

A MIXTURE MODEL APPROACH FOR ON-LINE CLUSTERING

Allou Samé, Christophe Ambroise and Gérard Govaert

Key words: EM algorithm, CEM algorithm, stochastic gradient, ICL criterion.

COMPSTAT 2004 section: Clustering.

Abstract: This article presents an original on-line algorithm dedicated to mixture model based clustering. The proposed algorithm is a stochastic gradient ascent which maximizes the expectation of the classification likelihood. This approach requires few calculations and exhibits a quick convergence. A strategy for choosing the optimal number of classes using the Integrated Classification Likelihood (ICL) is studied using simulated data. The results of the simulations show that the proposed method provides a fast and accurate estimation of the parameters (including the number of classes) when the mixture components are relatively well separated.

1 Introduction

Generally, stochastic gradient algorithms are used for on-line parameter estimation in signal processing and pattern recognition for their algorithmic simplicity. They have been shown to be faster than standard algorithms. In clustering, MacQueen on-line kmeans algorithm [8] is the one commonly used.

In the context of a flaw detection problem using acoustic emission, we have been brought to classify under real time constraints a set of points located in a plane. The solution provided by the so-called CEM algorithm [4] applied using a gaussian mixture model provides a satisfactory solution for this problem and is faster than the EM algorithm [5] one's. However, in spite of its speed, CEM algorithm is not able to react in real time when the number of acoustic emissions becomes too large (more than 10000 points). In this work, we aim to develop an on-line mixture model based clustering algorithm which also allows us to choose the appropriate number of classes.

Let us suppose that data are independent observations $\mathbf{x}_1, \dots, \mathbf{x}_n, \dots$ which are sequentially received and distributed following a mixture density of K components, defined on \mathbb{R}^p by

$$f(\mathbf{x}; \Phi) = \sum_{k=1}^K \pi_k f_k(\mathbf{x}; \theta_k),$$

with $\Phi = (\pi_1, \dots, \pi_K, \theta_1, \dots, \theta_K)$ where π_1, \dots, π_K denote the proportions of the mixture and $\theta_1, \dots, \theta_K$ the parameters of each density component.

We denote by z_1, \dots, z_n, \dots the classes associated to the observations, where $z_n \in \{1, \dots, K\}$ corresponds to the class of \mathbf{x}_n .

To estimate the parameter Φ , we choose to use a stochastic gradient algorithm. These algorithms generally allow to optimize the expectation of a criterion [2, 3]

$$C(\Phi) = E[J(\mathbf{x}, \Phi)],$$

where the expectation is computed using the unknown true parameter of the distribution function f . The criterion $J(\mathbf{x}, \Phi)$ measures the quality of the parameter Φ given the observation \mathbf{x} . The stochastic gradient algorithm aiming to maximize the criterion C is then written

$$\Phi^{(n+1)} = \Phi^{(n)} + \alpha_n \nabla_{\Phi} J(\mathbf{x}_{n+1}, \Phi^{(n)}) \quad (1)$$

where the learning rate α_n is a positive scalar or a positive definite matrix such that $\sum |\alpha_n| = \infty$ and $\sum |\alpha_n|^2 < \infty$.

In the second section, we present the Titterington on-line clustering approach [6]; the third section is devoted to the new stochastic gradient algorithm that we propose for on-line clustering; an experimental study is summarized in the fourth section.

2 Stochastic gradient algorithm derived from EM algorithm

Given the observed data $\mathbf{x}_n = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ and some initial parameter $\Phi^{(0)}$, the standard EM algorithm [5] maximizes the log-likelihood $\log p(\mathbf{x}_n; \Phi)$ by maximizing iteratively the expectation of the complete data conditionally to the available data:

$$Q(\Phi, \Phi^{(a)}) = E[\log p(\mathbf{x}_n, \mathbf{z}_n; \Phi) | \mathbf{x}_n, \Phi^{(a)}]$$

where $\mathbf{z}_n = (z_1, \dots, z_n)$.

