An Unsupervised Gaussian Mixture Classification Mechanism based on Statistical Learning Analysis

Rui Nian[1, 2], Guangrong Ji[2], Michel Verleysen[1]
[1] *Machine Learning Group, DICE, Université catholique de Louvain, Place du Levant, 3-B-1348 Louvain-la-Neuve, Belgium*
[2] *College of Information Science and Engineering, Ocean University of China, Qingdao, China, 266003*
nianrui_80@163.com, grji@mail.ouc.edu.cn, michel.verleysen@uclouvain.be

Abstract

This paper presents a scheme for unsupervised classification with Gaussian mixture models by means of statistical learning analysis. A Bayesian Ying-Yang harmony learning system acts as a statistical tool for the particular derivation and development of automatic joint parameter learning and model selection. The proposed classification mechanism roughly decides on the number of real classes, by earning activation for the winners and assigning penalty for the rivals, so that the most competitive center wins for possible prediction and the extra ones are driven far away when starting the algorithm from a too large number of classes without any prior knowledge. Simulation experiments prove the feasibility of the approach and show good performance for unsupervised classification and natural estimation on the number of classes.

1. Introduction

Statistics tools are always crucial for the analysis and development of knowledge discovery and data mining. Statistical learning theory is a machine learning principle with the ambition to explore the inherent distribution, dependence structure, and generalization ability in the learning model as well as possible [1, 2]. Vapnik first put forward statistical learning theory as a sound statistical basis for the assessment of the predictive model [1], and aroused Support Vector Machine as a popular practical method. Lei Xu also proposed a general statistical learning framework, Bayesian Ying-Yang (BYY) harmony learning theory, for simultaneous parameter learning and model selection [2].

Unsupervised classification attempts to partition and assign natural groups via a similarity measure based on features and properties [3-5]. As it starts from little prior knowledge about the data, without supervision as in a classification model and manual labels, potential difficulties exist: among the possible problems, one can anticipate that information from various sources could be mixed in a single class, and conversely that single information could be split among classes. The best choice for the number of classes is not always clear, so it is often an ad hoc decision to choose it; the problem becomes even more difficult in mixture distribution cases. Therefore we need to take some knowledge into consideration to select models and label classes during unsupervised classification.

In this paper we present an unsupervised classification mechanism for Gaussian mixture models, using statistical learning analysis on a Bayesian Ying-Yang harmony learning system. The proposed method includes automatic parameter learning and model selection in parallel. Both of the two key elements in unsupervised classification, i.e., the classification formation and the structure selection, are almost achieved at the same time in a single framework. Starting from a number of classes larger than the number of original mixtures, the unsupervised classification procedure proposed in this paper welcomes the real winners and banishes their rivals far away, so that the natural (effective) number of classes can be synchronously learned in most cases. Simulation experiments show good performance for unsupervised classification, including the determination of the number of classes.

2. Bayesian Ying-Yang harmony learning

Let $\boldsymbol{X}$ be the observation world and $\boldsymbol{Y}$ be the class representation domain; a Bayesian Ying-Yang system consists of two Bayesian perspectives of joint probability distribution decomposition from complementary paths [2]:

$$p(x,y) = p(y|x)p(x),\ q(x,y) = q(x|y)q(y) \quad (1)$$

In this equation, both $p(x,y)$ and $q(x,y)$ define the joint probability distribution of x and y, respectively coming from two paths called the Yang and the Ying paths; using q instead of p for the path is only a question of notation aimed to identify without ambiguity.

As $p(x,y)$ and $q(x,y)$ are first defined and estimated from two different perspectives, initially the two distributions are not necessarily equal. The

978-0-7695-3305-6/08 $25.00 © 2008 IEEE
DOI 10.1109/FSKD.2008.333

fundamental BYY harmony learning principle is to make the Ying and Yang machine be in best harmony in a twofold sense, i.e. to conform to the matching nature as well as the least complexity nature between the two paths p and q, so as to accomplish equality and simplicity of $p(x,y)$ and $q(x,y)$ at the end of learning. Mathematically, harmony learning is a continuous optimization task for parameter learning and a discrete optimization for model selection, both aiming to maximize the same cost function $H(p\|q)$ between p and q. Here $H(p\|q)$ is defined as [2]:

$$H(p\|q) = \int p(x,y)\ln q(x,y)dxdy - \ln z_q$$
$$z_q = \sum_{t=1}^{N} q(x_t, y_t) \quad (2)$$

where z_q is a sum of the joint probabilities $q(x_t, y_t)$ evaluated on all pairs x_t and y_t given an input dataset $\{x_t\}_{t=1}^{N}$.

