Enhancing Single-Objective Projective Clustering Ensembles

Francesco Gullo
DEIS Dept.
University of Calabria
87036 Rende (CS), Italy
fgullo@deis.unical.it

Carlotta Domeniconi
Department of Computer Science
George Mason University
22030 Fairfax - VA, USA
carlotta@cs.gmu.edu

Andrea Tagarelli
DEIS Dept.
University of Calabria
87036 Rende (CS), Italy
tagarelli@deis.unical.it

***Abstract*—*Projective Clustering Ensembles* (PCE) has recently been formulated to solve the problem of deriving a robust projective consensus clustering from an ensemble of projective clustering solutions [1]. PCE is formalized as an optimization problem with either a two-objective or a single-objective function, depending on whether the object-based and the feature-based representations of the clusters in the ensemble are treated separately. A major result in [1] is that single-objective PCE outperforms two-objective PCE in terms of efficiency, at the cost of lower accuracy in consensus clustering.**

In this paper, we enhance the single-objective PCE formulation, with the ultimate goal of providing more effective formulations capable of reducing the accuracy gap with the two-objective counterpart, while maintaining the efficiency advantages. We provide theoretical insights into the single-objective function, and introduce two heuristics that overcome the major limitations of the previous single-objective PCE formulation. Experimental evidence has demonstrated the significance of our proposed heuristics. In fact, results have not only confirmed a far better efficiency w.r.t. two-objective PCE, but have also shown the claimed improvements in accuracy of the consensus clustering obtained by the new single-objective PCE.

I. Introduction

Projective clustering and *clustering ensembles* represent two recent advances in data clustering. Clustering ensembles [2]–[5] are based on the idea of exploiting the information provided by a set of clustering solutions (the *ensemble*) in order to extract a *consensus clustering*, i.e., a clustering solution that summarizes the information available from the ensemble. The input ensemble is usually generated by varying one or more aspects of the clustering process, such as the clustering algorithm, the parameter setting, and the number of features, objects, or clusters. Projective clustering [6]–[9] aims to discover clusters that correspond to subsets of the input data and have different (possibly overlapping) dimensional subspaces associated with them. Projected clusters tend to be less noisy—each group of data is represented in a subspace which does not contain irrelevant dimensions—and more understandable—exploring a cluster is easier when few dimensions are involved.

In [1], projective clustering and clustering ensembles are treated for the first time in a unified framework. The underlying motivation of that study is related to the two major issues in data clustering, i.e., the high-dimensionality and the lack of a-priori knowledge, which usually co-exist in real-world applications. To address both issues simultaneously, the problem of *projective clustering ensembles* (PCE) is formalized in [1]. PCE is formulated as an optimization problem by exploiting the information available from the input ensemble, and a robust *projective consensus clustering* is sought as a solution to that problem. The PCE objective function meets desirable requirements, including the independence from the specific clustering algorithm and from any prior knowledge on the setup for ensemble generation, the capability of handling hard as well as soft data clustering, and of enabling feature weighting.

Two formulations of PCE are described in [1], namely *two-objective* and *single-objective* PCE. The former involves two objective functions, which separately consider the data object clustering and the feature-to-cluster assignment. This multi-objective optimization problem is solved by a well-founded heuristic, in which a *Pareto*-based *Multi-Objective Evolutionary Algorithm*, called *MOEA-PCE*, is used to avoid combining the two objective functions into a single one. However, although this strategy has been found to be particularly accurate [1], it may incur a number of issues that are intrinsic to a two-objective PCE formulation, such as inefficiency, hard parameter setting, and hard interpretation of results.

The single-objective PCE formulation has instead one objective function, which acts as an error criterion for the computation of a candidate cluster solution. It involves both the object-based and the feature-based representations of a candidate cluster. Based on this formulation, the EM-like heuristic *EM-PCE* [1] has been developed to overcome the major drawbacks of two-objective PCE, inefficiency in particular. Unfortunately, other issues arise with single-objective PCE, which result in a weaker formulation than the two-objective counterpart. As a result, single-objective PCE is outperformed by two-objective PCE in terms of effectiveness [1].

In this paper, we provide an insight into the objective function of single-objective PCE, and address its weaknesses. A major goal of our analysis is to develop enhanced formulations and heuristic algorithms for PCE that overcome the drawbacks of single-objective PCE while maintaining the advantages w.r.t. two-objective PCE. More specifically, we provide two formulations to enhance single-objective PCE:

- *Enhanced EM-based PCE*, which directly refines the EM-like single-objective PCE formulation. It introduces additional terms in the objective function to consider, not only the feature-based representations of the objects, but also the information about the object-to-cluster assignments. This is achieved via a maximization of the probability of co-membership of objects to clusters in the ensemble solutions.
- *Enhanced 2-Step-based PCE*, which is designed to overcome the mutual dependence between the object- and feature-to-cluster assignment functions. The key idea of this formulation is to perform a two-step scheme: the object-to-cluster assignments are optimized first, and independently of the feature-to-cluster assignments, which are then optimized in the second step.

Experimental results have shown that both the proposed enhanced formulations of single-objective PCE lead to a significant improvement in accuracy w.r.t. the basic EM-like single-objective PCE. Thus, our enhancements reduce the effectiveness gap w.r.t. the two-objective PCE formulation, while maintaining a large advantage in terms of efficiency.

