The Projective Clustering Ensemble problem for Advanced Data Clustering

EXTENDED ABSTRACT

Carlotta Domeniconi[1], Francesco Gullo[2], and Andrea Tagarelli[3]

[1] George Mason University, USA cdomenic@gmu.edu
[2] UniCredit, R&D Dept., Italy gullof@acm.org
[3] University of Calabria, Italy andrea.tagarelli@unical.it

Abstract. After more than five decades, a huge number of models and algorithms have been developed for data clustering. While most attention has been devoted to data types, algorithmic features, and application targets, in the last years there has also been an increasing interest in developing advanced data-clustering tools. In this respect, projective clustering and clustering ensembles represent two of the most important directions: the former is concerned with the discovery of subsets of the input data having different, possibly overlapping subsets of features associated with them, while the latter allows for the induction of a prototype consensus clustering from an available ensemble of clustering solutions.
In this paper we discuss the current state-of-the-art research in which the problems of projective clustering and clustering ensembles have been revisited and integrated in a unified framework, called Projective Clustering Ensemble (PCE). We discuss how PCE has originally been formalized as either a two-objective or a single-objective optimization problem, and how the limitations of such early approaches have been overcome by a metacluster-based formulation. We also summarize main empirical results, and provide pointers for future research.

1 Introduction

Given a set of data objects as points in a multi-dimensional space (or *feature space*), the problem of *clustering* consists in discovering a number of homogeneous subsets of data, called *clusters*, which are well-separated from each other [8]. Clustering is a key step for a myriad of data-management/data-mining tasks that require the discovery of unknown relationships and patterns in large datasets. Although most clustering approaches usually provide single clustering solutions and/or use the same (typically large) feature space, latest advances have focused on two main advanced aspects: (i) dealing with high dimensionality, and (ii) handling multiple clustering solutions.

Several problems of practical interest are inherently high-dimensional, i.e., they involve data objects represented by large sets of features. A common scenario with high-dimensional data is that several clusters may exist in different subspaces that correspond to different combinations of features. In fact, it is unlikely that all features of the data objects are equally relevant to form meaningful clusters.

Another challenge in the clustering process is that in many real-life domains multiple, and potentially meaningful groupings of the input data are available, hence providing different views of the data. For instance, in genomics, multiple clustering solutions

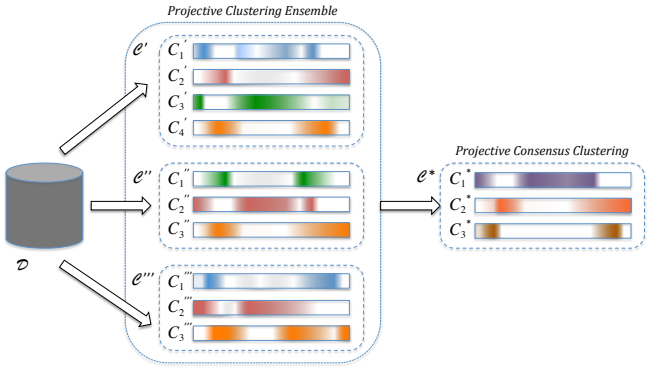

Fig. 1: Illustration of a projective clustering ensemble and derived consensus clustering. Each gradient refers to the cluster memberships over all objects. Colors denote different feature subspaces associated with the projective clusters [5].

would be needed to capture the multiple functional roles of genes. In text mining, documents discuss multiple topics, thus their grouping by content should reflect different informative views which correspond to multiple (possibly alternative) clusterings.

