CONSENSUS-BASED DISTRIBUTED EXPECTATION-MAXIMIZATION ALGORITHM FOR DENSITY ESTIMATION AND CLASSIFICATION USING WIRELESS SENSOR NETWORKS

*Pedro A. Forero, Alfonso Cano, and Georgios B. Giannakis**

Dept. of ECE, University of Minnesota, 200 Union Street SE, Minneapolis, MN, 55455

ABSTRACT

The present paper develops a decentralized expectation-maximization (EM) algorithm to estimate the parameters of a mixture density model for use in distributed learning tasks performed with data collected at spatially deployed wireless sensors. The E-step in the novel iterative scheme relies on local information available to individual sensors, while during the M-step sensors exchange information only with their one-hop neighbors to reach consensus and eventually percolate the global information needed to estimate the wanted parameters across the wireless sensor network (WSN). Analysis and simulations demonstrate that the resultant consensus-based distributed EM (CB-DEM) algorithm matches well the resource-limited characteristics of WSNs and compares favorably with existing alternatives because it has wider applicability and remains resilient to inter-sensor communication noise.

Index Terms— Expectation-Maximization, Mixture, Distributed Estimation, Sensor Networks , Distributed Consensus

1. INTRODUCTION

Nonlinear maximum-likelihood (ML) and maximum a posteriori (MAP) estimation problems are challenging and abundant in statistical modeling, classification and reconstruction tasks appearing in widespread applications such as computer vision, psychometrics, econometrics and computerized tomography, to name a few. In addition to being sensitive to initialization, gradient-based and Gauss-Newton iterative solvers require first- and second-order derivatives of the ML or MAP objective function, which may be impossible or computationally prohibitive [1, Ch. 7]. However, in cases where the underlying model exhibits a separable structure (e.g., when the underlying probability density function (pdf) comprises a finite mixture of densities), or when there are missing data and/or unobservable parameters whose knowledge could aid the estimation procedure, EM-based iterative estimators have well documented merits because: (i) they are computationally affordable; and (ii) guarantee convergence to a local optimum or a saddle point of the ML or MAP objective function [2].

Except for [3] and [4], existing batch and incremental algorithms of the EM class rely on data that are assumed to be available at a central processing unit. With the advent of WSNs however, there has been a growing interest towards decentralized detection, estimation and classification schemes for use in monitoring, surveillance and distributed learning applications. For such applications, distributed expectation-maximization (DEM) approaches are well motivated especially when the resultant algorithms have wide applicability, yet are simple and adhere to the stringent processing and communication constraints that resource-limited sensors are envisioned to operate with. In the WSN context, [3] has reported an incremental (I-) DEM scheme, while [4] has investigated a gossip-based (G-) DEM alternative. Inter-sensor communication links are assumed noise free and both [3] and [4] are confined to parameter estimation when the data pdf is modeled as a finite mixture of Gaussian pdfs – a case where local estimators are available in simple closed-form expressions and sufficient statistics obtained locally can be updated across sensors.

The present paper develops what is termed consensus-based (CB-) DEM algorithm for nonlinear ML or MAP parameter estimation based on data collected across spatially distributed sensors. The underlying data pdf is modeled as a finite mixture of (not necessarily Gaussian) pdfs, and inter-sensor links can be corrupted by additive (e.g., quantization or receiver) noise. Similar to I-DEM [3], the E-step in CB-DEM relies on local (per sensor) information. The key difference lies in re-formulating the M-step, where the average log-likelihood of EM's "complete data" is maximized. Specifically, CB-DEM decomposes the M-step into a set of local subproblems that individual sensors solve iteratively by communicating with their single-hop neighbors until global consensus is reached across the entire WSN. Such a consensus-based reformulation has been used recently to map centralized estimation criteria (other than EM) to equivalent criteria amenable to distributed implementation; see [5] and the references therein.

Without requiring closed-form expressions of local (i.e., per sensor) estimators in terms of sufficient statistics, the main novelty relative to [3] is that CB-DEM lends itself naturally

*Work in this paper was supported by the USDoD ARO Grant No. W911NF-05-1-0283; and also through collaborative participation in the C&N Consortium sponsored by the U. S. ARL under the CTA Program, Cooperative Agreement DAAD19-01-2-0011. The U. S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

to a general distributed clustering scheme where class-conditional pdfs are even allowed to be non-Gaussian. In addition, CB-DEM relies on "bridge sensors" (to be defined later) which offer a more desirable tradeoff between robustness and overhead when compared to the incremental looping through all nodes required by I-DEM [3], and the redundant communications involved in G-DEM [4].

