Probabilistic Evidence Accumulation for Clustering Ensembles

André Lourenço[1,2], Samuel Rota Bulò[3], Nicola Rebagliati[3], Ana Fred[2], Mário Figueiredo[2]
and Marcello Pelillo[3]

[1] *Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal*

[2] *Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal*

[3] *DAIS, Università Ca' Foscari Venezia, Venice, Italy*

Keywords: Clustering Algorithm, Clustering Ensembles, Probabilistic Modeling, Evidence Accumulation Clustering.

Abstract: Ensemble clustering methods derive a consensus partition of a set of objects starting from the results of a collection of base clustering algorithms forming the ensemble. Each partition in the ensemble provides a set of pairwise observations of the co-occurrence of objects in a same cluster. The evidence accumulation clustering paradigm uses these co-occurrence statistics to derive a similarity matrix, referred to as co-association matrix, which is fed to a pairwise similarity clustering algorithm to obtain a final consensus clustering. The advantage of this solution is the avoidance of the label correspondence problem, which affects other ensemble clustering schemes. In this paper we derive a principled approach for the extraction of a consensus clustering from the observations encoded in the co-association matrix. We introduce a probabilistic model for the co-association matrix parameterized by the unknown assignments of objects to clusters, which are in turn estimated using a maximum likelihood approach. Additionally, we propose a novel algorithm to carry out the parameter estimation with convergence guarantees towards a local solution. Experiments on both synthetic and real benchmark data show the effectiveness of the proposed approach.

1 INTRODUCTION

Clustering ensemble methods obtain consensus solutions from the results of a set of base clustering algorithms forming the ensemble. Several authors have shown that these methods tend to reveal more robust and stable cluster structures than the individual clusterings in the ensemble (Fred, 2001; Fred and Jain, 2002; Strehl and Ghosh, 2002). The leverage of an ensemble of clusterings is considerably more difficult than combining an ensemble of classifiers, due to the correspondence problem between the cluster labels produced by the different clustering algorithms. This problem is made more serious if additionally clusterings with different numbers of clusters are allowed in the ensemble.

In (Fred, 2001; Fred and Jain, 2002; Fred and Jain, 2005; Strehl and Ghosh, 2002), the clustering ensemble is summarized into a pair-wise *co-association matrix*, where each entry counts the number of clusterings in the ensemble in which a given pair of objects is placed in the same cluster, thus sidestepping the cluster label correspondence problem. This matrix, which is regarded to as a similarty matrix, is then used to feed a pairwise similarity clustering algorithm to deliver the final consensus clustering (Fred and Jain, 2005). The drawback of this approach is that the information about the very nature of the co-association matrix is not properly exploited during the consensus clustering extraction.

A first work in the direction of finding a more principled way of using the information in the co-association matrix is (Rota Bulò et al., 2010). There, the problem of extracting a consensus partition was formulated as a matrix factorization problem, under probability simplex constraints on each column of the factor matrix. Each of these columns can then be interpreted as the multinomial distribution that expresses the probabilities of each object being assigned to each cluster. The drawback of that approach is that the matrix factorization criterion is not supported on any probabilistic estimation rationale.

In this paper we introduce a probabilistic model for the co-association matrix, entitled PEACE - Probabilistic Evidence Accumulation for Clustering Ensembles, whose entries are regarded to as independent observations of binomial random variables counting the number of times two objects occur in a

Lourenço A., Rota Bulò S., Rebagliati N., Fred A., Figueiredo M. and Pelillo M. (2013).
Probabilistic Evidence Accumulation for Clustering Ensembles.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 58-67
DOI: 10.5220/0004267900580067

same cluster. These random variables are indirectly parametrized by the unknown assignments of objects to clusters, which are in turn estimated by adopting a maximum-likelihood approach. This translates into a non-linear optimization problem, which is addressed by means of a primal line-search procedure that guarantees to find a local solution. Experiments on real-world datasets from the UCI machine learning repository, on text-data benchmark datasets as well as on synthetic datasets show the effectiveness of the proposed approach.

The remainder of the paper is organized as follows. In Section 2, we describe our probabilistic model for the co-association matrix and the related maximum-likelihood estimation of the unknown cluster assignments. Section 3 is devoted to solving the optimization problem arising for the unknown cluster assignments estimation. Section 4 contextualizes this model on related work. Finally, Section 5 reports experimental results and Section 6 presents some concluding remarks.

2 PROBABILISTIC MODEL

Let $O = \{1,\dots,n\}$ be the indices of a set of objects to be clustered into K classes and let $\mathcal{E} = \{p_u\}_{u=1}^{N}$ be a clustering ensemble, *i.e.*, a set of N clusterings (partitions) obtained by different algorithms (*e.g.*, different parametrizations and/or initializations) on (possibly) sub-sampled versions of the object set. Each clustering $p_u \in \mathcal{E}$ is a function $p_u : O_u \to \{1,...,K_u\}$, where $O_u \subseteq O$ is a sub-sample of O used as input to the uth clustering algorithm, and K_u is the corresponding number of clusters, which can be different on each $p_u \in \mathcal{E}$. Let $\Omega_{ij} \subseteq \{1,...,N\}$ denote the set of clustering indices where both objects i and j have been clustered, *i.e.* , $(u \in \Omega) \Leftrightarrow \big((i \in O_u) \wedge (j \in O_u)\big)$, and let $N_{ij} = |\Omega_{ij}|$ be its cardinality. The ensemble of clusterings is summarized in the co-association matrix $\mathbf{C} = [c_{ij}] \in \{0,...,N\}^{n\times n}$. Each entry c_{ij} of this matrix having $i \neq j$ counts the number of times objects i and j are observed as clustered together in the ensemble $\mathcal{E}$, *i.e.*

$$c_{ij} = \sum_{l\in\Omega_{ij}} \mathbb{1}[p_l(i) = p_l(j)]$$

where $\mathbb{1}[\cdot]$ is an indicator function returning 1 or 0 according to whether the condition given as argument is true or false. Of course, $c_{ij} \in \{0,...,N_{ij}\}$.

Our basic assumption is that each object has an (unknown) probability of being assigned to each cluster independently of other objects. We denote by $\mathbf{y}_i = (y_{1i},...,y_{Ki})^\top$ the probability distribution over the set of class labels $\{1,\dots,K\}$, that is $y_{ki} = \mathbb{P}[i \in C_k]$, where C_k denotes the subset of O that constitutes the kth cluster. Of course, $\mathbf{y}_i$ belongs to the probability simplex $\Delta_K = \{\mathbf{x} \in \mathbb{R}_+^K : \sum_{j=1}^K x_j = 1\}$. Finally, we collect all the $\mathbf{y}_i$'s in a $K \times n$ matrix $\mathbf{Y} = [\mathbf{y}_1,\dots,\mathbf{y}_n] \in \Delta_K^n$.

