Proceeding of IC-ITS 2015 e-ISBN:978-967-0850-07-8
International Conference on Information Technology & Society
8-9 June 2015, Kuala Lumpur, MALAYSIA

Time Consuming Factors for Self-organizing Map Algorithm

Muhammad Firdaus Bin Mustapha[a], Noor Elaiza Binti Abd Khalid[b], Azlan Bin Ismail[c]

Faculty of Computer and Mathematical Sciences,
University of MARA Technology, Shah Alam, Selangor, Malaysia.

[a]*firdaus19@gmail.com,* [b]*elaiza@tmsk.uitm.edu.my,* [c]*azlanismail08@gmail.com*

Abstract. Self-organizing map (SOM) has been used as a tool in data exploratory in data mining. The SOM is very useful to visualize and explore the nature of data especially for large datasets by reducing high dimensionality of data into low dimensionality of data. SOM is differs from other techniques because it has learning nature and consists of projection and quantization methods. Despite its excellent performance, there is a major issues related to its slow processing time. The SOM algorithm consists of steps such as initialize neuron weights, find best matching unit (BMU), and update the weights. These steps involve a lot of calculations where the calculation of complexity depending on the circumstances. Both internal and external parameters of the algorithms should be analyzed with the interest to find consuming factors in SOM processing. This paper will examine factors that may affect SOM processing through several experiments. The experimental results are analyzed by comparing with different parameters. Thus, this paper discusses some of the factors to be considered to improve the processing of SOM.

Keywords: Self-organizing map, visualization.

INTRODUCTION

Self-organizing map (SOM) is an unsupervised neural network that has been used as a data analysis method. It is widely applied to clustering problem and data exploration in various areas of problems [1], with remarkable abilities to remove noise, outliers, and deal with missing values. The SOM algorithm is based on nonlinearly projection mapping that can reduce high dimensional data into low dimensional, normally two dimensional data (2D) through producing a network topology map. The SOM training process generates simultaneous clustering and projection of high-dimensional data which the learning nature of the algorithm allows us to constantly update information in order to improve the quality of the final results.

The SOM is varying from standard methods for exploratory data analysis because the technique combines clustering method and visualization method [2]. The clustering method performs through vector quantization and the visualization presented by a regular grid shape via projection of weights onto output space. In term of visualization, a graphical representation of an analyzed dataset is eligible for convenient analysis and interpretation and the SOM is an efficient and effective visualization technique [3]. SOM

also can be adapted with time which been used for exploratory temporal multivariate structure analysis [2].

The SOM algorithm consists five major steps [4]; initialize neurons on SOM map, calculate distance and find best matching unit (BMU), identify neighborhood region neuron, update the neurons located within the neighborhood region, and modify learning factor and neighborhood radius. SOM's measurement are used such as quantization errors (QE) and topographic errors (TE) with the interest to evaluate the SOM accuracy [5].

Despite its excellent performance, there are problems related to the large map, where it causes the workload on the processor especially when dealing with winner-search and updating weightage of neurons on the map [1]. Additionally, the larger mapping size also could burden the processing through the usage of high memory capacity which leads to high rate memory transfer [6], [7], [8]. On the other hand, datasets features also have a great influence to the SOM processing [8]. This research will evaluate factors that influence through computation experiments. The factors will be highlighted for future improvement with the interest to improve SOM processing.

SOM ALGORITHM

Basically SOM algorithm comprise of two layers; input layer and output layer. The input layer that contains vectors will project the values onto the output layer. The output layer consists of neurons that form a SOM map. Normally the map shapes are either square or rectangle. Different map size can be used such as 5 by 5 or 5*5 which contains 25 neurons in the map. Figure 1 depicts the SOM algorithm [9]. The algorithm start with the initialization of weights at the output layer.

