference on Neural Information Processing, ICONIP'95, Vol. 2, Beijing, China, 30 October–3 November 1995, pp. 977–988.

[3] 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, Internat. J. Neur. Syst. 11 (1) (2001) 43–69.

[4] L. Xu, Ying–Yang learning, in: M.A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, 2nd Edition, The MIT Press, Cambridge, MA, 2002, pp. 1231–1237.

[5] L. Xu, BYY harmony learning, structural RPCL, and topological self-organizing on mixture modes, Neur. Networks 15 (8–9) (2002) 1231–1237.

[6] L. Xu, A. Krzyzak, E. Oja, Rival penalized competitive learning for clustering analysis, RBF net, and curve detection, IEEE Trans. Neur. Networks 4 (4) (1993) 636–648.