In our model, the probability that objects i and j are co-clustered is

$$\sum_{k=1}^{K} \mathbb{P}[i \in C_k, j \in C_k] = \sum_{k=1}^{K} y_{ki}y_{kj} = \mathbf{y}_i^\top \mathbf{y}_j$$

Let C_{ij} be a Binomial random variable representing the number of times that objects i and j are co-clustered; from the assumptions above, we have that $C_{ij} \sim \text{Binomial}\big(N_{ij}, \mathbf{y}_i^\top \mathbf{y}_j\big)$, that is,

$$\mathbb{P}\big[C_{ij} = c|\mathbf{y}_i,\mathbf{y}_j\big] = \binom{N_{ij}}{c}\big(\mathbf{y}_i^\top \mathbf{y}_j\big)^c\big(1-\mathbf{y}_i^\top \mathbf{y}_j\big)^{N_{ij}-c}.$$

Each element c_{ij} of the co-association matrix is interpreted as a sample of the random variable C_{ij}, and the different C_{ij} are all assumed independent. Consequently,

$$\mathbb{P}[\mathbf{C}|\mathbf{Y}] = \prod_{\substack{i,j\in O\\ i\neq j}} \binom{N_{ij}}{c_{ij}} (\mathbf{y}_i^\top \mathbf{y}_j)^{c_{ij}}(1-\mathbf{y}_i^\top \mathbf{y}_j)^{N_{ij}-c_{ij}}.$$

The maximum log-likelihood estimate of $\mathbf{Y}$ is thus

$$\mathbf{Y}^* \in \arg\max_{\mathbf{Y}\in\Delta_K^n} f(\mathbf{Y}) \tag{1}$$

where

$$f(\mathbf{Y}) = \sum_{\substack{i,j\in O\\ i\neq j}} c_{ij}\log\left(\mathbf{y}_i^\top \mathbf{y}_j\right) + (N_{ij}-c_{ij})\log\left(1-\mathbf{y}_i^\top \mathbf{y}_j\right). \tag{2}$$

Hereafter, we use $\log 0 \equiv -\infty$, $0\log 0 \equiv 0$, and denote by $\mathbf{dom}(f) = \{\mathbf{Y} : f(\mathbf{Y}) \neq -\infty\}$ the domain of f.

3 OPTIMIZATION ALGORITHM

The optimization method described in this paper belongs to the class of primal line-search procedures. This method iteratively finds a direction which is *feasible*, *i.e.* satisfying the constraints, and *ascending*, *i.e.* guaranteeing a (local) increase of the objective function, along which a better solution is sought. The procedure is iterated until it converges or a maximum number of iterations is reached.

The first part of this section describes the procedure to determine the search direction in the optimization algorithm. The second part is devoted to determining an optimal step size to be taken in the direction found.

3.1 Computation of a Search Direction

Consider the Lagrangian of (1):

$$\mathcal{L}(\mathbf{Y},\lambda,\mathbf{M}) = f(\mathbf{Y}) + \mathrm{Tr}\left[\mathbf{M}^\top\mathbf{Y}\right] - \lambda^\top(\mathbf{Y}^\top\mathbf{e}_k - \mathbf{e}_n)$$

where $\mathrm{Tr}[\cdot]$ is the matrix trace operator, $\mathbf{e}_k$ is a k-dimensional column vector of all 1s, $\mathbf{Y} \in \mathbf{dom}(f)$ and $\mathbf{M} = (\mu_1,\dots,\mu_n) \in \mathbb{R}_+^{K\times n}$, $\lambda \in \mathbb{R}^n$ are the Lagrangian multipliers. By derivating $\mathcal{L}$ with respect to $\mathbf{y}_i$ and λ and considering the complementary slackness conditions, we obtain the first order Karush-Kuhn-Tucker (KKT) conditions (Luenberger and Ye, 2008) for local optimality:

$$\begin{cases} g_i(\mathbf{Y}) - \lambda_i\mathbf{e}_n + \mu_i &= \mathbf{0}, \quad \forall i \in O \\ \mathbf{Y}^\top\mathbf{e}_K - \mathbf{e}_n &= \mathbf{0} \\ \mathrm{Tr}\left[\mathbf{M}^\top\mathbf{Y}\right] &= 0, \end{cases} \tag{3}$$

where

$$g_i(\mathbf{Y}) = \sum_{j\in O\setminus\{i\}} c_{ij}\frac{\mathbf{y}_j}{\mathbf{y}_i^\top\mathbf{y}_j} - (N_{ij} - c_{ij})\frac{\mathbf{y}_j}{1-\mathbf{y}_i^\top\mathbf{y}_j},$$

and $\mathbf{e}_n$ denotes a n-dimensional column vector of all 1's. We can express the Lagrange multipliers λ in terms of $\mathbf{Y}$ by noting that

$$\mathbf{y}_i^\top\left[g_i(\mathbf{Y}) - \lambda_i\mathbf{e}_n + \mu_i\right] = 0,$$

yields $\lambda_i = \mathbf{y}_i^\top g_i(\mathbf{Y})$ for all $i \in O$.

Let $r_i(\mathbf{Y})$ be given as

$$r_i(\mathbf{Y}) = g_i(\mathbf{Y}) - \lambda_i\mathbf{e}_K = g_i(\mathbf{Y}) - \mathbf{y}_i^\top g_i(\mathbf{Y})\mathbf{e}_K,$$

and let $\sigma(\mathbf{y}_i)$ denote the support of $\mathbf{y}_i$, *i.e.* the set of indices corresponding to (strictly) positive entries of $\mathbf{y}_i$. An alternative characterization of the KKT conditions, where the Lagrange multipliers do not appear, is

$$\begin{cases} [r_i(\mathbf{Y})]_k = 0, & \forall i \in O, \forall k \in \sigma(\mathbf{y}_i), \\ [r_i(\mathbf{Y})]_k \le 0, & \forall i \in O, \forall k \notin \sigma(\mathbf{y}_i), \\ \mathbf{Y}^\top\mathbf{e}_K - \mathbf{e}_n = \mathbf{0}. \end{cases} \tag{4}$$

The two characterizations (4) and (3) are equivalent. This can be verified by exploiting the non negativity of both matrices $\mathbf{M}$ and $\mathbf{Y}$, and the complementary slackness conditions.

