COCE-SMART: CONSENSUS CLUSTERING BASED ON ENHANCED SPLITTING-MERGING AWARENESS TACTICS

Rui Fa[1]*, Basel Abu-Jamous*[1]*, David J. Roberts*[2,3] *and Asoke K. Nandi*[1,4]

[1] Department of Electronic and Computer Engineering, Brunel University, Uxbridge, UB8 3PH, United Kingdom.
[2] National Health Service Blood and Transplant, Oxford, United Kingdom
[3] The University of Oxford, John Radcliffe Hospital, Oxford, United Kingdom
[4] Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland.
Email: {Rui.Fa, Basel.AbuJamous, Asoke.Nandi}@brunel.ac.uk, david.roberts@ndcls.ox.ac.uk

ABSTRACT

In this paper, we propose a new consensus clustering algorithm, which is based on an existing clustering paradigm, called enhanced splitting merging awareness tactics (E-SMART). The problem of determining the number of clusters, which affects many state-of-the-art consensus clustering algorithms, is addressed by the proposed CoCE-SMART algorithm. The idea behind CoCE-SMART is that SMART is used repeatedly to one dataset, resulting in different clustering results, which might have different numbers of clusters. These SMART clustering results can be combined by clustering the centroids of all clusters as the estimate of real number of clusters can be determined from the SMART clustering results. Three benchmark datasets are utilised to assess the proposed algorithm. The experimental results strongly indicate that the proposed CoCE-SMART algorithm outperforms other state-of-the-art consensus clustering algorithms.

Index Terms— Consensus clustering, SMART, Gene expression analysis

1. INTRODUCTION

Clustering, or unsupervised learning, is one of fundamental problems in the machine learning field. The objective of clustering is to group a set of unlabelled data objects into clusters, within which similar objects are gathered. A number of industrial applications and scientific research areas exploit clustering as an exploratory tool to discover the structure in the unknown data, for example, data mining, image segmentation, information retrieval, and gene expression analysis [1–3]. Clustering is an extremely difficult problem because of its unsupervised nature as well as the uncertainty of the data. There have been many types of algorithms in the literature, trying to solve the same problem from different perspectives, for example, partitional clustering, hierarchical clustering, fuzzy clustering, neural-network based clustering, model-based clustering, *etc* [4–7]. However, these algorithms often provide very different results and no single algorithm is able to identify the structures of all sorts of data. Additionally, for some stochastic clustering algorithms, every single run on the same dataset will produce different results.

Consensus clustering, also known as ensemble clustering or cluster aggregation, is viewed as one of solutions to the above issue [8–16]. Consensus clustering formalises the idea that combining different clustering into a single representative or consensus, would emphasise the common organisation in the different clustering results. The pioneering work of consensus clustering was done by Fred and Jain [8, 10], who introduced the concept of evidence accumulation clustering (EAC) that maps the individual data partitions in a clustering ensemble into a new similarity measure between patterns, summarizing common structure perceived from these clusterings. The EAC algorithm is based on the assumption that similar data objects are more likely to be grouped into the same cluster than dissimilar ones, and conversely, those data objects that are often clustered in the same cluster should be regarded as being very similar. The EAC algorithm is designed to form a co-association matrix, which summarises a clustering ensemble by counting the number of co-occurrence between each pair of data objects. The co-association matrix is then clustered by using agglomerative hierarchical clustering. Based on the EAC algorithm, Lourenco and colleagues further developed probabilistic consensus clustering (PCC) [12]. Strehl and Ghosh [9] proposed three different combination methods, namely hypergraph-partitioning algorithm (HGPA), cluster-based similarity partitioning algorithm (CSPA), and meta-clustering algorithm (MCLA), exploring the concept of consensus between data partitions. Basel and colleagues [13, 14] proposed a consensus clustering algorithm named binarisation of consensus partition matrix (BiCoPaM), which is based on relabelling and voting strategy. All the above consensus clustering algorithms have their merits; however, they have a common weakness, which is that they all require the number of clusters as a parameter during clustering.

