Improving Cluster Analysis by Co-initializations (Supplemental Document)

He Zhang*, Zhirong Yang, Erkki Oja

Department of Information and Computer Science, Aalto University, Espoo, Finland

Abstract

In this document we give the following supplemental information: 1) detailed description of the datasets in the experiments, 2) the learning objective and constraints of the compared methods, 3) more experiment results that are complementary to those in the paper.

1. Datasets

Table 1 gives the information (sources and links) of 19 datasets used in the experiments. These datasets have been widely utilized by machine learning and data mining community, and are freely downloadable from the Internet. A brief description is given in the following (problems, samples, classes, and dimensionalities).

- ORL: the AT&T *ORL* database of face images. There are 10 different images for each of 40 distinct subjects with varying lighting conditions and facial expressions. The size of each image is 92×112 pixels, with 256 grey levels per pixel. In summary, the dataset contains 400 samples grouped in 40 classes, with each sample having a dimensionality of 10,304.

- MED: the *MED* database contains abstract text collections. There are 696 documents organized in 25 topics, with a dictionary containing 5,831 words. In summary, the dataset contains 696 samples grouped in 25 classes, with each sample having a dimensionality of 5,831.

- VOWEL: the LIBSVM *vowel* dataset, originally from UCI machine learning repository. The problem is specified by the accompanying data file, "vowel.data" which consists of a three dimensional array: voweldata [speaker, vowel, input]. The speakers are indexed by integers 0-89. The vowels are indexed by integers 0-10. For each utterance, there are ten floating-point input values, with array

*Corresponding author.

Email addresses: `he.zhang@aalto.fi` (He Zhang), `zhirong.yang@aalto.fi` (Zhirong Yang), `erkki.oja@aalto.fi` (Erkki Oja)

Table 1: Data information.

Dataset	Source	URL
ORL	ORL	http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
MED	LSI	http://web.eecs.utk.edu/research/lsi/
VOWEL	UCI	http://archive.ics.uci.edu/ml/
COIL20	COIL	http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
SEMEION	UCI	http://archive.ics.uci.edu/ml/
FAULTS	UCI	http://archive.ics.uci.edu/ml/
SEGMENT	UCI	http://archive.ics.uci.edu/ml/
CORA	LINQS	http://www.cs.umd.edu/projects/linqs/projects/lbc/index.html
CITESEER	LINQS	http://www.cs.umd.edu/projects/linqs/projects/lbc/index.html
7SECTORS	CMUTE	http://www.cs.cmu.edu/~TextLearning/datasets.html
OPTDIGITS	UCI	http://archive.ics.uci.edu/ml/
SVMGUIDE1	LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
ZIP	LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
USPS	UCI	http://archive.ics.uci.edu/ml/
PENDIGITS	UCI	http://archive.ics.uci.edu/ml/
PROTEIN	LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
20NEWS	CMUTE	http://www.cs.cmu.edu/~TextLearning/datasets.html
LET-REC	UCI	http://archive.ics.uci.edu/ml/
MNIST	MNIST	http://yann.lecun.com/exdb/mnist/

indices 0-9. The problem is to train the network as well as possible using only on data from "speakers" 0-47, and then to test the network on speakers 48-89, reporting the number of correct classifications in the test set. In summary, the dataset contains 990 samples grouped in 11 classes, with each sample having a dimensionality of 10.

- COIL20: the *COIL-20* dataset is a toy image database from Columbia University Image Library. It contains 1,440 grayscale images of 20 objects (72 images per object) with a vide variety of complex geometric and reflectance characteristics. Each image has a size of 128×128. The dataset has been widely used in image classification and retrieval tasks. In summary, the dataset contains 1,440 samples grouped in 20 classes, with each sample having a dimensionality of 16,384.

- SEMEION: the UCI *Semeion Handwritten Digit* dataset. Totally 1,593 handwritten digits from around 80 persons were scanned, stretched in a rectangular box 16×16 in a gray scale of 256 values.Then each pixel of each image was scaled into a bolean (1/0) value using a fixed threshold. In summary, the dataset contains 1,593 samples grouped in 10 classes, with each sample having a dimensionality of 256.

- FAULTS: the UCI *Steel Plates Faults* dataset consists of information of steel plate faults, classified into 7 different types, with 27 independent attributes. The goal was to train machine learning algorithms for automatic pattern recognition. In summary, the dataset contains 1,941 samples grouped in 7 classes, with each sample having a dimensionality of 27.

- SEGMENT: the UCI *Image Segmentation* dataset. The instances, represented by 19 high-level features, were drawn randomly from a database of 7 outdoor images, and the images were handsegmented to create a classification for every pixel. In summary, the dataset contains 2,310 samples grouped in 7 classes, with each sample having a dimensionality of 19.