The following proposition plays an important role in the selection of the search direction.

Proposition 1. *Assume* $\mathbf{Y} \in \mathbf{dom}(f)$ *to be feasible for* (1), i.e. $\mathbf{Y} \in \Delta_K^n \cap \mathbf{dom}(f)$. *Consider*

$$J \in \arg\max_{i\in O}\left\{[g_i(\mathbf{Y})]_{U_i} - [g_i(\mathbf{Y})]_{V_i}\right\},$$

where

$$U_i \in \arg\max_{k\in\{1\dots K\}}[g_i(\mathbf{Y})]_k \quad \text{and}$$

$$V_i \in \arg\min_{k\in\sigma(\mathbf{y}_j)}[g_i(\mathbf{Y})]_k\,.$$

Let $U = U_J$ *and* $V = V_J$. *Then the following holds:*

- $[g_J(\mathbf{Y})]_U \ge [g_J(\mathbf{Y})]_V$ *and*
- $\mathbf{Y}$ *satisfies the KKT conditions for* (1) *if and only if* $[g_J(\mathbf{Y})]_U = [g_J(\mathbf{Y})]_V$.

Proof. We prove the first point by simple derivations as follows:

$$[g_J(\mathbf{Y})]_U \ge \mathbf{y}_J^\top g_J(\mathbf{Y}) = \sum_{k\in\sigma(\mathbf{y}_J)} y_{kJ}[g_J(\mathbf{Y})]_k$$
$$\ge \sum_{k\in\sigma(\mathbf{y}_J)} y_{kJ}[g_J(\mathbf{Y})]_V = [g_J(\mathbf{Y})]_V\,.$$

By subtracting $\mathbf{y}_J^\top g_J(\mathbf{Y})$ we obtain the equivalent relation

$$[r_J(\mathbf{Y})]_U \ge 0 \ge [r_J(\mathbf{Y})]_V\,, \tag{5}$$

where equality holds if and only if $[g_J(\mathbf{Y})]_V = [g_J(\mathbf{Y})]_U$.

As for the second point, assume that $\mathbf{Y}$ satisfies the KKT conditions. Then $[r_J(\mathbf{Y})]_V = 0$ because $V \in \sigma(\mathbf{y}_J)$. It follows by (5) and (4) that also $[r_J(\mathbf{Y})]_U = 0$ and therefore $[g_J(\mathbf{Y})]_V = [g_J(\mathbf{Y})]_U$. On the other hand, if we assume that $[g_J(\mathbf{Y})]_V = [g_J(\mathbf{Y})]_U$ then by (5) and by definition of J we have that $[r_i(\mathbf{Y})]_{U_i} = [r_i(\mathbf{Y})]_{V_i} = 0$ for all $i \in O$. By exploiting the definition of U_i and V_i it is straightforward to verify that $\mathbf{Y}$ satisfies the KKT conditions. □

Given $\mathbf{Y}$ a non-optimal feasible solution of (1), we can determine the indices U, V and J as stated in Proposition 1. The next proposition shows how to build a feasible and ascending search direction by using these indices. Later on, we will point out some desired properties of this search direction. We denote by $\mathbf{e}_n^{(j)}$ the jth column of the n-dimensional identity matrix.

Proposition 2. *Let* $\mathbf{Y} \in \Delta_K^n \cap \mathbf{dom}(f)$ *and assume that the KKT conditions do not hold. Let* $\mathbf{D} = \left(\mathbf{e}_K^{(U)} - \mathbf{e}_K^{(V)}\right)\left(\mathbf{e}_n^{(J)}\right)^\top$, *where* J, U *and* V *are computed as in Proposition 1. Then, for all* $0 \le \varepsilon \le y_{VJ}$, *we have that* $\mathbf{Z}_\varepsilon = \mathbf{Y} + \varepsilon\mathbf{D}$ *belongs to* Δ_K^n, *and for all small enough, positive values of* ε, *we have* $f(\mathbf{Z}_\varepsilon) > f(\mathbf{Y})$.

Proof. Let $\mathbf{Z}_\varepsilon = \mathbf{Y} + \varepsilon\mathbf{D}$. Then for any ε,

$$\mathbf{Z}_\varepsilon^\top\mathbf{e}_K = (\mathbf{Y}+\varepsilon\mathbf{D})^\top\mathbf{e}_K = \mathbf{Y}^\top\mathbf{e}_K + \varepsilon\mathbf{D}^\top\mathbf{e}_K$$
$$= \mathbf{e}_n + \varepsilon\mathbf{e}_n^{(J)}\left(\mathbf{e}_K^{(U)} - \mathbf{e}_K^{(V)}\right)^\top\mathbf{e}_K = \mathbf{e}_n$$

As ε increases, only the (V,J)th entry of $\mathbf{Z}_\varepsilon$, which is given by $y_{VJ} - \varepsilon$, decreases. This entry is nonnegative for all values of ε satisfying $\varepsilon \le \mathbf{y}_{VJ}$. Hence, $\mathbf{Z}_\varepsilon \in \Delta_K^n$ for all positive values of ε not exceeding y_{VJ} as required.

As for the second point, the Taylor expansion of f at $\mathbf{Y}$ gives, for all small enough positive values of ε:

$$\begin{aligned} f(\mathbf{Z}_\varepsilon) - f(\mathbf{Y}) &= \varepsilon \left[\lim_{\varepsilon\to 0} \frac{d}{d\varepsilon} f(\mathbf{Z}_\varepsilon)\right] + O(\varepsilon^2) \\ &= \left(\mathbf{e}_K^{(U)} - \mathbf{e}_K^{(V)}\right)^\top g_J(\mathbf{Y}) + O(\varepsilon^2) > 0 \\ &= [g_J(\mathbf{Y})]_U - [g_J(\mathbf{Y})]_V + O(\varepsilon^2) > 0 \end{aligned}$$

The last inequality derives from Proposition 1 because if Y does not satisfy the KKT conditions then $[g_J(\mathbf{Y})]_U - [g_J(\mathbf{Y})]_V > 0$. □

3.2 Computation of an Optimal Step Size

Proposition 2 provides a direction $\mathbf{D}$ that is both feasible and ascending for $\mathbf{Y}$ with respect to (1). We will now address the problem of determining an optimal step ε^* to be taken along the direction $\mathbf{D}$. This optimal step is given by the following one dimensional optimization problem:

$$\varepsilon^* \in \arg\max_{0\le\varepsilon\le y_{VJ}} f(\mathbf{Z}_\varepsilon)\,, \tag{6}$$

where $\mathbf{Z}_\varepsilon = \mathbf{Y} + \varepsilon\mathbf{D}$. We prove this problem to be concave.