A common practice to tackle this problem is to estimate the number of clusters before the clustering procedure. It can be done by employing clustering validation criteria [17, 18]. However, most clustering validation criteria have to search exhaustively a range of numbers of clusters, with the hope that the real number of clusters is within the range. Moreover, their ability of correctly determining the number of clusters is limited in the noisy dataset [18]. Recently, A new clustering paradigm, named splitting merging awareness tactics (SMART), was proposed to cluster the dataset without specifying the number of clusters *a priori* [19–21]. The principle of the SMART algorithm relies on a splitting-while-merging (SWM) framework, where clusters split and merge in the same time constantly until a stopping criterion is satisfied. Later, the SMART algorithm was fur-

This article summarises independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Grant Reference Number RP-PG-0310-1004). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. A. K. Nandi would like to thank TEKES for their award of the Finland Distinguished Professorship. Professor Nandi is a Distinguished Visiting Professor of Tongji University, Shanghai, China.

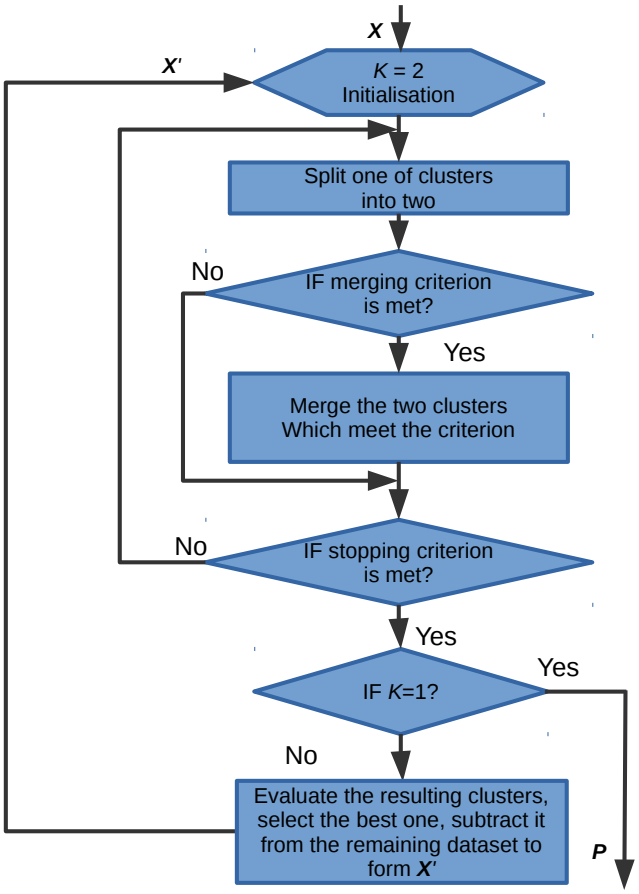

Fig. 1: The structure of E-SMART

ther enhanced by employing successive processing, called enhanced SMART (E-SMART) [22]. However, the issue in SMART and E-SMART is that they are stochastic algorithms, which start randomly and produce different results, sometimes with different numbers of clusters, depending on the noise level in the data.

In this paper, we propose a new consensus clustering algorithm based on the SMART paradigm, which is named CoCE-SMART, to produce consistent and accurate clustering results. By exploiting the greatest advantage of the SMART clustering, i.e. estimating the number of clusters fairly accurately while clustering, the problem of determining the number of clusters, which affects many state-of-the-art consensus clustering algorithms, is addressed by the proposed CoCE-SMART algorithm. The SMART algorithm, which starts randomly from one or two clusters and ends up with a clustering with a fairly accurate estimate of number of clusters, is used repeatedly to cluster the given dataset. The SMART clustering results, therefore, can be combined by grouping the centroids of all clusters as the estimate of real number of clusters can be obtained from the SMART clustering results. Three benchmark datasets are utilised to assess the proposed algorithm. The experimental results strongly indicate that the proposed CoCE-SMART algorithm outperforms other state-of-the-art consensus clustering algorithms on both accuracy and consistency.

2. SMART AND ENHANCED SMART

We are able to review only briefly the principle of the SMART and E-SMART algorithms here because of the limited space. The interested readers are referred to references [19–22].

