Efficient Self-Organizing Map Learning Scheme Using Data Reduction Preprocessing

Yang Xu [*], Tommy W. S. Chow [†], *Senior Member, IEEE*

Abstract—**The traditional Self-Organizing Map usually considers the whole data set in one go, whereas the dominative representative data are not well utilized. The learning process is found to be rigid and time-consuming when one is dealing with large data sets. In this paper, we propose to apply density based data reduction method as preprocessing. The proposed method extracts representative data preliminarily for the SOM training, and it is found to be particularly useful in terms of reducing the overall computational time. The accuracy of the SOM map is gradually increased according to the relationship between the remaining data and the representatives. In this paper, comparative studies between our proposed method and the basic SOM are included. Simulation results on three data sets demonstrate that the newly proposed method is an efficient approach and it consistently outperforms the conventional training method.**

Keywords: Self-Organizing Map, data reduction, classification

1 Introduction

Self-Organizing Map (SOM)[1][2][3] proposed by Kohonen has been widely and successfully applied in many areas, such as pattern recognition and data mining. The key idea is to map high-dimensional data into a low-dimensional space by competitive learning and the knowledge of topological neighborhood. The output space preserves the data topology and it facilitates data analysis. Nowadays, data sets from real-world applications have become inconceivably large, especially for those biomedical data. An efficient SOM algorithm is deemed as requirement for dealing with data set of huge size. There have been several techniques used for accelerating SOM algorithm, such as modifying the feature of neurons as preprocessing[4][5], or adjusting the learning parameters[6][7]. These methods focus on redefining neurons or developing different learning rules. But all these algorithms process the whole data set all together and they are apparently rigid and time-consuming when the data size is colossal. Since the whole data set is already on hand in the beginning, the most dominating data, which contain most of input features, can be used for building an interim model. Less significant data will be gradually added for training. In such a way, the overall training time is found to be significantly reduced. In [8], it studies the influences that changing the sequence of input data may bring to the overall performance. But the data features are not clearly extracted and well used.

When facing complex problems, the natural solution is to break the problem into smaller portions which is similar to the concept of Divide-and-Conquer[9]. The objective is to solve the core problem first, and subsequently take turns to solve the remaining smaller problems. In this paper, we propose a new efficient SOM training methodology. It firstly separates large input data set into representative data points and normal ones. Consequently, we are able to use the former ones for training the SOM map to obtain an interim SOM output map. The remaining data points are subsequently used to update the neurons according to the relationship with the representative ones. It is important to point out that a reasonable SOM map can be obtained by using only a relatively small portion of input data. The whole training time is significantly reduced as a result of the data reduction preprocessing. We have conducted the performance evaluation, and our simulation results demonstrate that the proposed method is able to deliver more or less the same classification performance compared with the conventional SOM, but cost significantly less computational time.

This paper is organized as follows. Section 2 illustrates the principle of the proposed method. In Section 3, the proposed method is studied with synthetic and real-world data, and comparisons with basic SOM are provided. The conclusion is presented in the last section.

2 Efficient SOM by Data Reduction Preprocessing

Separating complex problem into small portions is an effective way to increase the training efficiency, especially when one is dealing with a massive data set. First, we break the large data set into representative ones and normal ones by using a data reduction method. It is worth noting that using a very effective data reduction, which in fact can be relatively computationally complex, is not desirable, because it incurs substantial computational burden that results in lengthening the overall computational

[*] Yang Xu, Email: yangxu3@student.cityu.edu.hk
[†] Tommy W. S. Chow, Email: eetchow@cityu.edu.hk

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

time. In this study, we need to optimize the data reduction effectiveness with its computational burden. After an interim SOM map is obtained by adopting the dominating data, we fine tune the neuron weights according to the relationship between the remaining data and the selected data. Unlike the conventional SOM updating rule, which uses the whole large inputs to train the neurons, the proposed method preliminarily updates the map with small portion of data. And this new learning rule is a combination of two learning rules: SOM and incremental learning. Here we use density based data reduction method to extract representative data points for SOM algorithm and use the remaining data for incremental learning. Denote the whole data set as D, the representative portion of D as D^+, and the remaining data points as D^{++}. D^+ is chosen as follows:

1) Determine the ratio u that $u = M^+/M$, where M and M^+ are the number of D and D^+ respectively.