Titterington on-line clustering approach [6] consists in using a special stochastic gradient algorithm which can be derived from the standard EM algorithm. For this purpose, we define in the same way as for the EM algorithm the quantity

$$Q_{\mathbf{x}_{n+1}}(\Phi, \Phi^{(n)}) = E[\log p(\mathbf{x}_{n+1}, \mathbf{z}_{n+1}; \Phi) | \mathbf{x}_{n+1}, \Phi^{(n)}],$$

where, this time, the parameter $\Phi^{(n)}$ has been computed from the observations \mathbf{x}_n . The maximization of $\frac{1}{n+1} Q_{\mathbf{x}_{n+1}}(\cdot, \Phi^{(n)})$ using Newton method after replacing the hessian matrix term by its expectation which is the Fisher information matrix $I_c(\Phi^{(n)})$ associated to one complete observation (\mathbf{x}, z) results in the algorithm proposed by Titterington:

$$\Phi^{(n+1)} = \Phi^{(n)} + \frac{1}{n+1} \left(I_c(\Phi^{(n)}) \right)^{-1} \frac{\partial \log p(\mathbf{x}_{n+1}; \Phi^{(n)})}{\partial \Phi}. \quad (2)$$

Fisher information matrix $I_c(\Phi^{(n)})$ is positive definite for some density families like the regular exponential family and thus Titterington algorithm has the general form (1) of the stochastic gradient algorithms; which guarantees under some conditions [2, 3] that the criterion maximized by (2) is $E[\log p(\mathbf{x}; \Phi)]$.

3 Stochastic gradient algorithm derived from CEM algorithm

The criterion to be maximized in this section, by analogy with the classification likelihood maximized in the CEM algorithm [4] which can also be written $L_C(\Phi) = \max_{z_1, \dots, z_n} \log p(\mathbf{x}_1, \dots, \mathbf{x}_n, z_1, \dots, z_n; \Phi)$, is the expected criterion

$$C(\Phi) = E[\max_{1 \leq z \leq K} \log p(\mathbf{x}, z; \Phi)],$$

where $\log p(\mathbf{x}, z; \Phi)$ is the complete log-likelihood of the parameter Φ given the complete observation (\mathbf{x}, z) .

The application of algorithm (1) needs the gradient of the function $J(\mathbf{x}, \Phi) = \max_{1 \leq z \leq K} \log p(\mathbf{x}, z; \Phi)$ with respect to Φ to be computed. However, this gradient does not exist for some values of \mathbf{x} due to the well known non differentiability of the max function. In this situation which is very common, Bottou [2, 3] shows that it is sufficient to replace this gradient by a function $H(\mathbf{x}, \Phi)$ verifying $E[H(\mathbf{x}, \Phi)] = \nabla_{\Phi} C(\mathbf{x}, \Phi)$ on the one hand, and on the other hand that the functions $H(\mathbf{x}, \Phi)$ and $C(\Phi)$ verify certain conditions.

In the gaussian mixture case, we may consider the function $H(\mathbf{x}, \Phi)$ such that

$$H(\mathbf{x}, \Phi) = \begin{cases} \nabla_{\Phi} J(\mathbf{x}, \Phi) & \text{if } \nabla_{\Phi} J(\mathbf{x}, \Phi) \text{ exists} \\ 0 & \text{otherwise.} \end{cases}$$

The parameters to be updated in the gaussian mixture case are both the proportions π_1, \dots, π_k and the parameters $\theta_k = (\mu_k, \Sigma_k)$ of each gaussian of the mixture. The direct updating rule (1) applied to the proportions does not guarantee in practice that $0 < \pi_k < 1$ and $\sum_{k=1}^K \pi_k = 1$. Therefore to overcome this numerical instability, we use a logit parametrization [7] $w_k = \log \frac{\pi_k}{\pi_K}$. The resulting new variables w_1, \dots, w_{K-1} belong to \mathbb{R} . Finally the CEM stochastic gradient algorithm in the gaussian mixture model case can be defined as follows:

Step 0 initialization of the parameters $\pi_k^{(0)}$, $\mu_k^{(0)}$ and $\Sigma_k^{(0)}$

Step 1 (iteration $n + 1$) assignation of the new observation \mathbf{x}_{n+1} to the class k^* which maximizes the log-likelihood of the current parameter knowing this observation:

$$k^* = \underset{1 \leq k \leq K}{\operatorname{argmax}} \left(\log \pi_k^{(n)} - \frac{1}{2} \log \det(\Sigma_k^{(n)}) - \frac{1}{2} (\mathbf{x}_{n+1} - \boldsymbol{\mu}_k^{(n)})^T \Sigma_k^{(n)-1} (\mathbf{x}_{n+1} - \boldsymbol{\mu}_k^{(n)}) \right)$$