The rest of the paper is organized as follows. Modeling preliminaries are given in Section 2. The centralized EM is outlined leading to the development of the CB-DEM algorithm in Section 3. Finally, simulated tests and comparisons are presented in Section 4.

2. PRELIMINARIES

Consider a set $\mathcal{J} := \{1, \ldots, J\}$ of J fully-connected sensors, each (say the$j$-th) communicating with single-hop neighbors in its neighborhood $\mathcal{N}_j \in \mathcal{J}$. Supposing that sufficiently powerful error correcting codes are employed, links are considered error free. Node j records N independent data $x_{j,n}$, $n = 1, \cdots, N$, assumed to be drawn randomly from a set $\mathcal{K} := \{1, \ldots, K\}$ of K classes. Let $\mathbf{c}_{j,n}$ denote the class label indicating from which class $x_{j,n}$ is drawn from. For reasons that will become clear later, let $\mathbf{c}_{j,n}$ be a $K \times 1$ vector whose k-th entry $c_{j,n}(k) = 1$ if $x_{j,n}$ was drawn from the k-th class and zero otherwise. We clearly have K possible vectors $\mathbf{c}_{j,n}$, one per class. Data belonging to class k are distributed according to the pdf $f_k\left(x_{j,n}; \boldsymbol{\phi}_k\right)$ parameterized by the deterministic but unknown vector $\boldsymbol{\phi}_k$. Let $\pi_{j,k}$ denote the prior probability that data from the k-th class are drawn, which is also unknown but is allowed to be different from one sensor to another (thus the sub-index j). The pdf of $(x_{j,n}, \mathbf{c}_{j,n})$ which jointly describes a datum and its class label is

$$f\left(x_{j,n}, \mathbf{c}_{j,n}\right) = \prod_{k=1}^{K} \left[\pi_{j,k} f_k(x_{j,n}; \boldsymbol{\phi}_k)\right]^{c_{j,n}(k)} \tag{1}$$

where only one factor in (1) has exponent equal to 1 while all other factors have exponent equal to 0 and thus do not affect the product. Sensors do not know to which class each $x_{j,n}$ belongs to; hence, the only 'observable' random variable is $x_{j,n}$. The pdf of $x_{j,n}$ is obtained by marginalizing $f\left(x_{j,n}, \mathbf{c}_{j,n}\right)$ with respect to $\mathbf{c}_{j,n}$ to obtain

$$\begin{aligned} f\left(x_{j,n}\right) &= \sum_{\mathbf{c}_{j,n}} \prod_{k=1}^{K} \left[\pi_{j,k} f_k(x_{j,n}; \boldsymbol{\phi}_k)\right]^{c_{j,n}(k)} \\ &= \sum_{k=1}^{K} \pi_{j,k} f_k\left(x_{j,n}; \boldsymbol{\phi}_k\right). \end{aligned} \tag{2}$$

Let $\boldsymbol{\theta} := [\boldsymbol{\varphi}^T, \boldsymbol{\pi}^T]^T$ collect both global parameters $\boldsymbol{\varphi} := [\boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_K^T]^T$ and local ones $\boldsymbol{\pi} := [\boldsymbol{\pi}_1^T, \ldots, \boldsymbol{\pi}_J^T]^T$ with $\boldsymbol{\pi}_j := [\pi_{j,1}, \ldots, \pi_{j,K}]^T$. Given data $\mathbf{x} := [x_{1,1}, \cdots, x_{J,N}]^T$ distributed with pdf as in (2) across the WSN, the goal is to have all sensors form an estimate $\hat{\boldsymbol{\theta}}$ of the pdf parameters $\boldsymbol{\theta}$. Once $\hat{\boldsymbol{\theta}}$ is locally available, each sensor can construct a rule for classifying future data, e.g., using the MAP or the Neyman-Pearson criterion.