The process start with it acquires a record of data, x from the training set, X or dataset. Each neuron in the map will be calculated the distance with x, and find the nearest distance is called best matching unit (BMU). This step involves a lot of computation because it repeat the calculations for all records in the dataset. Many researchers agree that this step is the most complex computation in SOM algorithm [1], [6], [7], [8]. Next, to determine a subset of neurons close to the BMU using map radius. The neurons located within the radius will be migrated towards the BMU by updating neurons weight. The updating step also involves numerous calculation as calculate distance. The last step is to modify the learning factor and neighborhood radius. The algorithm will repeat previous steps for next record of data, x that acquires from the training set, X. After completion of whole records, it is counted as complete one epoch cycle. The algorithm will continue train the data until reach the number of iteration that pre-defined before the algorithm has executed.

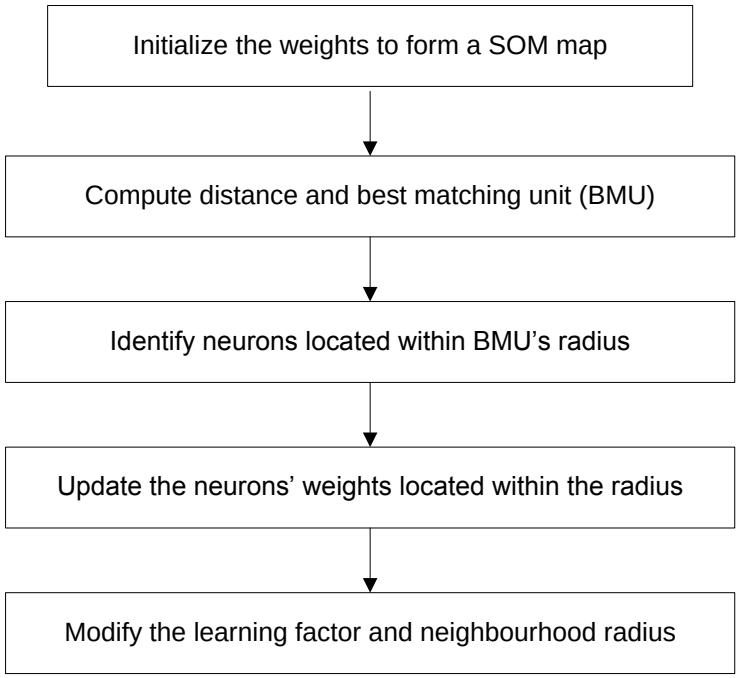

FIGURE 1. SOM algorithm [9].

SOM Measurements

There are two popular SOM measurement, QE and TE with two evaluation criteria: measuring the quality of the continuity of mapping and topology preservation or mapping resolution. These measures have been proposed for computing and comparing the quality of a SOM's projection [5]. The purpose of the measurements is to measure how better the topology is preserved in the projection onto the SOM's lower dimension map grid. Lower QE and TE values specify superior mapping quality.

SOM Map Size

Specification of map size (number of output neurons) in the SOM training process is very much important to detect the deviation of the data. If the map size is too small, outcome is more general patterns and it may not reveal some significant differences that should be detected. On the other hand, larger map sizes outcome into more detailed patterns where the differences are too small. Thus, controls run is repeated by changing the map size only during the training of the network, and select the optimum map size where the QE and TE resulted with minimum values.

EXPERIMENTAL SETUP

The experiments apply standard SOM algorithms from R package. Bank Marketing datasets are used for data testing that taken from UCI Machine Learning Repository. There are two different size of datasets whereas the first dataset contains 4522 samples and 17 dimensions. Meanwhile, the second dataset has 41188 samples and 21 dimensions. The following table indicate the information of both datasets.