The SMART paradigm starts the clustering procedure with one or two randomly generated clusters. Clusters that contain many distinct patterns split continuously and automatically; in the meanwhile, clusters that meet the merging criterion are joined together. Such a process enables SMART self-awareness to split and merge clusters automatically in iterations. To do so, many clustering tasks have to be performed. In the splitting task of each iteration (Task 2 in [19–21]), SMART splits one cluster into two. Then, the new clusters are tested by a merging criterion which is called merging task (Task 3 in [19–21]). If any pair of clusters meet the merging criterion, we merge the two clusters, otherwise skip the merging step. Then SMART goes through a termination-check, where a stopping criterion is applied. If the condition for termination is not satisfied, SMART goes to the next iteration and continues to split, otherwise, SMART finishes SWM process. The last step is clustering selection task (Task 4 in [19–21]). Minimum message length (MML) [23] was employed in clustering selection task. Note that these tasks in the SWM strategy can be done using many different clustering techniques, e.g., competitive learning paradigm was used for the splitting task in [19] while component-wise expectation maximisation (CWEM) for finite mixture models (FMM) was used for both splitting and merging tasks in [21]; in both original SMART and SMART-FMM, MML was employed in clustering selection task [19–21].

The SMART-FMM algorithm was enhanced by employing successive processing strategy [22]. This was the first attempt in the literature to employ successive processing in a clustering algorithm. Instead of selecting the best clustering from the results by using clustering selection criterion in original SMART framework, the successive processing strategy subtracted clusters one by one in iterations. In doing so, the silhouette index was employed to evaluate the intermediate clusters and order them according to their index values from high to low. Then the best cluster was subtracted from the original dataset and the remaining dataset was fed back to the SWM process to start a new iteration. The clustering and subtracting were repeated successively and terminated automatically, once no more splitting happened in the SWM process. The structure of E-SMART is shown in Fig. 1. In this paper, we will employ E-SMART as the main clustering algorithm.

3. COCE-SMART

Let the dataset that we are going to partition be $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_i | 1 \leq i \leq N\}$, where $\boldsymbol{x}_i \in \mathbb{R}^{M\times 1}$ denotes the i-th object, M is the dimension, and N is the number of objects. Suppose that R partitions generated by the E-SMART algorithm are denoted as $\boldsymbol{P} = \{\boldsymbol{P}_1, \boldsymbol{P}_2, ..., \boldsymbol{P}_R\}$. Since the E-SMART algorithm is a stochastic algorithm, the number of clusters in each partition may vary because of the noise, i.e., the numbers of clusters for R partitions can be denoted as $\boldsymbol{K} = \{K_1, K_2, ..., K_R\}$. We can obtain K_r centroids as representatives of K_r clusters from the r-th clustering results, and therefore, we can obtain a new dataset containing all centroids, called centroids set, with $N_c = \sum_{r=1}^{R} K_r$ centroids. We can also simply obtain the best number of clusters K_{best} as the mode of $\boldsymbol{K}$. Note that since the number of clusters has been given, any high-quality clustering algorithm can be used to cluster the new centroids set further at this stage. Here, we employ a model-based clustering algorithm [4]. Actually,

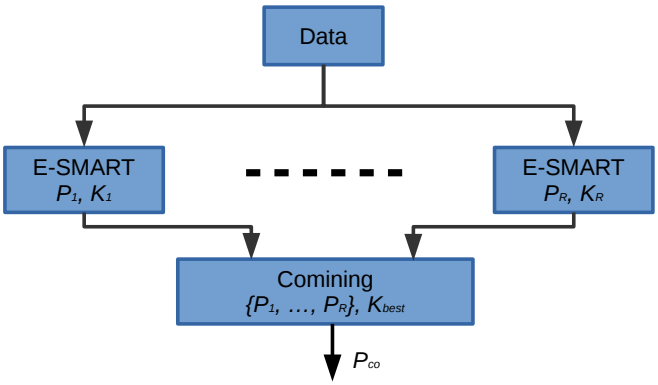

Fig. 2: The structure of CoCE-SMART.

such a re-clustering procedure is indeed a relabelling process. All clusters from different clustering partitions are relabelled into new partition labels in terms of the partition of the centroids set. Then we can generate a consensus partition matrix $\boldsymbol{C} \in \mathbb{Z}^{+}{}_{N\times K_{best}}$, whose element c_{nk} denotes the number of times that the n-th data object is assigned the k-th cluster. Once the consensus partition matrix is produced, we can either obtain the final partition by assigning the objects to the clusters that they are mostly clustered into, or send the consensus partition matrix into a BiCoPaM framework [13] to obtain different levels of tightness of clusters. The proposed CoCE-SMART algorithm, whose structure is shown in Fig. 2, can be summarised as follows:

1. To generate R clustering results by using the E-SMART algorithm;
2. To form a new set (centroids set) containing all centroids of clusters in all generated clustering results;
3. To estimate the number of clusters based on the generated clustering results;
4. To cluster the centroids set using the model-based clustering algorithm with the estimated number of clusters; and
5. To generate a consensus partition matrix based on the clustering of the centroids set.

