Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Grouped data clustering using a fast mixture-model-based algorithm

Allou SAMÉ
Laboratoire des Technologies Nouvelles
Institut National de Recherche sur les Transports et leur Sécurité
Noisy-le-Grand, France
same@inrets.fr

***Abstract*—Mixture-model-based clustering has become a popular approach in many data analysis problems for its statistical properties and the implementation simplicity of the EM algorithm. However the computation time of the EM algorithm and its variants increases significantly with the sample size. For large data sets, performing clustering on grouped data constitutes an efficient alternative to speed up the algorithms execution time. A rapid and effective algorithm dedicated to grouped data clustering is then proposed in this paper. Inspired by the Classification EM algorithm (CEM), the proposed approach estimates the missing sample at each iteration. An experimental study using simulated data and real acoustic emission data in the context of a flaw detection application on gas tanks reveals good performances of the proposed approach in terms of partitioning precision and computing time.**

***Index Terms*—Clustering, mixture models, EM algorithm, Classification EM algorithm, grouped Data, histogram**

I. Introduction

Clustering is an exploratory data analysis technique which consists in identifying homogeneous groups in data and thus constitutes a central problematic in many applications including web data mining, bio-informatics and image segmentation. Mixture-model-based clustering [1][6] continues to receive increasing attention for its statistical properties and its implementation simplicity. Its two commonly used approaches [6] for multidimensional data are the maximum likelihood (ML) approach and the classification maximum likelihood (CML) approach. The ML approach consists in estimating the model parameters by maximizing the log-likelihood by the Expectation Maximization (EM) algorithm and then to estimate a partition by the maximum a posteriori (MAP) rule. In the CML approach, the model parameters and the partition are simultaneously estimated by maximizing the classification log-likelihood using the Classification Expectation Maximization (CEM) algorithm [5].

With the improvement of measurement devices and computers, data sets with a large number of observations become frequent and the challenge is to develop clustering methods whose execution time does not depend on the number of observations. However the computation time of the EM and CEM algorithms increases significantly with the sample size. To speed up the EM algorithm, various attempts have been proposed: the incremental and sparse versions proposed by Neal and Hinton [10], the scalable version of the EM algorithm proposed by Bradley et al. [2] to handle large databases with a limited memory buffer, the Moore's approach that summarizes data using the so-called multiresolution *k*d-tree [9], the Ng and McLachlan approach [11] that combines the incremental EM and the multiresolution *k*d-tree approach. Another generic way of improving the clustering algorithms execution time is to operate on grouped data or histograms instead of original observations. This data structure also referred as binned data [4] simply results from the conversion of original observations into counts in disjoined subspaces of the overall observations space. Grouped data can also naturally occur when the measurement devices have finite precision. Figure 1 shows example of grouped data resulting from the discretization of real acoustic emission data. McLachlan and Jones [8], and Cadez et al. [4] have theoretically formulated the problem of estimating a mixture model from grouped data as a missing data problem, and have developed a dedicated EM algorithm. However, the multivariate formulation of this algorithm requires the numerical evaluation of multiple integrals which can be computationally expensive.

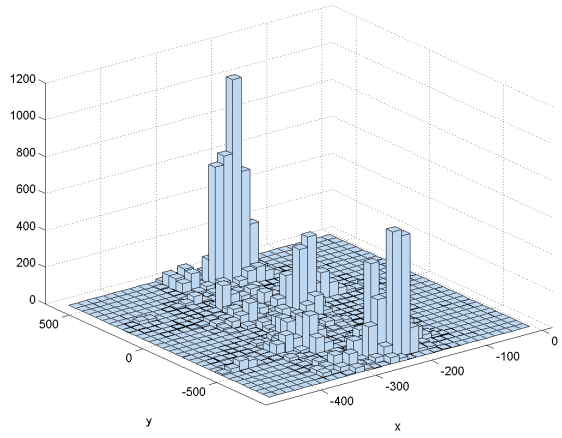

Fig. 1. Grouped data resulting from the discretization of real acoustic emission data

A faster mixture-model-based algorithm for grouped data clustering is presented in this paper. The proposed algorithm is inspired by the classification maximization likelihood approach and incorporates a sample estimation at each iteration.