Proposition 3. *The optimization problem in* (6) *is concave.*

Proof. The direction $\mathbf{D}$ is everywhere null except in the Jth column. Since the sum in (2) is taken over all pairs (i, j) such that $i \neq j$ we have that the argument of every log function (which is a concave function) is linear in ε. Concavity is preserved by the composition of concave functions with linear ones and by the sum of concave functions (Boyd and Vandenberghe, 2004). Hence, the maximization problem is concave. □

Let $\rho(\varepsilon')$ denote the first order derivative of f with respect to ε evaluated at ε', *i.e.*

$$\rho(\varepsilon') = \lim_{\varepsilon\to\varepsilon'} \frac{d}{d\varepsilon} f(\mathbf{Z}_\varepsilon) = \left(\mathbf{e}_K^{(U)} - \mathbf{e}_K^{(V)}\right)^\top g_J(\mathbf{Z}_{\varepsilon'})\,.$$

By the concavity of (6) and Kachurovskii's theorem (Kachurovskii, 1960) we derive that ρ is nonincreasing in the interval $0 \le \varepsilon \le y_{VJ}$. Moreover, $\rho(0) > 0$ since $\mathbf{D}$ is an ascending direction as stated by Proposition 2. In order to compute the optimal step ε^* in (6) we distinguish 2 cases:

- if $\rho(y_{VJ}) \ge 0$ then $\varepsilon^* = y_{VJ}$ for $f(\mathbf{Z}_\varepsilon)$ is nondecreasing in the feasible set of (6);
- if $\rho(y_{VJ}) < 0$ then ε^* is a zero of ρ that can be found by dichotomic search.

Suppose the second case holds, *i.e.* assume $\rho(y_{VJ}) < 0$. Then ε^* can be found by iteratively updating the search interval as follows:

$$\begin{aligned} \left(\ell^{(0)}, r^{(0)}\right) &= (0, y_{VJ}) \\ \left(\ell^{(t+1)}, r^{(t+1)}\right) &= \begin{cases} \left(\ell^{(t)}, m^{(t)}\right) & \text{if } \rho\left(m^{(t)}\right) < 0\,, \\ \left(m^{(t)}, r^{(t)}\right) & \text{if } \rho\left(m^{(t)}\right) > 0 \\ \left(m^{(t)}, m^{(t)}\right) & \text{if } \rho\left(m^{(t)}\right) = 0\,, \end{cases} \end{aligned} \tag{7}$$

for all $t > 0$, where $m^{(t)}$ denotes the center of segment $[\ell^{(t)}, r^{(t)}]$, *i.e.* $m^{(t)} = (\ell^{(t)} + r^{(t)})/2$.

We are not in general interested in determining a precise step size ε^* but an approximation is sufficient. Hence, the dichotomic search is carried out until the interval size is below a given threshold. If δ is this threshold, the number of iterations required is expected to be $\log_2(y_{VJ}/\delta)$ in the worst case.

3.3 Complexity

Consider a generic iteration t of our algorithm and assume $A^{(t)} = \mathbf{Y}^\top\mathbf{Y}$ and $g_i^{(t)} = g_i(\mathbf{Y})$ given for all $i \in O$, where $\mathbf{Y} = \mathbf{Y}^{(t)}$.

The computation of ε^* requires the evaluation of function ρ at different values of ε. Each function evaluation can be carried out in $O(n)$ steps by exploiting $\mathbf{A}^{(t)}$ as follows:

$$\begin{aligned} \rho(\varepsilon) = \sum_{i\in O\setminus\{J\}} c_{Ji} &\frac{\mathbf{d}_J^\top \mathbf{y}_i}{A_{Ji}^{(t)} + \varepsilon\mathbf{d}_J^\top\mathbf{y}_i} \\ &+ (N_{Ji} - c_{Ji}) \frac{\mathbf{d}_J^\top \mathbf{y}_i}{1 - A_{Ji}^{(t)} - \varepsilon\mathbf{d}_J^\top\mathbf{y}_i} \end{aligned}$$

where $\mathbf{d}_J = \left(\mathbf{e}_K^{(U)} - \mathbf{e}_K^{(V)}\right)$. The complexity of the computation of the optimal step size is thus $O(n\gamma)$ where γ is the average number of iterations needed by the dichotomic search.

Next, we can efficiently update $\mathbf{A}^{(t)}$ as follows:

$$\begin{aligned} \mathbf{A}^{(t+1)} &= \left(\mathbf{Y}^{(t+1)}\right)^\top \mathbf{Y}^{(t+1)} \\ &= \mathbf{A}^{(t)} + \varepsilon^* \left(\mathbf{D}^\top\mathbf{Y} + \mathbf{Y}^\top\mathbf{D} + \varepsilon^*\mathbf{D}^\top\mathbf{D}\right)\,. \end{aligned} \tag{8}$$

Indeed, since $\mathbf{D}$ has only two non-zero entries, namely (V,J) and (U,J), the terms within parenthesis can be computed in $O(n)$.

The computation of $\mathbf{Y}^{(t+1)}$ can be performed in constant time by exploiting the sparsity of $\mathbf{D}$ as $\mathbf{Y}^{(t+1)} = \mathbf{Y}^{(t)} + \varepsilon^*\mathbf{D}$.

