SCALABLE CLUSTERING BASED ON ENHANCED-SMART FOR LARGE-SCALE FMRI DATASETS

Chao Liu[1], Rui Fa[1], Basel Abu-Jamous[1], Elvira Brattico[2,3], Asoke Nandi[1,4,]*

1. Department of Electronic and Computer Engineering, Brunel University London, UK
2. Cognitive Brain Research Unit, Institute of Behavioural Science, University of Helsinki, Finland
3. Helsinki Collegium of Advanced Studies, University of Helsinki, Finland
4. Department of Mathematical Information Technology, University of Jyväskylä, Finland

ABSTRACT

In this paper, we propose a scalable clustering paradigm to address the problems of excessive computational load and limited clustering performance in large-scale data. The proposed method employs the enhanced splitting merging awareness tactics (E-SMART) algorithm. The large-scale dataset is divided into many sub-datasets sampled randomly from original data. These sub-datasets are clustered using E-SMART with the number of clusters K detected automatically and the resulting partitions are combined and re-clustered. We evaluate our method using synthetic fMRI datasets with different noise levels and one real fMRI dataset. Results show that the accuracy and execution time outperforms the traditional clustering algorithms in large-scale datasets.

Index Terms— large-scale data, scalable clustering, E-SMART, sampling

1. INTRODUCTION

The advances in data collection and storage have generated huge amount of data in industry and scientific domains. For example, the popular social network service Facebook stores millions of accounts and user data [1]. In research areas like biology, scientists have begun to grapple with the big dataset from gene analysis [2] and other medical imaging techniques such as fMRI [3]. Clustering has gained popularity in exploring these datasets and identifying interesting distributions in the datasets [4–8]. Clustering may complement current other methods, forming a more complete picture of the research.

Many algorithms have been developed to address the large-scale data clustering problems. Incremental clustering [9, 10] and divide-and-conquer clustering [11] reduce the time by operating in a single pass over the data. Sampling based methods [6, 7] obtain clusters based on a small subset randomly selected from the original data, thus reduce the computation time. Coreset algorithms [14], instead of choosing the samples randomly, find the cluster centres based on a small number of representative data points. Other methods like BIRCH [15] transform the data into structures such as graph or tree to accelerate the access speed. Nowadays, many parallel algorithms have been proposed with the advances in distributed computing [10-12], where a big task is divided into sub-tasks and they can be executed simultaneously on computing nodes. Then the results from sub-tasks are merged into the final solution. Despite the variety of clustering algorithms designed to meet different large-scale data clustering requirement, almost all of them still need the number of clusters K, which is often unknown in the real data, as input. Estimating the appropriate K would cost large amount of time by using the exhaustive search and evaluation. There are several algorithms that can cluster data without K, for example, the splitting merging awareness tactics (SMART) [19] and its enhanced version (E-SMART) [4]. However, they need large amount of computation time in the case of large-scale data.

In this paper we propose a solution that can reduce the time of clustering the large-scale dataset by utilizing sampling and combination, with the number of clusters K automatically detected. To achieve this, we use SMART and E-SMART, which can cluster the dataset without requiring the number of clusters *a priori*, as the main clustering method. Firstly E-SMART technique is applied on each randomly drawn sample set, whose size is much smaller than the original data, and the results from the samples are then combined and re-clustered to form the final partitions. We evaluate our method on simulated datasets and a real fMRI dataset. Then we compare the number of K specified, object membership accuracy and time from our proposed method with those obtained by k-means. Results show our method has great capability of clustering large-scale data in terms of correctly specifying the cluster number in the original data based on the samples as well as assigning the right object membership.

The rest of the paper is organized as follows. In Section 2, we detail the method we proposed. In Section 3, we describe the experiments that we setup to evaluate our algorithm. In Section 4, the experimental results are shown and finally we have a discussion about the results of this

* 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.

study and its potential benefits, and draw conclusions in Section 5.

