# SMART: NOVEL SELF SPLITTING-MERGING CLUSTERING ALGORITHM

*Rui Fa*[1] *and Asoke K Nandi*[1,2]

[1] Signal Processing and Communications Research Group, Department of Electrical Engineering and Electronics
The University of Liverpool, L69 3GJ, UK. {r.fa, a.nandi}@liverpool.ac.uk
[2] Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland.

## ABSTRACT

In this paper, we propose a new self splitting-merging clustering algorithm, named splitting-merging awareness tactics (SMART). The novel framework, which integrates many techniques, starts with one cluster and employs a splitting-while-merging process. The SMART has self-awareness to split and merge the clusters automatically in iterations. Both the framework and the techniques are detailed and illustrated by a good benchmark test. Furthermore, three microarray gene expression datasets are studied using our approach. The numerical results show that our proposal is automotive and effective.

***Index Terms***— self splitting-merging clustering algorithm, microarray

## 1. INTRODUCTION

Clustering, also known as unsupervised learning, has been used for decades in many fields, such as image processing, data mining and artificial intelligence [1, 2], and in recent years, has benefited microarray gene expression data analysis in genomic research [3,4]. The goal of the clustering analysis is to group individual genes or samples in a population within which the objects are more similar to each other than those in other clusters.

There are a lot of clustering algorithms in the literature, which generally are categorized into many different families, such as partitional clustering, hierarchical clustering, model-based clustering, density-based clustering, fuzzy clustering, neural networks based clustering, and so on [1–3]. For each different application, there are a number of algorithms providing relatively good clustering results. However, most of these successful clustering algorithms highly depend on the parameter setting and initialization, for example, the number of clusters and the initialization of the centroids (partitional clustering) or the weights (neural network based clustering). If the number of clusters is not set to the number of natural clusters or the centroids/weights are initialized randomly, the clustering results would be unreliable and inconsistent.

Recently, a variety of self splitting-merging clustering algorithms have been developed for both general purpose clustering [5, 6] and specific use, like gene expression data analysis [7]. A competitive learning paradigm, called one-prototype-take-one-cluster (OPTOC) [5], was proposed in the self-splitting clustering algorithm. There are two advantages of the OPTOC that, firstly, it is not sensitive to initialization, and secondly, in many cases, it is able to find natural clusters. However, its ability to find the natural clusters depends on the determination of suitable threshold, which is difficult [7]. Being aware of the shortcoming of the OPTOC, a self-splitting-merging competitive learning (SSMCL) algorithm [7] based on the OPTOC paradigm was developed for gene expression analysis. The SSMCL initially over-clusters the whole dataset using the OPTOC principle and then merge the groups based on the second order statistical characteristics. However, although the number of clusters can be initially set to any value larger than the number of natural clusters, the SSMCL still needs to properly set it as close to the number of natural clusters as possible, otherwise, too much computing power will be wasted due to the unnecessary over-clustering and merging. With the similar principle as the SSMCL, over-clustering and merging, a cohesion-based self-merging (CSM) algorithm, which was reported in [6] to combine the $k$-means and hierarchical clustering, also faces the same problem of setting the initial number of clusters.

In this paper, we propose a new self splitting-merging clustering algorithm, named splitting-merging awareness tactics (SMART). The novel approach, which integrates many techniques, starts with one cluster and employs a splitting-while-merging process. In such process, the SMART has self-awareness to split and merge the clusters automatically in iterations. The method employs the modified OPTOC technique to split the clusters. While splitting, a merging criterion is applied to indicate if a merging should take place. Finally, the process stops when a stopping criterion is satisfied, otherwise the splitting continues. A good benchmark test using quadrature phase shift keying (QPSK) data will demonstrate each step in the SMART flow. Furthermore, three microarray gene expression datasets are studied using our approach. The numerical results show that our proposed method is automotive and effective..

The project (Ref. NIHR-RP-PG-0310-1004-AN) is supported by National Institute for Health Research (NIHR), UK.