The section 2 introduces the essential concepts of grouped

data in the framework of mixture models, and briefly recalls the Cadez et al. approach [4] for parameters estimation via the EM algorithm. Our clustering approach adapted to binned data is described in the third section and an experimental study using simulated data and real acoustic emission data is summarized in Section 4.

II. Estimating a mixture model from grouped data

The original data are represented by an independent sample $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ resulting from a mixture density of K components, defined on $\mathbb{R}^p$ by

$$f(\boldsymbol{x}_i; \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k), \tag{1}$$

where $\boldsymbol{\Phi} = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, $\pi_1, \ldots, \pi_K$ denotes the proportions of the mixture and $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ the parameters of each component density. We denote by $\mathbf{z} = (z_1, \ldots, z_n)$ the set of classes associated to the sample $\mathbf{x}$, where $z_i \in \{1, \ldots, K\}$.

In addition, we consider a partition $(H_1, \ldots, H_v)$ of $\mathbb{R}^p$ into v hyperrectangular bins and we assume that the only information we have pertaining to the sample is the set of frequencies n_r of the observations $\boldsymbol{x}_i$ belonging to the bins H_r, summarized by a vector $\mathbf{n} = (n_1, \ldots, n_v)$ of counts, with $\sum_{r=1}^{v} n_r = n$. The observations of bin H_r can then be rewritten as $(\boldsymbol{x}_{rs})_{1 \leq s \leq n_r}$ and their associated classes labels by $(z_{rs})_{1 \leq s \leq n_r}$.

The estimation of the mixture model parameters from grouped data was first introduced by Dempster et al. [7] as a missing data problem in the framework of the EM algorithm, and then developed in the one-dimensional case by McLachlan and Jones [8]. It can be noticed that in this new situation, the missing data are the original observations and their cluster membership. The generalization to the multidimensional case can be attributed to Cadez et al. [4]. To maximize the observed log-likelihood defined by

$$L(\boldsymbol{\Phi}) = \log p(\mathbf{n}; \boldsymbol{\Phi}) = \sum_{r=1}^{v} n_r \log \int_{H_r} f(\boldsymbol{x}; \boldsymbol{\Phi}) d\boldsymbol{x} + c, \tag{2}$$

c being a contant term which does not depend on $\boldsymbol{\Phi}$, the EM algorithm starts from an initial parameter $\boldsymbol{\Phi}^{(0)}$ and then alternates the two following steps until convergence :

- **E-step (Expectation)** Computation of the expectation of the complete data log-likelihood conditionally on the available data and the current parameter vector $\boldsymbol{\Phi}^{(q)}$:

  $$Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(q)}) = E[L_c(\boldsymbol{\Phi}, \mathbf{x}, \mathbf{z}) | \mathbf{n}, \boldsymbol{\Phi}^{(q)}],$$

  where the complete data log-likelihood is defined by

  $$\begin{aligned} L_c(\boldsymbol{\Phi}, \mathbf{x}, \mathbf{z}) &= \log p(\mathbf{n}, \mathbf{x}, \mathbf{z}; \boldsymbol{\Phi}) \\ &= \sum_{k=1}^{K} \sum_{r=1}^{v} \sum_{s=1}^{n_r} z_{rsk} \log \pi_k f_k(\boldsymbol{x}_{rs}; \boldsymbol{\theta}_k). \end{aligned}$$

  with $z_{rsk} = 1$ if $z_{rs} = k$, and 0 otherwise.

- **M-step (Maximization)** Computation of the parameter vector $\boldsymbol{\Phi}^{(q+1)}$ maximizing $Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(q)})$.

Given the parameter vector estimated par the EM algorithm, a partition of the bins can easily be derived by assigning each bin H_r to the cluster which maximizes the posterior probability that an observation $\boldsymbol{x}_i$ of the bin H_r arises from the kth component of the mixture.

The main difficulty when implementing the EM algorithm applied to grouped data, is the numerical evaluation of multiples integrals at each iteration, which can be computationally expensive. The following section introduces a faster classification algorithm which does not require integrals calculations.

III. The proposed classification approach

The clustering method introduced here, in the context of grouped data, is inspired by the Classification EM algorithm (CEM) [5] applied to original observations. The standard CEM algorithm is an iterative clustering algorithm yielding simultaneously the parameters and the classification. It iteratively maximizes, with respect to the component membership vector $\mathbf{z} = (z_1, \ldots, z_n)$ and the parameter vector $\boldsymbol{\Phi}$, the classification log-likelihood criterion. The CEM algorithm also can be interpreted as a classification version of the EM algorithm [5].