Recent advances in data clustering have led to the definition of the problems of *projective clustering*—to deal with high dimensionality—and *clustering ensembles*—to handle multiple clustering solutions. The goal of projective clustering [9] is to discover *projective clusters*, i.e., subsets of the input data having different, and possibly overlapping subsets of features (subspaces) associated with them. Projective clustering aims to solve issues that typically arise in high-dimensional data, such as sparsity and concentration of distances [7]. The problem of clustering ensembles [12], also known as *consensus clustering* or *aggregation clustering*, is stated as follows: given a set of clustering solutions, or *ensemble*, one must derive a *consensus clustering* that properly summarizes the solutions of the ensemble. The input ensemble is typically 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 and clustering ensembles have recently been treated in a unified framework [2,3,4,5,6]. The underlying motivation of that study is that many real-world problems are high-dimensional *and* lack a-priori knowledge. Examples include clustering of multi-view data, privacy preserving clustering, news or document retrieval based on pre-defined categorizations, and distributed clustering of high-dimensional data. To address both issues simultaneously, the problem of *Projective Clustering Ensembles* (PCE) is hence formalized, whose goal is to compute a *projective consensus clustering* from an ensemble of projective clustering solutions. Intuitively, each projective cluster is characterized by a distribution of memberships of the objects as well as a distribution over the features that belong to the subspace of that cluster. Figure 1 illustrates an example of projective clustering ensemble with three projective clustering solutions, which are obtained according to different views over the same dataset. A projective cluster is graphically represented as a rectangle filled with a color gradient, where higher intensities correspond to larger membership values of objects to the cluster. Clusters of the same clustering may overlap with their gradient (i.e., objects can have multiple assignments with different degrees of membership), and colors change to denote that different groupings of objects are associated with different feature

subspaces. A projective consensus clustering is derived by suitably "aggregating" the ensemble members: the first projective consensus cluster is derived by summarizing C'_1, C''_2, and C'''_2, the second from C'_2, C''_3, and C'''_3, and the third from C'_4, C''_1, and C'''_1. Note that the resulting color in each projective consensus cluster would merge colors in the original projective clusters, which means that a projective consensus cluster is associated with a subset of features shared by the objects in the original clusters.

In this paper we provide an overview of the PCE problem. We discuss how PCE has been originally formalized as either a two-objective or a single-objective optimization problem, and how the limitations of the early approaches have been overcome by a metacluster-based method that is able to jointly consider the object-based and feature-based cluster representations in a single-objective optimization problem. We also summarize main empirical results obtained on benchmark datasets. We finally provide pointers for future research in such a challenging context.

2 Projective Clustering Ensembles (PCE)

Let $\mathcal{D}$ be a set of data objects, where each $\mathbf{o} \in \mathcal{D}$ is an $|\mathcal{F}|$-dimensional point defined over a feature space $\mathcal{F}$.[4] A *projective cluster* C defined over $\mathcal{D}$ is a pair $\langle \mathbf{\Gamma}_C, \mathbf{\Delta}_C \rangle$. $\mathbf{\Gamma}_C$, termed the *object-based* representation of C, is a $|\mathcal{D}|$-dimensional real-valued vector whose components $\mathit{\Gamma}_{C,\mathbf{o}} \in [0,1]$, $\forall \mathbf{o} \in \mathcal{D}$, represent the *object-to-cluster* assignment of $\mathbf{o}$ to C, i.e., the probability $\Pr(\mathbf{o}|C)$ that the object $\mathbf{o}$ belongs to C. $\mathbf{\Delta}_C$, termed the *feature-based* representation of C, is an $|\mathcal{F}|$-dimensional real-valued vector whose components $\mathit{\Delta}_{C,f} \in [0,1]$, $\forall f \in \mathcal{F}$, represent the *feature-to-cluster* assignments of the f-th feature to C, i.e., the probability $\Pr(f|C)$ that the feature f is *informative* for cluster C (f belongs to the subspace associated with C).

The object-based ($\mathbf{\Gamma}_C$) and the feature-based ($\mathbf{\Delta}_C$) representations of a projective cluster C implicitly define the *projective cluster representation matrix* (for short, *projective matrix*) X_C of C. X_C is a $|\mathcal{D}| \times |\mathcal{F}|$ matrix that stores, $\forall \mathbf{o} \in \mathcal{D}$, $f \in \mathcal{F}$, the probability of the intersection of the events "*object* $\mathbf{o}$ *belongs to* C" and "*feature* f *belongs to the subspace associated with* C". Under the assumption of independence between the two events, such a probability is equal to $\Pr(C|\mathbf{o}) = \mathit{\Gamma}_{C,\mathbf{o}}$ joint with $\Pr(f|C) = \mathit{\Delta}_{C,f}$. Hence, given $\mathcal{D} = \{\mathbf{o}_1, \ldots, \mathbf{o}_{|\mathcal{D}|}\}$ and $\mathcal{F} = \{1, \ldots, |\mathcal{F}|\}$, the matrix X_C can formally be defined as $\mathrm{X}_C = \mathbf{\Gamma}_C^{\mathrm{T}} \, \mathbf{\Delta}_C$.