II. Background

A. Projective Clustering Ensembles: Problem Definition

Let $\mathcal{D} = \{\vec{o}_1, \ldots, \vec{o}_N\}$ be a set of D-dimensional points (data objects), where $\vec{o}_n = (o_{n1}, \ldots, o_{nD})$, $\forall n \in \{1, \ldots, N\}$. A *projective clustering solution* C defined over $\mathcal{D}$ is a triple $\langle \mathcal{L}, \Gamma, \Delta \rangle$:

- $\mathcal{L} = \{\ell_1, \ldots, \ell_K\}$ is a set of cluster labels that identify the K clusters in the solution.
- $\Gamma : \mathcal{L} \times \mathcal{D} \to S_\Gamma$ is a function which stores the probability that object $\vec{o}_n$ belongs to the cluster labeled with ℓ_k, such that $\sum_{k=1}^{K} \Gamma(\ell_k, \vec{o}_n) = 1, \forall n \in \{1, \ldots, N\}$, where Γ_{kn} denotes $\Gamma(\ell_k, \vec{o}_n)$.
- $\Delta : \mathcal{L} \times \{1, \ldots, D\} \to [0,1]$ is a function that stores the probability that the d-th feature is a relevant dimension for the objects in the cluster labeled with ℓ_k, such that $\sum_{d=1}^{D} \Delta(\ell_k, d) = 1, \forall k \in \{1, \ldots, K\}$, where Δ_{kd} denotes $\Delta(\ell_k, d)$.

A *projective ensemble* defined over $\mathcal{D}$ is a set $\mathcal{E} = \{C^{(1)}, \ldots, C^{(M)}\}$, where each $C^{(m)} = \langle \mathcal{L}^{(m)}, \Gamma^{(m)}, \Delta^{(m)} \rangle$ is a projective clustering solution defined over $\mathcal{D}$, $\forall m \in \{1, \ldots, M\}$, and $\mathcal{L}^{(i)} \cap \mathcal{L}^{(j)} = \emptyset$, $\forall i, j \in \{1, \ldots, M\}, i \neq j$. Clustering solutions in $\mathcal{E}$ can contain in general a different number of clusters.

A projective ensemble $\mathcal{E}$ is associated with a (global) label set $\mathbf{L}$, defined as $\mathbf{L} = \{\mathbf{l}_1, \ldots, \mathbf{l}_H\} = \bigcup_{m=1}^{M} \mathcal{L}^{(m)}$. Each cluster labeled with $\mathbf{l}_h \in \mathbf{L}$ is associated with an *object-based representation* and a *feature-based representation* given by the vectors $\vec{\gamma}_h = (\gamma_{h1}, \ldots, \gamma_{hN})$ and $\vec{\delta}_h = (\delta_{h1}, \ldots, \delta_{hD})$, respectively. Vector $\vec{\gamma}_h$ contains the probabilities that the data objects in $\mathcal{D}$ belong to cluster $\mathbf{l}_h$, which are retrieved from the function Γ of the projective clustering solution that contains $\mathbf{l}_h$; the function Δ defines $\vec{\delta}_h$ similarly. Formally, let $\langle m, k \rangle$ be a pair corresponding to any $h \in \{1, \ldots, H\}$ such that the cluster ℓ_k of the solution $C^{(m)} = \langle \mathcal{L}^{(m)}, \Gamma^{(m)}, \Delta^{(m)} \rangle \in \mathcal{E}$ corresponds to the cluster $\mathbf{l}_h$; it holds that $\vec{\gamma}_h = (\Gamma_{k1}^{(m)}, \ldots, \Gamma_{kN}^{(m)})$ and $\vec{\delta}_h = (\Delta_{k1}^{(m)}, \ldots, \Delta_{kD}^{(m)})$.

B. Single-objective PCE

In [1], PCE is formulated as an optimization problem with a single objective function, which considers both the object-to-cluster and the feature-to-cluster assignments in $\mathcal{E}$:

$$C^* = \arg\min_{\hat{C}} Q(\hat{C}, \mathcal{E}) \quad (1)$$

$$s.t.$$

$$\sum_{k=1}^{K} \hat{\Gamma}_{kn} = 1, \ \forall n, \quad \sum_{d=1}^{D} \hat{\Delta}_{kd} = 1, \ \forall k \quad (2)$$

$$\hat{\Gamma}_{kn} \geq 0, \ \hat{\Delta}_{kd} \geq 0, \quad \forall k, \forall n, \forall d \quad (3)$$

where

$$Q(\hat{C}, \mathcal{E}) = \sum_{k=1}^{K} \sum_{n=1}^{N} \hat{\Gamma}_{kn}^{\alpha} \sum_{h=1}^{H} \gamma_{hn} \sum_{d=1}^{D} \left(\hat{\Delta}_{kd} - \delta_{hd} \right)^2 \quad (4)$$

and $\alpha > 1$ is an integer that guarantees that $\hat{\Gamma}_{kn}$ ranges within $[0,1]$ (instead of $\{0,1\}$).