The computation of $g_i^{(t+1)} = g_i(\mathbf{Y}^{(t+1)})$ for each $i \in O \setminus \{J\}$ can be efficiently accomplished in constant time (it requires $O(nK)$ to update all of them) as follows:

$$\begin{aligned} g_i^{(t+1)} &= g_i^{(t)} + c_{iJ}\left(\frac{\mathbf{y}_J^{(t+1)}}{A_{iJ}^{(t+1)}} - \frac{\mathbf{y}_J^{(t)}}{A_{iJ}^{(t)}}\right) \\ &\quad + (N_{iJ} - c_{iJ})\left(\frac{\mathbf{y}_J^{(t+1)}}{1 - A_{iJ}^{(t+1)}} - \frac{\mathbf{y}_J^{(t)}}{1 - A_{iJ}^{(t)}}\right) \end{aligned} \tag{9}$$

The complexity of the computation of $g_J^{(t+1)}$, instead, requires $O(nK)$ steps:

$$g_J^{(t+1)} = \sum_{i \in O \setminus \{J\}} c_{Ji} \frac{\mathbf{y}_i^{(t+1)}}{A_{Ji}^{(t+1)}} - (N_{Ji} - c_{Ji}) \frac{\mathbf{y}_i^{(t+1)}}{1 - A_{Ji}^{(t+1)}} . \tag{10}$$

By iteratively updating the quantities $A^{(t)}$, $g_i^{(t)}$'s and $\mathbf{Y}^{(t)}$ according to the aforementioned procedures, we can keep a per-iteration complexity of $O(nK)$, that is linear in the number of variables in $\mathbf{Y}$.

Iterations stop when KKT conditions of proposition (1) are satisfied under a given tolerance τ, i.e. $([g_J(\mathbf{Y})]_U - [g_J(\mathbf{Y})]_V) < \tau$.

Algorithm 1: PEACE.

```
Require: $\mathbf{Y}^{(0)} \in \Delta_K^n \cap \mathbf{dom}(f)$
  Initialize $g_i^{(0)} \leftarrow g_i(\mathbf{Y})$ for all $i \in O$
  Initialize $A_i^{(0)} \leftarrow \left(\mathbf{Y}^{(0)}\right)^\top \mathbf{Y}^{(0)}$
  $t \leftarrow 0$
  while termination-condition do
    Compute $U, V, J$ as in Prop. 1
    Compute $\varepsilon^*$ as described in Sec. 3.2/3.3
    Update $A^{(t+1)}$ as described in Sec. 3.3
    Update $Y^{(t+1)}$ as described in Sec. 3.3
    Update $g_i^{(t+1)}$ as described in Sec. 3.3
    $t \leftarrow t + 1$
  end while
  return $\mathbf{Y}^{(t)}$
```

4 RELATED WORK

The topic of clustering combination, also known as consensus clustering is completing its first decade of research. A very recent and complete survey can be found in (Ghosh and Acharya, 2011). Several consensus methods have been proposed in the literature (Fred, 2001; Strehl and Ghosh, 2002; Fred and Jain, 2005; Topchy et al., 2004; Dimitriadou et al., 2002; Ayad and Kamel, 2008; Fern and Brodley, 2004). Some of them are based on estimates of similarity between partitions, others cast the problem as a categorical clustering problem, and finally others on similarity between data points (induced by the clustering ensemble). Our work belongs to this last type, where similarities are aggregated on the co-association matrix.

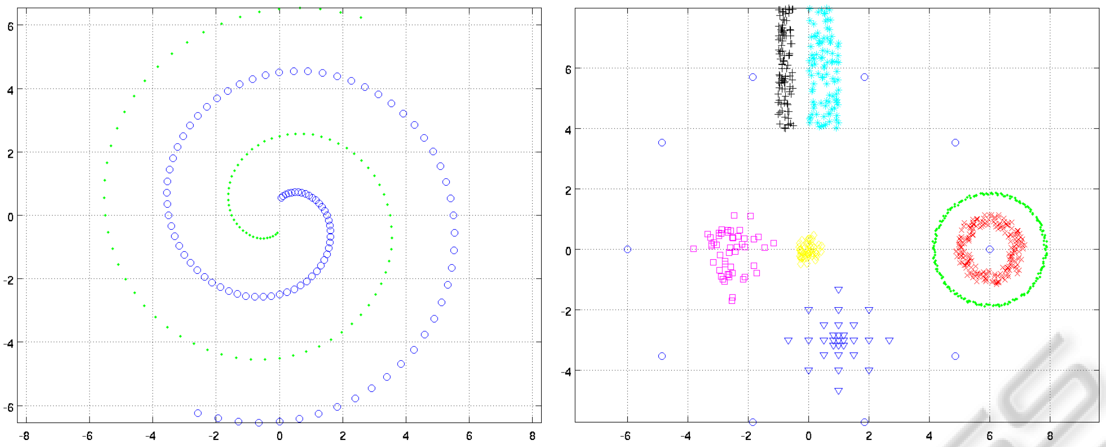
Moreover there are methods, that produce a crisp partition from the information provided by the ensemble, and methods that induce a probabilistic solution, as our work.

In (Lourenço et al., 2011) the entries of the co-association matrix are also exploited and modeled using a generative aspect model for dyadic data, and producing a soft assignment. The consensus solution is found by solving a maximum likelihood estimation problem, using the Expectation-Maximization (EM) algorithm.

In a different fashion, other probabilistic approaches to consensus clustering that do not exploit the co-association matrix are (Topchy et al., 2004) and (Topchy et al., 2005). There, the input space directly consists of the labellings from the clustering ensemble. The model is based on a finite mixture of multinomial distribution. As usual, the model's parameters are found according to a maximum-likelihood criterion by using an EM algorithm. In (Wang et al., 2009), the idea was extended introducing a Bayesian version of the multinomial mixture model, the *Bayesian cluster ensembles*. Although the posterior distribution cannot be calculated in closed-form, it is approximated using variational inference and Gibbs sampling, in a very similar procedure as in *latent Dirichlet allocation* model (Griffiths and Steyvers, 2004), (Steyvers and Griffiths, 2007), but applied to a different input feature space. Finally, in (Wang et al., 2010), a nonparametric version of this work was proposed.

5 EXPERIMENTS AND RESULTS

In this section we present the evaluation of our algorithm, using synthetic datasets, UCI data and two text-data benchmark datasets. We compare its performance against algorithms that rely on the same type of data, (the coassociation matrix) and on similar assumptions. The Baum-Eagon (BE) (Rota Bulò et al., 2010) algorithm, which also extracts a soft consensus partition from the co-association matrix, and against the classical EAC algorithm using as extraction criteria the hierarchical agglomerative single-link (SL) and average-link (AL) algorithms.



(a) (b)

Figure 1: Sketch of the Synthetic Datasets.

As in similar works, the performance of the algorithms is assessed using external criteria of clustering quality, comparing the obtained partitions with the known ground truth partition. Given O, the set of data objects to cluster, and two clusterings, $p_i = \{p_i^1,\dots,p_i^h\}$ and $p_j = \{p_l^1,\dots,p_l^k\}$, we chose one criterion based on cluster matching - Consistency Index (CI), and in F1-Measure (Baeza-Yates and Ribeiro-Neto, 1999).