A *projective clustering solution*, denoted by $\mathcal{C}$, is defined as a set of projective clusters that satisfy $\sum_{C \in \mathcal{C}} \mathit{\Gamma}_{C,\mathbf{o}} = 1$, $\forall \mathbf{o} \in \mathcal{D}$, and $\sum_{f \in \mathcal{F}} \mathit{\Delta}_{C,f} = 1$, $\forall C \in \mathcal{C}$. The semantics of any projective clustering $\mathcal{C}$ is that for each projective cluster $C \in \mathcal{C}$, the objects belonging to C are close to each other if (and only if) they are projected onto the subspace associated with C.

A *projective ensemble* $\mathcal{E}$ is defined as a set of projective clustering solutions. No information about the ensemble generation strategy (algorithms and/or setups), nor original feature values of the objects within $\mathcal{D}$ are provided along with $\mathcal{E}$. Moreover, each projective clustering solution in $\mathcal{E}$ may contain in general a different number of clusters.

The goal of PCE is to derive a *projective consensus clustering* that properly summarizes the projective clustering solutions within the input projective ensemble.

[4] Vectorial notation here denotes row vectors.

2.1 Single-objective PCE

The earliest PCE formulation proposed in [5] is based on a single-objective function:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \sum_{C \in \mathcal{C}} \sum_{\mathbf{o} \in \mathcal{D}} \Gamma_{C,\mathbf{o}}^{\alpha} \, X_{C,\mathbf{o}}, \tag{1}$$

$$\text{where} \qquad X_{C,\mathbf{o}} = \sum_{f \in \mathcal{F}} \left(\Delta_{C,f} - \Lambda_{\mathbf{o},f}\right)^2, \quad \Lambda_{\mathbf{o},f} = \frac{1}{|\mathcal{E}|} \sum_{\hat{\mathcal{C}} \in \mathcal{E}} \sum_{\hat{C} \in \hat{\mathcal{C}}} \Gamma_{\hat{C},\mathbf{o}} \, \Delta_{\hat{C},f}, \tag{2}$$

and $\alpha > 1$ is a positive integer. To solve the above optimization problem, the *EM-based Projective Clustering Ensembles (EM-PCE)* method is proposed in [5]. EM-PCE iteratively looks for the optimal values of $\Gamma_{C,\mathbf{o}}$ (resp. $\Delta_{C,f}$) while keeping $\Delta_{C,f}$ (resp. $\Gamma_{C,\mathbf{o}}$) fixed, until convergence.

Weaknesses of single-objective PCE. The objective function in (1) does not allow for a perfect balance between object- and feature-to-cluster assignments when measuring the error of a candidate projective consensus clustering solution. This weakness is formally shown in [3] and avoided by adjusting (1) with a corrective term. The resulting formulation of the problem based on the corrected objective function is the following:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \sum_{C \in \mathcal{C}} \sum_{\mathbf{o} \in \mathcal{D}} \Gamma_{C,\mathbf{o}}^{\alpha} \, \left(\frac{1}{2|\mathcal{E}|} X_{C,\mathbf{o}} + \frac{1}{|\mathcal{D}| - 1} X'_{C,\mathbf{o}} \right), \tag{3}$$

$$\text{where} \qquad X'_{C,\mathbf{o}} = \sum_{\mathbf{o}' \neq \mathbf{o}} \left(1 - \frac{\Gamma_{C,\mathbf{o}'}}{|\mathcal{E}|} \sum_{\hat{\mathcal{C}} \in \mathcal{E}} \sum_{\hat{C} \in \hat{\mathcal{C}}} \Gamma_{\hat{C},\mathbf{o}} \, \Gamma_{\hat{C},\mathbf{o}'} \right).$$

The above optimization problem is tackled in [3] by proposing two different methods. The first one, called *E-EM-PCE*, follows the same scheme as the EM-PCE algorithm for the early single-objective PCE formulation. The second method, dubbed *E-2S-PCE*, consists of two sequential steps that handle the object-to-cluster and the feature-to-cluster assignments separately.

