IMPROVING THE INITIALISATION AND RELIABILITY OF THE SELF ORGANISING OSCILLATOR NETWORK

S. A. Salem, L. B. Jack, and A. K. Nandi

Signal Processing and Communications Group,
Department of Electrical Engineering and Electronics, The University of Liverpool
Brownlow Hill, Liverpool, UK, L69 3GJ
(sameh.salem, a.nandi) @liv.ac.uk

ABSTRACT

The Self-Organising Oscillator Network (SOON) provides a novel way for data clustering [1, 2]. The SOON is a distance based algorithm, meaning that clusters are determined by a distance parameter, rather than by density distribution, or a pre-defined number of clusters. Repeated experiments have highlighted the sensitivity of this algorithm to the initial selection of phase values and prototypes. In repeated experiments, the SOON as proposed by Frigui is shown to have a number of shortfalls in terms of its performance over repeated clustering runs. This paper proposes improvements to the initialisation stage of the algorithm by comparing the difference between random initialisation of the phase curve and initialisation using the ordering obtained from a hierarchical clustering approach. This leads to improved convergence of the algorithm and more robust repeatability. When compared against random generation of phases and prototypes as published by Frigui originally, the changes in initialisation are shown to give significant improvements in the performance of the algorithm.

1. INTRODUCTION

Data overload is an increasing problem in many different areas of science and engineering. The creation of vast datasets in science, government and industry presents challenging analytical and statistical problems. Unfortunately, the ability of interested parties to analyse these datasets in a reasonable amount of time and at a reasonable cost has not kept pace. Clustering is one such area where the automated analysis of large datasets is important, while also becoming a problem. Numerous different unsupervised clustering techniques are commonly in use; the Kohonen Self Organising Map (SOM) has been perhaps one of the most popular unsupervised clustering algorithms, and is used in many different applications [3, 4], hierarchical clustering techniques, K-means clustering [5, 6], and k-medoids [7] are also widely used. Other alternative techniques, such as fuzzy clustering [7, 8] and many variations of vector quantisation [9] are density based and their main drawback is that they cannot map the distribution of data in areas where the density of the data is low [2]. Distance based clustering techniques attempt to alleviate this problem by utilising the premise that clusters are determined by a distance parameter, rather than by density distribution. The basis is simply that for any given cluster centre, all data points regarded as being members of the cluster will fall within a preset distance. This will allow clusters in sparsely populated areas of data space to be formed without affecting clustering in more dense areas of the data. This paper examines the applicability and the reliability of the SOON algorithm as a new clustering technique on a data set from a real-world communication data problem.

2. THEORY

The Self-Organising Oscillator Network (SOON) is a comparatively new clustering algorithm [1] that has received relatively little attention so far. SOON is a concept with roots in biology; the algorithm is modelled on biological principles: a good example of the synchronising oscillator phenomenon would be that of fireflies, which start by flashing at random initially, however the groups that are physically close to each other will synchronise their firing. Groups that are separated by distance will fire as disparate groups, each synchronised within itself.

The behaviour of self-organisation of components with an oscillatory nature gives rise to the name of the algorithm - the Self Organising Oscillator Network (SOON). With the SOON method, each object in the data, O_j, is represented as an “Integrate and Fire” oscillator, characterized by a phase ϕ_j and state χ_j, where:

$$\chi_j = f(\phi_j) = \frac{1}{b} ln[1 + (e^b - 1)\phi_j]. \qquad (1)$$

where $0 \leq \phi_j \leq 1$ and $0 \leq \chi_j \leq 1$ for $j = 1 \ldots\ldots n$; $f(\phi_j)$ is smooth, monotonically increasing function with $f(0) = 0$ and $f(1) = 1$; b is a constant determining how $f(\phi_j)$ is concave down (usually $b = 3$). Whenever an oscillator’s state reaches the threshold ($\chi_j = 1$), it “fires”, with the following consequenses:

- The oscillator phase and state, ϕ_j and χ_j are set to zero; and
- The phases of all the other oscillators change by an amount $\varepsilon(\phi_i)$, for $i = 1 \ldots\ldots n$; $i \neq j$.