2. METHODS

2.1. E-SMART

We briefly introduce the principles of SMART and E-SMART algorithms in this section. More details can be found in [4, 19].

The SMART algorithm initializes the clustering procedure with two randomly generated clusters. The core of the SMART algorithm is the splitting while merging (SWM) framework, where clusters will be split if they contain distinct patterns and merged if the merging criterion is satisfied. Thus, SMART can split and merge clusters automatically during iterations until the stopping criterion is met.

The E-SMART algorithm uses successive processing to enhance the standard SMART, which is the first attempt to employ successive processing in clustering literature. The flowchart of E-SMART is shown in Figure 1. The successive strategy extracts cluster one by one in iterations rather than selects the best clusters according to selection criterion. To do so, the intermediate clusters are ranked from high to low with respect to their silhouette index. Then the best cluster is chosen, while the rest of the dataset fed into the SWM process for a new iteration until there is no more splitting actions in the SWM procedure.

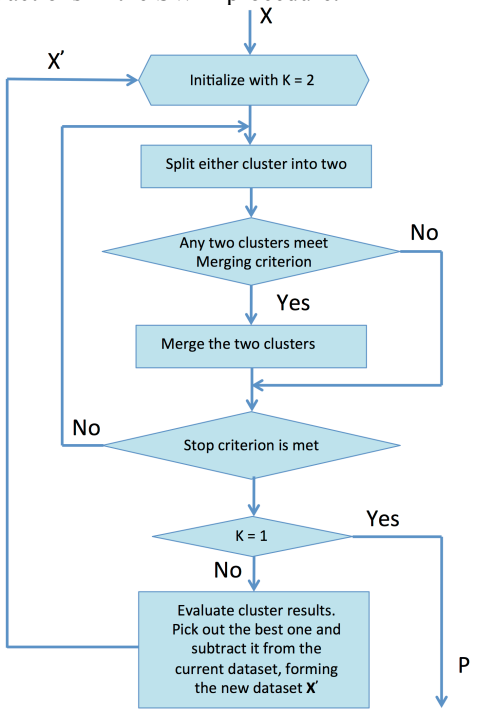

Figure 1. The flowchart of E-SMART algorithm

2.2. Sampling

Sampling is drawn randomly from the original dataset at the rate $1/s$. In this paper, we use sampling without replacement, so all s sub-datasets are generated with each one contains absolutely different data points from others and the union of them is the original dataset. Let $\boldsymbol{X}$ denote the whole data and all the subsets are $\{\boldsymbol{X}_i, i = 1, \dots, s\}$. Equation (1) describes the relationship among all the sub-datasets.

$$\boldsymbol{X} = \boldsymbol{X}_1 \cup \boldsymbol{X}_2 \cup \dots \cup \boldsymbol{X}_s, \boldsymbol{X}_i \cap \boldsymbol{X}_j = \emptyset, i \neq j, i, j \in [1, s]. \quad (1)$$

2.3. Combination

In total, s partitions $\{\boldsymbol{P}_i | i = 1, \dots, s\}$ are generated by E-SMART. Each partition $P_i (i = 1 \dots s)$ has its number of clusters detected as K_i, then there will be K_i cluster centroids which are denoted as $\{\boldsymbol{C}_k | k = 1, \dots, s\}$. To combine these intermediate results into a final partition, we use the following procedure.

1) Put all the cluster centroids $\{\boldsymbol{C}_k | k = 1, \dots, s\}$ into a new dataset $\boldsymbol{C}$, which is the assembly of all the cluster centroids detected.
2) Cluster centroid set $\boldsymbol{C}$ is further clustered by hierarchical clustering with Ward linkage.
3) Calculate the cluster number K for the whole dataset was calculated as the mode of $\{K_i\}$, denoted as K_m.
4) Retrieve the clustering results of step (2) by choosing cluster number K_m, yielding a new partition for the centre set C, denoted as $\boldsymbol{P}_c$.
5) Calculate the mean of each cluster in $\boldsymbol{P}_c$, which yielding K_m centre points.
6) For each datum in original dataset, assign it to its nearest centre from K_m centre point obtained in step (5) using Euclidian distance.