4. EXPERIMENTAL RESULTS

4.1. Datasets and Experiments Set-up

In this paper, we compare the proposed CoCE-SMART algorithm with several state-of-the-art consensus clustering algorithms, namely HGPA, CSPA, MCLA [9], PCC [12], BiCoPaM-kmeans (BK), and BiCoPaM-hybrid (BH) [13]. We investigate performance of accuracy and reducibility of each algorithm. The metrics measuring accuracy are adjusted Rand index (ARI) [24] when the ground truth is available. Silhoutte index (SI) [25] and Calinski-Harabasz (CH) index [26] will be used to evaluate the quality of clusterings when the ground truth is not available. To measure the reproducibility of the consensus clustering algorithms, we repeat the experiment many times, i.e. N_e. We define an average value of pair-wise ARI between any two experiment results as a metric to measure the reproducibility (REP), which can be mathematically expressed as

$$\text{REP} = \frac{\sum_{i=1}^{N_e}\sum_{j=1,j\neq i}^{N_e}\text{ARI}(\boldsymbol{P}_i,\boldsymbol{P}_j)}{N_e(N_e-1)}, \quad \text{REP}\in(-1,1]. \quad (1)$$

Table 1: Performance comparison of CoCE-SMART with other state-of-the-art algorithms, namely BK, BH [13], PCC [12], CSPA, HGPA, MCLA [9] in Iris flower dataset. Both mean and standard deviation of performance metrics are investigated.

	Known		Unknown	
	ARI	REP	ARI	REP
BK	0.73±1.1E-15	1.0±0.0	0.54±6.7E-16	1.0±0.0
BH	0.74±7.0E-3	0.99±1.0E-2	0.54±1.2E-15	1.0±0.0
PCC	0.55±7.1E-2	0.82±1.4E-1	0.57±7.6E-2	0.76±1.5E-1
CSPA	0.91±8.7E-2	0.95±1.3E-1	0.28±2.4E-2	0.54±7.4E-2
HGPA	0.92±1.7E-2	0.98±2.0E-2	0.32±2.1E-2	0.67±8.7E-2
MCLA	0.91±9.5E-2	0.94±1.4E-1	0.50±7.3E-2	0.76±0.1
CoCE-SMART	-	-	**0.92±1.0E-15**	**1.0±0.0**

We use three benchmark experiments to evaluate the performance of the CoCE-SMART algorithm and those compared algorithms. The first experiment use the Iris flower dataset, which is a collected multivariate dataset of three types ($k = 3$) of Iris flowers for a total of $N = 150$ data objects with $M = 4$ measured features (attributes) each [27]. The second experiment employs a mathematical model to simulate the periodic behaviour of yeast cell cycle gene expression [28]. The model of cyclic gene expression is given by

$$x_{ij}(p,q) = \gamma + (l+p\gamma)\cdot[\gamma + (l+p\gamma)]\sin(2\pi j/8 - \omega_i + q\gamma)] \quad (2)$$

where x_{ij} is the expression value of the i-th gene at the j-th time point, each instant of γ is an independent random number from the standard normal distribution $\mathcal{N}(0,1)$, the parameter l controls the magnitude of the sinusoid and it is fixed to three here. The parameter p controls the random component added to the magnitude and the parameter q controls the random component added to the phase. The parameter ω_i is the phase shift of the i-th gene and will determine which cluster the gene will be in. Since the noise in this model is not additive, we have to couple p and q to be a pair, and raise both their values to change the noise power. We simulate three scenarios with three levels of noise, where (p,q) are $(0.7, 0.07)$ (low), $(1.3, 0.13)$ (medium), and $(1.9, 0.19)$ (high) respectively. The third experiment employs real microarray yeast cell cycle α-38 dataset provided by Pramila and colleagues [29]. It consists of 500 genes with highest periodicity scores and each gene has 25 time samples. Additionally, their peaking times as percentages of the cell cycle have also been provided by Pramila and colleagues [29]. It is widely accepted that there are four phases in the cell cycle, namely, G1, S, G2 and M phases [29].