- CORA: the LINQS *Cora* dataset. It consists of 2,708 scientific publications classified into one of 7 classes. The citation network consists of 5,429 links. Each publication in the dataset is described by a 0/1-valued word vector indicating the absence/presence of the corresponding word from the dictionary. The dictionary consists of 1,433 unique words. In summary, the dataset contains 2,708 samples grouped in 7 classes, with each sample having a dimensionality of 1,433.

- CITESEER: the LINQS *CiteSeer* dataset. It consists of 3,312 scientific publications classified into one of 6 classes. The citation network consists of 4,732 links. Each publication in the dataset is described by a 0/1-valued word vector indicating the absence/presence of the corresponding word from the dictionary. The dictionary consists of 3,703 unique words. In summary, the dataset contains 3,312 samples grouped in 6 classes, with each sample having a dimensionality of 3,703.

- 7SECTORS: the *Industry Sector* dataset from CMU Text Learning group. The dataset is a collection of web pages belonging to companies from 7 economic sectors. In summary, the dataset contains 4,556 samples grouped in 7 classes, with each sample having a dimensionality of 10,000.

- OPTDIGITS: the UCI *optical recognition of handwritten digits* dataset was created by extracting normalized bitmaps of handwritten digits of 43 people from a preprinted form, which generates an input matrix of 8×8 with each element being an integer in the range $[0, 16]$. In summary, the dataset contains 5,620 samples grouped 10 classes, with each sample having a dimensionality of 64.

- SVMGUIDE1: the LIBSVM *svmguide1* dataset is obtained from an astroparticle application from Jan Conrad of Uppsala University, Sweden. There are 3,089 instances for training and 4,000 for testing, with each instance represented by 4 numerical features. In the paper, we used all the samples, i.e., 7,089 samples grouped in 2 classes, with each sample having a dimensionality of 4.

- ZIP: the LIBSVM *ZIP* handwritten digits dataset contains 9,298 samples grouped 10 classes, with each sample having a dimensionality of 256.

- USPS: the UCI *optical recognition of handwritten digits* dataset is used for optical character recognition, similar to OPTDIGITS. It contains 9,298 samples grouped 10 classes, with each sample having a dimensionality of 256.

- PENDIGITS: the UCI *pen-based recognition of handwritten digits* dataset was created as a digit database that contains 250 samples from 44 writers. It contains 10,992 samples grouped in 10 classes, with each sample having a dimensionality of 16.

- PROTEIN: the LIBSVM *protein* dataset from bioinformatics. The original dataset has 17,766 instances for training and 6,621 for testing, with each sample represented by 357 features. In the paper, we utilized the training subset, i.e., 17,766 samples grouped in 3 classes, with each sample having a dimensionality of 357.

- 20NEWS: text documents from *20 newsgroups*. This data set is a collection of 20,000 messages, collected from 20 different netnews newsgroups. One thousand messages from each of the twenty newsgroups were chosen at random and partitioned by newsgroup name. 10,000 words with maximum information gain are preserved. The dataset we used in the paper contains 19,938 samples grouped in 20 classes, with each sample having a dimensionality of 10,000.

- LET-REC: the UCI *letter recognition* dataset. The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15. In summary, the dataset contains 20,000 samples grouped 26 classes, with each sample having a dimensionality of 16.

- MNIST: the handwritten digit images database. The MNIST database has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a 28×28 image. In the paper, we used all the samples, i.e., 70,000 grouped in 10 classes, with each sample having a dimensionality of 784.

Note that, as a pre-processing step, the scattering features [1] have been extracted for each sample in the image datasets, and Tf-Idf features have been extracted for the text document datasets.

2. Optimization specifications

In this section we give the optimization specifications of the compared methods. Note that the specification of DCD has been presented in the paper already. Table 2 gives the frequently used notations.

Table 2: List of frequently used notations.

$G = (V, E)$	undirected graph G with vertices in V and edges in E
m, n, r	data dimensionality, sample size, reduced rank of matrix
$\mathbb{R}_+^{m\times n}$	space of nonnegative $m \times n$ matrices
X	data matrix of size $m \times n$, whose columns are n-dimensional vectors
A	similarity matrix of size $n \times n$
W	factorizing matrix of size $n \times r$, also called cluster indicator matrix
S	factorizing matrix of size $r \times r$
$\Pi = \{\pi_1, \ldots, \pi_M\}$	a cluster ensemble with M base clusterings (partitions)

- KM: the classical K-means algorithm [2]. Let $X = \{x_i\}, i = 1, \ldots, n$ be the set of m-dimensional points to be clustered into a set of K clusters, $C = \{c_k, k = 1, \ldots, K\}$. K-means algorithm finds a partition that minimizes the squared error between the empirical mean of a cluster and the points belonging to the cluster. The aim of K-means is to minimize the sum of the squared error over all clusters, given by the objective function:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - mu_k\|^2,$$

  where mu_k is the empirical mean of the cluster C_k. We directly utilized the Matlab function *kmeans* for implementation.