The rest of the paper is organized as follows: Sec. 2 describes the details of the SMART and demonstrates all the steps in the flow. Sec. 3 briefly introduces the datasets explored in the paper and the numerical results are presented. Finally, conclusions are drawn in Sec. 4.

## 2. SMART ALGORITHM

Suppose that we are going to partition the dataset $\mathbf{X} = \{\boldsymbol{x}_i | 1 \leq i \leq N\}$, 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. In this section, we describe the details of the splitting-merging awareness tactic (SMART). Since the SMART is an integration of many techniques, for the sake of convenience, we will firstly overview the SMART and present the flow in Sec. 2.1, then detail each technique in Sec. 2.2, at last, demonstrate each step in the flow using an example in Sec. 2.3.

### 2.1. Overview

The SMART is an integration of four techniques, which are going to be presented in the next subsection. The flowchart of the algorithm is illustrated in Fig. 1. The SMART starts with one cluster ($K = 1$, where $K$ is the number of clusters) and finds the highest local density center by using Technique 1. Subsequently, the data goes through a splitting-while-merging process, where splitting and merging are automatically conducted in iterations. In each iteration, the first step is that the SMART splits one of clusters into two using Technique 2. After splitting, in the second step, the new clustering is censored by a merging criterion, which is associated with the Technique 3. If the condition for merging is satisfied, then merge the two clusters which meet the criterion, otherwise skip the merging step. The last step of the SMART is termination-check, where a stopping criterion, which employs Technique 4, is applied. If the condition for termination is not satisfied, the SMART goes to the next iteration and continues to split, otherwise, the SMART is terminated. Note that the output of the SMART is the estimates of the number of clusters and the centroids of all clusters, rather than the clustering results. Depending on different applications, the data can be classified by simple nearest-neighbour method or $k$-means algorithm using the estimated centroids in the initialization.

### 2.2. Technical Details

The SMART employs four techniques to obtain self-awareness of splitting and merging. The techniques are organically integrated as presented in previous subsection. Here, their details are presented as follows:

***Technique 1:*** The purpose of Technique 1 is to find the highest local density center. We introduce a fraction $\alpha$ to control the number of local neighbouring objects for a given objects. Note that, to be *local*, $\alpha$ is preferably chosen in the range of (0.05, 0.2). Thus, there are $\alpha N$ neighbouring objects taken into account for local distance calculation and we may obtain the local distance $\mathcal{D}_\alpha$ for all the objects in the dataset. $\mathcal{D}_\alpha$ is mathematically given by

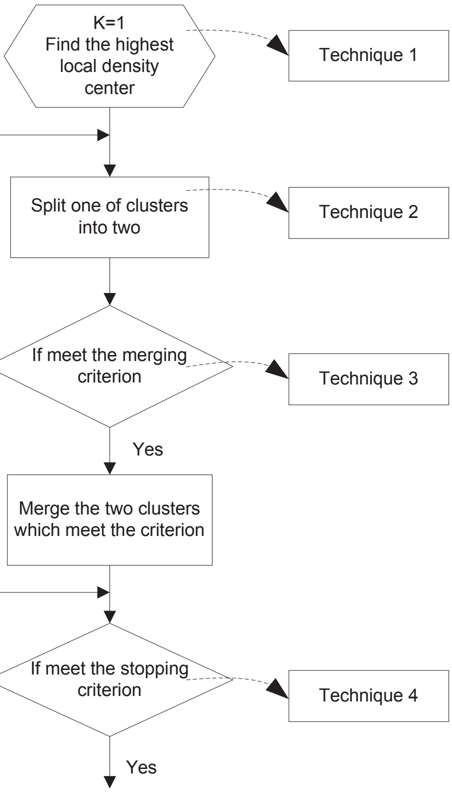
K=1
Find the highest local density center
Technique 1
Split one of clusters into two
Technique 2
If meet the merging criterion
Technique 3
No
Yes
Merge the two clusters which meet the criterion
If meet the stopping criterion
Technique 4
No
Yes