In the framework of grouped data, we propose an adaptation of the CEM algorithm which consists in estimating simultaneously the missing data $(\mathbf{x}, \mathbf{z})$ and the parameter vector $\boldsymbol{\Phi}$ by maximizing the complete data log-likelihood which can be rewritten as

$$L_c(\boldsymbol{\Phi}, \mathbf{x}, \mathbf{z}) = \sum_{r=1}^{v} \sum_{s=1}^{n_r} \log \pi_{z_{rs}} f_{z_{rs}}(\boldsymbol{x}_{rs}; \boldsymbol{\theta}_{z_{rs}}). \tag{3}$$

The proposed algorithm starts with an initial parameter $\boldsymbol{\Phi}^{(0)}$ and then alternates, at the qth iteration, the two following steps until convergence:

- **Step 1** Computation of

  $$(\mathbf{x}^{(q+1)}, \mathbf{z}^{(q+1)}) = \arg\max_{(\mathbf{x}, \mathbf{z})} L_c(\boldsymbol{\Phi}^{(q)}, \mathbf{x}, \mathbf{z}) \tag{4}$$

- **Step 2** Computation of

  $$\boldsymbol{\Phi}^{(q+1)} = \arg\max_{\boldsymbol{\Phi}} L_c(\boldsymbol{\Phi}, \mathbf{x}^{(q+1)}, \mathbf{z}^{(q+1)}) \tag{5}$$

As in the classical version of CEM applied to original observations, it can easily be proved that each iteration of this algorithms increases the complete data log-likelihood criterion and that the convergence is observed after a finite number of iterations.

The following sections detail the two steps of this algorithm under the assumption of a Gaussian mixture with diagonal covariance matrices which is a realistic hypothesis for many real applications and more particularly for the application that will be introduced in section 4.

A. Step 1: maximization of the complete data log-likelihood with respect to the sample and the partition

Under the Gaussian mixture assumption, the maximization of L_c with respect to $(\mathbf{x}, \mathbf{z})$ is equivalent to the minimization

of

$$h_1(\mathbf{x},\mathbf{z}) = \sum_{r=1}^{v}\sum_{s=1}^{n_r} \log|\mathbf{\Sigma}_{z_{rs}}^{(q)}| - 2\log\pi_{z_{rs}}^{(q)} + (\boldsymbol{x}_{rs}-\boldsymbol{\mu}_{z_{rs}}^{(q)})^T\mathbf{\Sigma}_{z_{rs}}^{-1}(\boldsymbol{x}_{rs}-\boldsymbol{\mu}_{z_{rs}}^{(q)}), \quad (6)$$

where the π_k, $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$ are respectively the proportions, mean vectors and covariance matrices of the mixture. Minimizing h_1 is equivalent to minimize, for all r and s,

$$h_2(\boldsymbol{x}_{rs},z_{rs}) = \log|\mathbf{\Sigma}_{z_{rs}}^{(q)}| - 2\log\pi_{z_{rs}}^{(q)} + (\boldsymbol{x}_{rs}-\boldsymbol{\mu}_{z_{rs}}^{(q)})^T\mathbf{\Sigma}_{z_{rs}}^{-1}(\boldsymbol{x}_{rs}-\boldsymbol{\mu}_{z_{rs}}^{(q)}) \quad (7)$$

under the constraints $\boldsymbol{x}_{rs}\in H_r$ and $z_{rs}\in\{1,\ldots,K\}$. Since the minimization of h_2 depends only on the bin label r, let us denote by $(\boldsymbol{x}_r^{(q+1)},z_r^{(q+1)})$ its solution. This solution can be obtained using the two following steps:

(a) For $k=1\ldots K$, compute

$$\boldsymbol{x}_{rk}^{(q+1)} = \arg\min_{\boldsymbol{x}\in H_r}(\boldsymbol{x}-\boldsymbol{\mu}_k^{(q)})^T{\Sigma_k^{(q)}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k^{(q)});$$