The (centralized) ML estimator of $\boldsymbol{\theta}$ can be formulated as the solution of the following optimization problem [cf. (2)]

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{ML} = \arg\max_{\boldsymbol{\theta}} & \prod_{n=1}^{N} \prod_{j=1}^{J} \left(\sum_{k=1}^{K} \pi_{j,k} f_k\left(x_{j,n}; \boldsymbol{\phi}_k\right) \right) \\ \text{s.t.} \quad & 0 \leq \pi_{j,k} \leq 1, \qquad \forall j \in \mathcal{J},\ k \in \mathcal{K} \\ & \sum_{k=1}^{K} \pi_{j,k} = 1. \end{aligned} \tag{3}$$

As mentioned in the introduction, the limitations of gradient and Gauss-Newton solvers, and the separable (with respect to k) structure of the objective in (3) motivate the EM approach.

In its centralized form, EM hinges on the idea that, if the class labels $\mathbf{c}_{j,n}$ for each $x_{j,n}$ were known, the ML problem in (3) would be easier. Specifically, instead of the observed $x_{j,k}$, consider the pdf of the partially observed (a.k.a. *complete*) data $(x_{j,n}, \mathbf{c}_{j,n})$, whose log-likelihood is given by

$$L(\mathbf{y}, \boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{j=1}^{J} \sum_{k=1}^{K} c_{j,n}(k) \log\left[\pi_{j,k} f_k\left(x_{j,n}; \boldsymbol{\phi}_k\right)\right] \tag{4}$$

where $\mathbf{y} := [(x_{1,1}, \mathbf{c}_{1,1}), \ldots, (x_{J,N}, \mathbf{c}_{J,N})]^T$ denotes the complete data. The algorithm starts with an initial guess $\boldsymbol{\theta}^{(0)}$ to build an estimate $\hat{c}_{j,n}^{(0)}(k)$ of each class.

Given $\boldsymbol{\theta}^{(i)}$ at iteration i, the E-step estimates class labels as $\hat{c}_{j,n}^{(i)}(k) = \mathbb{E}\{c_{j,n}(k)|\mathbf{x}, \boldsymbol{\theta}^{(i)}\} = P[c_{j,n}(k) = 1|\mathbf{x}, \boldsymbol{\theta}^{(i)}]$, where the last equality holds because $c_{j,n}(k)$ is a Bernoulli random variable. Application of Bayes' rule yields the closed form

$$\hat{c}_{j,n}^{(i)}(k) = \frac{\pi_{j,k}^{(i)} f_k^{(i)}\left(x_{j,n}; \boldsymbol{\phi}_k^{(i)}\right)}{\sum_{l=1}^{K} \pi_{j,l}^{(i)} f_l^{(i)}\left(x_{j,n}; \boldsymbol{\phi}_l^{(i)}\right)}. \tag{5}$$

With $\hat{c}_{j,n}^{(i)}(k)$ available, the M-step finds

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{c}}\left\{ L(\mathbf{y}, \boldsymbol{\theta}) | \mathbf{x}, \boldsymbol{\theta}^{(i)} \right\} \tag{6}$$

which is input to the E-step of iteration $i+1$. Such iterative passes of the E- and M-steps proceed until, for a prescribed tolerance ϵ, the condition $||\boldsymbol{\theta}^{(i+1)} - \boldsymbol{\theta}^{(i)}|| < \epsilon$ is satisfied.

3. THE CB-DEM ALGORITHM

In this section we develop the CB-DEM algorithm as an EM optimization problem that can be solved in a distributed fashion. To this end, it is useful to identify local and global variables needed at the E-step and M-step. The E-step in (5) entails only local information, but requires $\boldsymbol{\theta}^{(i)}$ to be known

from the previous M-step. In addition, the priors $\boldsymbol{\pi}_j$ estimated during the M-step can be locally estimated per iteration as the average of the class labels; i.e.,

$$\hat{\boldsymbol{\pi}}_j^{(i)} = \frac{1}{N}\sum_{n=1}^{N} \hat{\mathbf{c}}_{j,n}^{(i)} \quad \forall j \in \mathcal{J}. \tag{7}$$