Step 2 (iteration $n + 1$) updating of the parameters:

$$\begin{aligned} w_k^{(n+1)} &= w_k^{(n)} + \alpha_n (z_{n+1,k} - \pi_k^{(n)}) \quad \text{for } k = 1, \dots, K-1 \\ \pi_k^{(n+1)} &= \frac{\exp(w_k^{(n+1)})}{1 + \sum_{\ell=1}^{K-1} \exp(w_\ell^{(n+1)})} \quad \text{for } k = 1, \dots, K-1 \\ \pi_K^{(n+1)} &= \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(w_\ell^{(n+1)})} \\ \boldsymbol{\mu}_k^{(n+1)} &= \boldsymbol{\mu}_k^{(n)} + z_{n+1,k} \alpha_n \Sigma_k^{(n)-1} (\mathbf{x}_{n+1} - \boldsymbol{\mu}_k^{(n)}) \\ \Sigma_k^{(n+1)} &= \Sigma_k^{(n)} + z_{n+1,k} \alpha_n \cdot \\ &\quad \left(\frac{1}{2} \Sigma_k^{(n)-1} ((\mathbf{x}_{n+1} - \boldsymbol{\mu}_k^{(n)}) (\mathbf{x}_{n+1} - \boldsymbol{\mu}_k^{(n)})^T \Sigma_k^{(n)-1} - I) \right) \end{aligned}$$

where $z_{n+1,k}$ equals 0 if $k = k^*$ and 1 otherwise.

Particularly, an algorithm equivalent to MacQueen on-line kmeans algorithm [8] can be recovered if we consider a gaussian mixture with identical proportions and spherical covariance matrices (equal to the identity matrix) with a learning rate $\alpha_n = \frac{1}{n+1}$.

The proposed method for the choice of the number of classes consists in running the CEM stochastic gradient algorithm concurrently for models from 2 to K_{max} number of clusters and selecting the solution which maximizes the integrated classification likelihood criterion (ICL) proposed by Biernacki, Celeux and Govaert [1]. In our situation, the ICL criterion can be written as

$$ICL(m, K) = \log p(\boldsymbol{\Phi}^{(n)}; \mathbf{x}_1, \dots, \mathbf{x}_n, z_1, \dots, z_n) - \frac{\nu_{m,K}}{2} \log(n),$$

where $\boldsymbol{\Phi}^{(n)}$ is the parameter vector obtained at iteration n with the stochastic gradient algorithm, $\mathbf{x}_1, \dots, \mathbf{x}_n$ the data available at time n (or at iteration n), z_1, \dots, z_n the corresponding classes computed by applying the *maximum a posteriori* rule with the parameter $\boldsymbol{\Phi}^{(n)}$ and $\nu_{m,K}$ the number of free parameters of the model. This approach is possible because few calculations are required by the CEM stochastic gradient algorithm and this allows us to compare different runs.

4 Simulations

The adopted strategy for simulations consists in initially drawing n observations according to a mixture of two bi-dimensional gaussian distribution, to apply the standard CEM algorithm on a few points (n_0 points) and finally to apply the CEM stochastic gradient algorithm on the rest of the points. The main parameters which control the simulations are:

- the samples sizes: $n = 100, n = 300, n = 500, n = 1000, n = 3000, n = 5000$;
- the number n_0 of points initially processed with the CEM algorithm: $n_0 = 80$;
- the number of components of the mixture: $K_0 = 4$;
- the overlapping degree between the components of the mixture measured by the theoretical percentage of misclassified points which varies as a function of the distance between the class centers $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\mu}_4$; the retained overlapping degrees are:
 - 5% of theoretical error: $\boldsymbol{\mu}_1 = (0; 0), \boldsymbol{\mu}_2 = (4; 0), \boldsymbol{\mu}_3 = (0; 4), \boldsymbol{\mu}_4 = (4; 4)$
 - 14% of theoretical error: $\boldsymbol{\mu}_1 = (0; 0), \boldsymbol{\mu}_2 = (2.5; 0), \boldsymbol{\mu}_3 = (0; 2.5), \boldsymbol{\mu}_4 = (2.5; 2.5)$
 - 20% of theoretical error: $\boldsymbol{\mu}_1 = (0; 0), \boldsymbol{\mu}_2 = (2.2; 0), \boldsymbol{\mu}_3 = (0; 2.2), \boldsymbol{\mu}_4 = (2.2; 2.2)$
- the mixture proportions chosen equal: $\pi_k = \frac{1}{4}$ for $k = 1, \dots, 4$;
- the variance matrices fixed to the identity matrix.

In order to obtain optimal results, we have chosen the learning rate $\alpha_n = \frac{1}{0.3n}$. The maximal number of classes considered has been fixed to $K_{max} = 7$.