In this paper, we specify a specific Bayesian probability distribution architecture for unsupervised classification framework in the frame of BYY harmony learning; we then derive and explore the iterative learning procedure on Gaussian Mixture model.

3. Unsupervised classification mechanism for Gaussian mixture model

Within the statistical learning context detailed in Section 2, we first choose and settle a probability distribution model that might be proper for the unsupervised classification framework, and then derive and develop the relevant learning algorithm from the general probability learning process of BYY harmony learning system. First of all, suppose $p(x)$, $q(y)$ and $q(x|y)$ all take parametric form, while $p(y|x)$ is structure-free [2]. The task of machine learning is to decide the probability distribution form of $p(y|x)$ and to obtain in parallel all parameters in the probability distribution functions. The following selection and derivation of statistical learning analysis offers us a way to achieve the unsupervised classification mechanism for Gaussian Mixture model.

3.1. Gaussian mixture model

In statistics, a probability mixture model denotes the convex combination of multiple probability distributions. Assuming that each class adheres to a unimodal distribution, Gaussian mixture model tries to make each center capture and describe the truth of a single class around a single mode, and the membership of each observation is defined through unobserved latent variables. Here we also adopt the Gaussian mixture model for the Bayesian probability architecture as a start, and attempt to find centers of multiple natural classes. In order to simplify the derivation, suppose $q(x|y)$ follows one Gaussian distribution stated below,

$$q(x|y) = G(x|c_y, \Sigma_y),\ \Sigma_y = \sigma_y{}^2 I \quad (3)$$

Hereafter, for classification problems, we define y as the class label, $y = 1, 2, \cdots, k$, with k denoting the number of classes, and the covariance matrix Σ_y makes all classes isotropic.

3.2. Bayesian probability architecture

We then choose the probability distribution forms for $p(x)$ and $q(y)$, in order to set up a Bayesian probability architecture for unsupervised classification based on a Gaussian mixture model. As we have no a priori information about the relations and connections among inputs, $p(x)$ is derived from an empirical density, $p(x) = \frac{1}{N}\sum_{t=1}^{N}\delta(x - x_t)$. For an unsupervised classification problem, $q(y)$ takes a parametric form as,

$$q(y) = \sum_{\ell=1}^{k}\alpha_\ell \delta(y-\ell),\quad \sum_{\ell=1}^{k}\alpha_\ell = 1 \quad (4)$$

Here $q(y)$ is drawn from the Kronecker function distribution, and α_ℓ are the mixture proportions for Gaussian Mixture model.

3.3. General learning procedure

From the above framework of probability distributions, the general BYY harmony learning process for a structure-free $p(y|x)$ is then expressed as [2]:

$$p(y|x) = \delta(y - y(x)),$$
$$y(x) = \arg\min_y d(x,y), \quad (5)$$
$$d(x,y) = -\ln[q(x|y)q(y)] + \ln z_q$$

where $d(x,y)$ could be thought as one probability measure for the joint probability distribution $q(x,y)$.

3.4. Membership hypothesis

In practice, the most common case for unsupervised classification is that the number of classes is unknown

beforehand. However, many unsupervised classification algorithms, in particular K-means, strongly depend on the number of classes that has to be fixed in advance [4].

In order to perform classification and to choose the adequate number of classes in parallel, a basic idea for membership is introduced [6]. After each input presentation, not only the winner is modified to adapt to the input, but also its rival will receive a small penalty. Specifically, the following typical membership case is considered in the unsupervised classification framework,

$$Y_t = \begin{cases} y_t = \arg\min_y d(x_t, y(x_t)) \\ y_r = \arg\min_{y \neq y_t} d(x_t, y(x_t)) \end{cases} \quad (6)$$

where y_t refers to the class label of the winner, while y_r is the class label of its rival, representing the class with the second minimum measure $d(x, y)$ between the input except the winner. When starting from a number of classes that is larger than the natural number of groups in the data, the goal is to automatically adjust the effective number of classes during classification.