Note however, that $\boldsymbol{\varphi}$ is global. To estimate it across all sensors, we will rely on a subset of nodes that we term bridge sensors and denote as $\mathcal{B}$. The *bridge sensors* will impose consensus on the estimated parameters among the set of neighboring (one-hop) sensors. Sufficient conditions to determine the elements of $\mathcal{B}$ are given in [5]. Let $\boldsymbol{\varphi}_j$ denote the estimate of $\boldsymbol{\varphi}$ at sensor j, whose k-th component is $\boldsymbol{\phi}_{j,k}$. Vector $\boldsymbol{\varphi}_j$ together with the local $\boldsymbol{\pi}_j$ comprise a local estimate of the global parameter vector $\boldsymbol{\theta}$. If $\bar{\boldsymbol{\varphi}}_b$ denotes the estimate of $\boldsymbol{\varphi}$ at a bridge sensor b, the CB-DEM estimate can be found as the solution of the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\varphi}_j} \quad & -\sum_{j=1}^{J} \mathbb{E}_{\mathbf{c}} \left\{ L^c(\mathbf{y}, \boldsymbol{\theta}_j) | \mathbf{x}, \boldsymbol{\theta}_j^{(i)} \right\} \\ \text{s.t.} \quad & \boldsymbol{\varphi}_j = \bar{\boldsymbol{\varphi}}_b, \qquad j \in \mathcal{N}_j, \quad b \in \mathcal{B} \\ & 0 \le \pi_{j,k} \le 1, \qquad \forall j \in \mathcal{J},\ k \in \mathcal{K} \\ & \sum_{k=1}^{K} \pi_{j,k} = 1. \end{aligned} \tag{8}$$

The constraint $\boldsymbol{\varphi}_j = \bar{\boldsymbol{\varphi}}_b$ maintains consensus among global parameters across the network ensuring that at the end of the optimization algorithm $\boldsymbol{\varphi}_1 = \cdots = \boldsymbol{\varphi}_J = \boldsymbol{\varphi}$. The optimization problem in (8) not only guarantees global consensus on the optimal solution, but also ensures that the optimal solution coincides with the centralized one [5].

The problem in (8) can be solved in a distributed fashion using the alternating-direction method of multipliers (MoM) [6, Ch. 3]. For that matter, we construct the augmented Lagrangian

$$\begin{aligned} &\mathcal{L}\left(\mathbf{x}, \{\boldsymbol{\varphi}_j\}_{j=1}^{J}, \{\bar{\boldsymbol{\varphi}}_b\}_{b\in\mathcal{B}}\right) = \\ &-\sum_{j=1}^{J}\sum_{n=1}^{N}\sum_{k=1}^{K} \hat{c}_{j,n}(k) \log\left[f_k\left(x_{j,n}; \boldsymbol{\phi}_{j,k}\right)\right] \\ &+\sum_{j=1}^{J}\sum_{b\in\mathcal{B}} \boldsymbol{\lambda}_j^b \left(\boldsymbol{\varphi}_j - \bar{\boldsymbol{\varphi}}_b\right) + \sum_{j=1}^{J}\sum_{b\in\mathcal{B}} \frac{\xi_j}{2} \left\|\boldsymbol{\varphi}_j - \bar{\boldsymbol{\varphi}}_b\right\|_2^2 \end{aligned} \tag{9}$$

where $\boldsymbol{\lambda}_j^b$ denotes the Lagrange multipliers corresponding to the consensus constraint. The positive constants ξ_j are penalty coefficients which can be tuned to tradeoff convergence speed for steady-state error. Notice that the local variables $\boldsymbol{\varphi}_j$, $\boldsymbol{\pi}_j$ and $\boldsymbol{\lambda}_j^b$ are stored per sensor, whereas the consensus variables $\bar{\boldsymbol{\varphi}}_b$ are stored per bridge sensor. By combining the MoM with the EM algorithm, we obtain the following result (proof is omitted due to space limitations).

Proposition 1 *With iteration index i, consider iterations*

$$\boldsymbol{\lambda}_j^{b(i+1)} = \boldsymbol{\lambda}_j^{b(i)} + \xi_j \left(\boldsymbol{\varphi}_j^{(i)} - \bar{\boldsymbol{\varphi}}_b^{(i)}\right) \tag{10}$$

$$\begin{aligned} \boldsymbol{\varphi}_j^{(i+1)} = \arg\min_{\boldsymbol{\varphi}_j} & -\sum_{n=1}^{N}\sum_{k=1}^{K} \hat{c}_{j,n}(k) \log\left[f_k\left(x_{j,n}; \boldsymbol{\phi}_{j,k}\right)\right] \\ & +\sum_{b\in\mathcal{B}_j} \boldsymbol{\lambda}_j^b \left(\boldsymbol{\varphi}_j - \bar{\boldsymbol{\varphi}}_b\right) + \sum_{b\in\mathcal{B}_j} \frac{\xi_j}{2} \left\|\boldsymbol{\varphi}_j - \bar{\boldsymbol{\varphi}}_b\right\|_2^2 \end{aligned} \tag{11}$$