3. EXPERIMENT

3.1 Data preparation

We create the synthetic datasets to mimic the size of the real fMRI datasets by using first-order Markov model. Suppose that the synthetic fMRI dataset contains K patterns and each pattern has T time points. For each pattern, the time series are generated by using (2), where each time point is the sum of its previous status and a random factor having a Gaussian distribution (mean = 0, $\sigma^2 = 1$). The initial state is a variable having a uni-distribution in the interval $[-1, 1]$.

$$d_t = d_{t-1} + \epsilon_t, t = 1, \dots, T. \quad (2)$$

Repeating the process in formula (2) for K patterns form a $K \times T$ matrix $\boldsymbol{D}$, with each row representing a pattern. Each row of $\boldsymbol{D}$ was then normalized to zero mean and unit-variance. Then we create a $K \times N$ sparse matrix $\boldsymbol{S}$ where each column only has one non-zero element equal to one, indicating the membership of each data point. The final data is generated by using equation (3).

$$\boldsymbol{X} = \boldsymbol{D}^T \boldsymbol{S} + \boldsymbol{n}, \quad (3)$$

where $\boldsymbol{X}$ is a $T\times N$ matrix containing K clusters with each column representing a data point, and $\boldsymbol{n} \in \boldsymbol{R}^{T\times N}$ denotes additive white Gaussian noise $(0, \sigma^2)$. In this study, we choose σ equal to 0.01, 0.1, 0.2 and 0.3.

The real data come from an fMRI listening experiment related to the music emotions [20, 21] carried out in the University of Helsinki. The whole fMRI experiment for one participant has 450 scans (TR=2s) including 32 music categories with each one repeated twice and each scan contains 228,453 voxels after preprocessing. In this paper, we use only one condition from one random subject in the experiment and apply our paradigm to it. We remove the isolated small points from the results and visualize them in 3D space to see the brain areas that have highly correlated blood oxygen level dependent (BOLD) activities.

3.2. Clustering experiment

The experiments focus on three aspects listed below.

1) Test the ability of detecting correct number of clusters based on samples,
2) Test object membership accuracy, and
3) Test execution time.

We firstly use synthetic data for a quantitative evaluation of our method. We apply E-SMART on samples and the original synthetic data and compare the results of our method with those obtained by k-means. For k-means, we arbitrarily set an interval for number of clusters ranging from 45 to 55 and run k-means on each subset with all the K value. In the data with number of clusters unknown, the K needs to be chosen from a wider range. Silhouette index is used to evaluate the clustering results quality and determine the estimate K. The estimated K is chosen as the one that yields highest average Silhouette index. Then we compare the mode of these cluster numbers with the ground truth (K = 50) of the synthetic datasets. We also run k-means on the whole synthetic dataset to investigate its accuracy and performance. We find the calculation of Silhouette index for one result in this step cost huge amount of time (about 3 hours) in the pilot experiment. So in the whole dataset clustering, we only report the results obtained by k-means with K equal to 50.

For the real data, due to the fact that there is no ground truth of the number of clusters and the large amount of time needed for calculating Silhouette index for estimating number of clusters in k-means, we cannot set an arbitrary range for the number of clusters for k-means to reduce the experiment time. Thus we only apply our method on the real fMRI data to detect all the distinct BOLD patterns.

We use adjusted Rand index (ARI) [22] and normalized mutual information (NMI) [23] as the metrics to evaluate the clustering membership accuracy on the synthetic dataset.

The time we aim to compare includes two parts which are the time needed to specify appropriate cluster number and the actual execution time. The reason is that good clustering results can be obtained only if the appropriate cluster number is given. So it is important to include the time needed to specifying the cluster number together with the execution time.

4. RESULTS

4.1. Determining the number of clusters

In the low noise level condition, our proposed method can detect the number of clusters correctly on all samples while the results from k means fluctuate across the whole K range we set. With the increased noise level, our method still generate more stable estimation of the number of clusters than that from k-means.