TABLE(1). The information of datasets used in the experiment

Bank Marketing Datasets		
No. of samples	4522	41188
Data dimensions	17	21

Experimental Results

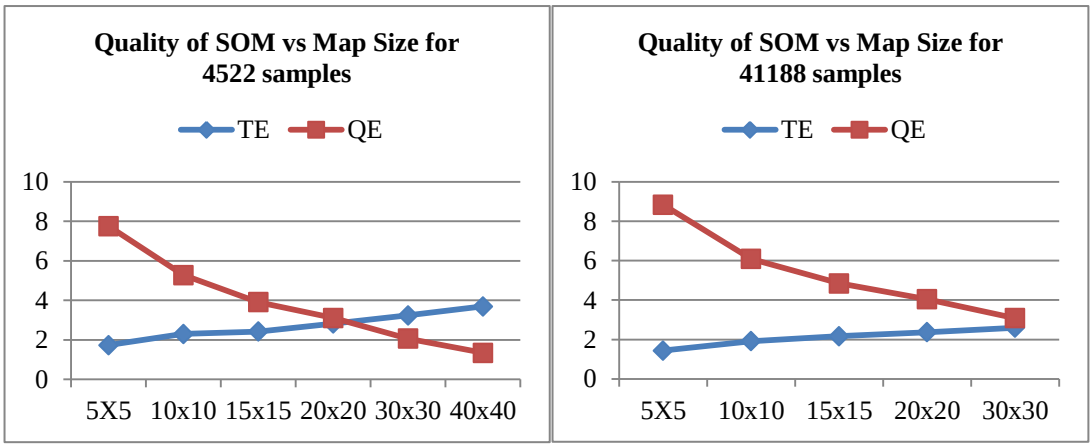

FIGURE 2. The quality measurement versus SOM map size for datasets 4522 samples at the left and 41188 at the right.

According to [1], map size (number of output neurons) can influent the SOM quality measurement that usually used by researchers; quantization error (QE) and topographic error (TE). From the experiment, both datasets achieve the similar pattern of results when applying larger map size. The QE reduces the errors when the map size growth while the TE expand the errors. The QE drive into lower values because when using a bigger map, the average of BMU become lower and the map can converge more efficiently.

In order to select the optimum map size, the QE and TE must resulted with minimum values [5]. During the training of SOM network, different of map sizes have uses for both datasets. Therefore, the 20x20 and 30x30 map sizes was chosen representatively for 4522 samples and 41188 samples. Both QE and TE scored optimum when they encounter each other in the graph shown in Figure 2.

SOM algorithms train the data to produce a SOM map. According to the map obtained, there are different feature maps for every parameters of the datasets. Figure 3 below shows the feature map of age, education, housing, and marital for 10x10 and

20x20 map sizes. Although the 20x20 map size has been chosen for 4522 samples, the pattern in 10x10 and 20x20 are quite identical each other where the different is the 20x20 shows more information such as data clusters in the pattern. In the map, a neuron has low values are coloured with darker (red) colour meanwhile a neuron has higher values are coloured with lighter colour (white).