Changing the phase of other oscillators has the effect of either exciting or inhibiting them. An oscillator is excited by having its phase increased, while it is inhibited by decreasing its phase. The precise amount of the change is determined by the coupling function $\varepsilon(\phi_i)$, which in turn depends on the dissimilarity between the two oscillators (equivalent objects). A typical coupling function would be as follows:

$$\varepsilon(\phi_i) = \begin{cases} C_E\,[1 - (\frac{\delta_{ij}}{\delta_o})^2] & \delta_{ij} \leq \delta_o \\ \\ C_I\,[(\frac{\delta_{ij}-1}{\delta_o-1})^2 - 1] & \delta_{ij} > \delta_o \end{cases} \qquad (2)$$

where $\delta_{ij} = d(O_i, O_j)$ is short-hand for the measure of dissimilarity between two objects i and j, and δ_o is a threshold dissimilarity that determines the cut off for what is deemed “similar”; δ_o can be viewed as a resolution parameter, as it affects the number of groups created. C_E and C_I are the maximum excitory and inhibitory couplings permitted. The firing of an oscillator tends to excite a few oscillators, whilst inhibiting many others. Thus an oscillator receives much more inhibition than excitation; hence $C_E \gg C_I$. Once a set of oscillators is synchronized, the way that members of the set interact with other oscillators not in the set must be made uniform.

For larger datasets, a set of prototypes can be used as a smaller number of initial cluster centres; these are normally either selected at random from the data, to create a subset of the data, alternatively they may be generated in order to cover the extents of data space, so that an even distribution of the centres over data space is made.

2.1 Stability and Robustness

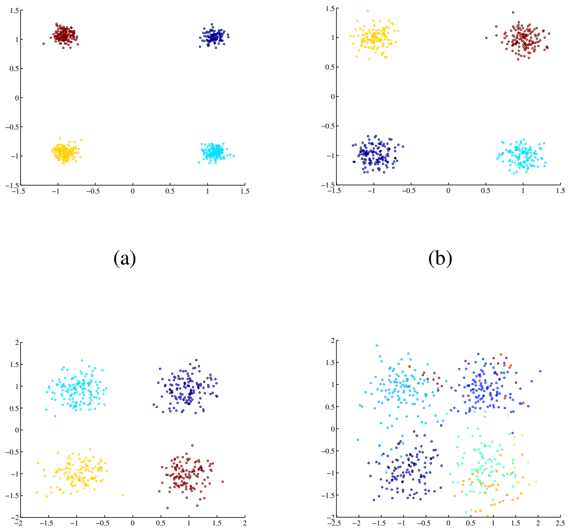
The algorithm, as originally published, called for the random initialisation of the phase values; this can potentially cause problems with the initial selection of cluster centres, as the first data point to fire within a certain group may not be best suited to be the centre of the cluster. Additionally, poor choice of the C_I and C_E parameters may cause incorrect formation of clusters, as data points are pushed off of the end of the phase curve. By changing the initialisation stages of both the phase and prototype values, it should be possible to improve this significantly. Part of the convergence time of the algorithm is based on the principle that over time, the random phases will take time to move from their random starting positions to clusters of the same phase values, attracted by the distance from the cluster centre. By using a hierarchical clustering technique to order the data points in terms of their proximity to each other by initialising the phase vector ϕ by the permutation vector of the node labels of the leaves of the dendrogram, it should be possible to achieve the same result in a much faster period of time, as the initial phase curve will much more closely match that of the actual distribution of the data in terms of distance. Thus, less time will be spent waiting for the algorithm to stabilise, as the algorithm will start off much closer to convergence than a random initialisation would allow. We propose to use the results of the hierarchical clustering to select the prototype centres which are expected to improve the clustering performance , as the centres will map the centres of the data more closely than selecting a random subset of the data as proposed by [1].