(b) Set $(\boldsymbol{x}_r^{(q+1)},z_r^{(q+1)}) = (\boldsymbol{x}_{rk^*}^{(q+1)},k^*)$, where

$$k^* = \arg\min_k \log|\mathbf{\Sigma}_k^{(q)}| - 2\log\pi_k^{(q)} + (\boldsymbol{x}_{rk}^{(q+1)}-\boldsymbol{\mu}_k^{(q)})^T{\mathbf{\Sigma}_k^{(q)}}^{-1}(\boldsymbol{x}_{rk}^{(q+1)}-\boldsymbol{\mu}_k^{(q)}).$$

The problem (a) is a convex optimization problem with linear constraints. It can be seen that $\boldsymbol{x}_{rk}^{(q+1)}$ remains in the boundary of the bin H_r if $\boldsymbol{\mu}_k^{(q+1)}$ is outside the bin H_r. For example, under the assumption of a 2-dimensional Gaussian mixture distribution with diagonal covariance matrices, if we set $H_r=\prod_{\ell=1}^{2}[a_r^\ell;b_r^\ell[$ and $\boldsymbol{\mu}_k^{(q)}=(\mu_{k1}^{(q)},\mu_{k2}^{(q)})^T$, it can be shown that $\boldsymbol{x}_{rk}^{(q+1)}=(x_1^*,x_2^*)^T$ with

$$x_1^*=\begin{cases} a_r & \text{if } \mu_{k1}^{(q)}<a_r \\ \mu_{k1}^{(q)} & \text{if } a_r\le\mu_{k1}^{(q)}\le b_r \\ b_r & \text{if } \mu_{k1}^{(q)}>b_r \end{cases} \quad x_2^*=\begin{cases} c_r & \text{if } \mu_{k2}^{(q)}<c_r \\ \mu_{k2}^{(q)} & \text{if } c_r\le\mu_{k2}^{(q)}\le d_r \\ d_r & \text{if } \mu_{k2}^{(q)}>d_r \end{cases}.$$

Figure 2 shows examples of locations of $\boldsymbol{x}_{rk}$ in relation to the current value of $\boldsymbol{\mu}_k$.

B. Step 2: maximization of the complete data log-likelihood with respect to the parameter vector

This step is similar to the M step of the standard CEM algorithm [5] applied to the sample $(\boldsymbol{x}_1^{(q+1)},\ldots,\boldsymbol{x}_v^{(q+1)})$ weighted by the frequencies $(n_1,\ldots,n_v)$. The proportions, mean vectors and covariance matrices which maximizes $L_c(\mathbf{\Phi},\mathbf{x}^{(q+1)},\mathbf{z}^{(q+1)})$ are given by

$$\pi_k^{(q+1)} = \frac{\sum_{r=1}^{v} n_r z_{rk}^{(q+1)}}{n}$$

$$\boldsymbol{\mu}_k^{(q+1)} = \frac{\sum_{r=1}^{v} n_r z_{rk}^{(q+1)}\boldsymbol{x}_r^{(q+1)}}{\sum_{r=1}^{v} n_r z_{rk}^{(q+1)}}$$

$$\mathbf{\Sigma}_k^{(q+1)} = \frac{diag\left(\sum_{r=1}^{v} n_r z_{rk}^{(q+1)}(\boldsymbol{x}_r^{(q+1)}-\boldsymbol{\mu}_k^{(q+1)})(\boldsymbol{x}_r^{(q+1)}-\boldsymbol{\mu}_k^{(q+1)})^T\right)}{\sum_{r=1}^{v} n_r z_{rk}^{(q+1)}}$$

where $z_{rk}=1$ if $z_r=k$, and 0 otherwise, and $diag(A)$ is the diagonal covariance matrix whose diagonal elements are those of matrix A.