**Fig. 1**. The flow chart of the SMART.

$$\mathcal{D}_\alpha(i) = \sum_{1\leq l\leq \alpha N} \mathcal{D}(\boldsymbol{x}_i, \boldsymbol{x}_l^i) \tag{1}$$

where $\boldsymbol{x}_l^i$ denotes the $l$-th object closest to the $i$-th object. $\mathcal{D}(\cdot,\cdot)$ denotes the calculation of the distance between two objects. The first initial centroid is the object that has the minimum local distance as

$$\boldsymbol{C}_1 = \boldsymbol{x}_I, \text{ where } I = \arg\min_{1\leq i\leq N} \mathcal{D}_\alpha(i). \tag{2}$$

***Technique 2:*** We combine the OPTOC paradigm [5] with the Kaufman Approach (KA) [8] in this technique. The KA successively selects the centroids, which, however, are centres with respect to the whole dataset, rather than individual group. Hence, we employ the KA as the initial step and OPTOC to refine the centroids. The pseudo-code of Technique 2 is presented in Table 1, where $\mathcal{P}$ denotes the centroid, $\mathcal{A}$ denotes the starting point of the learning process, $\delta$ is distance indicator and $n_A$ is the winning counter.

Note that the new centroid is put into a centroid pool whose members are always used in Technique 2 but not nec-

**Table 1**. The pseudo-code for Technique 2.

```
STEP 1:
for every non-centroid object $\boldsymbol{x}_i$ do
    for every non-centroid object $\boldsymbol{x}_j$ do
        Calculate $Q_{ji} = \max(D_j - d_{ji}, 0)$ where $d_{ji} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ and $D_j = \min_c d_{cj}$ being $c$ one of centroids;
    end for
    Calculate the gain of $\boldsymbol{x}_i$ by $\sum_j Q_{ji}$;
end for
Select the non-centroid object $\boldsymbol{x}_i$ which maximizes $\sum_j Q_{ji}$ as the new centroid $C_{new}$;
STEP 2:
Calculate the minimum distance between $C_{new}$ and other centroids $D$. Initialize $\mathcal{P}$ as $C_{new}$ and $\mathcal{A}$ as the mean of $C_{new}$ and the centroid closest to $C_{new}$. $i$ denotes the iteration index.
STEP 3:
Start competitive learning process as follows:
    $\mathcal{A}(i+1) = \mathcal{A}(i) + \frac{1}{n_A(i)} \cdot \delta \cdot (\boldsymbol{x} - \mathcal{A}(i)) \Theta(\mathcal{P}(i), \mathcal{A}(i), \boldsymbol{x})$
        $\Theta(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \begin{cases} 1 & \text{if } \mathcal{D}(\boldsymbol{a}, \boldsymbol{b}) \leq \mathcal{D}(\boldsymbol{a}, \boldsymbol{c}) \\ 0 & \text{otherwise} \end{cases}$
    $\delta(i) = \left( \frac{\mathcal{D}(\mathcal{P}(i), \mathcal{A}(i))}{\mathcal{D}(\mathcal{P}(i), \boldsymbol{x}) + \mathcal{D}(\mathcal{P}(i), \mathcal{A}(i))} \right)$
    $n_A(i+1) = n_A(i) + \delta(i) \cdot \Theta(\mathcal{P}(i), \mathcal{A}(i), \boldsymbol{x})$
    $\mathcal{P}(i+1) = \mathcal{P}(i) + \alpha(i) \cdot (\boldsymbol{x} - \mathcal{P}(i))$
    $\alpha(i) = (1 + \mathcal{D}(\mathcal{P}(i), \boldsymbol{x}) / \mathcal{D}(\mathcal{P}(i), \mathcal{A}(i)))^{-2}$
STEP 4:
Set $C_{new}$ to be resulting $\mathcal{P}$.
```

essarily to be the final output because some of centroids will be merged in the merging step. The reason why we still keep those merged centroids in the centroid pool is to avoid the case that they are repeatedly selected.