3. DATASETS

In order to examine the reliability of the SOON algorithm, we use a communication data set, which represents a real-world problem; in this specific case constellation diagrams from different communication modulation schemes are a good benchmark test problems for unsupervised clustering algorithms, primarily due to the clean delineation between different clusters within constellation at high SNR. However, by lowering the SNR of the signal, it is possible to gain less well-separated data, which can pose more of a problem for the clustering algorithm, as the data becomes non-separable.

4. EXPERIMENTS

The effect of the initialisation of the phase values of the data is investigated for both random generation of phase values and using the hierarchical clustering ordering to determine the initial phase values. A second set of experiments examine the effect of a randomly selected subset of the dataset for the initial prototypes, as opposed to using a subset of the clusters generated by the hierarchical clustering approach.



(a)

(b)

(c)

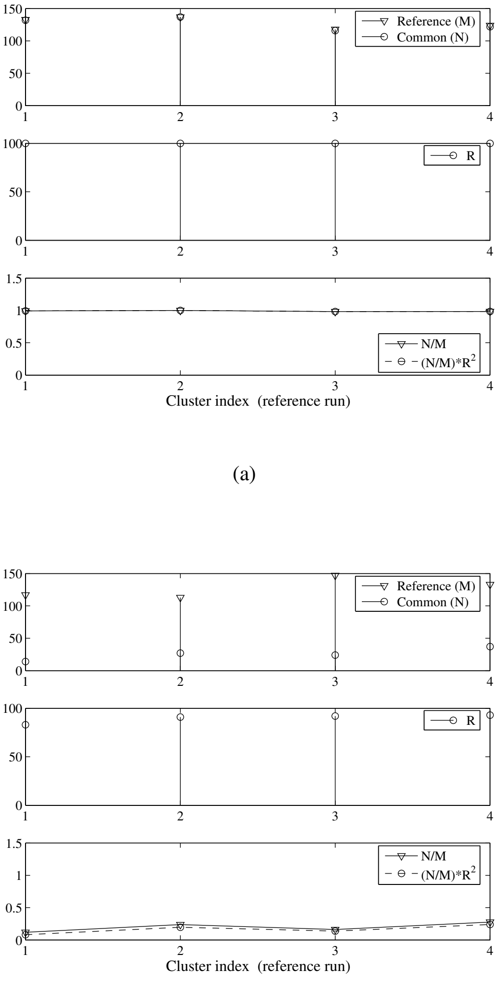
(d)

Figure 1: The clustering performance for SNR, (a) 25 dB, (b) 20 dB, (c) 15 dB, (d) 10 dB.

4.1 Results & Discussions

A series of experiments were carried out to examine the clustering performance on a QPSK signal for different SNR values using proper value of δ_o. Figure 1 depicts the clustering performance for different SNR values. The effect of random generation of prototypes is tested and evaluated 100 times using prototypes (set to one half of the total number of data points available for clustering). One reference run (best run) was selected from the 100 runs using a reliable validity index. In this paper I index [10] is used. Figure 2 shows the analysis for low SNR values (15 and 10 dB). The two plots in Figure 2 contain three subplots, these subplots show (top) the difference in magnitude between the size of reference run clusters (M) and the comparison clusters (N), showing the relative reproducibility of the results; (middle) the number of runs (R) that achieve the common cluster membership for each cluster in the reference run, (bottom) the “gain” per run (N/M) and the weighted gain per run $(N/M) * R^2$. As can be seen, examining the plots for 10 dB and 15 dB show that there is a marked deterioration in the reproducibility of the clustering result, as the fraction of cluster members maintained (i.e. the gain - N/M) across many runs drops significantly for the 10 dB SNR data.