The Consistency Index, also called H index (Meila, 2003), gives the accuracy of the obtained partitions and is obtained by matching the clusters in the combined partition with the ground truth labels:

$$CI(p_i, p_l) = \frac{1}{n} \sum_{k'=match(k)} m_{k,k'}, \qquad (11)$$

where $m_{k,k'}$ denotes the contingency table, *i.e.* $m_{k,k'} = |p_i^k \cap p_l^{k'}|$. It corresponds to the percentage of correct labellings when the number of clusters in p_i and p_l is the same.

5.1 UCI and Synthetic Data

We followed the usual strategy of producing clustering ensembles, and combining them on the co-association matrix. Two different types of ensembles were created: (1) using k-means with random initialization and random number of clusters (Lourenço et al., 2010); (2) combining multiple algorithms (agglomerative hierarchical algorithms: single, average, ward, centroid link; k-means(Jain and Dubes, 1988); spectral clustering (Ng et al., 2001)) applied over subsampled versions of the datasets (subsampling percentage 0.9).

Table 1 summarizes the main characteristics of the UCI and synthetic datasets used on the evaluation, and the parameters used for generating ensemble (2). Figure 1 illustrates the synthetic datasets used in the evaluation: (a) spiral; (b) image-c.

Figure 3 summarizes the average performance of both algorithms over ensembles (1) and (2), after several runs, accounting for possible different solutions due to initialization, in terms of Consistency Index (CI), and F-1 Measure.

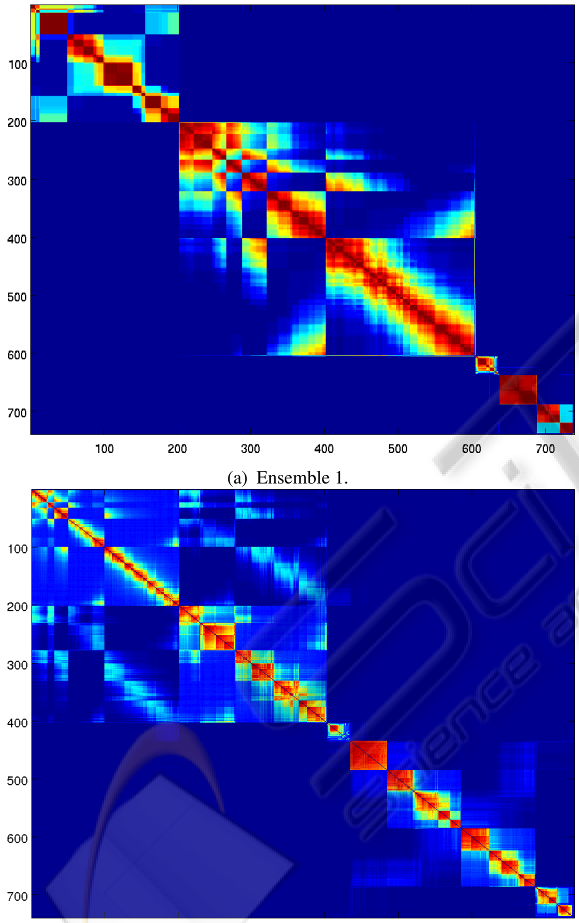
The performance of PEACE and BE is very different for the synthetic and UCI datasets. On the first, PEACE and BE have lower performance when compared with EAC-SL and EAC-AL (both on F1 and CI); while on the later have better performance on several examples. Comparing the performance of both ensembles: on ensemble (1), PEACE has better performance than other methods on 3 datasets (over 9), while on ensemble (2) it has better or equal performance that the other on 6 (over 9).

Ensembles (1) were generated using a split and merge strategy, which produces the splitting of natural clusters into smaller clusters inducing micro-blocks in the co-association matrix, as shown in figure 2, for the (s-2) dataset, which has 7 natural clusters. The results show that the proposed model is not so adequate to this type of block diagonal matrix, penalizing PEACE. Comparing it with the BE algorithm shows that in this complicated co-association matrices, it seems that PEACE is more robust.

Ensembles (2) are generated with a combination of several algorithms, inducing co-association matrices much more blockwise, as is shown in figure 2(b). The proposed model is much more suitable for this type of co-association matrices. The BE algorithm also has a better performance on this type of ensembles, leading to a similar performance.

Table 1: Benchmark datasets.

Data-Sets	K	n	Ensemble
			K_i
(s-1) spiral	2	200	2-8
(s-2) image-c	7	739	8-15,20,30
(r-1) iris	3	150	3-10,15,20
(r-2) wine	3	178	4-10,15,20
(r-3) house-votes	2	232	4-10,15,20
(r-4) ionosphere	2	351	4-10,15,20
(r-5) std-yeast-cell	5	384	5-10,15,20
(r-6) breast-cancer	2	683	3-10,15,20
(r-7) optdigits	10	1000	10, 12, 15, 20, 35, 50



(a) Ensemble 1.

(b) Ensemble 2.

Figure 2: Example of co-association Matrices obtained with ensemble (1) and (2) - reordered using VAT (Bezdek and Hathaway, 2002) - on the (s-2) data-set.

5.2 Text Data

We also evaluated the proposed algorithm over two well known text-data benchmark datasets: the KDD mininewsgroups[1] and the WebKD dataset[2]. The mininewsgroups dataset, is composed by usenet articles from 20 newsgroups. After removing three newsgroups not corresponding to a clear concept ('talk.politics.misc', 'talk.religion.misc', 'comp.os.ms-windows.misc'), we ended up analyzing 17 newsgroups, grouped in 7 macro-categories ('rec','comp','soc','sci','talk','alt','misc'). In this collection there are only 100 documents on each newsgroups, totalizing 1700 documents.