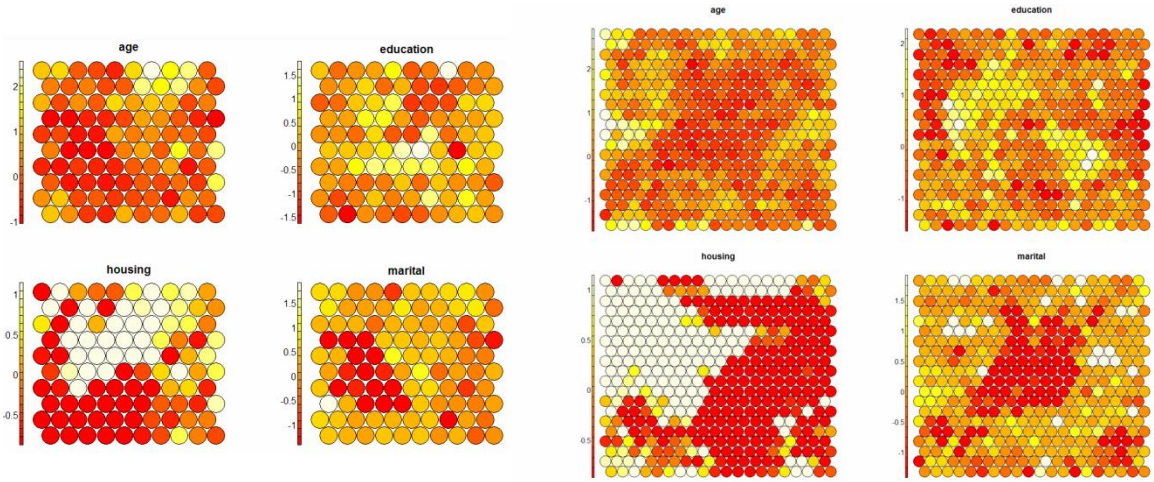

FIGURE 3. (a) SOM 10x10 pattern and (b) SOM 20x20 pattern of bank marketing 4522 samples

Patterns in Figure 4 were generated from 41188 samples data. As we can see the pattern in 30x30 map size well converge and we clearly can determine the cluster of data. For the example, housing data has 3 values yes, no, and unknown. The lower values indicate unknown which the neurons with darker colour, yes indicate middle values between unknown and no, which the neurons has coloured with lighter colour than unknown, and no indicates higher value which coloured with white. From the pattern we can make an analysis that there are many of bank customers not having any housing loan.