- NCUT: Normalized Cut [3]. NCUT partitions the graph G into two disjoint sets A and B by minimizing the cost as a fraction of the total edge connections to all the nodes in the graph, given by the objective function:

$$NCUT(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)},$$

  where $cut(A, B) = \sum_{u\in A, v\in B} w(u, v)$ denotes the degree of dissimilarity between the subgraphs A and B computed as total weight of the edges removed for separating the two subgraphs, and $assoc(A, V) = \sum_{u\in A, t\in V} w(u, t)$ is the total connection from nodes in A to all nodes in the graph and $assoc(B, V)$ is similarly defined. The objective given above can be minimized by solving a generalized eigenvalue problem (see details in [3]). For implementation, we utilized the

NCUT Matlab package[1] downloaded from the author's website.

- 1-SPEC: 1-Spectral Ratio Cheeger Cut [4]. The ratio Cheeger cut (RCC) of a partition $(C, \overline{C})$, where $C \subset V$ and $\overline{C} = V \setminus C$, is to minimize the objective function:

$$RCC(C, \overline{C}) = \frac{cut(C, \overline{C})}{min\{|C|, |\overline{C}|\}}.$$

  We adopted the 1-SPEC software[2] by Hein and Bühler with its default setting in our implementation.

- PNMF: Projective NMF [5]. Given the nonnegative input data matrix $X \in \mathbb{R}_+^{n \times m}$, PNMF [6, 7] based on Frobenius norm aims to find a factorizing matrix $W \in \mathbb{R}_+^{n \times r}$ for the optimization problem: $\underset{W \geq 0}{\text{minimize}} \|X - WW^T X\|_F^2$. In our paper, we utilized the kernelized version of PNMF, i.e., to replace XX^T by the similarity matrix A between samples.

- NSC: Nonnegative Spectral Clustering [8]. NSC solves the normalized cut by using the multiplicative update algorithm of NMF, i.e., to solve the following optimization problem:

$$\underset{H^T DH = I,\ H \geq 0}{\text{minimize}} -trace(H^T AH),$$

  where $D = diag(d_1, \ldots, d_n)$ with $d_i = \sum_{j=1}^n A_{i,j}$.

- ONMF: Symmetric Tri-Factor Orthogonal NMF [9]. ONMF is the special case of orthogonal tri-factor NMF, when the given input is a matrix of pairwise similarities. ONMF thus solves the following optimization problem:

$$\underset{W \geq 0, S \geq 0}{\text{minimize}} \|A - WSW^T\|_F^2, \text{ subject to } W^T W = I,$$

  where $S \in \mathbb{R}_+^{r \times r}$.

- LSD: Left Stochastic Matrix Decomposition [10]. LSD is a probabilistic clustering method. It estimates a scaling factor c^* and a cluster probability matrix W^*

---

[1] http://www.cis.upenn.edu/~jshi/software/

[2] http://www.ml.uni-saarland.de/code/oneSpectralClustering/oneSpectralClustering.html

that solves the following optimization problem:

$$\underset{c\in\mathbb{R}_+}{\text{minimize}}\left\{\underset{W\geq 0}{\text{minimize}}\,\|cA - WW^T\|_F^2,\ \text{subject to}\ \sum_{k=1}^{r} W_{ik} = 1\right\}.$$

Note that minimizing the scaling factor c^* is given in a closed form and does not depend on a particular solution W^*, which means that only W needs to be updated.

- PLSI: Probabilistic Latent Semantic Indexing [11]. PLSI assumes that the data is generated from a multinomial distribution, and maximizing the PLSI likelihood function is equivalent to minimizing the Kullback-Leibler divergence. We utilized the symmetric version of PLSI in our context, i.e., to solve the following optimization problem:

$$\underset{W\geq 0, S\geq 0}{\text{minimize}}\left\{A_{ij}\log\frac{A_{ij}}{(WSW^T)_{ij}}\right\},\ \text{subject to}\ \sum_{i=1}^{n} W_{ik} = 1 \text{ and } \sum_{k=1}^{r} S_{kk} = 1,$$

  where S is a diagonal matrix.

- BEST: BESTCLUSTERING ensemble algorithm [12]. Given a set of n samples $X = \{x_i\}$ and a cluster ensemble of M base clusterings $\Pi = \{\pi_1, \ldots, \pi_M\}$, the following simple 0/1 distance function checks if two clusterings π_1 and π_2 place x_i and x_j in the same clusters:

$$d_{(x_i,x_j)}(\pi_1,\pi_2) = \begin{cases} 1 & \text{if } \pi_1(x_i) = \pi_2(x_j) \text{ and } \pi_1(x_i) \neq \pi_2(x_j), \\ & \text{or } \pi_1(x_i) \neq \pi_2(x_j) \text{ and } \pi_1(x_i) = \pi_2(x_j), \\ 0 & \text{otherwise.} \end{cases}$$