To solve the above problem, the *EM-based Projective Clustering Ensembles (EM-PCE)* heuristic is defined. To find the optimal values of $\hat{\Gamma}_{kn}$ (resp. $\hat{\Delta}_{kd}$), while keeping $\hat{\Delta}_{kd}$ (resp. $\hat{\Gamma}_{kn}$) fixed, EM-PCE iterates over two main EM-like steps using the equations:

$$\Gamma_{kn}^{*} = \left[\sum_{k'=1}^{K} \left(\frac{X_{kn}}{X_{k'n}} \right)^{\frac{1}{\alpha - 1}} \right]^{-1} \quad \text{and} \quad \Delta_{kd}^{*} = \frac{Z_{kd}}{Y_k}$$

where

$$X_{kn} = \sum_{h=1}^{H} \gamma_{hn} \sum_{d=1}^{D} \left(\hat{\Delta}_{kd} - \delta_{hd} \right)^2 \quad (5)$$

$$Y_k = \sum_{n=1}^{N} \hat{\Gamma}_{kn}^{\alpha} \sum_{h=1}^{H} \gamma_{hn} = M \sum_{n=1}^{N} \hat{\Gamma}_{kn}^{\alpha} \quad (6)$$

$$Z_{kd} = \sum_{n=1}^{N} \hat{\Gamma}_{kn}^{\alpha} \sum_{h=1}^{H} \gamma_{hn} \ \delta_{hd} \quad (7)$$

III. Single-objective PCE: enhancements

A. Issues in single-objective PCE

To illustrate the issues that affect the single-objective PCE formulation, we first provide an alternative explanation of the function Q defined in (4). For this purpose, let us consider each object $\vec{o}_n$ as a *multi-valued* instance described by the set of vectors $\{\gamma_{1n} \times \vec{\delta}_1, \ldots, \gamma_{Hn} \times \vec{\delta}_H\}$, i.e., the set $\{\vec{\delta}_1, \ldots, \vec{\delta}_H\}$ of feature-based representations of the projective clusters of all the solutions in the ensemble (cf. Sect. II-A), weighted by the probabilities $\gamma_{1n}, \ldots, \gamma_{Hn}$ that $\vec{o}_n$ belongs to any cluster $\mathbf{l}_h$, $\forall h$. We refer to $\{\gamma_{1n} \times \vec{\delta}_1, \ldots, \gamma_{Hn} \times \vec{\delta}_H\}$ as the *feature-based* representation of object $\vec{o}_n$. Within this view, function Q can be interpreted as a special version of the function optimized by the K-means clustering algorithm, in which the following aspects hold. Each cluster labeled with $\hat{\ell}_k$ within the candidate projective solution $\hat{C}$ has as centroid the vector of features $(\hat{\Delta}_{k1}, \ldots, \hat{\Delta}_{kD})$, i.e., the values that are eventually used as representative of the feature-to-cluster assignment of cluster $\hat{\ell}_k$. The distance between any object $\vec{o}_n$ and the centroid of any cluster $\hat{\ell}_k$ is computed by summing the Euclidean distances between the centroid $(\hat{\Delta}_{k1}, \ldots, \hat{\Delta}_{kD})$ and the vectors $\{\vec{\delta}_1, \ldots, \vec{\delta}_H\}$ belonging to the multi-representation of $\vec{o}_n$, where each distance is weighted by the corresponding γ_{hn} (indeed, the distance is computed as $X_{kn} = \sum_{h=1}^{H} \gamma_{hn} \sum_{d=1}^{D} \left(\hat{\Delta}_{kd} - \delta_{hd}\right)^2$). Since the membership of any object $\vec{o}_n$ to any cluster $\hat{\ell}_k$ may be soft, the distances between $\vec{o}_n$ and the centroid of cluster $\hat{\ell}_k$ are multiplied by $\hat{\Gamma}_{kn}^{\alpha}$, which denotes the probability that $\vec{o}_n$ belongs to $\hat{\ell}_k$.

According to the above interpretation, we can state that:

1) the feature-to-cluster assignments of any cluster $\hat{\ell}_k$ (i.e., the values $(\hat{\Delta}_{k1}, \ldots, \hat{\Delta}_{kD})$) are given by the vector that minimizes the (weighted) squared Euclidean distance from all the vectors that compose the feature-based representation of the objects in $\hat{\ell}_k$;
2) the compactness of a cluster $\hat{\ell}_k$ is measured only according to the feature-based multi-representation of the data objects; indeed, it is inversely proportional to the X_{kn} values of any pair of objects $\vec{o}_n$ within $\hat{\ell}_k$.

The first statement is reasonable since, for each cluster labeled with $\hat{\ell}_k$, the probability that a feature d represents well cluster $\hat{\ell}_k$ is directly proportional to how well the objects in $\hat{\ell}_k$ are represented by the feature d in the ensemble. On the other hand, a major issue arises from statement 2, as discussed in the following example.

Example: Given a projective ensemble $\mathcal{E}$, let us first consider two objects $\vec{o}_i$ and $\vec{o}_j$ which, according to the information available from $\mathcal{E}$, are always or often clustered together (i.e., $\sum_{h=1}^{H} |\gamma_{hi} - \gamma_{hj}|$ is equal or close to 0). It is desirable that any projective consensus function guarantees that $\vec{o}_i$ and $\vec{o}_j$ are clustered together in the output projective consensus clustering as well (i.e., it should be guaranteed that $\sum_{k=1}^{K} |\hat{\Gamma}_{ki} - \hat{\Gamma}_{kj}|$ is equal or close to 0). This requirement is satisfied by function Q; indeed, the fact that $\vec{o}_i$ and $\vec{o}_j$ always or often belong to the same cluster in the solutions of the ensemble clearly implies that they also share similar γ_{hi} and γ_{hj} values and, therefore, similar X_{ki} and X_{kj} distances from any given vector $(\hat{\Delta}_{k1}, \ldots, \hat{\Delta}_{kD})$.