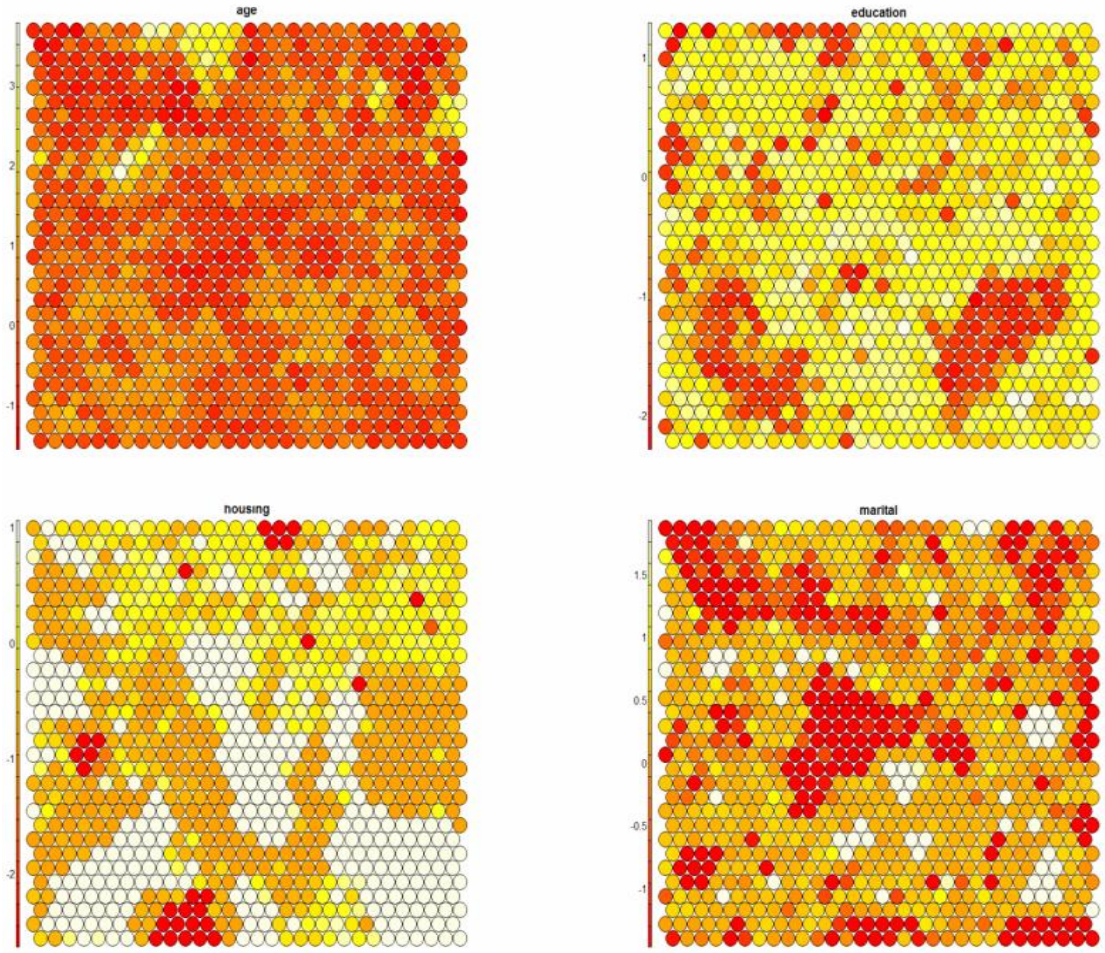

FIGURE 4. SOM 30x30 pattern of bank marketing 41188 samples.

Experimental Analysis

The experimental analysis concerns on SOM map size because the map size significantly influence the computation time, topographic error(TE), and quantization error (QE). Besides, the analysis also based on SOM map size over memory utilized when training process being held. Figure 5 below shows the SOM map sizes make different over both datasets. The used of bigger SOM map size will also lead to increase the computation time for both datasets. Obviously, the processing of 41188 samples drastically consume more time when using SOM larger map size compared to 4522 samples. The larger map size is something that cannot be avoided and must take into account when to identify the optimum SOM map size. The bigger size of map will burden the computer processing and it is the interesting issue to solve in SOM algorithm.

Besides the map size factor, the number of samples of datasets also can has a high effect to computation time as depicted in the Figure 5. By comparing the samples of 4522 and 41188, the bigger samples drastically rise when increases the map size. Another

factor may affect the computation time is dimension of dataset. The combination of number of samples and number of dimensions of datasets will make massively influence the computation time.

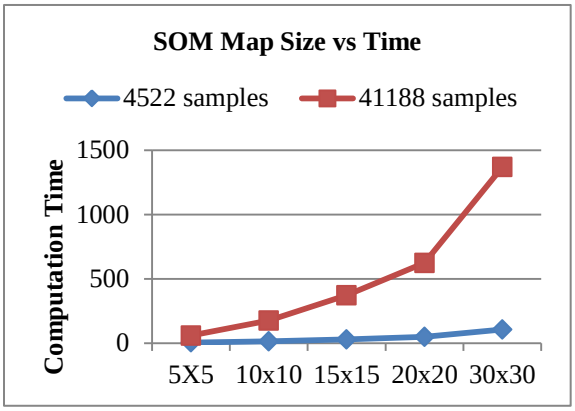

FIGURE 5. SOM map size versus computation time for both datasets.

Apart from, the memory utilization to the train SOM map may important to be discussed. Figure 6 depict the comparison of SOM map size versus memory utilization between the two samples of datasets. The memory utilization increase drastically when 41188 samples of datasets trained compared to 4522 samples of dataset. The increment of memory consumption is also caused by the size of the map as larger map used, the bigger memory required.