Table 1 presents ICL criterion as a function of the number K of classes taken into account by the CEM stochastic gradient algorithm and the sample size n , for an overlap leading to 14% of theoretical error. We observe in this situation that the number of classes found by our method, that is the one for which the ICL criterion is greater, corresponds to the true simulated number of clusters which is 4 clusters. This observed behavior is the same, even for small values of n (100, 300). The situation corresponding to 5% of theoretical error gives also good results. However, the true number of classes is not recovered in the situation leading to 20% of theoretical error (see table 2), even for the relatively large sample sizes n (3000, 5000). This behavior is not surprising because CEM algorithm is known to provide biased estimations when the classes are not well separated.

	$n = 100$	$n = 300$	$n = 500$	$n = 1000$	$n = 3000$	$n = 5000$
$K = 2$	-0.0447	-0.1303	-0.2135	-0.4230	-1.2705	-2.1175
$K = 3$	-0.0440	-0.1275	-0.2115	-0.4190	-1.2570	-2.0925
$K = 4$	-0.0437	-0.1250	-0.2052	-0.4065	-1.2135	-2.0175
$K = 5$	-0.0459	-0.1298	-0.2115	-0.4185	-1.2465	-2.0600
$K = 6$	-0.0464	-0.1366	-0.2153	-0.4255	-1.2870	-2.1575
$K = 7$	-0.0487	-0.1335	-0.2180	-0.4325	-1.3200	-2.1175

Table 1: ICL criteria (divided by 10^4) as a function of the number of classes K and the sample sizes n , for an overlapping leading to 14% of theoretical error.

	$n = 100$	$n = 300$	$n = 500$	$n = 1000$	$n = 3000$	$n = 5000$
$K = 2$	-0.0415	-0.1205	-0.2023	-0.4050	-1.2075	-2.0200
$K = 3$	-0.0418	-0.1208	-0.2033	-0.4045	-1.2015	-2.0025
$K = 4$	-0.0430	-0.1263	-0.2122	-0.4165	-1.2480	-2.0950
$K = 5$	-0.0422	-0.1209	-0.2035	-0.4015	-1.2045	-1.9975
$K = 6$	-0.0457	-0.1266	-0.2120	-0.4155	-1.2180	-2.0900
$K = 7$	-0.0449	-0.1286	-0.2077	-0.4195	-1.2900	-2.0300

Table 2: ICL criteria (divided by 10^4) as a function of the number of classes K and the sample sizes n , for an overlapping leading to 20% of theoretical error.

5 Conclusion

This paper proposes an on-line estimation of a mixture model parameters. The proposed stochastic gradient algorithm is a generalization of the on-line kmeans algorithm introduced by MacQueen [8]. The few required computations allows several models to be estimated concurrently, defining an inexpensive strategy of model choice.

Although the proposed method provides reasonably good results, the convergence analysis of the CEM stochastic gradient algorithm toward a local maxima of the expected classification likelihood $E[\max_{1 \leq z \leq K} \log p(\mathbf{x}, z; \Phi)]$ is met only under some conditions [2, 3] which are often difficult to prove. The verification of these conditions at least for some particular models remains the main prospect of this work.

References

- [1] C. Biernacki, G. Celeux, and G. Govaert (2000). *Assessing a mixture model for clustering with the integrated completed likelihood*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(7):719-725.
- [2] L. Bottou (1991). *Une approche théorique de l'apprentissage connexionniste; applications à la reconnaissance de la parole*, thèse de doctorat, université d'Orsay.

- [3] L. Bottou (1998). *Online learning and stochastic approximations*. In online learning in neural networks, D. Saad, Ed., Cambridge: Cambridge University Press.
- [4] G. Celeux and G. Govaert (1992). *A classification EM algorithm for clustering and two stochastic versions*. *Computation Statistics and Data Analysis*, 14, 315-332.
- [5] A. P. Dempster, N. M. Laird and D. B. Rubin (1977). *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*. *J. Royal Stat. Soc. B*, 39(1):1-38.
- [6] D.M. Titterton. (1984). *Recursive parameter estimation using incomplete data*. *J. Royal Statist. Soc., B*, vol. 46, pp. 257-267.
- [7] J.-F. Yao (2000). *On recursive estimation of incomplete data models*. *Statistics*, 34(1):27-51.
- [8] MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. In *Proceedings of 5th Berkeley Symposium on Mathematics, Statistics and Probability*, 1:281- 298.

Address: Université de Technologie de Compiègne
HEUDIASYC, UMR CNRS 6599
BP 20529, 60205 Compiègne Cedex, France
E-mail: {same,ambroise,govaert}@utc.fr