Unfortunately, wrong decisions may be taken using function Q if the opposite situation happens, i.e., when $\vec{o}_i$ and $\vec{o}_j$ are never or seldom clustered together according to the information available from $\mathcal{E}$ (i.e, when $\sum_{h=1}^{H} |\gamma_{hi} - \gamma_{hj}|$ is equal or close to $2M$). This case should ideally lead to the assignment of $\vec{o}_i$ and $\vec{o}_j$ to different clusters in the output consensus clustering (i.e.,$\sum_{k=1}^{K} |\hat{\Gamma}_{ki} - \hat{\Gamma}_{kj}|$ should be equal or close to $2K$). However, function Q does not guarantee this: it may happen, in fact, that the (distinct) clusters of $\mathcal{E}$ to which $\vec{o}_i$ and $\vec{o}_j$ belong, share similar feature-based representations, which is sufficient to make $X_{ki} = X_{kj}, \forall k$. In this case, $\vec{o}_i$ and $\vec{o}_j$ will be clustered together with high probability in the projective consensus clustering, which conflicts with the information available from $\mathcal{E}$. ■

The issue illustrated above arises because function Q measures the distance between any pair of objects using only their corresponding feature-based representations.

B. Enhancing Single-objective PCE

1) The E-EM-PCE algorithm: Our first approach to overcome the issues explained in the previous subsection aims to "complete" function Q, by adding a term which takes into account the dissimilarity of any two objects measured according to how often they are clustered together in the various solutions $\{C^{(1)}, \ldots, C^{(M)}\}$ of the input ensemble $\mathcal{E}$. For each cluster $\hat{\ell}_k$, given that object $\vec{o}_n$ belongs to $\hat{\ell}_k$, it should be guaranteed that any other object $\vec{o}_{n'}$ $(n' \neq n)$ belongs to $\hat{\ell}_k$ if and only if $\vec{o}_{n'}$ is often clustered with $\vec{o}_n$ in the ensemble. In other words, our goal is to maximize the probability that both the events "$\vec{o}_n$ *and* $\vec{o}_{n'}$ *are clustered together in* $\mathcal{E}$" (denoted as $A_{nn'}$) and "$\vec{o}_{n'}$ *belong to* $\hat{\ell}_k$" (denoted as $B_{n'}$) occur. Maximizing $\Pr(A_{nn'} \cap B_{n'})$ $\forall n' \neq n$ is equivalent to minimize $1 - \Pr(A_{nn'} \cap B_{n'})$ $\forall n' \neq n$. Since Q is minimized, the latter terms are added.

Therefore, we define $(\forall k, \forall n)$ $X'_{kn} = \sum_{\forall n' \neq n}(1 - \Pr(A_{nn'} \cap B_{n'}))$, which, assuming independence between $A_{nn'}$ and $B_{n'}$, can be computed as:

$$X'_{kn} = \sum_{\forall n' \neq n}(1 - \Pr(A_{nn'})\ \Pr(B_{n'})) = \sum_{\forall n' \neq n}(1 - \Pr(A_{nn'})\,\hat{\Gamma}_{kn'})$$

It can easily be shown that

$$\Pr(A_{nn'}) = \sum_{m=1}^{M} \Pr(A_{nn'} | C^{(m)}) \Pr(C^{(m)}) = \frac{1}{M} \sum_{h=1}^{H} \gamma_{hn} \gamma_{hn'} \tag{8}$$

Thus, the final expression for X'_{kn} is

$$X'_{kn} = \sum_{\forall n' \neq n} 1 - \frac{\hat{\Gamma}_{kn'}}{M} \sum_{h=1}^{H} \gamma_{hn}\ \gamma_{hn'} \tag{9}$$

The new function to be optimized, comprised of the terms X'_{kn}, becomes

$$Q_E(\hat{C}, \mathcal{E}) = \sum_{k=1}^{K} \sum_{n=1}^{N} \Gamma_{kn}^{\alpha}\ \widetilde{X}_{kn} \tag{10}$$

Algorithm 1 E-EM-PCE

Input: a projective ensemble $\mathcal{E}$ defined over a set $\mathcal{D}$ of data objects; the number K of clusters in the output projective consensus clustering
Output: the projective consensus clustering C^*

```
L* ← {1, ..., K}
⟨Γ*, Δ*⟩ ← randomGen(E, K)
repeat
    compute Γ* according to (12)
    compute Δ* according to (13)
until convergence
C* = ⟨L*, Γ*, Δ*⟩
```

where

$$\widetilde{X}_{kn} = \frac{1}{2M} X_{kn} + \frac{1}{N-1} X'_{kn} \tag{11}$$

subject again to the constraints listed in (2)-(3). The two terms X_{kn} (defined in (5)) and X'_{kn} are normalized in order to make them comparable (indeed, it can be easily proved that $X_{kn} \in [0, 2M]$ and $X'_{kn} \in [0, N-1]$).