  The distance between two clusterings π_1 and π_2 is defined as the number of pairs of objects on which the two clusterings disagree, that is,

$$d_X(\pi_1,\pi_2) = \sum_{(x_i,x_j)\in X\times X} d_{(x_i,x_j)}(\pi_1,\pi_2)$$

  Therefore, BEST algorithm finds the clustering π^* from Π that minimizes the total number of disagreements with all the given clusterings, i.e., to solve the following optimization problem:

$$\underset{\pi^*}{\text{minimize}}\left\{\sum_{i=1}^{M} d_X(\pi_i, \pi*), \pi^* \neq \pi_i\right\}.$$

- CO: evidence accumulation method based on CO-association matrix [13]. The assumption is that patterns belonging to a “natural” cluster are very likely to be co-located in the same cluster in different data partitions π_i. Taking the co-occurrences of pairs of patterns in the same cluster as votes for their association, the M data partitions of n patterns are mapped into a $n \times n$ co-association matrix:

$$CO(i, j) = \frac{n_{ij}}{M},$$

where where n_{ij} is the number of times the pattern pair (x_i, x_j) is assigned to the same cluster among the M partitions. The evidence accumulation mechanism thus maps the partitions in the clustering ensemble into a new similarity measure (the co-association matrix CO) between patterns. Therefore, any similarity-based clustering algorithms can be used to produce the final clustering result. In this paper, we obtained the final partition by using the complete-linkage hierarchical clustering algorithm (implemented by the authors in [14] as a Matlab package[3]).

- CTS: link-based ensemble clustering with Connected-Triple based Similarity matrix [15]. CTS method can better handle the unknown relations between data samples than the CO method does. Briefly, the weight assigned to edge connecting clusters i and j is estimated in accordance with the proportion of their overlapping members $w_{ij} = \frac{|X_i \cap X_j|}{|X_i \cup X_j|}$, where X_C denotes set of data points belonging to cluster C. The count of all triples $1, \ldots, q$ between cluster i and cluster j can be calculated as $C_{ij} = \sum_{k=1}^{q} \{min(w_{ik}, w_{jk})\}$. The similarity between clusters i and j can be estimated as follows:

$$S_{WT}(i, j) = \frac{C_{ij}}{C_{max}},$$

where C_{max} is the maximum C_{ij} value of any two clusters i and j. As for the CTS method, for the m-th ensemble member, the similarity of data sample x_i and x_j is estimated as:

$$S_m(x_i, x_j) = \begin{cases} 1 & \text{if } C(x_i) = C(x_j), \\ S_{WT}(C(x_i), C(x_j)) \times DC & \text{otherwise.} \end{cases}$$

where DC is a constant decay factor ranging in $[0, 1]$ (i.e. the confidence level of accepting two non-identical objects as being similar). Finally, each entry in the CTS similarity matrix can be computed as:

$$CTS(x_i, x_j) = \frac{1}{M} \sum_{m=1}^{M} S_m(x_i, x_j).$$

Note that we directly ran the CTS algorithm from the Matlab package mentioned

[3] `http://www.jstatsoft.org/v36/i09`

in [14], and produced the final partition using the complete-linkage algorithm as well.

3. Experiments

Table 3 gives the complete clustering results of the first group of experiments (see Section 5.3 of the original paper). Table 4 shows the clustering results of the second group of experiments (see Section 5.3 of the original paper), where the ensemble bases are created by the classical *k*-means algorithm.

References

[1] S. Mallat, Group invariant scattering, Communications on Pure and Applied Mathematics 65 (10) (2012) 1331–1398.

[2] S. Lloyd, Last square quantization in pcm, IEEE Transactions on Information Theory, special issue on quantization 28 (1982) 129–137.

[3] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 888–905.

[4] M. Hein, T. Bühler, An inverse power method for nonlinear eigenproblems with applications in 1-Spectral clustering and sparse PCA, in: Advances in Neural Information Processing Systems (NIPS), 2010, pp. 847–855.

[5] Z. Yang, E. Oja, Linear and nonlinear projective nonnegative matrix factorization, IEEE Transactions on Neural Networks 21 (5) (2010) 734–749.

[6] Z. Yuan, E. Oja, Projective nonnegative matrix factorization for image compression and feature extraction, in: Proceedings of 14th Scandinavian Conference on Image Analysis (SCIA), Joensuu, Finland, 2005, pp. 333–342.

[7] Z. Yang, Z. Yuan, J. Laaksonen, Projective non-negative matrix factorization with applications to facial image processing, International Journal on Pattern Recognition and Artificial Intelligence 21 (8) (2007) 1353–1362.

[8] C. Ding, T. Li, M. Jordan, Nonnegative matrix factorization for combinatorial optimization: Spectral clustering, graph matching, and clique finding, in: IEEE International Conference on Data Mining (ICDM), IEEE, 2008, pp. 183–192.

