Clustering High-Dimensional Data
Using Growing SOM*

Junlin Zhou and Yan Fu
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, Sichuan 610054, China
{jlzhou,fuyan}@uestc.edu.cn

Abstract. The self-organizing map (SOM) is a very popular unsupervised neural-network model for analyzing of high-dimensional input data as in scientific data mining applications. However, to use the SOM, the network structure must be predetermined, this often leads constrains on potential applications. When the network is unfit to the data model, the resulting map will be of poor quality. In this paper, an intuitive and effective SOM is proposed for mapping high-dimensional data onto the two-dimensional SOM structure with a growing self-organizing map. In the training phase, an improved growing node structure is used. In the procedure of adaptive growing, the probability distribution of sample data is also a criterion to distinguish where the new nodes should to be added or deleted besides the maximal quantization error (mqe) of a unit. The improved method is demonstrated on a data set with promising results and a significantly reduced network size.

1 Introduction

The Self-Organizing Map (SOM in short), originated by Kohonen[1], is an unsupervised, competitive learning algorithm that maps high dimensional data onto a discrete network structure of lower dimensions. Since the mapping of data from a high-dimensional space to a two or three-dimensional grid makes the inter-relations among the data points perceptible, it provides a better insight into the data structure and clustering tendency. This mapping retains the relationship between input data as faithfully as possible, thus describing a topology-preserving representation of input similarities in terms of distances in the output space. It is then possible to visually identify clusters on the map. This feature capability has made the SOM an important tool in a wide range of applications such as data mining, and or more generally, pattern recognition and knowledge acquisition.

However, some difficulties in SOM utilization remained largely untouched, even though a large number of research papers on applications of the SOM were presented over the years[2]. First, the SOM uses a fixed network architecture

* This work was supported by National Science Foundation of China under Grant 10476006.

© Springer-Verlag Berlin Heidelberg 2005
in terms of number and arrangement of neural processing elements, which has to be defined prior to training. In most cases, the user does not have much knowledge of the inherent data structure. So it is difficult to predefine a proper layout of the network. This often leads to a significant limitation on potential applications\cite{3}. It is very likely that a predetermined size of the network is either too small or too big. In either case, the resulting map will be of poor quality. Thus, it certainly is worth considering neural-network models that determine the number and arrangement of units during their unsupervised training process. We refer to\cite{4} for recently proposed models that are based on the SOM, yet allow for adaptation of the network architecture during training.

Another disadvantage of the fixed grid is that though data vectors are mapped to corresponding bestmatching neurons, it is usually difficult to provide much information about the global distribution of the data by observing the raw map. In this paper, a growing SOM based on probability distribution is proposed to help overcome the constraints imposed by the traditional SOM.

2 Related Work

The principal goal of the SOM is to transform an incoming signal pattern of arbitrary dimension into a one or two-dimensional discrete map, and perform this transformation adaptively in topologically ordered fashion. There are three essential processes involved in the formation of the self-organizing map as below.

Competition. For each input vector, the neurons in the map compute their respective values of a discriminant function. The particular neuron with the largest value of discriminant function is declared winner of the competition. Let \( m \) denote the dimension of the input (data) space. Let an input pattern selected at random from the input space be denoted by \( \mathbf{x} = [x_1, x_2, ..., x_m] \). Let the synaptic weight vector of neuron \( j \) be denoted by \( \mathbf{w}_j = [w_1, w_2, ..., w_m]^T \). To find the best match of the input vector \( \mathbf{x} \) with the synaptic weight vectors \( \mathbf{w}_j \), use the best matching criterion, based on maximizing the inner product \( \mathbf{w}_j^T \cdot \mathbf{x} \), is mathematically equivalent to minimizing the Euclidean distance between vectors \( \mathbf{x} \) and \( \mathbf{w}_j \). If we use the index \( i(\mathbf{x}) \) to identify the neuron that best matches the input vector \( \mathbf{x} \), we may then determine \( i(\mathbf{x}) \) by applying the condition \( i(\mathbf{x}) = \arg\min_j ||\mathbf{x} - \mathbf{w}_j||, j = 1, 2, ..., l \). The particular neuron \( i \) that satisfies this condition is called the winning neuron for the input vector \( \mathbf{x} \).

Cooperation. The winning neuron determines the spatial location of a topological neighborhood of excited neurons, thereby providing the basis for cooperation among such neighboring neurons. Let \( h_{ij} \) denote a typical topological neighborhood centered on the winning neuron \( i \), and \( j \) denote a typical neuron of a set of excited neurons around winning neuron \( i \). Let \( d_{ij} \) denote the lateral distance between winning neuron \( i \) and excited neuron \( j \). A typical choice of \( h_{ij} \) that satisfies these two requirements is the Gaussian function

\[
h_{j,i(\mathbf{x})}(n) = e^{-\frac{d_{ij}^2}{2\sigma(n)^2}}, n = 0, 1, 2, ...
\]