It is straightforward to see that the new objective function Q_E mitigates the problem of single-objective PCE explained in the previous subsection. Indeed, the additional terms X'_{kn} in Q_E compare any pair of objects considering not only their feature-based representations, but also the information on how often they are clustered together in the ensemble.

In order to provide a heuristic solution for the PCE formulation involving function Q_E, we resort to the EM paradigm and propose an algorithm called *Enhanced EM-based Projective Clustering Ensembles (E-EM-PCE)* (Alg. 1). Like EM-PCE, the proposed E-EM-PCE comprises two main steps, which are iterated until convergence. However, the algorithm now exploits the function Q_E defined in (10) (instead of Q). The objective is again to find an optimal solution for $\hat{\Gamma}_{kn}$, while keeping $\hat{\Delta}_{kd}$ fixed, and vice versa. The basic equations for the two steps are as follows ($\forall k, \forall n, \forall d$):

$$\Gamma^*_{kn} = \left[\sum_{k'=1}^{K} \left(\frac{\widetilde{X}_{kn}}{\widetilde{X}_{k'n}} \right)^{\frac{1}{\alpha-1}} \right]^{-1} \tag{12}$$

$$\Delta^*_{kd} = \frac{Z_{kd}}{Y_k} \tag{13}$$

where $\widetilde{X}_{kn}$, Y_k and Z_{kd} are defined in (11), (6) and (7), respectively. The way these expressions are derived implies that Alg. 1 converges to a local minimum of the function Q_E in a finite number of steps.

2) The E-2S-PCE algorithm: Looking at (12), it can be noted that the object-to-cluster assignments (i.e., Γ^*_{kn} values) computed by E-EM-PCE still depend on X_{kn} (indeed Γ^*_{kn} is inversely proportional to $\widetilde{X}_{kn} = X_{kn}/2M + X'_{kn}/(N-1)$). This is a weakness of E-EM-PCE. Although the dependence is mitigated by the presence of X'_{kn}, the reasoning explained in Sect. III-A would suggest that Γ^*_{kn} should not depend on X_{kn} at all (see Example in Sect. III-A). Taking this into account, we propose an alternative heuristic, called *Enhanced 2-Step-based Projective Clustering Ensembles (E-2S-PCE)* (Alg. 2). The proposed E-2S-PCE discards the EM-like optimization paradigm to embrace a scheme consisting of two main steps which are executed sequentially. Since Γ^*_{kn} values should not be influenced by Δ^*_{kd} values (whereas the vice versa should not hold), the key idea is to first compute Γ^*_{kn} independently of Δ^*_{kd} (first step), and, once the optimal Γ^*_{kn} are available, choose Δ^*_{kd} consistently with Γ^*_{kn} (second step). This technique removes the undesired "dependences" (i.e., Γ^*_{kn} on Δ^*_{kd}) while maintaining the desirable ones (i.e., Δ^*_{kn} on Γ^*_{kn}).

Algorithm 2 E-2S-PCE

Input: a projective ensemble $\mathcal{E}$ defined over a set $\mathcal{D}$ of data objects; the number K of clusters in the output projective consensus clustering
Output: the projective consensus clustering C^*

```
1: P ← pairwiseObjectDistances(E)          {(14)}
2: Γ* ← objectPartitioning(D, P, K)
3: Δ* ← deltaValues(Γ*, E)                 {(18)}
4: L* ← {1, ..., K}
5: C* = ⟨L*, Γ*, Δ*⟩
```

In Alg. 2, the first step (Lines 1-2) is carried out by resorting to a well-established idea in standard clustering ensembles, i.e., performing a clustering task over the input set $\mathcal{D}$ of data objects based on pairwise distances derived from the so-called *co-occurrence* matrix $\mathbf{P}$. In order to consider (possibly soft) projective clustering solutions, we define any entry $\mathbf{P}_{nn'}$ ($n \neq n'$) of $\mathbf{P}$ as one minus the probability (cf. (8)) that o_n and $o_{n'}$ are clustered together according to the information available in $\mathcal{E}$:

$$\mathbf{P}_{nn'} = 1 - \frac{1}{M} \sum_{h=1}^{H} \gamma_{hn} \gamma_{hn'} \tag{14}$$

Once the objects in $\mathcal{D}$ have been clustered according to the pairwise distances in $\mathbf{P}$, the feature-to-cluster assignments are computed exploiting the information available from both the results obtained by the clustering task and the ensemble (Line 3). The objective is to represent the set of objects in each cluster discovered in the first step to reflect the feature-to-cluster assignments as accurately as possible. To this end, we define the following optimization problem:

$$\Delta^* = \arg\min_{\hat{\Delta}} Q_{2S}(\hat{\Delta}, \Gamma^*, \mathcal{E}) \tag{15}$$

$$s.t. \quad \sum_{k=1}^{K} \hat{\Delta}_{kd} = 1, \forall k \quad \text{and} \quad \hat{\Delta}_{kd} \geq 0, \forall k, \forall d \tag{16}$$

where

$$Q_{2S}(\hat{\Delta}, \Gamma^*, \mathcal{E}) = \sum_{k=1}^{K} \sum_{n=1}^{N} \Gamma^*_{kn} \sum_{h=1}^{H} \gamma_{hn} \sum_{d=1}^{D} \left(\hat{\Delta}_{kd} - \delta_{hd} \right)^2 \tag{17}$$