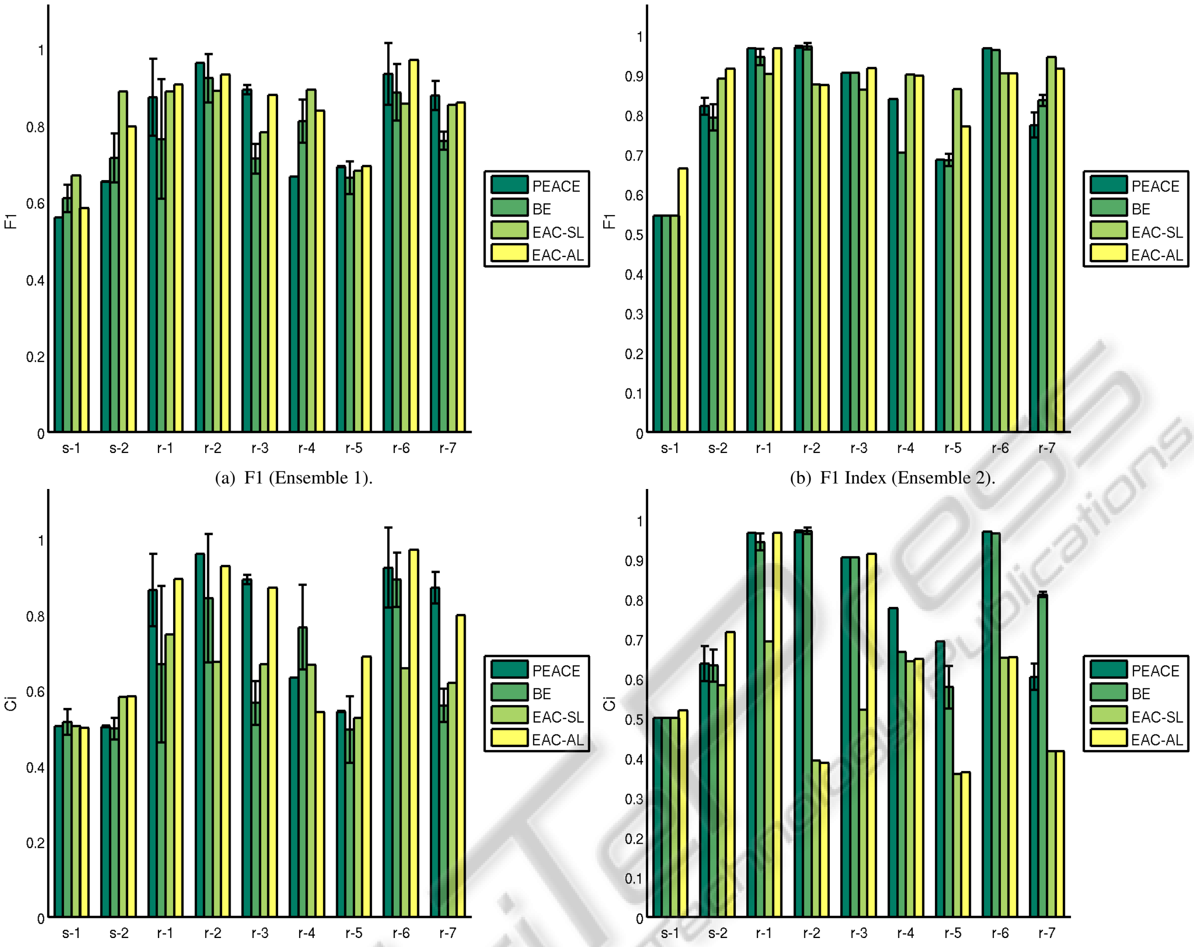
The WebKD dataset corresponds to WWW-pages collected from computer science departments of various universities in January 1997. We concentrated our analysis on 4 categories ('project', 'student', 'course', 'faculty'). For each, we analyzed only the documents belonging to universities ('texas','washington','wisconsin','cornell'), totalizing 1041 documents.

The analysis followed the usual steps for text-processing (Manning et al., 2008): tokenization, stopword-removal, stemming (Porter Stemmer), feature weighting (using Tf-Idf) and feature removal. For feature removal, we removed tokens that appeared in less than 40 documents and words that had low variance of occurrence, following Cui et al. ("Non-Redundant Multi-View Clustering Via Ortogonalization"). On the mininewsgroups dataset this feature removal step, led to 500-dimension term frequency vector, while on the WebKD led to 335-dimension term frequency vector.

We build the clustering ensembles based on the split and merge strategy (ensemble (1)) using: K-means with cosine similarity - ensemble 1a; and Mini-Batch K-means algorithm (Sculley, 2010), a variant of the classical algorithm using mini-batches (random subset of the dataset), to compute the centroids - en-

[1] http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html

[2] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/



(a) F1 (Ensemble 1).

(b) F1 Index (Ensemble 2).

(c) *CI* (Ensemble 1).

(d) *CI* (Ensemble 2).

Figure 3: Performance Evaluation in terms of F1 and Consistency Index.

semble 1b. For the generation we assumed that each partition had a random number of clusters, chosen in the interval $K = \{\sqrt{ns}/2; \sqrt{ns}\}$, where ns is the number of samples.

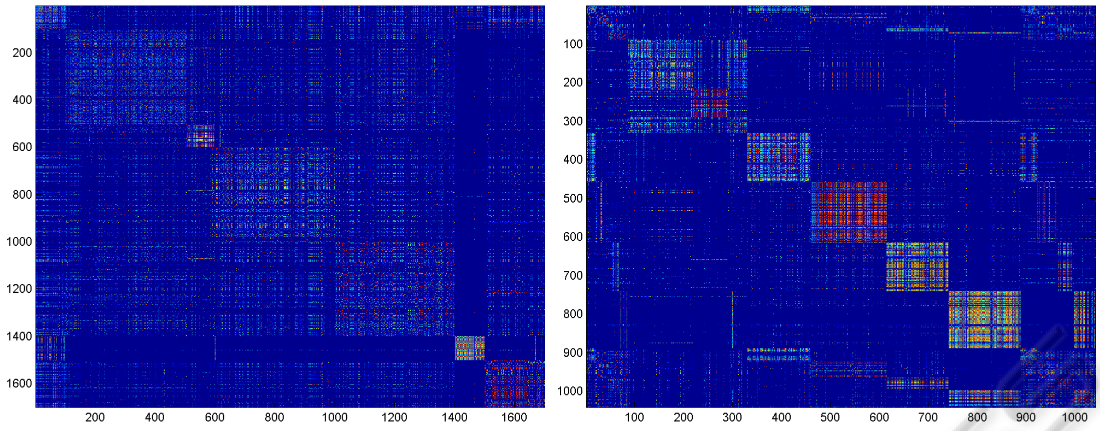
Figure 4 illustrates an example of the obtained coassociation matrices. To allow a better understanding of obtained co-association matrices, samples are aligned according to ground truth. The block-diagonal structure of the co-association of webKD dataset is much more evident than on the miniNewsgroups.