2.2 Two-objective PCE

PCE is formulated in [5] also as a two-objective problem, whose functions account for the object-based (function Ψ_o) and the feature-based (function Ψ_f) representations:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \, \{\Psi_o(\mathcal{C}, \mathcal{E}), \, \Psi_f(\mathcal{C}, \mathcal{E})\}, \tag{4}$$

$$\text{where} \qquad \Psi_o(\mathcal{C}, \mathcal{E}) = \sum_{\hat{\mathcal{C}} \in \mathcal{E}} \overline{\psi}_o(\mathcal{C}, \hat{\mathcal{C}}), \qquad \Psi_f(\mathcal{C}, \mathcal{E}) = \sum_{\hat{\mathcal{C}} \in \mathcal{E}} \overline{\psi}_f(\mathcal{C}, \hat{\mathcal{C}}). \tag{5}$$

Functions $\overline{\psi}_o$ and $\overline{\psi}_f$ are defined as $\overline{\psi}_o(\mathcal{C}', \mathcal{C}'') = \frac{1}{2}(\psi_o(\mathcal{C}', \mathcal{C}'') + \psi_o(\mathcal{C}'', \mathcal{C}'))$ and $\overline{\psi}_f(\mathcal{C}', \mathcal{C}'') = \frac{1}{2}\left(\psi_f(\mathcal{C}', \mathcal{C}'') + \psi_f(\mathcal{C}'', \mathcal{C}')\right)$, respectively, where

$$\psi_o(\mathcal{C}', \mathcal{C}'') = \frac{1}{|\mathcal{C}'|} \sum_{C' \in \mathcal{C}'} (1 - \max_{C'' \in \mathcal{C}''} J(\mathbf{\Gamma}_{C'}, \mathbf{\Gamma}_{C''})),$$

$$\psi_f(\mathcal{C}', \mathcal{C}'') = \frac{1}{|\mathcal{C}'|} \sum_{C' \in \mathcal{C}'} (1 - \max_{C'' \in \mathcal{C}''} J(\mathbf{\Delta}_{C'}, \mathbf{\Delta}_{C''})),$$

and $J(\mathbf{u}, \mathbf{v}) = (\mathbf{u}\,\mathbf{v}^{\mathrm{T}}) / (\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 - \mathbf{u}\,\mathbf{v}^{\mathrm{T}})$ denotes the Tanimoto coefficient.

The problem defined in (4) is solved by the *MOEA-PCE* method, in which a *Pareto*-based *Multi-Objective Evolutionary Algorithm* is exploited to avoid combining the two objective functions into a single one.

Weaknesses of two-objective PCE. Experimental evidence in [5] has shown that the two-objective PCE formulation is more accurate than the single-objective one. Nevertheless, the original two-objective PCE still suffers from an important conceptual issue: it does not take into consideration the interrelation between the object-based and the feature-based cluster representations. This can be overcome by employing a *metacluster-based* PCE formulation, which we discuss next.

2.3 Metacluster-based PCE

Metacluster-based PCE is defined in terms of the following single-objective function [4]:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \; \Psi_{of}(\mathcal{C}, \mathcal{E}), \tag{6}$$

where Ψ_{of} is a function designed to measure the "distance" of a projective clustering solution $\mathcal{C}$ from $\mathcal{E}$ in terms of both data clustering and feature-to-cluster assignment: $\Psi_{of}(\mathcal{C}, \mathcal{E}) = \sum_{\hat{\mathcal{C}} \in \mathcal{E}} \overline{\psi}_{of}(\mathcal{C}, \hat{\mathcal{C}})$, where $\overline{\psi}_{of}(\mathcal{C}', \mathcal{C}'') = \frac{1}{2}(\psi_{of}(\mathcal{C}', \mathcal{C}'') + \psi_{of}(\mathcal{C}'', \mathcal{C}'))$, $\psi_{of}(\mathcal{C}', \mathcal{C}'') = \frac{1}{|\mathcal{C}'|} \sum_{C' \in \mathcal{C}'} (1 - \max_{C'' \in \mathcal{C}''} \hat{J}(\mathrm{X}_{C'}, \mathrm{X}_{C''}))$, and $\hat{J}$ is the Tanimoto coefficient between the linearized versions of the matrices to be compared.

The PCE formulation reported in (6) is well-suited to measure the quality of a candidate consensus clustering in terms of both object-to-cluster and feature-to-cluster assignments as a whole. As shown in [4], this enables us to overcome the conceptual disadvantages of both early single-objective and two-objective PCE.

To solve the metacluster-based PCE problem, a two-step is defined in [4]. The method, called CB-PCE, first discovers a suitable metacluster structure, and then computes optimal object- and feature-based representations of each metacluster.