Note that Q_{2S} is the same as Q (cf. (4)) in which $\hat{\Gamma}$ values are fixed (i.e., they are given by function Γ^* computed in the first step of E-2S-PCE). This choice is essentially motivated by the reasoning reported in Sect. III-A. Indeed, according to function Q, the feature-to-cluster assignments $\Delta^*_{k1}, \ldots, \Delta^*_{kD}$ of any cluster $\hat{\ell}_k$ are computed by considering the "average" feature-to-cluster assignments of the objects within cluster $\hat{\ell}_k$, i.e., the final Δ^* represents a kind of centroid for the objects assigned to $\hat{\ell}_k$ according to the Γ^* function. The optimal solution of the problem (15)-(16) is given by $(\forall k, \forall d)$:

$$\Delta^*_{kd} = \frac{\hat{Z}_{kd}}{\hat{Y}_k} \tag{18}$$

where

$$\hat{Y}_k = M \sum_{n=1}^{N} \Gamma^*_{kn} \quad \text{and} \quad \hat{Z}_{kd} = \sum_{n=1}^{N} \Gamma^*_{kn} \sum_{h=1}^{H} \gamma_{hn}\ \delta_{hd}$$

3) Computational remarks: Let $\mathcal{D}$ be a set of N D-dimensional objects, $\mathcal{E}$ be a projective ensemble of size M defined over $\mathcal{D}$, and K be the number of clusters in the output projective consensus clustering; also, let us assume that H is $\mathcal{O}(K\ M)$. It can be proved that the computational complexities of the proposed E-EM-PCE and E-2S-PCE are $\mathcal{O}(MKN(IK+D))$ and $\mathcal{O}(MKN(N+D))$, respectively. Note that MOEA-PCE and EM-PCE have costs $\mathcal{O}(ItMK^2(N+D))$ and $\mathcal{O}(MKND)$, respectively.

IV. Experimental evaluation

The objective of the experimental evaluation was to assess accuracy and efficiency of the consensus clusterings obtained by the proposed E-EM-PCE and E-2S-PCE, and to compare them against the MOEA-PCE and EM-PCE algorithms. To this end, we followed the methodology described in [1] for generating ensembles, for setting the parameters of MOEA-PCE and EM-PCE, and for selecting the test datasets. We used eight benchmark datasets from the UCI Machine Learning Repository [10], namely Iris, Wine, Glass, Ecoli, Yeast, Segmentation, Abalone and Letter, and two time-series datasets from the UCR Time Series Classification/Clustering Page [11], namely Tracedata and ControlChart.

A. Assessment criteria

We assessed the quality of a consensus clustering $\check{C} = \langle \check{\mathcal{L}}, \check{\Gamma}, \check{\Delta} \rangle$, with $|\check{\mathcal{L}}| = \check{K}$, by evaluating the similarity of $\check{C}$ w.r.t. a reference classification.

Let $\widetilde{C} = \langle \widetilde{\mathcal{L}}, \widetilde{\Gamma}, \widetilde{\Delta} \rangle$ denote a reference classification, where $\widetilde{\mathcal{L}} = \{\widetilde{\ell}_1, \ldots, \widetilde{\ell}_{\widetilde{K}}\}$ and $\widetilde{\Gamma}$ are provided with $\mathcal{D}$, and $\widetilde{\Delta}$ is computed as suggested in [8], i.e., as $\widetilde{\Delta}_{kd} = \exp(-\mathbf{X}_{kd}/h) / \sum_{d'=1}^{D} \exp(-\mathbf{X}_{kd'}/h)$, $\forall k \in \{1, \ldots, \widetilde{K}\}, d \in \{1, \ldots, D\}$, where the LAC's parameter h was set equal to 0.2 and:

$$\mathbf{X}_{kd} = \left(\sum_{n=1}^{N} \widetilde{\Gamma}_{kn} \right)^{-1} \sum_{n=1}^{N} \widetilde{\Gamma}_{kn} \big(\overline{c}_{kd} - o_{nd} \big)^2$$

$$\overline{c}_{kd} = \left(\sum_{n=1}^{N} \widetilde{\Gamma}_{kn} \right)^{-1} \sum_{n=1}^{N} \widetilde{\Gamma}_{kn} o_{nd}$$

Similarity between $\check{C}$ and $\widetilde{C}$ was computed in terms of the *Normalized Mutual Information*, by taking into account their object-based (NMI_o) representations, feature-based representations (NMI_f), or both (NMI_{of}), and by adapting the original definition given in [2] to handle soft solutions. Due to space limitations, here we report the formal definition of NMI_{of} (NMI_o and NMI_f can be derived in a similar way):

$$NMI_{of}(\check{C}, \widetilde{C}) = \frac{\sum_{k=1}^{\check{K}} \sum_{k'=1}^{\widetilde{K}} \frac{c_{kk'}}{T_{of}} \times \log\left(\frac{N^2 \times c_{kk'}}{T_{of} \times \check{c}_k \times \tilde{c}_{k'}} \right)}{\sqrt{H_{of}(\check{C}) \times H_{of}(\widetilde{C})}}$$