In tables 2 and 3 we summarize the obtained results for the PEACE and BE algorithm, indicating minimum, maximum, average and standard deviation of the validation indexes. In addition, the first column ("selected") refers to the value of the validation index selected according to the intrinsic optimization criterion, i.e highest value of $\mathbb{P}[\mathbf{C}|\mathbf{Y}]$. Highest values for each data set are highlighted in bold.

From the analysis of tables 2 and 3 it is apparent that the PEACE algorithm has better performance in ensembles exhibiting higher compactness properties. However, in situations where the co-association matrices have a less evident structure, with a lot of noise connecting clusters, its performances tend to decrease.

6 CONCLUSIONS

In this paper we introduced a new probabilistic approach, based on the EAC paradigm, for consensus clustering. In our model, the entries of the co-association matrix are regarded as realizations of binomial random variables parameterized by probabilistic assignments of objects to clusters, and we estimate such parameters by means of a maximum likelihood approach. The resulting optimization problem is non-linear and non-convex and we addressed it using a primal line-search algorithm. Evaluation on both



(a) miniNewsgroups.

(b) webKD.

Figure 4: Examples of obtained co-associations for miniNewsgroups and webKD datasets using an ensemble of K-means with cosine similarity.

Table 2: Consistency indices of consensus solutions for the clustering ensemble.

Ensemble	Data Set	PEACE					BE				
		selected	av	std	max	min	selected	av	std	max	min
E1a	miniN	0.425	0.431	0.028	**0.468**	0.385	**0.433**	0.439	0.020	0.459	0.418
	webkd	**0.414**	0.423	0.046	**0.492**	0.339	0.405	0.396	0.010	0.406	0.387
E1b	miniN	0.242	0.242	0.001	0.242	0.241	**0.356**	0.356	0.000	**0.356**	0.356
	webkd	0.297	0.369	0.067	**0.419**	0.294	**0.320**	0.320	0.000	0.320	0.320

Table 3: F1 of consensus solutions for the clustering ensemble.

Ensemble	Data Set	PEACE					BE				
		selected	av	std	max	min	selected	av	std	max	min
E1a	miniN	0.551	0.541	0.021	0.565	0.494	**0.559**	0.583	0.009	**0.595**	0.559
	webkd	**0.616**	0.618	0.046	0.678	0.528	0.580	0.636	0.059	**0.693**	0.580
E1b	miniN	**0.853**	0.845	0.015	**0.861**	0.822	0.769	0.774	0.006	0.778	0.769
	webkd	**0.663**	0.698	0.032	**0.723**	0.663	0.530	0.532	0.004	0.539	0.530

synthetic and real benchmarks data assessed the effectiveness of our approach. As future work we want to develop methods for exploiting the uncertainty of information given by the probabilistic assignments, as well as exploiting the possibility of having overlapping groups in the co-association matrix.

ACKNOWLEDGEMENTS

This work was partially funded by FCT under grants SFRH/PROTEC/49512/2009, PTDC/EIACCO/103230/2008 (EvaClue project), and by the ADEETC from Instituto Superior de Engenharia de Lisboa, whose support the authors gratefully acknowledge.

REFERENCES

Ayad, H. and Kamel, M. S. (2008). Cumulative voting consensus method for partitions with variable number of clusters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(1):160–173.

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Bezdek, J. and Hathaway, R. (2002). Vat: a tool for visual assessment of (cluster) tendency. In *Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on*, volume 3, pages 2225 – 2230.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, first edition edition.

Dimitriadou, E., Weingessel, A., and Hornik, K. (2002). A combination scheme for fuzzy clustering. In *AFSS'02*, pages 332–338.

Fern, X. Z. and Brodley, C. E. (2004). Solving cluster ensemble problems by bipartite graph partitioning. In *Proc ICML '04*.

Fred, A. (2001). Finding consistent clusters in data partitions. In Kittler, J. and Roli, F., editors, *Multiple Classifier Systems*, volume 2096, pages 309–318. Springer.

Fred, A. and Jain, A. (2002). Data clustering using evidence accumulation. In *Proc. of the 16th Int'l Conference on Pattern Recognition*, pages 276–280.

Fred, A. and Jain, A. (2005). Combining multiple clustering using evidence accumulation. *IEEE Trans Pattern Analysis and Machine Intelligence*, 27(6):835–850.

Ghosh, J. and Acharya, A. (2011). Cluster ensembles. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 1(4):305–315.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5228–5235.

Jain, A. K. and Dubes, R. (1988). *Algorithms for Clustering Data*. Prentice Hall.

Kachurovskii, I. R. (1960). On monotone operators and convex functionals. *Uspekhi Mat. Nauk*, 15(4):213–215.

Lourenço, A., Fred, A., and Figueiredo, M. (2011). A generative dyadic aspect model for evidence accumulation clustering. In *Proc. 1st Int. Conf. Similarity-based pattern recognition*, SIMBAD'11, pages 104–116, Berlin, Heidelberg. Springer-Verlag.

Lourenço, A., Fred, A., and Jain, A. K. (2010). On the scalability of evidence accumulation clustering. In *20th International Conference on Pattern Recognition (ICPR)*, pages 782 –785, Istanbul Turkey.

Luenberger, D. G. and Ye, Y. (2008). *Linear and Nonlinear Programming*. Springer, third edition edition.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Meila, M. (2003). Comparing clusterings by the variation of information. In Springer, editor, *Proc. of the Sixteenth Annual Conf. of Computational Learning Theory (COLT)*.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856. MIT Press.

Rota Bulò, S., Lourenço, A., Fred, A., and Pelillo, M. (2010). Pairwise probabilistic clustering using evidence accumulation. In *Proc. 2010 Int. Conf. on Structural, Syntactic, and Statistical Pattern Recognition*, SSPR&SPR'10, pages 395–404.

Sculley, D. (2010). Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1177–1178, New York, NY, USA. ACM.

Steyvers, M. and Griffiths, T. (2007). *Probabilistic topic models*, chapter Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *J. of Machine Learning Research 3*.

Topchy, A., Jain, A., and Punch, W. (2004). A mixture model of clustering ensembles. In *Proc. of the SIAM Conf. on Data Mining*.

Topchy, A., Jain, A. K., and Punch, W. (2005). Clustering ensembles: Models of consensus and weak partitions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12):1866–1881.

Wang, H., Shan, H., and Banerjee, A. (2009). Bayesian cluster ensembles. In *9th SIAM Int. Conf. on Data Mining*.

Wang, P., Domeniconi, C., and Laskey, K. B. (2010). Nonparametric bayesian clustering ensembles. In *ECML PKDD'10*, pages 435–450.