[9] C. Ding, T. Li, W. Peng, H. Park, Orthogonal nonnegative matrix t-factorizations for clustering, in: International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2006, pp. 126–135.

[10] R. Arora, M. Gupta, A. Kapila, M. Fazel, Clustering by left-stochastic matrix factorization, in: International Conference on Machine Learning (ICML), 2011, pp. 761–768.

[11] T. Hofmann, Probabilistic latent semantic indexing, in: International Conference on Research and Development in Information Retrieval (SIGIR), 1999, pp. 50–57.

[12] A. Gionis, H. Mannila, P. Tsaparas, Clustering aggregation, in: International Conference on Data Engineering (ICDE), IEEE, 2005, pp. 341–352.

[13] A. Fred, A. Jain, Combining multiple clusterings using evidence accumulation, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (6) (2005) 835–850.

[14] N. Iam-On, S. Garrett, Linkclue: A matlab package for link-based cluster ensembles, Journal of Statistical Software 36 (9) (2010) 1–36.

[15] N. Iam-On, T. Boongoen, S. Garrett, Refining pairwise similarity matrix for cluster ensemble problem with cluster relations, in: International Conference on Discovery Science (DS), Springer, 2008, pp. 222–233.

DATASET	KM	NCUT	1-SPEC	PNMF	NSC	ONMF	LSD	PLSI	DCD1	DCD1.2	DCD2	DCD5
ORL	0.70	0.81	0.81	0.81 0.81 0.81 0.81	0.81 0.81 0.81 0.81	0.53 0.78 0.80 0.80	0.82 0.82 0.82 0.82	0.65 0.81 0.83 0.83	0.67 0.81 0.83 0.83	0.74 0.81 0.82 0.82	0.13 0.23 0.21 0.22	0.03 0.03 0.03 0.03
MED	0.59	0.57	0.53	0.57 0.57 0.56 0.56	0.57 0.59 0.57 0.57	0.51 0.57 0.56 0.56	0.57 0.57 0.57 0.57	0.57 0.57 0.57 0.58	0.57 0.57 0.57 0.58	0.58 0.57 0.57 0.58	0.53 0.57 0.60 0.60	0.36 0.36 0.36 0.36
VOWEL	0.40	0.35	0.34	0.38 0.35 0.37 0.38	0.35 0.35 0.35 0.35	0.36 0.35 0.37 0.38	0.34 0.37 0.38 0.40	0.34 0.36 0.38 0.40	0.28 0.36 0.36 0.40	0.31 0.35 0.37 0.40	0.34 0.35 0.35 0.40	0.16 0.24 0.29 0.32
COIL20	0.63	0.71	0.67	0.67 0.71 0.71 0.71	0.73 0.71 0.72 0.72	0.63 0.71 0.72 0.72	0.71 0.68 0.68 0.68	0.58 0.75 0.69 0.70	0.62 0.75 0.69 0.70	0.65 0.75 0.69 0.70	0.64 0.75 0.69 0.70	0.23 0.73 0.70 0.73
SEMEION	0.68	0.64	0.66	0.60 0.66 0.60 0.60	0.66 0.66 0.66 0.66	0.62 0.60 0.60 0.60	0.76 0.72 0.74 0.75	0.67 0.65 0.74 0.77	0.68 0.65 0.75 0.77	0.77 0.65 0.75 0.77	0.66 0.64 0.72 0.77	0.23 0.68 0.71 0.73
FAULTS	0.42	0.40	0.40	0.42 0.45 0.45 0.45	0.44 0.39 0.38 0.38	0.42 0.45 0.42 0.42	0.39 0.44 0.43 0.43	0.35 0.41 0.44 0.44	0.35 0.40 0.43 0.44	0.35 0.39 0.43 0.44	0.35 0.39 0.43 0.44	0.46 0.41 0.44 0.44