For HGPA, CSPA, MCLA, PCC, and BK algorithms, we employ k-means randomly initialised as the clustering algorithm. Each of these algorithms combine $R = 100$ partitions. BH employs k-means with Kauffman approach (KA) initialisation, self-organising map (SOM), hierarchical clustering average linkage, complete linkage and Ward linkage [6], $R = 5$ clustering algorithms. We also notice that all compared algorithms require the number of clusters as an input parameter, however, such information is not available *a priori*. For a fair comparison, for each compared algorithm, we investigate performance for both perfectly known number of clusters and estimated number of clusters using CH. All above algorithms in all datasets with known and unknown number of clusters are repeated $N_e = 100$ times.

4.2. Results

In the first place, we investigate the Iris flower dataset. Both accuracy and reproducibility performance comparison of CoCE-SMART with other state-of-the-art algorithms, namely BK, BH, PCC, CSPA,

Table 2: Performance comparison of CoCE-SMART with other state-of-the-art algorithms in synthetic cell cycle gene expression dataset with low noise.

	Known		Unknown	
	ARI	REP	ARI	REP
BK	1.0±5.0E-4	1.0±7.4E-4	0.83±1.9E-1	0.81±1.9E-1
BH	1.0±1.2E-3	1.0±1.7E-3	1.0±8.9E-3	1.0±3.5E-3
PCC	1.0±7.1E-4	1.0±9.9E-4	0.87±8.4E-2	0.78±9.7E-2
CSPA	0.99±5.0E-2	0.98±6.8E-2	0.33±7.8E-3	0.18±8.6E-3
HGPA	1.0±5.0E-4	1.0±7.0E-4	0.38±1.3E-2	0.23±1.0E-2
MCLA	1.0±5.0E-4	1.0±7.0E-4	0.63±1.0E-1	0.47±1.1E-1
CoCE-SMART	-	-	**1.0±1.2E-3**	**1.0±1.7E-3**

Table 3: Performance comparison of CoCE-SMART with other state-of-the-art algorithms in synthetic cell cycle gene expression dataset with medium noise.

	Known		Unknown	
	ARI	REP	ARI	REP
BK	1.0±3.3E-3	1.0±4.6-3	0.54±9.8E-2	0.69±1.8E-1
BH	1.0±5.8E-3	0.99±8.1E-3	0.45±1.1E-1	0.83±1.8E-1
PCC	1.0±3.4E-3	0.99±4.8E-3	0.83±1.1E-1	0.7±1.2E-1
CSPA	0.99±2.4E-2	0.99±3.4E-2	0.33±1.1E-2	0.19±1.0E-2
HGPA	1.0±3.6E-3	0.99±5.0E-3	0.37±1.7E-2	0.22±1.2E-2
MCLA	1.0±3.3E-3	0.99±4.6E-3	0.55±1.1E-1	0.38±9.7E-2
CoCE-SMART	-	-	**1.0±7.5E-3**	**0.99±1.1E-2**

HGPA, MCLA, is shown in Table 1. The results reveal that all state-of-the-art algorithms work well with perfectly known number of clusters, while work poorly without predetermined number of clusters, except PCC, which works poorly in both scenarios. Among them, CSPA, HGPA, and MCLA provide best performance, with acceptably high accuracy, only when number of clusters is known. We may note that their standard deviation of the REP performance is relatively high, which means that they are less stable than BK. Compared with these state-of-the-art algorithms, the proposed CoCE-SMART algorithm provide performance with high mean values and low standard deviation values in both accuracy and reproducibility measures without the need of specifying the number clusters.

In the second experiments, we test the proposed CoCE-SMART algorithm in synthetic cell cycle gene expression datasets with different noise levels. The performance of investigated algorithms in these three scenarios, i.e. low, medium, and high noise, is shown in Table 2, Table 3, and Table 4 respectively. Interestingly, all state-of-the-art algorithms perform very well in all three noise levels when the number of clusters is perfectly known. However, when the number of clusters is unknown, only BH produces good performance in the low noise scenario, and degrades rapidly when the noise goes high. Other state-of-the-art algorithms cannot provide acceptable performance in all three scenarios when the number of clusters is unknown, which emphasizes that the number of clusters is critical to the clustering algorithms. The results indicate that the proposed CoCE-SMART algorithm is able to produce high quality clustering results consistently, even in the environments with high noise, without the number of clusters as a parameter.