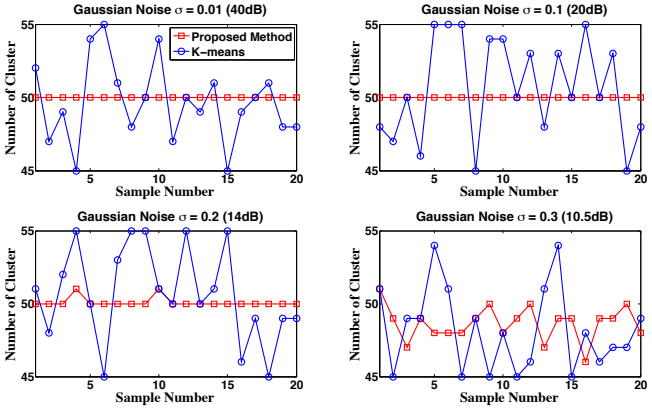

Figure 2. Number of clusters detected by E-SMART and k-means on each sample under different noise levels.

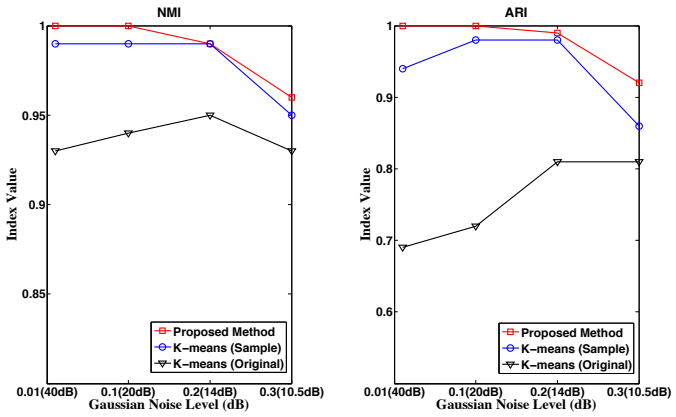

Figure 3. The normalized mutual information (NMI) and adjusted Rand index (ARI) comparison under different noise levels.

4.2. Object membership accuracy

The comparisons of the accuracies of assigning object membership are shown in Figure 3. We note that the proposed method has the perfect accuracy in the high SNR situation and very competent results compared to k-means on samples and k-means on original dataset under different noise level for the middle SNR situation, especially compared with the k-means on original data. Even when the SNR is low, the proposed method still achieves the highest accuracy both in NMI and ARI.

4.3. Execution time

Noise Level	Method	Mean Time (sec)
$\sigma = 0.01$ $(40dB)$	Proposed method	277
	K-means (Sample)	5600
	K-means (Original)	9400
$\sigma = 0.1$ $(20dB)$	Proposed method	1700
	K-means (Sample)	6000
	K-means (Original)	9700
$\sigma = 0.2$ $(14dB)$	Proposed method	3050
	K-means (Sample)	6000
	K-means (Original)	9650
$\sigma = 0.3$ $(10.5dB)$	Proposed method	3400
	K-means (Sample)	6400
	K-means (Original)	10100

Table 1. Execution time on subsets under different noise level.

Note the time for proposed method is the mean of the duration of the experiment on each of the 20 subsets. The K-means (Sample) is the estimated time of applying k-means with all the possible K values which should approximately range from 1 to $\sqrt{n/2}$ (~300 in this study) in this experiment. And the number in K-means (Original) is the time for single run and evaluation on the whole synthetic dataset.

4.4. Results from real data and visualization

On the real fMRI data with no ground truth of the number of clusters , our method detects stable estimation of K which is around 170. The execution time increases due to the large K in the data, compared with the time on samples with K equal to 50 (Figure 4). But the speed is still competent compared with the case that use k-means to do the exhaustive search on real data. The 3D mapping of the results covers many large areas of the brain indicating highly synchronized BOLD activities as shown in the time series of the corresponding clusters in Figure 5.