SEGMENT	0.59	0.61	0.55	0.49 0.54 0.49 0.53	0.39 0.61 0.69 0.71	0.49 0.51 0.53 0.53	0.30 0.64 0.61 0.65	0.26 0.62 0.64 0.65	0.26 0.61 0.61 0.65	0.26 0.63 0.61 0.65	0.30 0.63 0.61 0.65	0.17 0.16 0.16 0.16
CORA	0.53	0.39	0.36	0.41 0.36 0.41 0.41	0.36 0.36 0.36 0.36	0.34 0.36 0.43 0.43	0.48 0.51 0.51 0.54	0.41 0.45 0.52 0.55	0.41 0.45 0.52 0.55	0.49 0.44 0.50 0.55	0.47 0.53 0.51 0.55	0.51 0.49 0.51 0.52
CITESEER	0.61	0.30	0.31	0.28 0.29 0.29 0.28	0.26 0.28 0.25 0.25	0.31 0.32 0.28 0.28	0.38 0.43 0.45 0.47	0.35 0.44 0.44 0.48	0.37 0.44 0.44 0.48	0.39 0.43 0.45 0.49	0.40 0.43 0.45 0.48	0.49 0.39 0.46 0.46
7SECTORS	0.39	0.25	0.25	0.29 0.29 0.29 0.29	0.26 0.25 0.25 0.25	0.24 0.29 0.29 0.29	0.27 0.37 0.40 0.35	0.27 0.37 0.40 0.35	0.30 0.37 0.40 0.35	0.28 0.36 0.40 0.35	0.32 0.37 0.40 0.36	0.34 0.39 0.35 0.35
OPTDIGITS	0.72	0.74	0.76	0.70 0.68 0.68 0.68	0.66 0.77 0.77 0.77	0.68 0.68 0.68 0.68	0.71 0.76 0.82 0.87	0.51 0.72 0.76 0.85	0.57 0.76 0.71 0.85	0.74 0.76 0.82 0.85	0.71 0.77 0.83 0.85	0.20 0.77 0.80 0.87
SVMGUIDE1	0.71	0.75	0.93	0.68 0.68 0.68 0.68	0.82 0.77 0.79 0.81	0.68 0.68 0.68 0.68	0.70 0.78 0.91 0.91	0.59 0.77 0.90 0.91	0.58 0.78 0.90 0.91	0.57 0.82 0.90 0.91	0.64 0.87 0.90 0.91	0.56 0.90 0.90 0.91
ZIP	0.49	0.74	0.74	0.54 0.70 0.72 0.68	0.65 0.74 0.74 0.74	0.55 0.70 0.67 0.68	0.72 0.84 0.83 0.85	0.33 0.74 0.84 0.84	0.36 0.76 0.84 0.84	0.56 0.74 0.84 0.84	0.63 0.74 0.84 0.84	0.17 0.17 0.17 0.17
USPS	0.74	0.74	0.74	0.67 0.80 0.75 0.68	0.72 0.74 0.74 0.74	0.62 0.80 0.75 0.68	0.80 0.79 0.84 0.85	0.48 0.73 0.80 0.85	0.51 0.75 0.80 0.85	0.80 0.75 0.81 0.85	0.84 0.80 0.85 0.85	0.28 0.78 0.80 0.85
PENDIGITS	0.72	0.80	0.73	0.79 0.79 0.79 0.79	0.49 0.79 0.73 0.73	0.63 0.79 0.79 0.79	0.66 0.88 0.89 0.89	0.24 0.82 0.88 0.89	0.25 0.84 0.88 0.89	0.36 0.85 0.88 0.89	0.40 0.85 0.88 0.89	0.11 0.11 0.11 0.11
PROTEIN	0.46	0.46	0.46	0.46 0.46 0.46 0.46	0.46 0.46 0.46 0.46	0.46 0.46 0.46 0.46	0.46 0.46 0.48 0.50	0.46 0.51 0.51 0.50	0.47 0.46 0.51 0.50	0.50 0.46 0.50 0.50	0.49 0.46 0.50 0.49	0.49 0.51 0.50 0.48
20NEWS	0.07	0.43	0.36	0.39 0.39 0.39 0.38	0.39 0.43 0.40 0.40	0.27 0.38 0.38 0.38	0.41 0.48 0.48 0.49	0.22 0.44 0.49 0.49	0.23 0.45 0.49 0.50	0.29 0.45 0.49 0.50	0.30 0.45 0.49 0.50	0.07 0.19 0.20 0.23
LET-REC	0.29	0.21	0.15	0.36 0.37 0.34 0.35	0.17 0.21 0.21 0.21	0.29 0.35 0.35 0.34	0.31 0.31 0.37 0.37	0.16 0.26 0.32 0.38	0.17 0.25 0.32 0.38	0.18 0.27 0.32 0.38	0.21 0.30 0.32 0.38	0.04 0.04 0.04 0.04
MNIST	0.60	0.77	0.88	0.57 0.87 0.73 0.57	0.74 0.79 0.79 0.79	0.57 0.75 0.65 0.57	0.93 0.75 0.97 0.97	0.46 0.79 0.97 0.98	0.55 0.81 0.97 0.98	0.96 0.89 0.97 0.98	0.97 0.96 0.97 0.98	0.25 0.96 0.97 0.97