where

$$c_{kk'} = \sum_{n=1}^{N} \sum_{d=1}^{D} \check{\Gamma}_{kn} \check{\Delta}_{kd} \widetilde{\Gamma}_{k'n} \widetilde{\Delta}_{k'd}$$

$$\check{c}_k = \sum_{n=1}^{N} \sum_{d=1}^{D} \check{\Gamma}_{kn} \check{\Delta}_{kd}, \quad \tilde{c}_{k'} = \sum_{n=1}^{N} \sum_{d=1}^{D} \widetilde{\Gamma}_{k'n} \widetilde{\Delta}_{k'd}$$

$$H_{of}(\check{C}) = -\sum_{k=1}^{\check{K}} \frac{\check{c}_k}{N} \log \frac{\check{c}_k}{N}, \quad H_{of}(\widetilde{C}) = -\sum_{k'=1}^{\widetilde{K}} \frac{\tilde{c}_{k'}}{N} \log \frac{\tilde{c}_{k'}}{N}$$

$$T_{of} = \sum_{n=1}^{N} \sum_{d=1}^{D} \left(\sum_{k=1}^{\check{K}} \check{\Gamma}_{kn} \check{\Delta}_{kd} \right) \left(\sum_{k'=1}^{\widetilde{K}} \widetilde{\Gamma}_{k'n} \widetilde{\Delta}_{k'd} \right)$$

This evaluation stage was devised to demonstrate that the consensus clusterings computed by the proposed methods are generally closer to the reference classification than any clustering solution randomly chosen from the ensemble. The ultimate goal was to assess the gain/loss in similarity between $\check{C}$ and $\widetilde{C}$ w.r.t. the average similarity between $\widetilde{C}$ and the solutions in the ensemble. Formally, we were interested in evaluating the quantities Θ_{of}, Θ_o, and Θ_f (the larger each of these quantities, the better the quality of $\check{C}$); Θ_{of} is defined as $NMI_{of}(\check{C}, \widetilde{C}) - \text{avg}_{C \in \mathcal{E}} NMI_{of}(C, \widetilde{C})$ (Θ_o and Θ_f are defined similarly).

B. Results

1) Accuracy: For each algorithm, dataset and ensemble, we performed 50 different runs. We report clustering results obtained by the proposed E-EM-PCE and E-2S-PCE, and the earlier EM-PCE and MOEA-PCE in Table I.

Both E-2S-PCE and E-EM-PCE were able to produce consensus clusterings with higher Θ_{of} than EM-PCE (first 3-column groups in Table I), on average by 0.137 and 0.129, respectively. In particular, compared to EM-PCE, E-EM-PCE obtained an average improvement of 0.018, with a maximum gain of 0.071 (Ecoli), whereas E-2S-PCE obtained an average improvement of 0.026, with peaks above 0.050 on three datasets up to a maximum of 0.084 (Trace). Comparing E-2S-PCE with E-EM-PCE, the former achieved higher quality on nearly all datasets, with an average gain of about 0.010 and peaks on Trace (0.064) and Wine (0.058).

Table I
EVALUATION W.R.T. THE REFERENCE CLASSIFICATION

data	Θ_{of}				Θ_o				Θ_f			
	MOEA PCE	*EM PCE*	*E-EM PCE*	*E-2S PCE*	*MOEA PCE*	*EM PCE*	*E-EM PCE*	*E-2S PCE*	*MOEA PCE*	*EM PCE*	*E-EM PCE*	*E-2S PCE*
Iris	+.146	+.168	+.167	+.098	+.319	+.228	+.252	+.169	+.198	-.095	-.092	-.017
Wine	+.136	+.083	+.090	+.148	+.201	+.130	+.134	+.294	+.152	+.030	+.030	-.006
Glass	+.105	+.162	+.173	+.165	+.092	+.134	+.134	+.141	+.048	+.060	+.059	+.224
Ecoli	+.164	+.086	+.157	+.144	+.245	+.125	+.159	+.144	+.042	+.042	+.044	+.219
Yeast	+.049	+.021	+.036	+.057	+.090	+.066	+.076	+.057	+.006	+.090	+.092	+.080
Segm.	+.137	+.144	+.158	+.159	+.102	+.206	+.207	+.182	+.075	+.079	+.079	+.055
Abal.	+.116	+.111	+.129	+.137	+.141	+.116	+.129	+.138	+.093	+.092	+.093	+.120
Letter	+.111	+.107	+.128	+.137	+.146	+.122	+.134	+.143	+.092	+.097	+.087	+.125
Trace	+.097	+.019	+.039	+.103	+.032	+.026	+.033	+.087	-.007	+.114	+.115	-.013
Contr.	+.091	+.204	+.209	+.220	+.050	+.011	+.034	+.027	+.233	+.416	+.416	+.416
min	+.049	+.019	+.036	+.057	+.032	+.011	+.033	+.027	-.007	-.095	-.092	-.017
max	+.164	+.204	+.209	+.220	+.319	+.228	+.252	+.294	+.233	+.416	+.416	+.416
avg	+.115	+.110	+.129	+.137	+.142	+.116	+.129	+.138	+.093	+.093	+.092	+.120

The superior performance of E-2S-PCE and E-EM-PCE w.r.t. EM-PCE was also confirmed in terms of object-based representations, with average Θ_o equal to 0.138 and 0.129, and average improvement w.r.t. EM-PCE of 0.022 and 0.013, respectively. In terms of feature-based representations, the two outperforming methods led to an average Θ_f equal to 0.12 and 0.092; compared to EM-PCE, E-2S-PCE obtained an average improvement of 0.028 and E-EM-PCE was substantially comparable to EM-PCE.

