Modified Kohonen networks for complex cluster-analysis problems

Marian B. Gorzałczany and Filip Rudziński

Department of Electrical and Computer Engineering
Kielce University of Technology
Al. 1000-lecia P.P. 7, 25-314 Kielce, Poland
m.b.gorzalczany@tu.kielce.pl, f.rudzinski@tu.kielce.pl

Abstract. The paper presents a modification of the self-organizing Kohonen networks for more efficient coping with complex, multidimensional cluster-analysis problems. The essence of modification consists in allowing the neuron chain - as the learning progresses - to disconnect and later to reconnect again. First, the operation of the modified approach has been illustrated by means of synthetic data set. Then, this technique has been tested with the use of real-life, complex, multidimensional data set (*Pen-Based Recognition of Handwritten Digits Database*) available from the FTP server of the University of California at Irvine (ftp.ics.uci.edu).

1 Introduction

This paper is a continuation of work [1] included in this volume. The paper [1] presents two methods offering flexible solutions to cluster-analysis problems by providing the user with an image (in two-dimensional space) of the cluster distribution in a given, multidimensional data set. The first method employs a genetic-algorithm-based solution of the Travelling Salesman Problem, and the second method - a self-organizing Kohonen network with one-dimensional neighbourhood (the neurons are arranged in a chain). The main idea behind the first method is to find a minimal route coming through all the data points. Then, a histogram of nearness of the neighbouring points along the route is determined. This histogram reveals an image of the data structure and cluster distribution in a given data set.

In order to get a proper image of the cluster distribution, the following condition should be fulfilled by both methods: the route determined by either method must cross only once a given data group being a candidate for a final cluster. If this condition is not fulfilled (e.g., due to complexity of the cluster distribution in a given data set), both methods overestimate the number of clusters.

This paper, first, illustrates the above-mentioned issue by means of an example. Then, a modification of the self-organizing Kohonen networks is formulated. It allows the networks to generate proper image of cluster distributions even in very complex problems; examples illustrating this issue are also provided. Finally, the benefits of using the modified self-organizing networks are demonstrated with the use of a real-life, complex and multidimensional data set (*Pen-Based Recognition of Handwritten Digits Database*) [2].
2 Illustration of drawback of conventional Kohonen networks in some cluster-analysis problems

Fig. 1 shows a synthetic data set (1024 points on the plane) with three visible "parallel" clusters and the route in this set determined by 100-neuron self-organizing network (the neuron chain). Due to complexity of the cluster distribution, the route comes through some clusters more than once which results in an overestimation of the number of actual clusters. It is confirmed by the histogram of nearness (see [1] for details) for the route of Fig. 1. This histogram - shown in Fig. 2 - suggests the occurrence of four or five clusters.

3 Modification of self-organizing Kohonen networks

The main goal of modification of self-organizing networks is to allow the neuron chain of the network to fit the data structure (the cluster distribution) as good as possible. In order to achieve this goal, the following three-stage learning scenario can be employed:

Stage 1. Train the network in a conventional way.

Stage 2. Allow the network - under some conditions - to disconnect (in one or more points) its neuron chain. Continue the training of the network (to be more specific, two or more neuron sub-chains) in a conventional way.

Stage 3. Allow some of the neuron sub-chains - under some conditions - to reconnect and continue the final phase of training in a conventional way.
The first stage of the proposed scenario brings the original neuron chain (with neurons connected by the neighbourhood mechanism) to the point beyond which the network is not able to better fit the data structure. What’s even worse, if the network encounters in the data space two comparable in size areas of data points (two candidates for clusters), the neurons from the first area are attracted to the second area, and vice-versa. Allowing the neuron chain to disconnect at some point between these clusters (the second stage of the proposed scenario) is a natural solution to the problem at hand. In turn, two or more sub-chains of the original neuron chain are subject to learning in a conventional way. These sub-chains are better suited to more precisely fit the data structure. In the course of further learning, it may happen that some sub-chains approach each other (they may represent different parts of the same data cluster). Therefore, it is reasonable to allow some of these sub-chains to reconnect (the third stage). After such actions, the conventional learning of the whole system should be continued for some time.

As far as the second stage of the afore-outlined learning scenario is concerned, the following four conditions - based on experimental investigations - allowing the neuron chain to disconnect have been formulated. Possible disconnection takes place between neuron no. \(i\) and neuron no. \(i + 1\), where \(i \in \{1, 2, \ldots, r - 1\}\), where \(r\) represents the number of neurons either in an original neuron chain, or in a given sub-chain subject to further disconnection.

**Condition 1:**

\[
d_{i,i+1} > \alpha_1 \frac{\sum_{j=1}^{r-1} d_{j,j+1}}{r},
\]

where: \(d_{i,i+1}\) - distance between neurons no. \(i\) and no. \(i + 1\) (see [1] for details) and \(\alpha_1\) - experimentally selected coefficient (usually \(\alpha_1 \in [1.5, 2]\)).

The first condition prevents the excessive disconnection of the neuron chain or sub-chain by allowing to disconnect only relatively remote neurons.