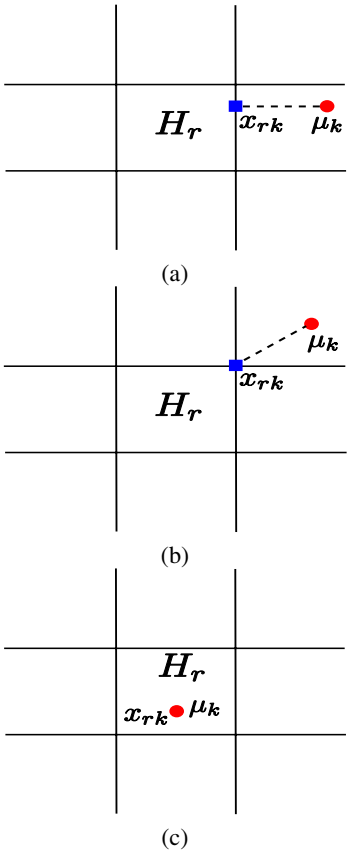

Fig. 2. Exemples of localisations of $\boldsymbol{x}_{rk}$: for cases (a) and (b) the Gaussian mean is outside the bin H_r and for case (c) the Gaussian mean is inside the bin

IV. Experimental study

This section evaluates the proposed algorithm, that we shall call Bin-CEM, in terms of precision and computing time using simulated data and real world data.

A. Experiments using simulated data

The protocol of the simulations is as follows: n observations are generated according to a mixture of 2 bivariate Gaussian densities; these observations are used to compute an histogram containing v bins (grouped data).

Two different mixture models with 15% of theoretical Bayes error rate have been considered:

- mixture model A: $\pi_1=\pi_2=1/2$, $\boldsymbol{\mu}_1=(-2,0)$, $\boldsymbol{\mu}_2=(0,0)$, $\mathbf{\Sigma}_1=\mathbf{\Sigma}_2=diag(1,1)$;
- mixture model B: $\pi_1=\pi_2=1/2$, $\boldsymbol{\mu}_1=(1.6,0)$, $\boldsymbol{\mu}_2=(0,0)$, $\mathbf{\Sigma}_1=diag(1,1/8)$, $\mathbf{\Sigma}_2=diag(1/8,1)$.

Figure 3 shows, examples of simulations of grouped data according to mixtures A and B.

Three algorithm are compared in all these simulations: the standard CEM algorithm applied to the original data (individual observations), the Bin-CEM and Bin-EM (EM applied to grouped data) algorithms applied to grouped data. The

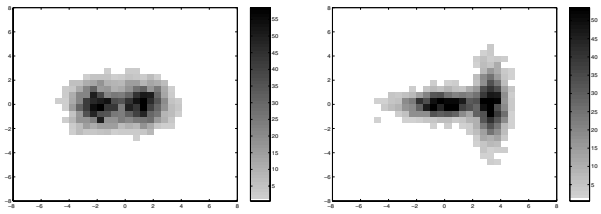

Fig. 3. Example of a simulation of the grouped data according to model A (left) and B (right) at 40×40 bins

quality of a partition provided by each algorithm is measured by computing its misclassification rate with respect to the simulated partition. For each simulated original data set or grouped data set, the misclassification rates and CPU times are averaged over 25 different samples (Monte-Carlo simulation scheme). Figure 4 reports the misclassification rate obtained for mixtures A and B in relation to the number of bins per dimension, when $n = 5000$. We observe a rapid decrease until the number of bins per dimension reaches 40. T Thereafter, the misclassification rate is almost constant and close to that of CEM: the three algorithms become similar.

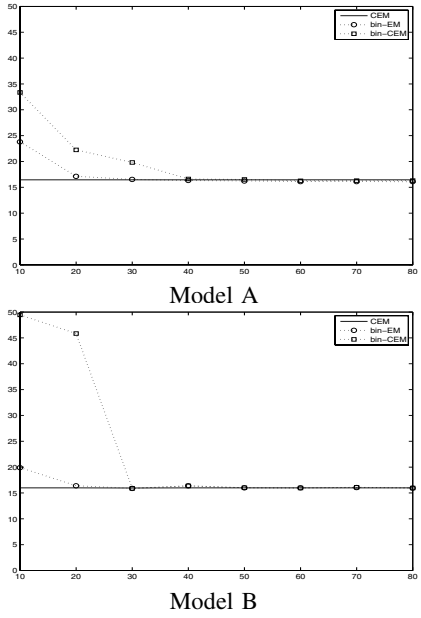

Fig. 4. Misclassification rate (%) in relation to the number of bins per dimension, obtained for $n = 5000$ with CEM (no mark), bin-EM (circle) and bin-CEM (square) for mixtures A (top) and B (bottom)

Figure 5 represents CPU times in second for the three algorithms with respect to the sample size for a fixed number of bins per dimension (40 bins per dimension). The execution time of Bin-CEM and Bin-EM algorithms does not vary while that of the CEM algorithm grows considerably with the sample size. Bin-CEM outperforms the CEM and bin-EM algorithms.