A further important remark is that the proposed algorithms in general reduce the gap between MOEA-PCE and the basic EM-PCE. Indeed, looking at Table I, E-EM-PCE and E-2S-PCE obtained average Θ values which are comparable or even better than MOEA-PCE. More precisely, E-EM-PCE and E-2S-PCE improved MOEA-PCE by 0.014 and 0.022, respectively, in terms of Θ_{of}. Furthermore, E-2S-PCE outperformed MOEA-PCE by 0.027 in terms of Θ_f, and was still comparable in terms of Θ_o.

2) Efficiency: Table II reports the runtimes of the algorithms MOEA-PCE, EM-PCE, E-EM-PCE, and E-2S-PCE. EM-PCE maintained its advantage in terms of efficiency w.r.t. E-2S-PCE and E-EM-PCE; nevertheless, the advantage w.r.t. E-EM-PCE and E-2S-PCE was noticeable only when the ratios K/D (i.e., the number of clusters to the number of features) and N/D (i.e., the number of objects to the number of features) increase, respectively. As an example, the relative times of EM-PCE, E-EM-PCE and E-2S-PCE were close to each other on Tracedata and ControlChart, which are the datasets having the minimum K/D and N/D ratios. Nevertheless, the major claim of this work is confirmed: both the proposed E-2S-PCE and E-EM-PCE algorithms maintain a large efficiency gain w.r.t. MOEA-PCE, like the basic EM-PCE.

V. CONCLUSION

The projective clustering ensembles (PCE) problem was originally introduced in [1] to provide a robust projective consensus clustering from a given ensemble of projective clustering solutions. In this paper, we have focused on the formulation of PCE as a single-objective optimization problem, and proposed new, well-founded enhancements to single-objective PCE in order to overcome major limitations of the early formulation in terms of effectiveness. As a result, we have developed two heuristics, namely E-EM-PCE and E-2S-PCE, which follow different approaches to embedding both object-based and feature-based cluster representations in the objective function. Experimental evidence has shown that the new single-objective PCE algorithms achieve significant improvements in accuracy. At the same time, they still maintain a large advantage in terms of efficiency w.r.t. the two-objective PCE.

Table II
EXECUTION TIMES (MILLISECONDS)

data	*MOEA-PCE*	*EM-PCE*	*E-EM-PCE*	*E-2S-PCE*
Iris	17,223	55	250	353
Wine	21,098	184	477	522
Glass	61,700	281	1,257	939
Ecoli	94,762	488	2,354	2,291
Yeast	1,310,263	1,477	5,459	80,158
Segm.	1,250,732	11,465	37,048	154,720
Abal.	13,245,313	34,000	312,485	1,875,968
Letter	7,765,750	54,641	451,453	2,057,187
Trace	86,179	4,880	4,138	2,285
Contr.	291,856	2,313	2,900	9,874

ACKNOWLEDGEMENTS

This work is in part supported by NSF CAREER Award IIS-0447814.

REFERENCES

[1] F. Gullo, C. Domeniconi, and A. Tagarelli, "Projective Clustering Ensembles," in *Proc. ICDM Conf.*, 2009, pp. 794–799.

[2] A. Strehl and J. Ghosh, "Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions," *JMLR*, vol. 3, pp. 583–617, 2002.

[3] A. P. Topchy, A. Jain, and W. Punch, "Clustering Ensembles: Models of Consensus and Weak Partitions," *TPAMI*, vol. 27, no. 12, pp. 1866–1881, 2005.

[4] X. Z. Fern and C. Brodley, "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning," in *Proc. ICML Conf.*, 2004, pp. 281–288.

[5] C. Domeniconi and M. Al-Razgan, "Weighted Cluster Ensembles: Methods and Analysis," *TKDD*, vol. 2, no. 4, 2009.

[6] M. L. Yiu and N. Mamoulis, "Iterative Projected Clustering by Subspace Mining," *TKDE*, vol. 17, no. 2, pp. 176–189, 2005.

[7] E. Achtert, C. Böhm, H. P. Kriegel, P. Kröger, I. Müller-Gorman, and A. Zimek, "Finding Hierarchies of Subspace Clusters," in *Proc. PKDD Conf.*, 2006, pp. 446–453.

[8] C. Domeniconi, D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, and D. Papadopoulos, "Locally Adaptive Metrics for Clustering High Dimensional Data," *Data Mining and Knowledge Discovery*, vol. 14, no. 1, pp. 63–97, 2007.

[9] G. Moise, J. Sander, and M. Ester, "Robust projected clustering," *Knowl. Inf. Syst.*, vol. 14, no. 3, pp. 273–298, 2008.

[10] A. Asuncion and D. Newman, "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml/.

[11] E. Keogh, X. Xi, L. Wei, and C. A. Ratanamahatana, "The UCR Time Series Classification/Clustering Page," http://www.cs.ucr.edu/∼eamonn/time_series_data/.