**Condition 2:**

\[
\text{win}_{i}, \text{win}_{i+1} < \alpha_2 \frac{\sum_{j=1}^{r} \text{win}_j}{r},
\]

where: \(\text{win}_j\) - number of wins of \(i\)-th neuron and \(\alpha_2\) - experimentally selected coefficient (usually \(\alpha_2 \in [0.1, 0.2]\)).

The second condition prevents the disconnection of the neuron chain (sub-chain) within the group of close-to-each-other data, that is, between the neurons characterized by a relatively high level of wins.

**Condition 3:**

\[
r_{s1}, r_{s2} > \alpha_3 r,
\]

where: \(r_{s1}, r_{s2}\) - number of neurons in first (s1) and second (s2) sub-chain, respectively, of the \(r\)-element neuron chain to be disconnected (\(r_{s1} + r_{s2} = 1\)) and \(\alpha_3\) - experimentally selected coefficient (usually \(\alpha_3 \in [0.1, 0.2]\)).

The third condition prevents the disconnection of short sub-chains from a given chain. Experiments show a tendency of chain-end neurons to escape from data groups they are in (chain-end neurons have fewer neighbours and are weakly attracted by the winning neurons). Without the third condition, undesirable, short chain-end sub-chains would occur.

**Condition 4:**

\[
\frac{\sum_{j=1}^{r_{s1}} \text{win}_{s1,j}}{r_{s1}} > \alpha_4 \frac{\sum_{j=1}^{r_{s2}} \text{win}_{s2,j}}{r_{s2}},
\]

where: \(\text{win}_{s1,j}, \text{win}_{s2,j}\) - number of wins of \(j\)-th neuron from first (s1) and second (s2) sub-chain, respectively, and \(\alpha_4\) - experimentally selected coefficient (usually \(\alpha_4 \in [0.1, 0.2]\)).
The fourth condition prevents the disconnection of sub-chains of rarely winning neurons corresponding to areas of data loosely scattered outside more compact data groups.

Experimental investigations leading to formulation of criteria for the third stage of the earlier-outlined learning scenario (reconnection of some neuron sub-chains) are underway.

Fig. 3 shows the performance of the modified self-organizing network - at several steps - applied to the synthetic data set of Fig. 1. Conventional self-organizing network has not been able to correctly cope with these data, as shown in Figs. 1 and 2. The present approach - as it can be seen in Fig. 3d and Fig. 4 showing the nearness histogram for the route of Fig. 3d - provides a clear and correct image of the cluster distribution in the problem considered.

![Fig. 3. Performance of the modified self-organizing network for the synthetic data set of Fig. 1](image)

![Fig. 4. Histogram of nearness for the route of Fig. 3d](image)
4 Application to complex, multidimensional cluster-analysis problem

The technique that has been presented in previous section of the paper, now will be tested with the use of a real-life, complex, multidimensional data set (Pen-Based Recognition of Handwritten Digits Database) [2]. The number of classes (equal to 10: digits 0 through 9) and the class assignments are known here, which allows us for direct verification of the results obtained.

The database under consideration contains as many as 7494 records (handwritten digits); each record is described by 16 nominal attributes. Due to high dimensionality of the attribute space, for the graphical presentation of these data - as in [1] - a well-known Sammon’s mapping has been used. Fig. 5 presents the Sammon’s planar mapping of multidimensional attribute space of the data set considered. Fig. 6 shows the envelope of the nearness histogram for the route determined in this set by 500-neuron self-organizing Kohonen network. It is easy to see that - due to data complexity - the conventional self-organizing network simply is not able to cope with this problem. Based on Fig. 6 it is difficult to determine both the number of clusters and the cluster boundaries in a given data set.

Fig. 5. Sammon’s planar mapping of multidimensional attribute space (Pen-Based Recognition of Handwritten Digits Database)

Fig. 6. Envelope of nearness histogram for the route in attribute space of Fig. 5 determined by 500-neuron conventional self-organizing network

In turn, Fig. 7 presents a route in this data set determined by the modified self-organizing network of Section 3 of this paper, and Fig. 8 - the envelope of
the nearness histogram for the route of Fig. 7. This time a perfectly clear image of the cluster distribution, including the number of the clusters and the cluster boundaries (determined by 9 local minima on the plot of Fig. 8), is revealed. The percentage of correct decisions, equal to 85.7%, regarding the class assignments is also very high.

![Envelope of nearness histogram for the route of Fig. 7](image)

Fig. 8. Envelope of nearness histogram for the route of Fig. 7

5 Conclusions

The modification of the self-organizing Kohonen network with one-dimensional neighbourhood for efficient coping with complex, multidimensional cluster-analysis problems has been presented in this paper. The concept of modification consists in allowing - under some conditions - the neuron chain to disconnect in the course of learning. As the learning progresses, some sub-chains of neurons - under some other conditions - could be allowed to reconnect again. This enables the generalized network to fit much better the structures "encoded" in data. First, the operation of the modified self-organizing network has been illustrated by means of synthetic data set. Then, this technique has been tested with the use of real-life, complex, multidimensional data set (Pen-Based Recognition of Handwritten Digits Database) [2]. The superiority of the modified approach in comparison to conventional self-organizing network has been clearly demonstrated.

References