***Technique 3:*** In [6], a similarity measure, namely cohesion, was proposed. The cohesion was defined as follows:

$$chs(\mathcal{C}_i, \mathcal{C}_j) = \frac{\sum_{\boldsymbol{x} \in \mathcal{C}_i, \mathcal{C}_j} join(\boldsymbol{x}, \mathcal{C}_i, \mathcal{C}_j)}{|\mathcal{C}_i| + |\mathcal{C}_j|}, \tag{3}$$

where $\mathcal{C}_i$ is the cluster with the centroid $\boldsymbol{C}_i$, $|\mathcal{C}_i|$ is the size of the cluster of $\boldsymbol{C}_i$. $join(\boldsymbol{x}, \mathcal{C}_i, \mathcal{C}_j)$ defines the similarity of the two clusters referring to the existence of an object $\boldsymbol{x}$, which is defined as

$$join(\boldsymbol{x}, \mathcal{C}_i, \mathcal{C}_j) = \min\left(f_i(\boldsymbol{x}), f_j(\boldsymbol{x})\right), \tag{4}$$

where $f_i(\boldsymbol{x})$ and $f_j(\boldsymbol{x})$ are the probability density function (pdf) of the distributions in clusters $\mathcal{C}_i$ and $\mathcal{C}_j$. Assume that an object in each cluster follows a multivariate normal distribution, whose pdf is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{M/2} (\det \boldsymbol{\Psi})^{1/2}} \exp\left[ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Psi}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right], \tag{5}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$ are mean and covariance matrix, respectively.

Given any cluster $\mathcal{C}$ we may obtain the values of ($\boldsymbol{\mu}$, $\boldsymbol{\Psi}$) of the cluster using the following formulae:

$$\hat{\boldsymbol{\mu}} = \frac{1}{|\mathcal{C}|} \sum_{\boldsymbol{x} \in \mathcal{C}} \boldsymbol{x}, \quad \boldsymbol{\Psi} = \frac{1}{|\mathcal{C}|} \sum_{\boldsymbol{x} \in \mathcal{C}} (\boldsymbol{x} - \hat{\boldsymbol{\mu}})(\boldsymbol{x} - \hat{\boldsymbol{\mu}})^T. \tag{6}$$

Once a new centroid is found, we can obtain the cohesions of the new cluster to the other clusters. If the maximum of the new cohesions is $T_{chs}$ times larger than that of the old ones, the two clusters with maximal cohesion should be merged. We also have a counter counting the number of merging. If the number of merging is greater than a threshold $T_m$, we will terminate the SMART regarding the case as no significant cluster existing.

***Technique 4:*** This technique is used, co-acting with Technique 3, to define a stopping criterion to terminate the SMART. At the end of previous stage, we obtain a centroid pool which has $K'$ members but only $K$ members are final results. We then obtain the minimum spinning tree (MST) for these $K$ centroids and get an array $\boldsymbol{D}_{mst}$ containing $(K-1)$ branches of the MST. Thus, an average MST distance can be obtained by

$$d_{mst}^{ave}(K) = \frac{1}{K} \sum_{d_i \in \boldsymbol{D}_{mst}} d_i. \tag{7}$$

We find that while splitting, $d_{mst}^{ave}(K)$ will increase with the increase of $K$ until $K$ meets the number of natural clusters. To this end, we define the stopping criterion as

```
if $d_{mst}^{ave}(K-1) - d_{mst}^{ave}(K) > \epsilon$ then
    SMART stops,
else
    SMART continues splitting,
end if
```

where $\epsilon$ is a small positive value.