Figure 4. The number of clusters detected in real fMRI data and the execution time on each sample.

5. DISCUSSIONS AND CONCLUSIONS

In this paper, we utilized the feature of E-SMART algorithm and sampling to propose a solution to speed up the clustering of large-scale data with the number of clusters automatically detected based on random sampling and the clustering results combination. E-SMART was applied on all samples randomly drawn from the original data, yielding one partition per sample. Then these partitions of the samples were further clustered and combined into a final partition for the original data. From the results shown in Figure 2, we could see our method has great capabilities to estimate the cluster number in original data from the random sampled sub-datasets. On the contrary, the traditional k-means could not stably detect the cluster number very well even under the low noise condition. From Figure 3, we could see the object membership accuracy obtained by our method could achieve very high value, indicating excellent clustering results, outperforming k-means applied on both samples and original dataset. In terms of time, E-SMART avoided exhaustive search in determining the appropriate number of clusters, and due to the small size of samples, our method gave competent speed performance. Our method could also detect an unknown number of clusters on the real fMRI dataset as shown in Figure 4. The clusters obtained reveal all the brain areas having highly correlated BOLD activities during the fMRI experiment (Figure 5). The capability of extracting all the distinct patterns in the fMRI data by our method could provide very fine information for studying the whole brain functional connectivity [24].

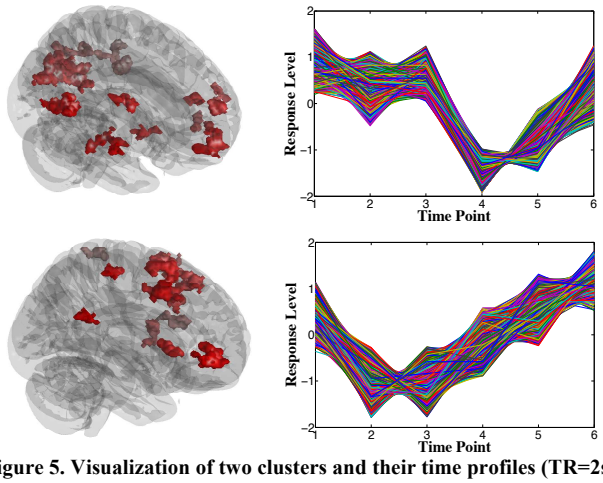

Figure 5. Visualization of two clusters and their time profiles (TR=2s).

Another good feature of our proposed method is the experiments can be run simultaneously on multiple machines, as it does not need data communications between different subsets before the combination. With the help of the power of distributed computation technique, each computing element can handle more than one sample clustering tasks. So in ideal case, no matter how big the data is, the completion time for clustering the whole data is equal to the time needed by the algorithm for one sample. This algorithm is able to be extended in the future to do the random sampling repeatedly and combine all these clustering results, which would benefit from the diversity of the sampling, yielding more sound clustering results.

6. REFERENCES

[1] D. Boyd and K. Crawford, "Critical Questions for Big Data," *Information, Commun. Soc.*, vol. 15, no. 5, pp. 662–679, Jun. 2012.

[2] V. Marx, "Biology: The big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, Jun. 2013.

[3] S. A. Huettel, A. W. Song, and G. McCarthy, *Functional magnetic resonance imaging*. 2004, p. 542.

[4] R. Fa, B. Abu-Jamous, D. J. Roberts, and A. K. Nandi, "Enhanced SMART framework for gene clustering using successive processing," *2013 IEEE Int. Work. Mach. Learn. Signal Process.*, pp. 1–6, Sep. 2013.

[5] B. Abu-Jamous, R. Fa, D. J. Roberts, and A. K. Nandi, "Paradigm of tunable clustering using Binarization of Consensus Partition Matrices (Bi-CoPaM) for gene discovery.," *PLoS One*, vol. 8, no. 2, p. e56432, Jan. 2013.