3.5. Iterative algorithm derivation

On the basis of the above assumptions and definitions made on the probability distributions for the Bayesian architecture, we can detail the adaptive update algorithm for unsupervised classification. First set the initial state as $z_q(t) = 0$, $\alpha_y = 1/k$; $z_q(t)$ can be cumulatively computed like this:

$$z_q(t) = z_q(t-1) + \sum_y \alpha_y G(x_t \,|\, c_y, \sigma_y{}^2 I) \quad (7)$$

A set of update factors in the iterative steps is obtained by computing all derivatives of $H(p\|q)$ with respect to the corresponding variables c_y , σ_y , α_y and so on.

$$\begin{aligned} c_y{}^{new} &= c_y{}^{old} + \gamma_{t,y}(x_t - c_y{}^{old}) \\ \sigma_y{}^{new} &= \sigma_y{}^{old} + \gamma_{t,y}\sigma_y{}^{old}(D(x_t, c_y{}^{old}) - \sigma_y{}^{old\,2}) \\ \alpha_y{}^{new} &= e^{\alpha_y{}^{\sim new}} \Big/ \sum_{\ell=1}^{k} e^{\alpha_\ell{}^{\sim new}} \\ \alpha_y{}^{\sim new} &= \alpha_y{}^{old} + \gamma_{t,y} \begin{cases} 1 - \alpha_y{}^{old} & y = y_t \\ -\alpha_y{}^{old} & y = y_r \end{cases} \\ \tau^{new} &= \tau^{old} + 1 \end{aligned} \quad (8)$$

with

$$\gamma_{t,y} = \begin{cases} \gamma_0 (1/\tau^{new} - \alpha_y G(x_t \,|\, c_y, \sigma_y{}^2 I)) \big/ z_q(t)) & y = y_t \\ -\gamma_0 \alpha_y G(x_t \,|\, c_y, \sigma_y{}^2 I) \big/ z_q(t) & y = y_r \end{cases} \quad (9)$$

Here τ is introduced as a counter, $D(x_t, c_y{}^{old})$ is the similarity measure between the input x_t and the previous class mean $c_y{}^{old}$. $\gamma_{t,y}$ is a rate for machine learning specific to the input x_t and the class label y , and γ_0 is a given constant. In virtue of the measure $d(x, y)$, $\gamma_{t,y}$ from the winner y_t stands for the learning rate, while its counterpart $\gamma_{t,y}$ from the rival y_r refers as the penalty rate.

After taking iterative steps, the unsupervised classification framework roughly decides on the number of classes. The new input is assigned to the nearest class with a winning label based on the minimum measure $d(x, y)$ in the competitive condition that denotes the similarity and correlation between data and classes.

3.6. Simplification

Knowledge discovery and data mining always refer to large scale dataset. In practice, a large scale dataset inevitably leads to the huge computation load. Though the above statistical learning idea is not too heavy on the computational point of view, it might still be necessary to simplify the computations as much as possible. The Cityblock distance is first adopted instead of the direct computation of Euclidean metric for an easy realization in the large scale dataset with only additions, subtractions, and arithmetic comparisons.

Secondly for any random input, one easy way to realize the above iterative steps is to replace $\gamma_{t,y}$ by,

$$\gamma_{t,y} = \begin{cases} \gamma_t \eta_{t,y} & y = y_t \\ -\gamma_r \eta_{t,y} & y = y_r \end{cases} \quad (10)$$

where γ_t and γ_r are both constant rates that are positive and lower than 1, one for learning, and the other one for penalty, with $\gamma_r \leq \gamma_t$; $\eta_{t,y}$ is made up of the variant part of $\gamma_{t,y}$ with regard to the input x_t and the class label y , and can be simply implemented as follows:

$$\eta_{t,y} = \begin{cases} 2d_{t,y} / d_t & y = y_t \\ (d_{t,y_t} + d_{y_t,y}) / d_t & y = y_r \end{cases} \quad (11)$$

Here d_t takes the sum form, $d_t = d_{t,y_t} + d_{t,y_r} + d_{y_t,y_r}$, where d_{t,y_t} , d_{t,y_r} and d_{y_t,y_r} all refer to the similarity measures with regard to the relevant variables: $d_{t,y} = \{D(x_t, c_y) \,|\, y \in Y_t\}$ and $d_{y_t,y_r} = D(c_{y_t}, c_{y_r})$.

Every input will in turn be used to update the mean of the class as follows:

$$c_y^{new} = c_y^{old} + \Delta c_y, \quad \Delta c_y = \gamma_{t,y}(x_t - c_y^{old}) \qquad (12)$$

The steps are repeated until one of the two following conditions is fulfilled: either each extra mixture proportion α_y is pushed towards zero (and consequently each extra mean c_y is pushed far away from the data), or if the classification results remains roughly fixed for all inputs.

