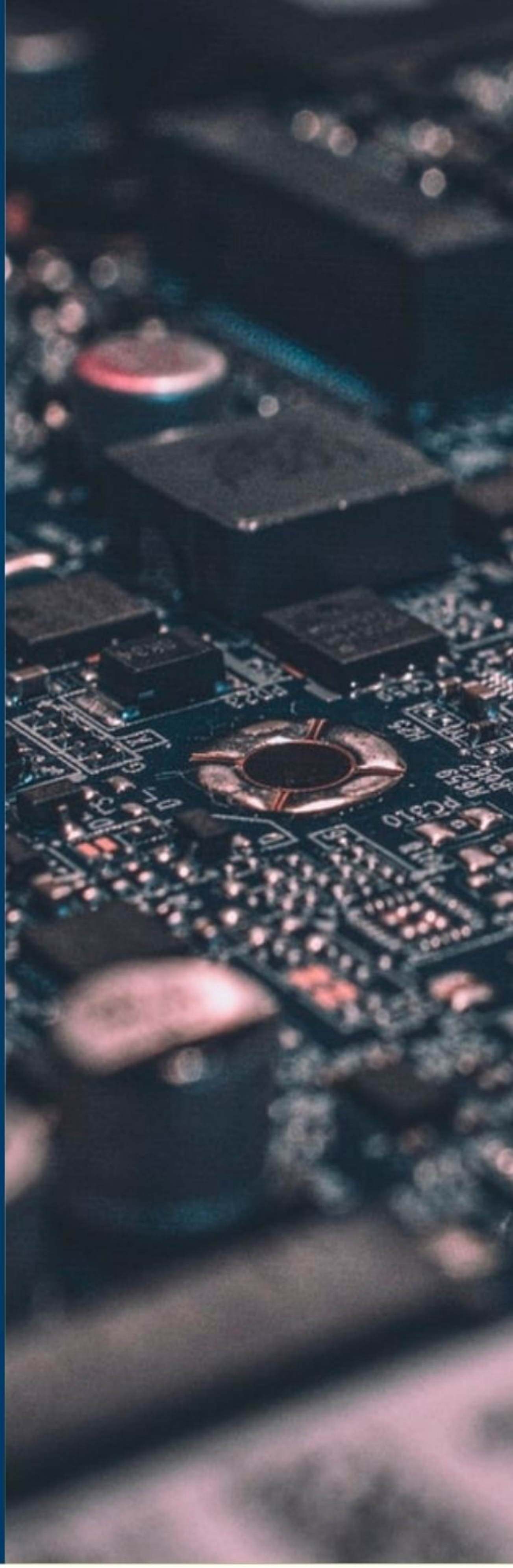

Inductive ensemble clustering using kernel support matching

S. Park, J. Hah and J. Lee✉

A novel inductive ensemble clustering method is proposed. In the proposed method, kernel support matching is applied to a co-association matrix that aggregates arbitrary basic partitions in order to detect clusters of complicated shape. It also has the advantage of naturally detecting the number of clusters and assigning clusters for out-of-sample data. In the proposed method, a new similarity is learned from various clustering results of the basic partition, and a kernel support function capable of clustering learning data and test data is constructed. Experimental results demonstrated that the proposed method is effective with respect to clustering quality and has the robustness to induce clusters of out-of-sample data.

Introduction: Dividing data into non-convex clusters of the same type is very difficult in unsupervised learning and is susceptible to noise inherent in original data. Also since the necessity to protect confidential personal information is increasing, access to raw data becomes impossible, and only basic partitioning results reporting the relation between objects can be obtained. To solve these problems, recent ensemble clustering, also called consensus clustering, attracts increasing attention as it combines basic partitions and provides robust clustering results by capturing clusters of more complex shapes [1–6]. However, most of the existing ensemble clustering methods need to pre-fix the number of anticipated clusters and cannot perform inductive reasoning on out-of-sample data.

In order to deal with these problems, we propose a new inductive ensemble clustering algorithm using basic partitions and dissimilarity between objects instead of original data. The proposed method utilises and refines a co-association matrix (rCM) that combines the several basic partitions from the k-means algorithm as in [5]. With kernel support matching, this approach approximates support for data distribution described by the rCM. The method then finds the representative points of each cluster and cluster out-of-sample data by analysing the phase characteristics of the constructed support.