2.4 Enhanced metacluster-based PCE

Although the early metacluster-based PCE formulation in (6) mitigates the issues of two-objective PCE, such a formulation has still some limitations. By some rearrangement and omitting constant terms, the objective function in (6) can be rewritten as:

$$\Psi_{of}(\mathcal{C}, \mathcal{E}) = \underbrace{\sum_{\hat{\mathcal{C}} \in \mathcal{E}} \sum_{\hat{C} \in \hat{\mathcal{C}}} \min_{C \in \mathcal{C}} \left(1 - \hat{J}\left(\mathrm{X}_C, \mathrm{X}_{\hat{C}}\right)\right)}_{\Psi'_{of}(\mathcal{C}, \mathcal{E})} + \underbrace{\sum_{C \in \mathcal{C}} \sum_{\hat{\mathcal{C}} \in \mathcal{E}} \min_{\hat{C} \in \hat{\mathcal{C}}} \left(1 - \hat{J}\left(\mathrm{X}_C, \mathrm{X}_{\hat{C}}\right)\right)}_{\Psi''_{of}(\mathcal{C}, \mathcal{E})}. \tag{7}$$

As formally shown in [6], both Ψ'_{of} and Ψ''_{of} suffer from conceptual drawbacks: the optimization of Ψ'_{of} favors projective clustering solutions composed of clusters coming all from the same input ensemble solution, while the optimization of Ψ''_{of} favors the replication of the same projective cluster within the output consensus projective clustering.

Table 1: Datasets used in the experimental evaluation

dataset	*objects*	*features*	*classes*
Iris	150	4	3
Wine	178	13	3
Glass	214	10	6
Ecoli	327	7	5
Yeast	1,484	8	10
Multiple-Features	2,000	585	10
Segmentation	2,310	19	7
Abalone	4,124	7	17
Waveform	5,000	40	3
Letter	7,648	16	10
Isolet	7,797	617	26
Gisette	13,500	5,000	2
p53-Mutants	300	5,409	2
Amazon	120	10,000	4
Arcene	200	10,000	2

dataset	*objects*	*features*	*classes*
Shapes	160	1,614	9
Tracedata	200	275	4
ControlChart	600	60	6
Twopat	800	128	4
N30	1,356	20	8
D75	1,365	75	7
S2500	2,262	20	8

To solve these issues, an enhanced metacluster-based PCE formulation is proposed in [6]: Given a projective ensemble $\mathcal{E}$ defined over objects $\mathcal{D}$ and features $\mathcal{F}$, and an integer $K > 0$, find a projective clustering solution $\mathcal{C}^*$ such that $|\mathcal{C}^*| = K$ and

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \sum_{C \in \mathcal{C}} \sum_{\hat{\mathcal{C}} \in \mathcal{E}} \sum_{\hat{C} \in \hat{\mathcal{C}}} x(\hat{C}, C)\, T(\hat{C}, C) \tag{8}$$

$$s.t. \quad \sum_{C \in \mathcal{C}} \Gamma_{C,\mathbf{o}} = 1, \forall \mathbf{o} \in \mathcal{D}, \quad \sum_{f \in \mathcal{F}} \Delta_{C,f} = 1, \forall C \in \mathcal{C},$$

$$x(\hat{C}, C) \in \{0, 1\}, \quad \forall \hat{\mathcal{C}} \in \mathcal{E},\ \forall \hat{C} \in \hat{\mathcal{C}},\ \forall C \in \mathcal{C},$$

$$\sum_{C \in \mathcal{C}} x(\hat{C}, C) \geq 1, \quad \forall \hat{\mathcal{C}} \in \mathcal{E},\ \forall \hat{C} \in \hat{\mathcal{C}}, \tag{9}$$

$$\sum_{\hat{C} \in \hat{\mathcal{C}}} x(\hat{C}, C) \geq 1, \quad \forall \hat{\mathcal{C}} \in \mathcal{E},\ \forall C \in \mathcal{C}. \tag{10}$$

The above ($\mathbf{NP}$-hard [6]) problem is solved by E-CB-PCE, i.e., a more effective and more efficient variant of the method devised for early metacluster-based PCE.

3 Experiments

In this section we report the main experimental findings on the various state-of-the-art PCE methods, i.e., MOEA-PCE [2,5], EM-PCE [2,5], E-EM-PCE [3], E-2S-PCE [3], CB-PCE [4], and E-CB-PCE [6].