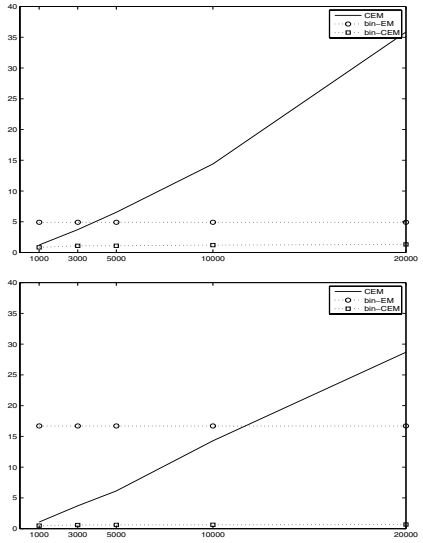

Fig. 5. CPU time in second for the CEM (no mark), bin-EM (circle) and bin-CEM (square) algorithms in relation to the sample size at 40×40 bins for mixture B

B. Experiments using real acoustic emission data

A motivation of this work was to develop a computer-aided decision procedure to assist the detection, in real time, of damaged zones on the surface of a gas tank, through the use of acoustic emissions. Experts are in agreement that spatial concentrations of acoustic emissions, identified using spatial coordinates, are of primary importance in the detection of damaged zones. Other features of acoustic emissions are useful in distinguishing between major and minor flaws. Our method, therefore, consists of two steps :

- Identification of zones where acoustic emissions are concentrated (clustering step) which is the object of the present study;
- Separation of the identified clusters into different categories according to the severity of the imperfection: these categories are termed *minor*, *active* and *critical*. This step is done using standard discrimination methods such as Bayesian discrimination with Gaussian densities.

According to specialists in the field, the method we have described produces satisfactory results using the CEM algorithm for the first step, so long as the number of acoustic emissions does not exceed 10000. When there are more than 10000 emissions the CEM clustering step becomes too slow (more than a few seconds' delay) for a real-time application.

The assumption of Gaussian mixture with diagonal covariance matrices can be attributed to the fact that the damage zones are usually to be found horizontally and vertically along welding lines.

To evaluate our new strategy, we performed a comparison of CEM and Bin-CEM on a real data sets of 18843 acoustic emissions. The number of clusters is selected by maximizing the integrated classification likelihood criterion (ICL) [3]

which is the complete likelihood penalized by the number of free parameters of the mixture model. Both strategies, using CEM and bin-CEM, select a model with 11 clusters. Table I shows the misclassification rate between CEM and Bin-CEM in relation to the number of bins. Figure 6 shows the resulting partitions given by CEM and Bin-CEM at 50 bins per dimension and 80 bins per dimension.

TABLE I
MISCLASSIFICATION RATES BETWEEN CEM AND BIN-CEM FOR REAL ACOUSTIC EMISSION DATA

Number of bins	CEM Vs. Bin-CEM
50×50	5.70%
60×60	5.08%
70×70	3.19%
80×80	2.68%
90×90	2.32%

This shows that results obtained with Bin-CEM are close to those of the CEM algorithm applied to the original sample. In this application, we have therefore achieved our aim of obtaining partitions similar to CEM partitions but in less time.

V. CONCLUSION

A new clustering algorithm for grouped data has been proposed in this paper. The main objective was to obtain a rapid and effective algorithm for clustering low-dimensional massive data sets. The proposed Bin-CEM clustering approach produces partitions from grouped data, which are close to those computed by the CEM algorithm applied to the original observations. Almost no differences in results are observed between CEM and Bin-CEM partitions when the number of bins is sufficiently large (greater than 40 in the simulations we have presented). The execution time of the proposed algorithm is almost constant, since it depends only on the number of bins, while the execution time for CEM increases significantly as the number of available observations increases. The application of the proposed algorithm on real acoustic emission data has also reveal good performances of the proposed algorithm which therefore represents an efficient alternative for clustering large data sets.