Consensus ensemble: In this Letter, we aggregate the clustering results using co-association matrix (CM), whose elements represent the number of co-occurrences in basic partitions. Let $\mathcal{X} = \{o_1, o_2, \ldots, o_n\}$ be the set of n observations (or objects). Suppose that we have p clustering results from the base partitions $P_1, \ldots, P_p$ where $P_k:\mathcal{X} \mapsto \{1, 2, \ldots, b_k\}$ is the partition function. The original CM, $\mathbf{C} \in \mathbb{R}^{n\times n}$, is defined as

$$C_{i,j} = \frac{1}{p}\sum_{k=1}^{p} \delta(P_k(o_i), P_k(o_j)), \quad \delta(a, b) = \begin{cases} 1 & \text{if} \quad a = b \\ 0 & \text{if} \quad a \neq b \end{cases}.$$

This CM can be regarded as a similarity matrix or a kernel matrix, but it gives the same weight for all pairs in the same cluster. In [3], Zhong *et al.* differentiated the weights depending on the distances of the pairs. We generalise these weights so as to enable to use any dissimilarity measure instead of the distances of the original data. They specified the weight directly from a dissimilarity measure $\mathrm{DS}(o_i, o_j)$:

$$w_{i,j}(k) = \begin{cases} 1 - \dfrac{\mathrm{DS}(o_i, o_j)}{L^k_{i,j}} & \text{if} \quad P_k(o_i) = P_k(o_j) \\ 0 & \text{if} \quad P_k(o_i) \neq P_k(o_j) \end{cases}, \tag{1}$$

where $L^k_{i,j}$ is the maximum dissimilarity of two points in the same cluster with o_i and o_j for the partition k. From this weight, the rCM, $\tilde{C}_{i,j}$, is

$$\tilde{C}_{i,j} = \frac{1}{p}\sum_{k=1}^{p} w_{i,j}(k). \tag{2}$$

Proposed method: Our work is motivated from the drawbacks of most consensus clustering methods that suffer from choosing the number of clusters and assigning cluster labels for new data points. We propose a new clustering ensemble framework using support vector domain description in [7]. First, an rCM is calculated from the results of basic partitions. Then with a non-linear transformation Φ from the input space to some high-dimensional space, we find the smallest enclosing sphere of radius R with soft constraints:

$$\|\Phi(o_j) - \boldsymbol{a}\|^2 \le R^2 + \xi_j, \quad \xi_j \ge 0, \ \forall j, \tag{3}$$

where $\boldsymbol{a}$ is the centre and ξ_j are some slack variables. In order to solve (3), its dual problem is obtained as

$$\begin{aligned} &\max_{\beta_j} \quad \sum_j \beta_j - \sum_{i,j} \beta_i\beta_j\tilde{C}_{i,j} \\ &\text{s.t.} \quad 0 \le \beta_j \le C, \quad \sum_j \beta_j = 1, \ \forall j. \end{aligned} \tag{4}$$

where the inner products of $\Phi(o_i)\cdot\Phi(o_j)$ are replaced with an rCM value $\tilde{C}_{i,j}$. By solving (4), the following trained kernel support function can be can be used for estimating support of a data distribution:

$$\begin{aligned} R^2(o_*) &= \|\Phi(o_*) - \boldsymbol{a}\|^2 \\ &= 1 - 2\sum_{j\in \mathrm{SV}} \beta_j\tilde{C}_{*,j} + \sum_{i,j\in \mathrm{SV}} \beta_i\beta_j\tilde{C}_{i,j}, \end{aligned} \tag{5}$$

Here, those optimal points corresponding to the optimal objective value $\hat{R}^2$ with $0 < \beta_j < C$ are called support vectors (SVs), those points with $\beta_j = C$ are called bounded SVs (BSVs), and SV is an index set of SVs and BSVs. The level set $L_{R^2}(\hat{R}^2) = \{o_*{:}R^2(o_*) \le \hat{R}^2\}$ divides the dataset into a number of connected components from which the number of clusters is naturally determined. It also captures arbitrary clusters by providing a new similarity relationship between given data points aggregated from the primary basic partitions. To assign the cluster labels of training objects o_j, however, we need to evaluate $R^2(o_*)$ at any object o_*. Since the information of $\tilde{C}_{*,j}$ is not available, it is not possible to apply the support-based clustering methods as in [6, 8–10] directly to ensemble clustering.

To overcome this problem, we propose a kernel support matching method for inductive ensemble clustering. Our method starts from the metric representations of the data objects. Otherwise, we can apply any available non-metric multi-dimensional scaling method or its extensions to transform the data object o_j to its metric representation $x_j \in \mathbb{R}^d$. We next approximate the support function R^2 with the following kernel support function of the form:

$$\tilde{f}(x) = 1 - 2\sum_j \beta'_j k(x_j, x) + \sum_{i,j\in SV} \beta_i\beta_j\tilde{C}_{i,j}, \tag{6}$$

