A Self-Growing and Self-Organizing Batch Map with Automatic Stopping Condition

Se Won Kim and Tang Van To
Department of Computer Science, Faculty of Science & Technology
Assumption University
Bangkok 10240, Thailand
kim,tvto@scitech.au.edu

Abstract— **This paper proposes a model of self-growing and self-organizing feature map designed to alleviate the difficulty of predetermining an appropriate size and shape of the feature map suitable for the input data in the applications of the Self-Organizing Map. The proposed model progressively builds a feature map by incremental growing of the network in a way that maintains two-dimensional regular grid structure and gradual adaptation of the reference vectors by coordinated competitive learning dynamics of the Batch Map algorithm. Experimental results based on iris data set and Italian olive oil data set show that the proposed model is effective in discovering an appropriate size and shape of the network grid to manifest a suitable feature map for the input data and that the resultant feature maps are comparable to feature maps produced by the standard SOM algorithm in their quality.**

Keywords- Self organizing feature maps; unsupervised learning; neural networks; data mining

I. INTRODUCTION

The Self-Organizing Map (SOM) [1], also known as Kohonen network, is a model of unsupervised artificial neural network that is capable of capturing statistical relationships that exist in the principal components of a high-dimensional input data manifold and then map them onto a low-dimensional network structure for easier visualization of the topology of the data manifold.

A typical Kohonen network consists of a set of neural nodes arranged in a two-dimensional rectangular grid. Associated with each node is a feature vector or its reference vector of the same dimension as the input data space. The SOM's learning or training process involves for each given input data vector determining a winner node whose reference vector is most similar to the input data vector and then selectively adjusting reference vectors of the nodes that are determined to be in the neighborhood of the winner node in the grid space in order to systematically learn from the features present in each input data. The extent of learning adopted by the nodes in the neighborhood, reflected by the amount of adjustments in the reference vectors, is inversely proportional to the distance of a node from the winner node in the grid space, with the winner node learning the most. The magnitude of changes in the reference vectors is controlled by time dependent parameters *radius of the neighborhood* and *learning rate*, both of which are made to decrease monotonically over the period of the training process.

Upon repeated presentation of input data vectors over a long period of training epochs, the reference vectors will adapt to the distribution of the input data for which they have become winners forming a *quantized approximation* of the distribution of the input data. Furthermore due to the coordinated updating of the reference vectors in the neighborhood, nodes that are near one another in the grid space will develop similar feature vectors giving rise to the important property of *topology preservation* [2] meaning that data samples that are near from one another in the input space will be mapped to nodes that are also close to one another in the space of the network grid. Thus the learning dynamics of the SOM can be described as a *coordinated competitive learning* performing the combination of two concurrent tasks: topology preserving dimensionality reduction and vector quantization.

Regrettably, no formal theory has been established that can describe how the learning dynamics of the SOM works to generate topologically correct feature maps and remains an elusive goal of the theoretical analysis of the SOM; the complexity of the mathematical theory of the SOM is best summed up by the Kohonen's description of the SOM as belonging to the "ill-posed" problems in mathematics [3]. Notwithstanding the lack of success in establishing the mathematical formalism, the SOM has been widely embraced as an effective tool for visualization of high-dimensional complex systems and data mining especially for classification and clustering tasks as evidenced by myriads of applications [4,5] found in various fields of interests.

Nevertheless, practical applications of the SOM entail a time-consuming trial and error method of generating numerous feature maps of different sizes and shapes using varied learning parameters and evaluating the feature maps based on subjective criteria to determine a feature map that suits the intended purpose of the application. Experimental analyses [6] suggest that variations in learning parameters have nominal effects on the learning dynamics of the SOM as long as they are made to decrease monotonically over the training period except for possible differences in the training time. However, appropriate size and shape of the network grid to yield a good feature map depend on the statistical characteristics of the data and cannot be made into a generic parameter independent of the input data. Therefore, the requirement in the SOM algorithm to specify the

978-1-4673-4853-9/13/$31.00 ©2013 IEEE