## 2.3. Example

Here, we employ quadrature phase shift keying (QPSK) data at the signal-to-noise ratio (SNR) level of 15 dB as a benchmark test to demonstrate each step in the flow of the SMART. Fig. 2 illustrate the whole process of the clustering. As shown in Fig. 2 (a), the first centroid, which is marked by red $\times$, is found using Technique 1. Fig. 2 (b) shows the initial $\mathcal{P}$ and $\mathcal{A}$ labelled by $\star$ and $\circ$, respectively, following the STEP 1 of Technique 2. Fig. 2 (c) shows that the centroid is pulled to the center by the OPTOC competitive learning. Fig. 2 (d) shows the clustering results with two clusters, which does not satisfy the merging and stopping criteria, thus, the splitting continues. Fig. 2 (e) and (f) show the clustering for three and four clusters, respectively. Since the splitting continues, as sown in Fig. 2 (g), the top-right cluster is splitted into two, which triggers the merging process. The merging step moves the old centroid of the cluster to a new location which is labelled by $*$

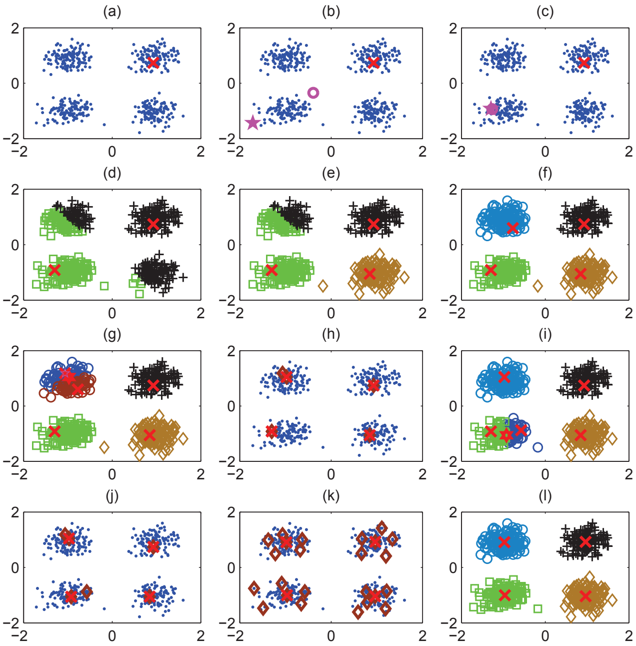

**Fig. 2**. The demonstration of the SMART using a benchmark test.

in Fig. 2 (g). Meanwhile, the new centroid is put into the centroid pool for the use of splitting but ignored in the clustering, as shown in Fig. 2 (h), where the members of the centroid pool are labelled by $\diamondsuit$ and the members for final clustering are labelled by $\times$. Fig. 2 (i) and Fig. 2 (j) show one more iteration with merging process and the resulting centroids, respectively. In this case, the SMART continues splitting and merging until the number of the times of merging exceeds the threshold $T_m$, without using Technique 4. It means that Technique 4 is not necessarily triggered every time. Fig. 2 (k) shows the final centroids and Fig. 2 (l) shows the final clustering results. This example indicates that the SMART is more intelligent than other self spitting and merging algorithms.

## 3. DATASETS AND NUMERICAL RESULTS

Here, three microarray gene expression datasets are studied by using the SMART. There are two parameters to be set: one is $T_{chs}$, which indicates if a merging should take place, and the other is the maximum number of merging $T_m$. Note that since the SMART is not sensitive to the parameters, $T_{chs}$ can be in $[5\ 10]$ and $T_m$ can be in $[10\ 20]$. For our experiments, including the QPSK data, we employ $T_{chs} = 5$ and $T_m = 20$.