For improvement of the clustering performance of the SOON algorithm, we propose to use the hierarchical clustering to set up and order the phases on the concave function curve that express the SOON model which leads to better performance as well as speeding up the algorithm. Figure 3 shows the effect of hierarchical clustering for phase genera-



(a)

(b)

Figure 2: Analysis of the effect of noise on clustering behaviour, over 100 runs for SNR, (a) 15 dB, (b) 10 dB.

tion on the clustering performance at 10 dB SNR. As shown in Figure 3, the clustering performance using hierarchical clustering for phase generation gives a significant improvement, which is shown by the gain per run, as well as the improvement in the weighted gain.

For further improvement of the repeatability of the repeatability of the clustering performance, the results from the hierarchical clustering can also be used for the generation of prototypes, in order to achieve better distribution of the prototypes among the dataset. This reduces the effect of noise within the data, while also improving the initial cluster positions that the algorithm accepts, by choosing centres that are implicitly well distributed throughout the data. The effect of this is very clear when using hierarchical clustering for phase generation. Figure 4 depicts the clustering performance of hierarchical prototype generation using random and hierarchical phases. As can be seen, using the hierarchically derived prototypes significantly improves the repeatability of the algorithm, while also unifying the gain across all clusters. Additionally, the performance remains substantially the same as the number of prototypes increases or or decreases.

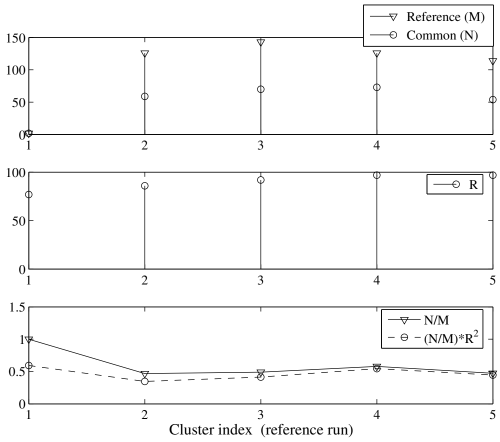


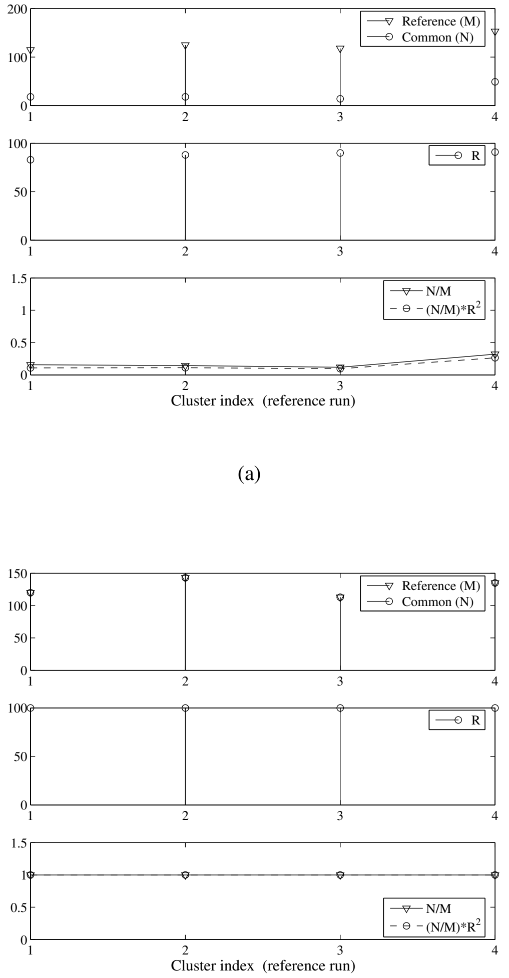
Figure 3: Analysis of the effect of hierarchical phase generation on repeatability and performance over 100 runs.

In order to rely on the conclusion obtained above, clustering validity methods [10] can be used for evaluating and assessing the results of the proposed improvements of the SOON algorithm. From Table 1, validation indices [10] values, *I, CH*, and *DBnc* indicate that the hierarchical phase with hierarchical prototype selection is better than the hierarchical phase with random prototype selection, and the latter is better than the random phase with random prototype selection. However, *Dunn* index [10] values do not agree with the above conclusion, where it is very sensitive to the presence of noise in datasets [11].