$$\bar{\boldsymbol{\varphi}}_b^{(i+1)} = \sum_{j\in\mathcal{N}_b} \frac{1}{\sum_{\beta\in\mathcal{N}_b}\xi_\beta} \left(\boldsymbol{\lambda}_j^{b(i+1)} + \xi_j \boldsymbol{\varphi}_j^{(i+1)}\right) \tag{12}$$

along with the parameters $c_{j,n}^{(i)}(k)$ and $\boldsymbol{\pi}_j^{(i)}$ estimated as in (5) *and* (7)*, respectively, with $j \in \mathcal{J}$ and $b \in \mathcal{B}$. Initializing with $\boldsymbol{\pi}^{(0)}$, $\{\boldsymbol{\varphi}_j^{(0)}\}_{j=1}^{J}$, $\{\bar{\boldsymbol{\varphi}}_b^{(0)}\}_{b\in\mathcal{B}}$, and $\{\boldsymbol{\lambda}_j^{b(0)}\}_{b\in\mathcal{B}_j}$ the iterates in* (10)-(12) *together with* (5) *and* (7) *converge to either a global maximum, local maximum, or saddle point of the ML function in* (3)*.*

Remark 1 Following the steps in [5], it is possible to prove that Proposition 1 applies even when inter-sensor links are noisy. This extension is ommited due to lack of space, but is tested through simulations in Section 4.

The CB-DEM algorithm can be summarized as follows: at iteration i, the CB-DEM algorithm begins when the bridge sensors transmit their consensus variables $\bar{\boldsymbol{\varphi}}_b$ to all their neighboring nodes. In particular, node j may receive consensus variables from more than one node. Each node j proceeds to calculate locally the current class labels $\hat{c}_{j,n}^{(i)}(k)$ via (5), thus completing the local E-Step for each sample taken per sensor. Next, $\hat{c}_{j,n}^{(i)}(k)$ is used to obtain $\boldsymbol{\pi}_j^{(i+1)}$ via (7). Each of the Lagrange multipliers $\boldsymbol{\lambda}_j^{b(i+1)}$ at node j is updated using (10) and subsequently used to find $\boldsymbol{\varphi}_j^{(i+1)}$ via (11). The sums $\boldsymbol{\lambda}_j^{b(i+1)} + \xi_j \boldsymbol{\varphi}_j^{(i+1)}$ are sent back from each sensor j to their corresponding neighboring bridge nodes where they are used to compute $\bar{\boldsymbol{\varphi}}_b^{(i+1)}$ via (12). Now the bridge sensors have completed the i-th iteration and are ready to transmit again their updated consensus variables to begin iteration $i+1$.

Remark 2 The merits of CB-DEM relative to [3] and [4] are: i) it does not require finding a, preferably minimal, path traversing across all sensors; ii) it is robust to rounding errors and additive noise present in inter-sensor communications (a major limitation of [3] and [4]); iii) it only requires one-hop connectivity among sensors; iv) it allows parallel processing of the data per node; and, v) it does not require $\boldsymbol{\varphi}$ estimators to be available in closed form and thus applies to general mixtures whose class-conditional pdfs are differentiable and (log) concave.

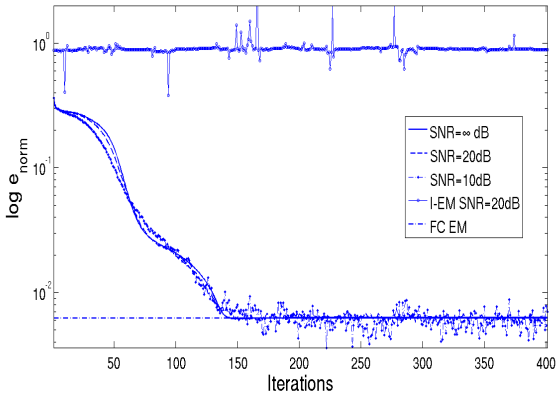

Fig. 1. Error mean comparison among FC-EM, I-EM with SNR=20dB, CB-DEM with SNR=∞, 20dB, and 10dB.