Here $k(\cdot\,,\,\cdot)$ is a kernel function and we have used in this Letter the Gaussian radial basis function (rbf) kernel given by $k(\mathbf{x}_i, \mathbf{x}_j) = k(\|\mathbf{x}_i - \mathbf{x}_j\|) = \mathrm{e}^{-q\|\mathbf{x}_i - \mathbf{x}_j\|^2}$. To this end, we first fit the rbf kernel q to preserve the kernel similarity with that of rCM for each pair of $(\mathbf{x}_i, \mathbf{x}_j)$. Specifically, we use the least-squares method to find q that minimises the replace the Gram matrix $\min_q \|\boldsymbol{K} - \tilde{\boldsymbol{C}}\|$ where $\tilde{\boldsymbol{C}}$ is the Gram matrix of the rCM with $\tilde{C}_{i,j}$ and $\boldsymbol{K}$ is the kernel matrix with $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. The new kernel makes it possible to get the similarity between a given data object and a new data object. With this fitted kernel, we then approximate the kernel support function of (6) to that of (5) by calibrating the coefficients of the second term in (6) to solve

$$\min_{\beta} \sum_{i\in \mathrm{SV}} \left(\sum_{j\in \mathrm{SV}} (\beta_j\tilde{C}_{i,j} - \beta'_j K_{i,j}) \right)^2, \tag{7}$$

Finally, we locate the stable equilibrium vectors (SEVs) of the dynamical system associated with this matched kernel support $\tilde{f}$ as in [9]. The SEVs in the same connected component of the level set $L_{\tilde{f}}(\hat{R}^2) = \{\mathbf{x}{:}\tilde{f}(\mathbf{x}) \le \hat{R}^2\}$ will be assigned to the same cluster label. When the system is applied to, each data object converges to one of the SEVs and the same cluster label will be assigned to it.

One of the salient features of this method is inductive clustering. Any unknown new data object belongs to one of the SEV basins and can be assigned to the same cluster label of the corresponding SEV. This labelling process can be expedited by adopting the fast phase as in [10]. With this procedure, the entire data sample space can be divided into several cluster regions, allowing for inductive clustering processing. The procedure of the proposed method (IECS) can be summarised as follows:

(i) (co-association) Perform clustering with p basic partitions, and get the rCM using (2).

(ii) (fitting the kernel parameter) Optimise (4) with the rCM, and fit a kernel parameter (q in an rbf kernel) via the least-squares method.
(iii) (kernel support matching) Obtain the kernel support (6) matching to that of (5) by solving (7).
(iv) (cluster labelling) Locate the SEVs of the dynamical system associated with (6) and assign the same labels to the SEVs belonging to the same connected component of the level set $L_{\tilde{f}}$.
(v) (inductive clustering) Assign to each data object (given training data or unknown test data) the same cluster label of its corresponding SEV.

Experiments: To evaluate the performance of the proposed method, we used a number of real-word datasets from the UCI repository [11]. The detailed descriptions of datasets are given in Table 1. K-means algorithm, one of the widely-used clustering algorithms, is used to generate the basic partitions. We set a dissimilarity measure to a distance between data. We compared the proposed method with state-of-the arts ensemble clustering methods such as cluster-based similarity partitioning algorithm (CSPA), hyper-graph partitioning algorithm (HGPA), meta-clustering algorithm (MCLA), and spectral clustering of the rCM (SPC) [1, 3, 5]. We obtained 100 basic partitions where the number of clusters ranges from the true cluster number k_{true} to $\sqrt{n}$ since the compared methods need to set the number of final clusters before the experiments whereas our method does not.

Table 1: Descriptions of the datasets

Datasets	Clusters (k_{true})	Instances (n)	Dimensions (d)
Orange	9	140	2
Two circles	2	300	2
Iris (UCI)	3	150	4
Glass (UCI)	6	214	9
Zoo (UCI)	7	101	16
WPBC (UCI)	2	198	32
Satimage (UCI)	6	6435	36

In our experiments, we implemented our algorithm in MATLAB and adopted the adjusted Rand index (ARI) as the cluster evaluation measure [11]. The higher the ARI is, the better a quality of clustering is. Table 2 reports the clustering performance results of the compared methods. For each dataset, the first row reports the ARI of the compared method when the true cluster number is known (the best result is underlined) and the second row when it is not known (the best result is bold typed). Our proposed method outperformed the other ensemble clustering algorithms in most cases, and it had comparable performances to MCLA with true cluster number is given a priori in Zoo and Smimage datasets. The results show that our method performs very well in real-world problems when the number of true clusters is unknown. Fig. 1 shows a typical example that IECS can detect non-convex shaped clusters well compared with the other ensemble algorithm even though the basic partitions can only address convex shaped data.