2) Calculate the distances among all input data in D.

3) For each data, calculate the sum of distances between itself and its $(M^+ - 1)$ nearest neighbors.

4) Find out the smallest sum; mark this data as a representative, and its $(M^+ - 1)$ nearest neighbors are not taken into account in the following steps.

5) If all the data in D have been chosen, stop. Otherwise, go to step 3.

Let F be a SOM model trained by some data. Thereby, $F(D)$ defines a topology of data set D. I is the updated model by using the relationship between D^+ and D^{++}. If $F(D) \approx F(D^+) \cup I(D^{++})$, the goal to find a satisfactory approximate model is achieved. The efficient SOM algorithm by using density based data reduction is as follows:

Step 1 Find the representative portion of the whole data set by using the method mentioned above.

Step 2 Use the representative data set to train the map, which process is the same as the basic SOM.

Step 3 Randomly select a data v from the remaining points, find its nearest representative data p_1, p_2, p_3, p_4 in the input space. Their corresponding winning neurons w_v and $w_{p_1}, w_{p_2}, w_{p_3}, w_{p_4}$ are found by best matching unit search.

Step 4 Update the weight of w_v by

$$w_v(t+1) = w_v(t) + l(i)(w_{p_i} - w_v(t)), i \in \{1, 2, 3, 4\}, \quad (1)$$

where $l(i)$ is the learning rate reverse to the distance between data v and data p_i.

Step 5 If the maximum iteration reaches, stop. Otherwise, go to Step 3.

After completing the training process, the neurons are well trained and the projected map that preserves the data topology is created. In order to assure the proposed method does not trade computational time with classification performance, we compare the accuracy of the output SOM map between using the conventional approach and our proposed approach. It is interesting that in some of our study cases, the proposed method can even deliver more accurate performance compared with the conventional training method. In some other study cases, the proposed method can deliver comparative accuracy despite being less accurate than the conventional training method.

3 Simulation results

The performance of the proposed efficient SOM algorithm is demonstrated by three data sets. These studied data sets include a synthetic data set and two UCI benchmark data sets[10]. Their properties are shown in Table 1. The synthetic data set has clear characteristics, in which three classes of data are centralized by $[0 \ \ 0]^T$, $[-2 \ \ -2]^T$ and $[2 \ \ 2]^T$ respectively. These data sets with relatively large number of data points are used to validate whether or not the classification performance would be degraded by employing the proposed training methodology. In this study, we define the accuracy to be the proportion of correctly classified points.

The ratio of the representative data to the whole data set is set to be 0.2. Thus, 20 percents of the whole input space, selected by density based data reduction method, are used to train the neurons and get an interim SOM map. The remaining data are used subsequently (Method 1). For better comparative study, we add another simulation that 10 percents are selected by data reduction and another 10 percents are randomly chosen from the input space (Method 2). The comparative results shown in Table 2 are the average learning time and accuracy after 10 different runs. It can be seen that the classification accuracies of Method 1 are a slightly lower than those of the basic SOM, but the running time are substantially shortened by at least one half. Not surprisingly, the accuracies of Method 2 are higher than those of the Method 1 and the required computational time is even reduced. This is because using data other than representatives can avoid neglecting outlines and sparse data points. The excellent performance is more obvious for the synthetic data, because its features are quite explicit, and its representatives include almost all features. The results of the other two data sets show that this method is also effective for large data sets with high dimensions.