Additionally, the convergence time (on Xeon 2.8 GHZ CPU with 512 MB ram using C code) is experimentally tested for 100 runs. As illustrated in Table 1, the convergence time of hierarchical phase with hierarchical prototype generations is the lowest one compared against random prototype generations with random phases, or hierarchical phases.

Validation indices	**Random phase generation**	**Hierarchical phase generation**	**Hierarchical phase & protototypes**
I [10]	1.7	1.8	**2.8**
CH [10]	467	471	**690**
DBnc [10]	0.98	0.89	**0.62**
Dunn [10]	**0.0071**	0.0070	0.0031
Convergence time in sec	75	72	**36**

Table 1: Results of different validity indices & convergence times



(a)

(b)

Figure 4: Analysis of the effect of structured prototype selection on repeatability for, (a) Random phase with hierarchical selection of prototypes, (b) Hierarchical phase with hierarchical selection of prototypes.

One of the main advantages of SOON algorithm is its robustness in terms of the input parameters, where all data points regarded as being members of the cluster will fall within a preset distance which actually describes the degree of the similarity between objects compared against the hierarchical clustering algorithm or any partitional clustering algorithms that depend on the number of clusters K as one of the input parameters. Therefore, it is very difficult to get a fair comparison between radius based clustering algorithms and partitional clustering algorithms.

5. CONCLUSIONS

Two improvements to the initialisation of the SOON algorithm have been proposed. As can be seen, the proposed use of hierarchically derived initial phase and prototype values causes significant improvements in the repeatability of the clustering performance of the algorithm. Additionally, the convergence of the algorithm is also significantly increased as the phase curve is already ordered in terms of similarity, meaning that fewer iterations are required for the algorithm to reach convergence.

REFERENCES

[1] H. Frigui and M. B. H. Rhouma, "Self-Organization of Pulse-Coupled Oscillators with Application to Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, pp. 180-195, Feb. 2001.

[2] L. B. Jack and A. K. Nandi, "Microarray Data using The Self Organising Oscillator Network," in *Proc. EUSIPCO 2004*, Vienna, Austria, September 6-10. 2004, pp. 2183-2186.

[3] T. Kohonen, *Self-Organising Maps*, Springer-Verlag, 1997.

[4] S. Zhang, R. Ganesan, and G. D. Xistris, "Self-organizing neural neworks for automated machinery monitoring systems," *Mechanical systems and Signal Processing*, vol. 10, pp. 517-532, 1996.

[5] J. MacQueen. "Some methods for classification and analysis of multivariate observations," in *Proc. the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1967, vol. 1, pp. 281-297.

[6] C. W. Chen, J. B. Luo, and K. J. Parker, "Image Segmentation Via Adaptive K-Mean Clustering And Knowledge-Based Morphological Operations With Biomedical Applications," *IP*, vol. 7, pp. 1673-1683, Dec. 1998.

[7] L. Kaufman and P. Rousseeuw, *Finding Groups in data: an introduction to cluster analysis*, John Wiley and Sons, New York, 1990.

[8] N. J. Pizzi, M. Alexander, R. Vivanco, and R. Somarjaj. "Fast non iterative registration of magnetic resonance images," in *Proc. SPIE: Medical Imaging 2001*, vol. 4322, pp. 1599-1608.

[9] W. Xu, A. K. Nandi, and J. Zhang, "Novel fuzzy reinforcement learning vector quantization algorithm and its application in image compression," *IEE Proceedings on Image and Signal Processing*, vol. 150, pp. 292-298, Oct. 2003.

[10] U. Maulik and S. Bandyopadhyay, "Performance Evaluation of some Clustering Algorithms and Validity Indices," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1650-1654, Dec. 2002.

[11] M. Haliki, Y. Batistakis, and M. Vazirgiannis, "Cluster Validity Methods: Part II," *SIGMOD Record*, vol. 31, Issue 3, Sep. 2002.