In the last experiment, we evaluate all algorithms in the real gene expression dataset, i.e. Yeast cell cycle α-38 dataset. The performance comparison is shown in Table 5. Since the ground truth is not available in this case, we employ SI and CH to evaluate the accuracy of algorithms. Consistent to the previous two experiments, all state-of-the-art algorithms work well in terms of their SI and CH performance when the number of clusters is set, where BK and BH provide the clustering results with the highest SI and CH performance, while they are more consistent than other algorithms. However, when the number of clusters is unknown, their performance drops dramatically. The proposed CoCE-SMART algorithm is supported by the results that it provides higher quality and more reliable clustering results more consistently than other state-of-the-art algorithms, especially when the number of clusters is not available.

Table 4: Performance comparison of CoCE-SMART with other state-of-the-art algorithms in synthetic cell cycle gene expression dataset with high noise.

	Known		Unknown	
	ARI	REP	ARI	REP
BK	0.97±1.2E-2	0.95±1.6E-2	0.36±1.0E-2	0.76±0.2
BH	0.93±2.4E-2	0.86±3.0E-2	0.37±3.3E-2	0.78±0.2
PCC	0.97±1.3E-2	0.94±1.7E-2	0.79±1.3E-1	0.64±1.4E-1
CSPA	0.97±1.3E-2	0.95±1.7E-2	0.32±1.3E-2	0.18±1.0E-2
HGPA	0.97±1.7E-2	0.93±2.3E-2	0.35±1.8E-2	0.20±1.2E-2
MCLA	0.97±1.5E-2	0.94±1.9E-2	0.46±8.0E-2	0.30±5.8E-1
CoCE-SMART	-	-	**0.97±4.4E-2**	**0.91±5.8E-2**

Table 5: Performance comparison of CoCE-SMART with other state-of-the-art algorithms in Yeast cell cycle α-38 dataset.

		Known	Unknown
BK	SI	0.41±7.3E-16	0.29±1.8E-3
	CH	7.0±3.9E-15	5.7±7.0E-3
	REP	1.0±0.0	0.98±4.1E-2
BH	SI	0.42±8.2E-3	0.26±3.3E-2
	CH	7.0±2.2E-1	5.2±3.3E-1
	REP	0.88±8.9E-2	0.78±1.3E-1
PCC	SI	0.29±3.3E-2	0.26±2.6E-2
	CH	5.6±4.8E-1	4.8±4.1E-1
	REP	0.77±1.3E-1	0.78±9.9E-2
CSPA	SI	0.27±8.9E-3	0.12±1.3E-2
	CH	5.5±2.4E-1	2.4±1.2E-1
	REP	0.74±2.1E-1	0.60±6.9E-2
HGPA	SI	0.28±3.5E-2	0.05±2.7E-2
	CH	5.6±5.6E-1	2.0±1.3E-1
	REP	0.59±1.5E-1	0.36±4.1E-2
MCLA	SI	0.31±1.2E-2	0.18±2.7E-2
	CH	5.8±2.9E-1	2.6±1.3E-1
	REP	0.65±1.6E-1	0.61±5.2E-2
CoCE-SMART	SI	-	**0.40±1.6E-3**
	CH	-	**6.9±2.2E-2**
	REP	-	**0.96±3.2**

5. DISCUSSIONS AND CONCLUSIONS

In this paper, we proposed a consensus clustering algorithm based on E-SMART (CoCE-SMART). SMART is an existing clustering paradigm, which is able to split and merge clusters automatically to reach a high quality clustering results without specifying the number of clusters *a priori*. E-SMART is an enhanced version employing successive processing. The problem of determining the number of clusters, which affects many state-of-the-art consensus clustering algorithms, was addressed by the CoCE-SMART algorithm. The idea behind CoCE-SMART is that SMART is used repeatedly to one dataset, resulting different clustering results, which might have different numbers of clusters. These SMART clustering results, therefore, can be combined simply by grouping the centroids of all clusters as the estimate of real number of clusters can be determined easily.

Three benchmark datasets, including Iris flower dataset, synthetic cell cycle gene expression datasets, and real microarray Yeast cell cycle dataset, were investigated to assess the proposed algorithm. We measured both accuracy and reproducibility performance of CoCE-SMART and many other state-of-the-art algorithms, namely BK, BH [13], PCC [12], CSPA, HGPA, MCLA [9]. The experimental results strongly indicate that the proposed CoCE-SMART algorithm outperforms other state-of-the-art consensus clustering algorithms, especially when the number of clusters is not available.

6. REFERENCES

[1] Christopher M. Bishop, *Pattern recognition and machine learning*, vol. 1, springer New York, 2006.