4. Simulation experiment

Simulation experiments were carried out to verify the performance of the proposed unsupervised classification framework. The experimental dataset consists of a set of samples drawn from a mixture of no more than five Gaussian distributions with different locations, mixture proportions and degrees of overlap among classes inside the [-1, 1] domain in a 2-dimensional space. Some examples of datasets are shown in Figure 1.

Given a hypothetical number of classes larger than the original number of mixtures, both K-means and the proposed mechanism were respectively employed for unsupervised classification.

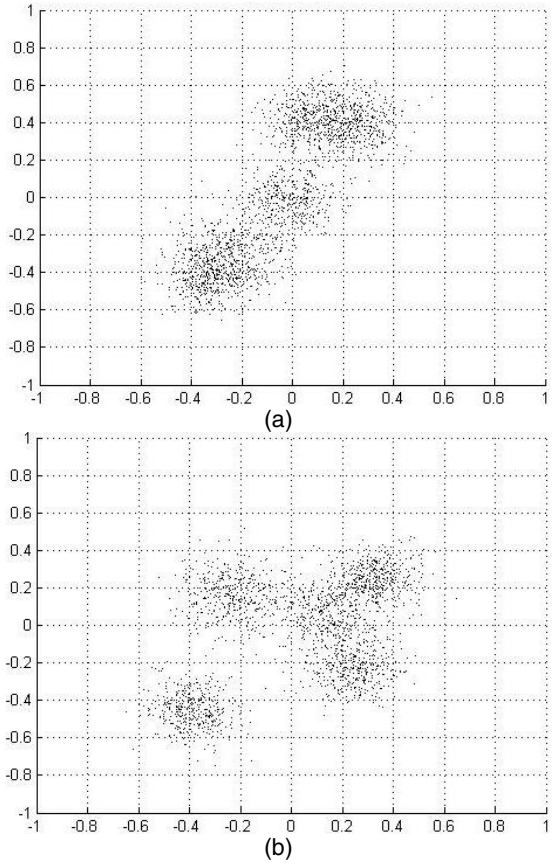
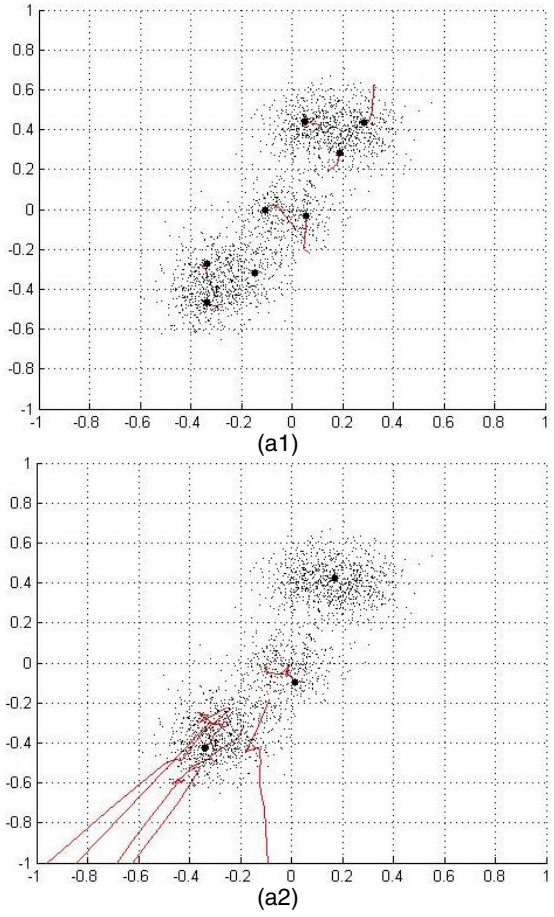


Figure 1. Dataset examples

5. Result analysis

Starting from a too large number of classes (set to eight here), the classification performances as well as the paths of centers in both K-means and the proposed algorithm for the above example databases are shown as Figure 2. Figure 2 (a1) and (b1) are the results of the K-means algorithm, and (a2) and (b2) are the results of the classification mechanism proposed in this paper; (a1) and (a2) refer to dataset (a), and (b1) and (b2) to dataset (b). The unsupervised classification framework proposed in this paper earned activation for the winners and assigned penalty for their rivals, so that the winners concentrate more around the natural centers of the classes and their rivals are driven far away from the datasets. Samples from unknown classes are then assigned to the most competitive classes, whose centers are representatives of the datasets. The effective number of classes could be easily observed, while the extra ones can be identified and removed after or even during learning. On the contrary, the K-means algorithm maintains the originally given number of classes, some of them turning out to be meaningless at the end if the correct number was not guessed before learning.