(a) Purity

DATASET	KM	NCUT	1-SPEC	PNMF	NSC	ONMF	LSD	PLSI	DCD1	DCD1.2	DCD2	DCD5
ORL	0.85	0.90	0.92	0.89 0.90 0.89 0.89	0.89 0.90 0.90 0.90	0.76 0.88 0.89 0.89	0.90 0.90 0.90 0.90	0.84 0.90 0.91 0.91	0.83 0.90 0.91 0.91	0.87 0.90 0.91 0.91	0.27 0.39 0.36 0.38	0.00 0.00 0.00 0.00
MED	0.55	0.57	0.52	0.56 0.55 0.55 0.55	0.57 0.58 0.57 0.57	0.51 0.57 0.56 0.56	0.56 0.57 0.57 0.57	0.56 0.56 0.57 0.58	0.56 0.57 0.57 0.58	0.58 0.57 0.57 0.59	0.57 0.58 0.59 0.59	0.01 0.04 0.04 0.04
VOWEL	0.43	0.40	0.38	0.39 0.36 0.37 0.39	0.37 0.38 0.37 0.37	0.37 0.36 0.37 0.39	0.35 0.38 0.40 0.40	0.32 0.40 0.38 0.41	0.28 0.39 0.38 0.40	0.31 0.38 0.38 0.41	0.32 0.38 0.40 0.41	0.08 0.21 0.25 0.30
COIL20	0.77	0.79	0.77	0.75 0.79 0.79 0.79	0.81 0.79 0.80 0.80	0.74 0.79 0.79 0.79	0.79 0.78 0.77 0.77	0.71 0.80 0.80 0.80	0.74 0.80 0.80 0.80	0.75 0.80 0.80 0.80	0.75 0.81 0.80 0.80	0.31 0.80 0.80 0.81
SEMEION	0.57	0.61	0.62	0.58 0.62 0.58 0.58	0.63 0.63 0.63 0.63	0.59 0.57 0.57 0.57	0.67 0.63 0.65 0.66	0.59 0.61 0.66 0.68	0.61 0.61 0.67 0.68	0.69 0.61 0.67 0.69	0.64 0.61 0.67 0.68	0.14 0.60 0.61 0.64
FAULTS	0.10	0.08	0.09	0.09 0.11 0.11 0.11	0.10 0.07 0.07 0.07	0.10 0.11 0.09 0.09	0.06 0.11 0.11 0.11	0.03 0.08 0.12 0.11	0.02 0.08 0.11 0.11	0.03 0.08 0.11 0.11	0.03 0.08 0.10 0.11	0.14 0.09 0.12 0.11
SEGMENT	0.58	0.55	0.58	0.43 0.48 0.43 0.49	0.25 0.56 0.62 0.63	0.38 0.46 0.44 0.44	0.13 0.53 0.51 0.58	0.08 0.55 0.53 0.58	0.07 0.55 0.53 0.58	0.07 0.55 0.53 0.58	0.13 0.55 0.53 0.58	0.01 0.01 0.01 0.01
CORA	0.34	0.16	0.14	0.14 0.13 0.17 0.17	0.14 0.14 0.14 0.14	0.11 0.13 0.17 0.17	0.22 0.24 0.23 0.25	0.15 0.20 0.24 0.25	0.15 0.20 0.24 0.25	0.22 0.20 0.23 0.26	0.22 0.23 0.24 0.26	0.22 0.20 0.22 0.22
CITESEER	0.34	0.10	0.12	0.07 0.07 0.07 0.07	0.07 0.08 0.07 0.07	0.10 0.12 0.07 0.07	0.13 0.18 0.20 0.20	0.10 0.17 0.19 0.21	0.11 0.17 0.18 0.21	0.10 0.18 0.20 0.21	0.13 0.17 0.19 0.21	0.17 0.15 0.19 0.19
7SECTORS	0.17	0.04	0.05	0.05 0.04 0.04 0.04	0.05 0.04 0.05 0.05	0.01 0.04 0.04 0.04	0.04 0.10 0.14 0.11	0.04 0.07 0.13 0.11	0.04 0.08 0.13 0.11	0.04 0.08 0.13 0.11	0.05 0.09 0.13 0.11	0.10 0.12 0.10 0.10
OPTDIGITS	0.70	0.72	0.80	0.67 0.68 0.68 0.67	0.66 0.77 0.78 0.78	0.67 0.68 0.67 0.67	0.69 0.72 0.78 0.83	0.40 0.71 0.73 0.82	0.51 0.74 0.69 0.82	0.71 0.74 0.77 0.82	0.66 0.75 0.78 0.82	0.07 0.74 0.77 0.83
SVMGUIDE1	0.31	0.35	0.65	0.27 0.27 0.27 0.27	0.34 0.39 0.41 0.44	0.27 0.27 0.27 0.27	0.12 0.25 0.60 0.60	0.02 0.38 0.59 0.59	0.02 0.40 0.59 0.59	0.01 0.45 0.59 0.59	0.06 0.54 0.59 0.59	0.01 0.59 0.60 0.59
ZIP	0.40	0.78	0.79	0.54 0.67 0.65 0.64	0.61 0.78 0.78 0.78	0.56 0.66 0.62 0.64	0.66 0.78 0.80 0.81	0.18 0.77 0.79 0.81	0.21 0.78 0.79 0.81	0.46 0.77 0.79 0.81	0.56 0.75 0.79 0.81	0.01 0.01 0.01 0.01
USPS	0.62	0.77	0.80	0.66 0.75 0.71 0.66	0.71 0.78 0.78 0.78	0.62 0.75 0.71 0.66	0.75 0.77 0.81 0.82	0.40 0.75 0.77 0.81	0.46 0.76 0.77 0.81	0.74 0.76 0.78 0.81	0.77 0.77 0.82 0.81	0.13 0.77 0.76 0.82
PENDIGITS	0.68	0.81	0.78	0.78 0.78 0.78 0.78	0.51 0.79 0.79 0.78	0.63 0.78 0.77 0.77	0.61 0.83 0.86 0.86	0.10 0.81 0.83 0.86	0.10 0.81 0.83 0.86	0.21 0.81 0.83 0.86	0.27 0.81 0.83 0.86	0.01 0.01 0.01 0.01
PROTEIN	0.00	0.01	0.01	0.02 0.01 0.01 0.02	0.01 0.01 0.01 0.01	0.00 0.01 0.01 0.00	0.01 0.00 0.02 0.04	0.01 0.04 0.04 0.04	0.02 0.01 0.04 0.04	0.03 0.01 0.04 0.04	0.02 0.03 0.03 0.02	0.02 0.04 0.03 0.02
20NEWS	0.05	0.54	0.52	0.36 0.36 0.36 0.34	0.48 0.54 0.52 0.52	0.24 0.34 0.34 0.34	0.36 0.43 0.44 0.44	0.14 0.44 0.44 0.45	0.14 0.45 0.44 0.45	0.19 0.44 0.44 0.45	0.22 0.44 0.44 0.45	0.01 0.19 0.19 0.24
LET-REC	0.35	0.38	0.26	0.43 0.43 0.42 0.43	0.21 0.38 0.37 0.37	0.35 0.43 0.43 0.42	0.39 0.41 0.45 0.45	0.17 0.37 0.42 0.46	0.18 0.36 0.42 0.46	0.22 0.37 0.42 0.46	0.25 0.39 0.42 0.46	0.00 0.00 0.00 0.00
MNIST	0.51	0.81	0.89	0.59 0.82 0.72 0.59	0.73 0.84 0.84 0.84	0.58 0.75 0.64 0.59	0.87 0.76 0.93 0.93	0.34 0.81 0.92 0.94	0.48 0.80 0.92 0.93	0.91 0.87 0.93 0.93	0.93 0.92 0.93 0.94	0.16 0.92 0.93 0.93