### 3.1. Datasets

**Table 2**. Description of the datasets employed in the experiments.

| Dataset | Clusters | Objects (Genes) | Samples |
|---|---|---|---|
| Synthetic gene data [9] | 5 | 450 | 24 |
| Stanford Yeast data [10, 11] | 5/6 | 384 | 17 |
| $\alpha$-38 Yeast data [12] | 4 | 500 | 25 |

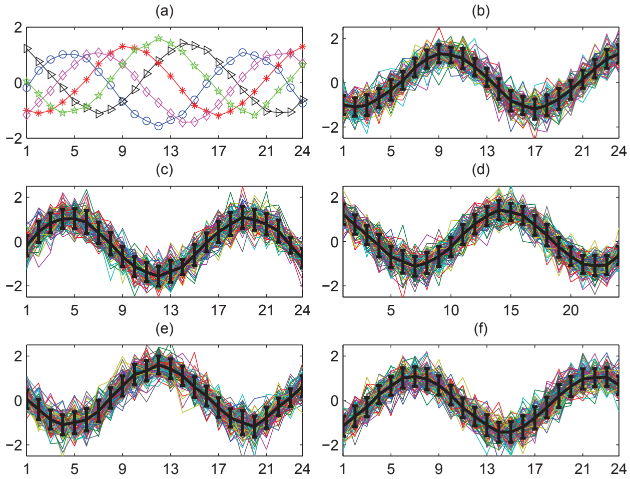

**Fig. 3**. The clustering results for the synthetic gene expression dataset.

The three gene expression datasets are explored in this paper, including a synthetic gene dataset and two Yeast cell cycle datasets, are listed in Table 2. The synthetic dataset, which we investigate here, models gene expression data with cyclic behaviour. Classes are modelled as genes that have peak times over the times course as presented in [9]. In this work, we generate a dataset with 450 genes and 24 samples, which has five clusters. The yeast cell cycle data, which is available at http://faculty.washing ton.edu/kayee/model/ [10], is also investigated. It consisted of 384 genes over 17 time points taken at 10 minutes intervals. It was commonly believed that the time course was divided into five phases including early G1, late G1, S, G2, and M phases biologically. But recently, our study reveals a numerical sixth phase rather than a biological one, which is called Q phase and should be the boundary between previous M phase and the present G1 phase [11]. Another yeast cell cycle dataset is $\alpha$-38 dataset presented in [12]. It consists of 500 genes with highest periodicity scores and each gene has 25 time samples. In this dataset, four phases, namely, G1, S, G2 and M phases, are accepted.

### 3.2. Numerical Results

In Fig. 3, the final clustering results for the synthetic gene expression dataset are presented. The SMART automatically stops at $K = 5$, which is the exact number of natural clusters. The means of members in five different clusters are shown in Fig. 3 (a), where it is easy to discern the five clusters. All members in each cluster and the errorbar plot of the cluster are