Datasets. The evaluation was performed on 22 publicly-available benchmark datasets, whose main characteristics are shown in Table 1.[5]

Projective-ensemble generation. Projective ensembles were generated by running the well-established projective-custering LAC algorithm [1] on the selected datasets. The diversity of the projective clustering solutions was ensured by randomly choosing the initial centroids and varying the LAC's parameter h. For each dataset, we generated 10 different projective ensembles; all results were averaged over these 10 ensembles.

[5] The first 15 datasets are from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml), the next four ones are from the UCR Time Series Classification/Clustering Page (http://www.cs.ucr.edu/ eamonn/timeseries data), whereas the last three ones are synthetic datasets from http://dme.rwth-aachen.de/en/OpenSubspace/evaluation.

Table 2: Average gains of E-CB-PCE w.r.t. the other methods

criterion	*MOEA-PCE*	*EM-PCE*	*E-EM-PCE*	*E-2S-PCE*	*CB-PCE*
$F1_{of}$	0	.014	.021	.025	.020
$F1_o$	.029	.049	.054	.108	.023
$F1_f$	.005	.056	.055	.060	.078
$\overline{F1}_{of}$	.147	.186	.194	.237	.019
$\overline{F1}_o$	.021	.046	.055	.100	.021
$\overline{F1}_f$	.022	.066	.067	.076	-.010
avg	.037	.070	.074	.101	.025

Assessment criteria. The quality of a projective consensus clustering $\mathcal{C}$ was assessed by using both external and internal evaluations: the former compares $\mathcal{C}$ against a reference classification, whereas the latter is based on the average similarity w.r.t. the solutions of the input projective ensemble. Three variants of the classic *F1-measure* were used as external criteria, aiming at comparing projective clustering solutions in terms of their object-based representation ($F1_o$), feature-based representation ($F1_f$), or both ($F1_{of}$), respectively. The same criteria were used for internal evaluation too. The difference in this case is that the target consensus clustering $\mathcal{C}$ is compared against all solutions in the input ensemble, and an aggregated score is ultimately derived from all individual scores. Again, three variants were employed, accounting for object-based representation ($\overline{F1}_o$), feature-based representation ($\widehat{F1}_f$), or both ($\overline{F1}_{of}$), respectively. For a more detailed definition of all $F1_o$, $F1_f$, $F1_{of}$, $\overline{F1}_o$, $\overline{F1}_f$, and $\overline{F1}_{of}$ criteria, please refer to [6].

Results. Table 2 shows the accuracy of all methods for each assessment criterion, averaged on all the selected datasets.[6] In particular the table shows the average gain achieved by the E-CB-PCE method, which was recognized as the most accurate one. E-CB-PCE was more accurate than both MOEA-PCE and CB-PCE on 16 out of 22 datasets on average, while achieving average gains up to 0.147 ($\overline{F1}_{of}$ assessment criterion) and 0.078 ($F1_f$ assessment criterion) w.r.t. MOEA-PCE and CB-PCE, respectively. Larger improvements were produced by E-CB-PCE w.r.t. the early single-objective PCE methods, i.e., EM-PCE, E-EM-PCE, and E-2S-PCE.

Table 3 shows the runtimes (in milliseconds) of the various algorithms involved in the comparison. E-CB-PCE was slower than EM-PCE on most datasets, while clearly outperforming MOEA-PCE. The runtimes of E-CB-PCE were one or two orders of magnitude smaller than those of MOEA-PCE on average, up to four orders on Isolet. Only on one dataset, MOEA-PCE was more efficient than E-CB-PCE (Glass), even though the runtimes of the two methods remained of the same order of magnitude. Compared to CB-PCE, E-CB-PCE was faster on 11 out of 22 datasets.

4 Conclusions and Future Work

We provided an overview of research work on projective clustering ensembles (PCE). We discussed the major reasons behind the formulation of PCE as a new topic in the data clustering field, which required novel computational approaches and algorithmic solutions. We formally presented a summary of existing formulations of PCE, and reported major findings learned from experimental evaluation studies.

Our research work paved the way for the development of data clustering frameworks that can support a variety of current and emerging applications in several domains of data management and analysis. For instance, in the complex network systems

[6] Dataset-by-dataset scores are available in [6].