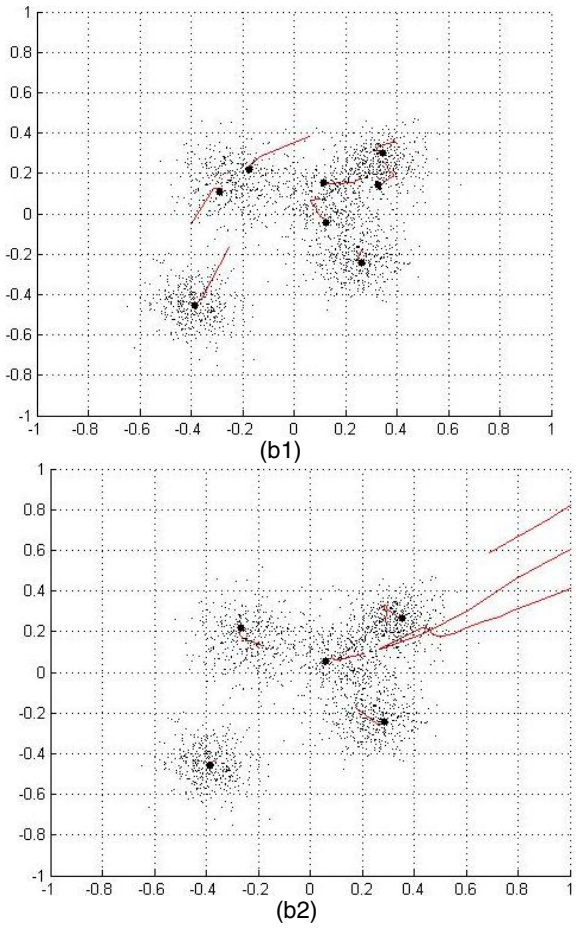


Figure 2. Paths of class centers during learning. The final position is identified by a dot. Figures (a1) and (b1) show the results of the K-means algorithm; (a2) and (b2) show the results of the classification mechanism proposed in this paper. (a1) and (a2) correspond to dataset (a), and (b1) and (b2) to dataset (b) in Figure 1.

6. Conclusions

In this paper, an unsupervised classification mechanism for Gaussian mixture models is presented based on a general statistical learning tool, the Bayesian Ying-Yang harmony learning system. The model is specified by probability distribution hypotheses, and the learning mechanism is derived from an objective function. The main feature of the model is that it performs parameter learning and model selection in parallel: the proposed classification mechanism roughly decides on the number of real classes, prompts the winner by activation and obstructs its rival by penalty, so that the most competitive center wins for possible prediction and the extra ones are driven far away from the distribution. The only prerequisite is to start with a number of classes that exceeds the natural number of classes in the data. Simulation experiments achieve good performance for the unsupervised classification of two sample datasets, and show how the number of effective classes is automatically extracted during learning.

7. Acknowledgements

Rui Nian is sponsored by the China Scholarship Council for her overseas study in Belgium. This research was fully supported by the Natural Science Foundation of China (60572064).

8. References

[1] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, Berlin, 1995.

[2] L. Xu, " Bayesian Ying Yang harmony learning" , *The handbook of brain theory and neural networks*, Arbib, M.A., Cambridge, MA, the MIT Press, 2002.

[3] J. A. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", *Technical Report, University of Berkeley,* ICSI-TR-97-021, 1997.

[4] S.Z.Selim, M.A.Ismail, "*K*-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality", *IEEE Trans. On PAMI-6*, 1, 1984.

[5] B. Zhang, "Comparison of the Performance of Center-Based Clustering Algorithms", *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science*, Volume 2637/2003, pp. 569, 2003.

[6] L. Xu, A. Krzyzak, E. Oja, "Rival Penalized Competitive Learning for Clustering Analysis, RBF Net, and Curve Detection Bayesian Ying Yang harmony learning", *IEEE Trans. Neural Networks*, 4, pp. 636-649, 1993.