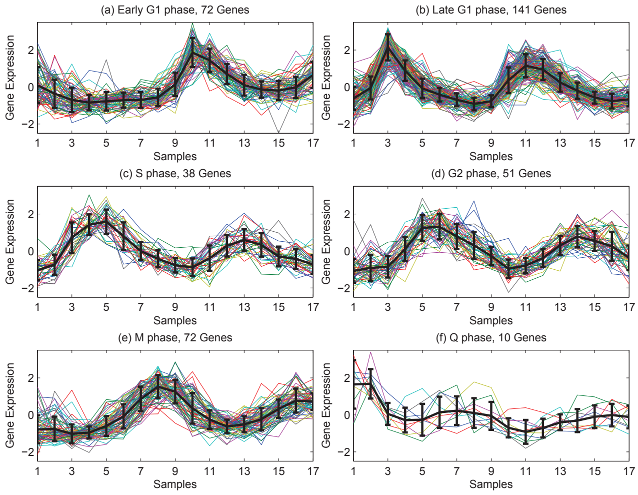

**Fig. 4**. The clustering results for the Stanford Yeast cell cycle dataset.

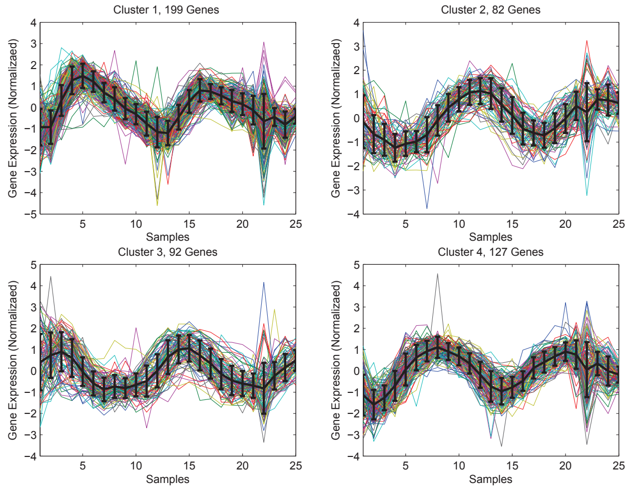

**Fig. 5**. The clustering results for the $\alpha$-38 Yeast cell cycle dataset.

shown individually in Fig. 3 (b)-(f), where the bolder black lines, like spines, are the errorbar plots.

The results for the Stanford Yeast cell cycle dataset are shown in Fig. 4, which also illustrates the members and errorbar plots for individual cluster, from (a) to (f) respectively. It is worth noting that Fig. 4 (f) indicates the Q phase reported in [11], which clearly has distinguishable pattern to other clusters although the number of members is small.

Fig. 5 shows the clustering results for the $\alpha$-38 Yeast cell cycle dataset, where four clusters are clearly displayed. The most attractive point of the proposed SMART scheme is that for above three different datasets, it has awareness of splitting, merging and terminating without the need of tuning the parameters and setting the number of clusters.

## 4. CONCLUSIONS

In this paper, we proposed a new self splitting-merging clustering algorithm, named splitting-merging awareness tactics (SMART). The novel framework, which integrates many techniques, starts with one cluster and employs a splitting-while-merging process. The SMART has self-awareness to split and merge the clusters automatically in iterations. Both the framework and the techniques were detailed and illustrated by a good benchmark test. Furthermore, three microarray gene expression datasets were studied using our approach. The numerical results show that the SMART is automotive and effective..

## 5. REFERENCES

[1] R. Duin A. Jain and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 4C37, 2000.

[2] R. Xu and D. II Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645-678, 2005.

[3] D. X. Jiang, C. Tang, and A. D. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Trans. Know. and Data Eng.*, vol. 16, no. 11, pp. 1370-1386, 2004.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[5] Ya-Jun Zhang and Zhi-Qiang Liu, "Self-splitting competitive learning: a new on-line clustering paradigm," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 369-380, 2002.

[6] Cheng-Ru Lin and Ming-Syan Chen, "Combining partitional and hierarchical algorithms for robust and efficient data clustering with cohesion self-merging," *IEEE Trans. Know. and Data Eng.*, vol. 17, no. 2, pp. 145-159, 2005.

[7] S. H. Wu, A. W.-C. Liew, H. Yan, and M. S. Yang, "Cluster analysis of gene expression data based on self-splitting and merging competitive learning," *IEEE Trans. Inf. Tech. in Biomed.*, vol. 8, no. 1, pp. 5-15, 2004.

[8] P. J. Rousseeuw L. Kaufman, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, 1990.

[9] R. Presntice L. P. Zhao and L. Breeden, "Statistical modelling of large microarray data sets to identify stimulus-response profiles," *Proc. Natl. Acad. Sci. (PNAS)*, vol. 98, no. 10, pp. 5631–5636, 2001.

[10] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, no. 10, pp. 977-987, 2001.

[11] R. Fa and A. K. Nandi, "Comparisons of validation criteria for clustering algorithms in microarray gene expression data analysis," in *The second Int. Workshop on Genomic Sig. Proc. (GSP2011)*, 2011.

[12] T. Pramila, W. Wu, S. Miles, W. S. Noble, and L. 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 & Dev*, 2006. 20: 2266-2278.