Table 3: Execution times (milliseconds)

dataset	*MOEA-PCE*	*EM-PCE*	*E-EM-PCE*	*E-2S-PCE*	*CB-PCE*	**E-CB-PCE**
Iris	2,056	37	109	253	74	1,492
Wine	2,558	29	88	163	144	1,223
Glass	7,712	56	615	248	500	9,177
Ecoli	14,401	59	685	625	1,147	9,337
Yeast	227,878	757	20,438	21,560	30,384	83,008
Mult.-Feat.	490,602	116,334	151,582	88,713	1,562,139	114,368
Segmentation	233,951	1,854	17,385	40,014	25,692	31,780
Abalone	3,411,116	4,105	156,959	430,974	275,053	857,164
Waveform	125,247	4,912	15,905	91,844	10,487	2,179
Letter	2,248,695	9,591	126,222	772,883	70,458	135,832
Isolet	20,676,754	10,100	10,000	8,488	66,136	1,447
Gisette	966,108	34,260	38,216	29,700	93,148	1,450
p53-Mutants	58,695	22,209	22,168	19,347	65,501	1,490
Amazon	395,988	21,556	21,619	20,914	135,446	12,737
Arcene	120,537	21,557	21,499	17,405	81,433	2,413
Shapes	211,654	17,752	19,152	12,473	282,857	106,180
Tracedata	12,777	1,062	1,108	960	9,000	4,716
ControlChart	50,798	1,522	5,397	2,801	13,900	20,708
Twopat	31,850	2,946	5,706	4,606	9,788	5,344
N30	164,969	1,340	13,781	16,243	22,904	44,517
D75	135,297	4,558	13,414	15,477	32,938	30,631
S2500	290,408	2,717	29,223	44,008	39,039	55,607

area, one interesting direction is to exploit PCE for addressing problems of dimensionality reduction and community detection in time-evolving attributed networks. Another problem that can benefit from our framework is outlier detection in high dimensional data, which has many of the same challenges as clustering. Some work has been done on outlier ensembles (e.g., [10]) and on subspace outlier detection (e.g., [11]), but a unified framework encompassing both does not exist.

References

1. Domeniconi, C., Gunopulos, D., Ma, S., Yan, B., Al-Razgan, M., Papadopoulos, D.: Locally Adaptive Metrics for Clustering High Dimensional Data. Data Mining and Knowledge Discovery 14(1), 63–97 (2007)
2. Gullo, F., Domeniconi, C., Tagarelli, A.: Projective Clustering Ensembles. In: Proc. IEEE Int. Conf. on Data Mining (ICDM). pp. 794–799 (2009)
3. Gullo, F., Domeniconi, C., Tagarelli, A.: Enhancing Single-Objective Projective Clustering Ensembles. In: Proc. IEEE Int. Conf. on Data Mining (ICDM). pp. 833–838 (2010)
4. Gullo, F., Domeniconi, C., Tagarelli, A.: Advancing data clustering via projective clustering ensembles. In: Proc. ACM SIGMOD Int. Conf. on Management of Data. pp. 733–744 (2011)
5. Gullo, F., Domeniconi, C., Tagarelli, A.: Projective Clustering Ensembles. Data Min. Knowl. Discov. 26(3), 452–511 (2013)
6. Gullo, F., Domeniconi, C., Tagarelli, A.: Metacluster-based projective clustering ensembles. Machine Learning 98(1), 181–216 (2015)
7. Hinneburg, A., Aggarwal, C.C., Keim, D.A.: What Is the Nearest Neighbor in High Dimensional Spaces? In: Proc. VLDB Conf. pp. 506–515 (2000)
8. Jain, A., Dubes, R.: Algorithms for Clustering Data. Prentice-Hall (1988)
9. Kriegel, H.P., Kröger, P., Zimek, A.: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. TKDD 3(1), 1:1–1:58 (2009)
10. Rayana, S., Akoglu, L.: Less is more: Building selective anomaly ensembles. TKDD 10(4), 42:1–42:33 (2016)
11. Sathe, S., Aggarwal, C.C.: Subspace outlier detection in linear time with randomized hashing. In: Proc. IEEE Int. Conf. on Data Mining (ICDM). pp. 459–468 (2016)
12. Strehl, A., Ghosh, J.: Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions. Journal of Machine Learning Research 3, 583–617 (2002)