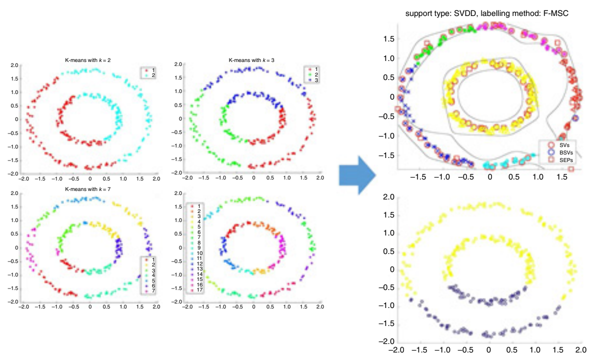

Fig. 1 *K-means and ensemble clustering results for two circles data. Left four graphs are basic partitions from k-means, top right is result of IECS, and bottom right is result of SPC*

In order to verify whether IECS can perform inductive clustering, we used the dataset consisting of five Gaussians with 1000 instances. We split the dataset into a training set and a test set and changed the ratio between them to compare the results. Table 3 shows that IECS works very well despite a small amount of training data.

Table 2: Performance comparisons in ensemble clustering's by ARI

	SPC ($\sqrt{n}$)	CSPA ($\sqrt{n}$)	HGPA ($\sqrt{n}$)	MCLA ($\sqrt{n}$)	IECS
Orange	0.5605	0.5969	0.6380	0.8549	**0.8951**
	0.8026	0.5607	0.6741	0.8318	
Two Circles	0	0	0	0.0021	**0.6179**
	0.0912	0.0151	0.0167	0.0221	
Iris	0.5923	0.6004	0.6530	0.5667	**0.6949**
	0.3699	0.2067	0.2620	0.4454	
Glass	0.1459	0.1726	0.1797	0.1663	**0.2100**
	0.1566	0.1138	0.0937	0.1287	
Zoo	0.5603	0.4430	0.4166	0.8135	**0.7781**
	0.4348	0.3489	0.3653	0.6109	
Smimage	0.5415	0.3716	0.3841	0.5774	**0.5656**
	0.4140	0.3427	0.3414	0.4595	
WPBC	0.0164	0.0136	0.0063	0.0110	**0.0524**
	0	0.0011	0.0038	0	

Table 3: Inductive performances of IECS changing the test ratio

Test ratio (%)	20	40	60	80	100
ARI	0.8925	0.9120	0.8881	0.8997	0.8533

Conclusion: In this Letter, we have presented a new clustering ensemble algorithm. The method aggregates basic clustering results using kernel support matching and automatically decides the effective number of clusters. Experimental results show that the proposed method not only effectively captures non-convex clusters but also makes inductive clustering very well for out-of-sample data.

Acknowledgment: This work was supported by the National Research Foundation of Korea (NRF) grand funded by the Korean government (MEST) (no. 2016R1A2B3014030).

Submitted: *8 June 2017* E-first: *6 November 2017*
doi: 10.1049/el.2017.2159
One or more of the Figures in this Letter are available in colour online.

S. Park, J. Hah and J. Lee (*Department of Industrial Engineering, Seoul National University, Seoul, Republic of Korea*)

✉ E-mail: jaewook@snu.ac.kr

References

1 Strehl, A., and Ghosh, J.: 'Cluster ensembles – a knowledge reuse framework for combining multiple partitions', *J. Mach. Learn. Res.*, 2002, **3**, pp. 583–617
2 Fred, A.L., and Jain, A.K.: 'Combining multiple clusterings using evidence accumulation', *Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (6), pp. 835–850
3 Zhong, C., Yue, X., Zhang, Z., *et al.*: 'A clustering ensemble: two-level-refined co-association matrix with path-based transformation', *Pattern Recognit.*, 2015, **48**, (8), pp. 2699–2709
4 Liu, H., Liu, T., Wu, J., *et al.*: 'Spectral ensemble clustering'. Proc. of the 21st ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2015, pp. 715–724
5 Liu, H., Shao, M., Li, S., *et al.*: 'Infinite ensemble for image clustering'. Proc. of ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016
6 Ben-Hur, A., Horn, D., Siegelmann, H.T., *et al.*: 'Support vector clustering', *J. Mach. Learn. Res.*, 2001, **2**, pp. 125–137
7 Tax, D.M.J., and Duin, R.P.W.: 'Support vector domain description', *Pattern Recognit. Lett.*, 1999, **43**, pp. 1191–1199
8 Jung, K.H., Lee, D., and Lee, J.: 'Fast support-based clustering method for large-scale problems', *Pattern Recognit.*, 2010, **43**, (5), pp. 1975–1983
9 Kim, K., Son, Y., and Lee, J.: 'Voronoi cell-based kernel support clustering', *Trans. Knowl. Data Eng.*, 2015, **27**, (4), pp. 1146–1156
10 Lichman, M.: 'UCI machine learning repository' (University of California, School of Information and Computer Science, California, 2013). Available at http://archive.ics.uci.edu/ml
11 Hubert, L., and Arabie, P.: 'Comparing partitions', *J. Classif.*, 1985, **2**, pp. 193–218