[6] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg, "A whole brain fMRI atlas generated via spatially constrained spectral clustering.," *Hum. Brain Mapp.*, vol. 33, no. 8, pp. 1914–28, Aug. 2012.

[7] A. Venkataraman, K. R. A. Van Dijk, R. L. Buckner, and P. Golland, "Exploring functional connectivity in fMRI via clustering," in *ICASSP 2009*, 2009, pp. 441–444.

[8] B. Thirion, G. Varoquaux, E. Dohmatob, and J.-B. Poline, "Which fMRI clustering gives good brain parcellations?," *Front. Neurosci.*, vol. 8, no. July, pp. 1–13, Jul. 2014.

[9] F. Can, E. A. Fox, C. D. Snavely, and R. K. France, "Incremental clustering for very large document databases: Initial MARIAN Experience," *Information Sciences*, vol. 84. pp. 101–114, 1995.

[10] F. Can, "Incremental clustering for dynamic information processing," *ACM Transactions on Information Systems*, vol. 11. pp. 143–164, 1993.

[11] C. C. Aggarwal, T. J. Watson, R. Ctr, J. Han, J. Wang, and P. S. Yu, *A Framework for Clustering Evolving Data Streams*. 2003, pp. 81–92.

[12] S. Guha, R. Rastogi, and K. Shim, "CURE: an efficient clustering algorithm for large databases," *Inf. Syst.*, vol. 26, no. 1, pp. 35–58, 2001.

[13] G. Kollios, D. Gunopulos, N. Koudas, and S. Berchtold, "Efficient biased sampling for approximate clustering and outlier detection in large data sets," *IEEE Trans. Knowl. Data Eng.*, vol. 15, pp. 1170–1187, 2003.

[14] S. Har-peled and M. Soham, "Coresets for k -Means and k -Median Clustering and their Applications," pp. 1–21, 2003.

[15] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," in *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, 1996, pp. 103–114.

[16] W. Y. Chen, Y. Song, H. Bai, C. J. Lin, and E. Y. Chang, "Parallel spectral clustering in distributed systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, pp. 568–586, 2011.

[17] W. Zhao, H. Ma, and Q. He, "Parallel K-means clustering based on MapReduce," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, vol. 5931 LNCS, pp. 674–679.

[18] M. Wang, W. Zhang, W. Ding, D. Dai, H. Zhang, H. Xie, L. Chen, Y. Guo, and J. Xie, "Parallel clustering algorithm for large-scale biological data sets," *PLoS One*, vol. 9, 2014.

[19] R. Fa, D. J. Roberts, and A. K. Nandi, "SMART: unique splitting-while-merging framework for gene clustering.," *PLoS One*, vol. 9, no. 4, p. e94141, Jan. 2014.

[20] E. Brattico, V. Alluri, B. Bogert, T. Jacobsen, N. Vartiainen, S. Nieminen, and M. Tervaniemi, "A Functional MRI Study of Happy and Sad Emotions in Music with and without Lyrics.," *Front. Psychol.*, vol. 2, no. December, p. 308, Jan. 2011.

[21] V. Alluri, P. Toiviainen, T. E. Lund, M. Wallentin, P. Vuust, A. K. Nandi, T. Ristaniemi, and E. Brattico, "From vivaldi to beatles and back: Predicting lateralized brain responses to music," *Neuroimage*, vol. 83, pp. 627–636, 2013.

[22] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *J. Am. Stat. Assoc.*, vol. 66, pp. 846–850, 1971.

[23] A. F. McDaid, D. Greene, and N. Hurley, "Normalized Mutual Information to evaluate overlapping community finding algorithms," *Arxiv Prepr. arXiv1110.2515*, pp. 1–3, 2011.

[24] K. Li, L. Guo, J. Nie, G. Li, and T. Liu, "Review of methods for functional brain connectivity detection using fMRI.," *Comput. Med. Imaging Graph.*, vol. 33, no. 2, pp. 131–9, Mar. 2009.