The comparative classification accuracy due to using dif-

Table 1: Evaluated data sets

Data Sets Name	Attributes	Classes	Training Data	Test Data
Synthetic Data	2	3	3999	2001
Wine Quality	11	2	5849	648
Handwritten Recog.	16	10	7494	3498

Table 2: Comparisons of average learning time (seconds) and accuracy of three data sets

Data Set	Synthetic Data		Wine Quality		Handwritten Recog.	
Algorithm	Time	Accuracy	Time	Accuracy	Time	Accuracy
Basic SOM	32.63	99.97%	107.34	94.49%	169.74	88.7%
Method 1	11.66	99.95%	52.66	89.4%	94.53	78.27%
Method 2	11.80	99.89%	48.63	90.2%	88.03	85.99%

ferent ratios of Method 2 is shown in Figure 1. The accuracy of synthetic data is stable irrespective of different ratio. On the contrary, the accuracies of small ratios for other two data sets are deteriorated. However, if the requirement is not strict, this result could be satisfying and the time cost is shorter. The value of ratio should balance the time cost and the accuracy simultaneously according to the required performance.

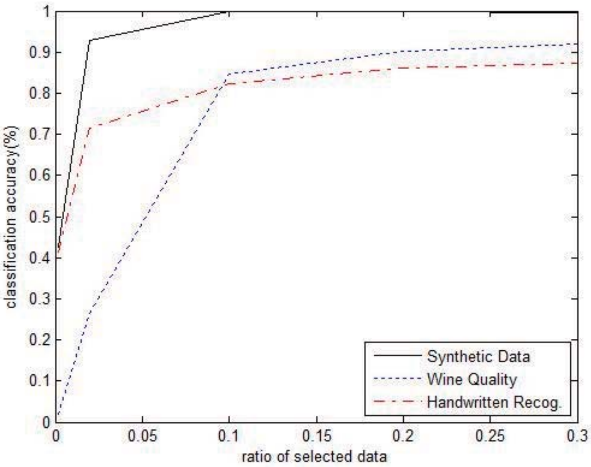

Figure 1: The influence of different ratios on classification accuracy.

In conclusion, from the simulation results, it is found that the proposed efficient SOM algorithm is much faster than basic SOM when dealing with large data sets.

4 Conclusion

In this paper, a new efficient SOM is developed. The design of the proposed method was motivated by the idea of Divide-and-Conquer that breaks a complex problem into smaller manageable data size. The proposed method considers the characteristic of data so that the most representative data points are extracted for providing an interim SOM output map. The relationship between remaining data and representative data increases the accuracy of map during incremental learning. The simulation result demonstrates that the proposed algorithm can significantly speed up the overall training but without sacrificing the classification performances.

References

[1] Kohonen, T., *Self-Organizing Maps*, Springer, Berlin, 1997.

[2] Kohonen, T., “Self-organized formation of topologically correct feature maps,” *Biological Cybernetics*, V43, pp. 59-69, 1982.

[3] Vesanto, J., Alhoniemi,E., “Clustering of the self-organizing map,” *IEEE Transactions on neural networks*, V11, pp. 586-600, 2000.

[4] Rizzo, R., Chella, A., “A Comparison between Habituation and Conscience Mechanism in Self-Organizing Maps,” *IEEE Transactions on neural networks*, V17, pp. 807-810, 2006.

[5] DeSieno, D., “Adding a conscience to competitive learning,” *In: Proc. ICNN’88, International Conference on Neural Networks*, IEEE Computer Society Press, Piscataway, NJ, pp. 117-124, 1988.

[6] Berglund, E., Sitte, J., “The parameterless self-organizing map algorithm,” *IEEE Transactions on neural networks*, V17, pp. 305-316, 2006.

[7] Haese, K., “Auto-SOM: Recursive Parameter Estimation for Guidance of Self-Organizing Feature Maps,” *Neural Computation*, V13, pp. 595-619, 2001.

[8] Miyoshi, T., Kawai, H., Masuyama, H., “Efficient SOM Learning by Data Order Adjustment,” *Proceedings of 2002 IEEE World Congress on Computational Intelligence (WCCI2002)*, USA, pp.784-784, 2002.

[9] Brassard, G., Bratley P., *Fundamentals of Algorithmics*, Prentice-Hall, 1996.

[10] Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J., Uci Repository of Machine Learning Databases, University of California, Irvine, Dept. of Information and Computer Sciences, http://archive.ics.uci.edu/ml/, 1998.