[2] Rui Xu and Donald C. Wunsch, *Clustering*, vol. 10, John Wiley & Sons, 2008.

[3] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14863–14868, 1998.

[4] Ka Yee Yeung, Chris Fraley, Alejandro Murua, Adrian E. Raftery, and Walter L. Ruzzo, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, pp. 977–987, 2001.

[5] Daxin Jiang, Chun Tang, and Aidong Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, pp. 1370–1386, 2004.

[6] Rui Xu, Donald Wunsch, et al., "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[7] Rui Xu and Donald C. Wunsch, "Clustering algorithms in biomedical research: a review," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 120–154, 2010.

[8] Ana LN Fred and Anil K Jain, "Data clustering using evidence accumulation," in *2002. Proceedings. 16th International Conference on Pattern Recognition,* IEEE, 2002, vol. 4, pp. 276–280.

[9] Alexander Strehl and Joydeep Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *The Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.

[10] Ana LN Fred and Anil K Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 835–850, 2005.

[11] Sandro Vega-Pons and Jos Ruiz-Shulcloper, "A survey of clustering ensemble algorithms," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, pp. 337–372, 2011.

[12] André Lourenço, Samuel Rota Bulò, Nicola Rebagliati, Ana LN Fred, Mário AT Figueiredo, and Marcello Pelillo, "Probabilistic consensus clustering using evidence accumulation," *Machine Learning*, pp. 1–27, 2013.

[13] Basel Abu-Jamous, Rui Fa, David J Roberts, and Asoke K Nandi, "Paradigm of tunable clustering using binarization of consensus partition matrices (bi-copam) for gene discovery," *PLOS ONE*, vol. 8, no. 2, pp. e56432, 2013.

[14] Basel Abu-Jamous, Rui Fa, David J Roberts, and Asoke K Nandi, "Yeast gene cmr1/ydl156w is consistently co-expressed with genes participating in dna-metabolic processes in a variety of stringent clustering experiments," *Journal of The Royal Society Interface*, vol. 10, no. 81, pp. 20120990, 2013.

[15] Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi, "Application of the bi-copam method to five escherichia coli datasets generated under various biological conditions," *Journal of Signal Processing Systems*, pp. 1–8, 2014.

[16] Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi, "Comprehensive analysis of forty yeast microarray datasets reveals a novel subset of genes (APha-RiB) consistently negatively associated with ribosome biogenesis," *BMC Bioinformatics*, vol. 15, no. 1, pp. 322, 2014.

[17] Peter J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[18] Rui Fa and Asoke K. Nandi, "Noise resistant generalized parametric validity index of clustering for gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 4, pp. 741–752, July 2014.

[19] Rui Fa and Asoke K. Nandi, "SMART: Novel self splitting-merging clustering algorithm," *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 2198–2202.

[20] Rui Fa, David J. Roberts, and Asoke K. Nandi, "SMART: Unique splitting-while-merging framework for gene clustering," *PLOS ONE*, vol. 9, pp. e94141, 2014.

[21] Rui Fa, and Asoke K. Nandi, "An enhanced splitting-while-merging algorithm with finite mixture models," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3332–3336.

[22] Rui Fa, Basel Abu-Jamous, David J. Roberts, and Asoke K. Nandi, "Enhanced smart framework for gene clustering using successive processing," *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–6.

[23] Chris S. Wallace and David L. Dowe, "Minimum message length and kolmogorov complexity," *The Computer Journal*, vol. 42, no. 4, pp. 270–283, 1999.

[24] William M Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, pp. 846–850, 1971.

[25] Lawrence Hubert and Phipps Arabie, "Comparing partitions," *Journal of classification*, vol. 2, pp. 193–218, 1985.

[26] Tadeusz Caliński and Jerzy Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[27] Ronald A Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[28] Lue Ping Zhao, Ross Prentice, and Linda Breeden, "Statistical modeling of large microarray data sets to identify stimulus-response profiles," *Proceedings of the National Academy of Sciences*, vol. 98, no. 10, pp. 5631–5636, 2001.

[29] Tata Pramila, Wei Wu, Shawna Miles, William Stafford Noble, and Linda L Breeden, "The forkhead transcription factor hcm1 regulates chromosome segregation genes and fills the s-phase gap in the transcriptional circuitry of the cell cycle," *Genes & development*, vol. 20, pp. 2266–2278, 2006.