(b) NMI

Table 2: The complete clustering results of the first group of experiments.

Table 4: Clustering performance comparison of DCD using *heterogeneous co-initialization* with three ensemble clustering methods. Rows are ordered by dataset sizes. Boldface numbers indicate the best. The ensemble bases are created by the classical *k*-means algorithm.

	Purity				NMI			
DATASET	BEST	CO	CTS	DCD	BEST	CO	CTS	DCD
ORL	0.77	0.75	0.76	**0.83**	0.89	0.88	0.88	**0.91**
MED	0.60	**0.64**	0.56	0.58	0.58	0.58	**0.59**	0.58
VOWEL	**0.40**	0.33	**0.40**	**0.40**	0.43	0.34	**0.44**	0.40
COIL20	**0.74**	0.63	0.56	0.70	**0.80**	0.77	0.74	**0.80**
SEMEION	0.69	0.58	0.62	**0.77**	0.59	0.54	0.57	**0.68**
FAULTS	0.42	0.42	0.42	**0.44**	0.10	0.10	0.10	**0.11**
SEGMENT	0.59	0.57	0.57	**0.65**	**0.58**	0.57	0.56	**0.58**
CORA	**0.60**	0.57	0.54	0.55	**0.39**	0.36	0.37	0.25
CITESEER	0.67	**0.68**	**0.68**	0.48	0.38	**0.39**	**0.39**	0.21
7SECTORS	0.28	0.30	0.28	**0.35**	0.07	0.09	0.07	**0.11**
OPTDIGITS	0.80	0.65	0.74	**0.85**	0.75	0.70	0.72	**0.82**
SVMGUIDE1	0.71	0.71	0.71	**0.91**	0.31	0.31	0.31	**0.59**
ZIP	0.52	0.41	0.51	**0.84**	0.42	0.37	0.44	**0.81**
USPS	0.73	0.73	0.69	**0.85**	0.61	0.61	0.61	**0.81**
PENDIGITS	0.77	0.57	0.68	**0.89**	0.69	0.62	0.68	**0.86**
PROTEIN	0.46	0.46	0.46	**0.50**	0.01	0.01	0.01	**0.04**
20NEWS	0.46	0.42	0.38	**0.50**	**0.58**	0.52	0.50	0.45
LET-REC	0.28	0.23	0.27	**0.38**	0.36	0.29	0.34	**0.46**
MNIST	0.58	0.48	0.52	**0.98**	0.49	0.38	0.43	**0.93**