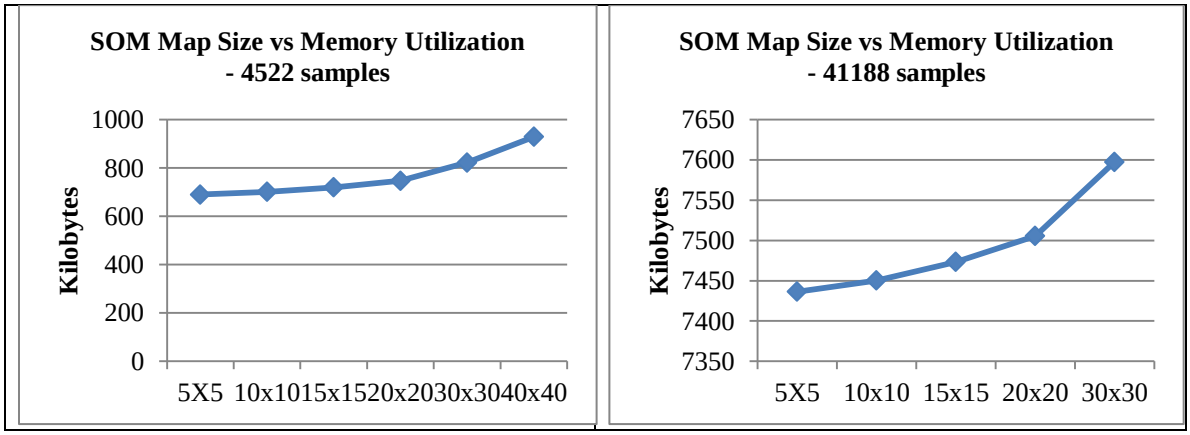

FIGURE 6. SOM map size versus memory utilization for both datasets.

CONCLUSION

Some factors need to be taken into account with the interest to improve the processing of SOM. The main factor is the quality of SOM map which indicate the result of the computation. The quality must be sustain in order to acquire a good result. Secondly the SOM map size play the important role as it influence the computation time of SOM training, the quality of map and as well as the memory utilization when training the SOM map. Besides, dataset size and number of feature dimension of dataset also could burden the processing. Table 2 shows the details of factors that could influence the SOM processing.

TABLE (2). Factors that are considered to affect the processing of SOM.

Factors	Details
Map size	As agreed by researchers, map size is important as bigger map size will increase processing workload. The memory utilization also increase when using bigger map size. The quality of SOM visualization is depending on map size as the suitable map size could deliver better visualization.
SOM quality	Both of QE and TE could impact the SOM quality. In order to achieve good in quality the both measurement should with minimum scored.
Size of datasets	Bigger samples of dataset will increase the computation time.
Number of features dimension of dataset	The increment of dataset's dimension also could increase the computation time.

In conclusion, the improvement should consider these parameters in order to resolve the issues in reducing consumption time in SOM training and to improve SOM visualization. The use of large map can lead high computation time which at the same time maintaining the visualization quality. Meanwhile, the SOM processing should be improve in order to handle large dataset that has many features.

REFERENCES

1. Kohonen, T.: Essentials of the self-organizing map. Neural Netw. 37, 52–65 (2013).
2. Sarlin, P., Yao, Z.: Clustering of the Self-Organizing Time Map. Neurocomputing. 121, 317–327 (2013).
3. Olszewski, D.: Fraud detection using self-organizing map visualizing the user profiles. Knowledge-Based Syst. 70, 324–334 (2014).
4. Astudillo, C. a., Oommen, B.J.: Self-organizing maps whose topologies can be learned with adaptive binary search trees using conditional rotations. Pattern Recognit. 47, 96–113 (2014).
5. Chattopadhyay, M., Dan, P.K., Mazumdar, S.: Application of visual clustering properties of self organizing map in machine–part cell formation. Appl. Soft Comput. 12, 600–610 (2012).

6. McConnell, S., Sturgeon, R., Henry, G., Mayne, A., Hurley, R.: Scalability of Self-organizing Maps on a GPU cluster using OpenCL and CUDA. J. Phys. Conf. Ser. 341, 012018 (2012).
7. Gajdoš, P., Platoš, J.: Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic, August, 2011. 179, (2013).
8. Hasan, S., Shamsuddin, S.M., Lopes, N.: Machine Learning Big Data Framework and Analytics for Big Data Problems. Int. J. Adv. Soft Comput. Appl. 6, 1–17 (2014).
9. Astudillo, C. a., Oommen, B.J.: Topology-oriented self-organizing maps: a survey. Pattern Anal. Appl. 17, 223–248 (2014).