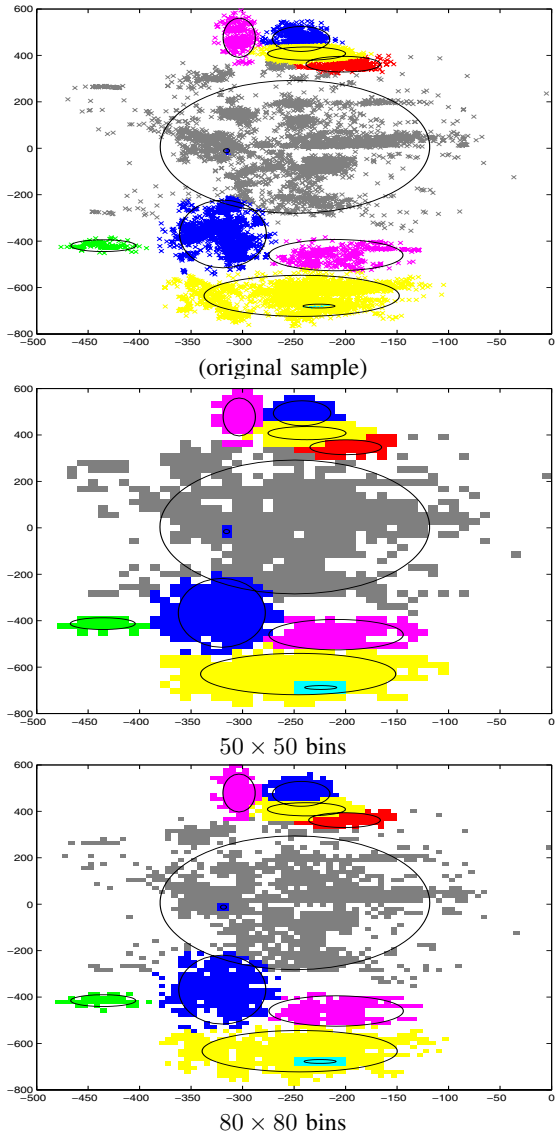

Fig. 6. Results obtained with CEM (top) and Bin-CEM (middle and bottom) on real acoustic emission data

REFERENCES

[1] J.D. Banfield and A.E. Raftery, *Model-based Gaussian and non Gaussian clustering*. Biometrics, 49, 803-821, 2003.

[2] P.S. Bradley, M.U. Fayyad and C.A. Reina, *Scaling EM (Expectation Maximization) clustering to large databases*. Technical Repport MSR-TR-98-35, Microsoft Research, 1998.

[3] C. Biernacki and G. Celeux and G. Govaert, *Assessing a mixture model for clustering with the integrated completed likelihood*. IEEE Transactions on Pattern Analysis and Machine Intelligence. 22, 7, 719-725, 2000.

[4] I.V. Cadez and P. Smyth and G.J. McLachlan and C.E. McLaren, *Maximum likelihood estimation of mixture densities for binned and truncated multivariate data*. Machine Learning, 47, 7-34, 2000.

[5] G. Celeux and G. Govaert, *A classification EM algorithm for clustering and two stochastic versions*. Computational Statistics and Data Analysis, 14, 315-332, 1992.

[6] G. Celeux and G. Govaert, *Gaussian parsimonious clustering models*. Pattern Recognition, 28(5), 781793, 1995.

[7] A.P. Dempster and N.M. Laird and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, Series B, 39(1), 1-38, 1977.

[8] G.J. McLachlan and P.N. Jones, *Fitting mixture models to grouped and truncated data via the EM algorithm*. Biometrics, 44(2), 571-578, 1988.

[9] A. W. Moore, *Very fast EM based mixture model clustering using multiresolution kd-trees*. In Solla M.S. Kearns and D.A Cohn, editors, Advances in Neural Information Processing Systems, 11, 543-549, MIT Press, 1999.

[10] R.M. Neal and Hinton, *A view of the EM algorithm that justifies incremental, sparse and other variants*. In M.I. Jordan, editor, Learning in Graphical Models, 355-368, Kluwer, Dordrecht, 1998

[11] S.-K. Ng and G. J. McLachlan, *On some variants of the EM algorithm for the finite mixture models*. Autrian Journal of Statistics, 32(1-2),143-161, 2003.