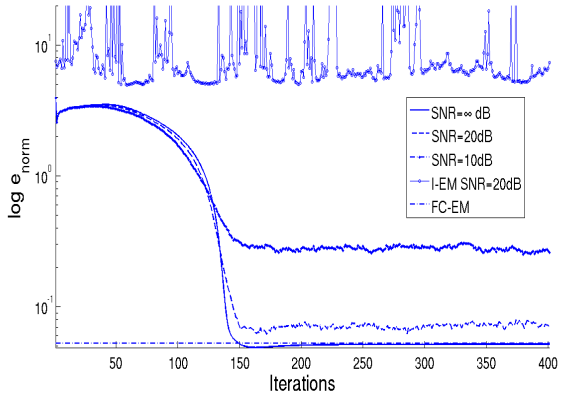

Fig. 2. Error variance comparison among FC-EM, I-EM with SNR=20dB, CB-DEM with SNR=∞, 20dB, and 10dB.

4. SIMULATIONS

In this section, we test convergence of CB-DEM and compare with [3] via simulations. We simulate a WSN with $J = 27$ nodes, each collecting $N = 100$ samples drawn from $K = 3$ classes whose class conditional pdfs are Gaussian with mean and variance $\phi_k := [\mu_k, \sigma_k^2]$, $k = 1, 2, 3$. Zero-mean additive white Gaussian noise is present at every iteration. The signal-to-noise ratio is defined as SNR:=$10\log(\varphi^2_{\text{Tmin}}/\sigma_n^2)$, where φ^2_{Tmin} is the value of the smallest parameter in the true $\boldsymbol{\varphi}$ and σ_n^2 denotes the noise variance. The penalty coefficients are set to $\xi_j = 10$. Parameters $\boldsymbol{\varphi}_j^{(0)}$, $\boldsymbol{\pi}^{(0)}$, and $\bar{\boldsymbol{\varphi}}_b^{(0)}$ are all initialized identically to those in the centralized EM. Notice that for the same initialization, both centralized EM and CB-DEM converge to the same estimate, thus demonstrating that CB-DEM is a valid decentralization of the EM algorithm. As error performance metric, we adopt $e^{(i)}_{\text{norm}} = \sum_{j=1}^{J} ||\hat{\boldsymbol{\gamma}}^{(i)}(j) - \boldsymbol{\gamma}_{\text{true}}||/\,\|\boldsymbol{\gamma}_{\text{true}}\|$, where $\boldsymbol{\gamma}_{\text{true}}$ is the vector of true mean or variance parameters, and $\hat{\boldsymbol{\gamma}}^{(i)}(j)$ denotes the corresponding estimate at node j in iteration i.

Once the parameters describing the class conditional distributions are specified, we construct a MAP classifier to separate new incoming data per sensor. At this point, the classification task can be performed locally. Fig. 3 shows the classification results of a test data set once the parameters describing the class conditional pdfs have been learned. The two-dimensional synthetic data come from three different classes whose class-conditional pdfs are Gaussian. The classification result coincides with that of the centralized MAP classifier since the parameters found by CB-DEM coincide with those of the centralized EM for this specific test.

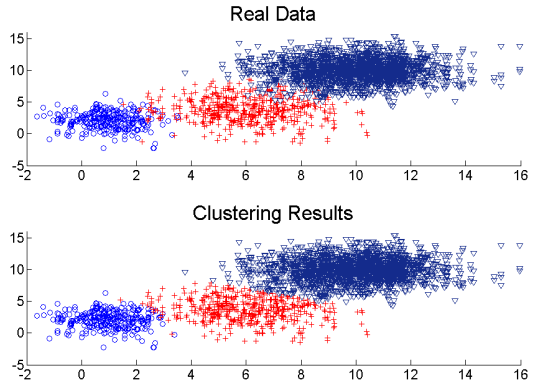

Fig. 3. 2D classification results after CB-DEM for 3 classes with Gaussian class-conditional pdfs.

5. REFERENCES

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, 1993.

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

[3] R. D. Nowak, "Distributed EM algorithms for density estimation and clustering in sensor networks," *IEEE Trans. on Signal Processing*, pp. 2245–2253, Aug. 2003.

[4] W. Kowalczyk and N. Vlassis, "Newscast EM," *Advances in Neural Info. Processing Systems*, pp. 713–720, 2005.

[5] A. Ribeiro, I. D. Schizas and G. B. Giannakis, "Consensus in ad hoc WSNs with noisy links - Part I distributed estimation of deterministic signals," *IEEE Tran. on Signal Proc.*, 2007; http://spincom.ece.umn.edu/papers04/tsp07det-cons.pdf.

[6] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, 1997.