Ensemble Methods for Boosting Visualization Models

Bruno Baruque[1], Emilio Corchado[1], Aitor Mata[2], and Juan M. Corchado[2]

[1] Department of Civil Engineering, University of Burgos, Spain
escorchado@ubu.es, bbaruque@ubu.es
[2] Department of Computer Science and Automatics, University of Salamanca, Spain
aitor@usal.es, corchado@usal.es

Abstract. Topology preserving mappings are great tools for data visualization and inspection in large datasets. This research presents a study of the combination of different ensemble training techniques with a novel summarization algorithm for ensembles of topology preserving models. The aim of these techniques is the increase of the truthfulness of the visualization of the dataset obtained by this kind of algorithms and, as an extension, the stability conditions of the former. A study and comparison of the performance of some novel and classical ensemble techniques, using well-known datasets from the UCI repository (Iris and Wine), are presented in this paper to test their suitability, in the fields of data visualization and topology preservation when combined with one of the most widespread of that kind of models such as the Self-Organizing Map.

Keywords: topology preserving mappings, boosting, bagging, unsupervised learning.

1 Introduction

From the range of tools that can be used to treat the high amounts of data that industrial and business operations processes, one of the most useful is the unsupervised leaning, in the field of artificial neural networks (ANNs). For unsupervised learning only the input and the network's internal dynamics are the two elements required. No external mechanism is used to obtain the results. The present work is centred on one of the major methods of unsupervised learning: competitive learning, where the output neurons of a neural network compete among themselves for being the one to be active.

The Self-Organising Map (SOM) [1] is probably the most widely used algorithm making use of this kind of learning. It is based on an adaptive process in which the neurons in a neural network gradually become sensitive to different input categories, or sets of samples in a specific domain of the input space. The SOM was conceived as a visualization tool to enable the representation of high-dimensional datasets on 2-dimensional maps and thereby facilitating data interpretation tasks for human experts.

The principal problem of the models based on competitive learning is, as happens with all ANNs, their instability. This means that even running the same algorithm several times with the same parameters can lead to rather different results. The present research is focused on the comparison and study of some novel and classical ensemble extension versions of the two competitive learning models based on the topology preserving concept. A novel summarization of topology preserving

J. Cabestany et al. (Eds.): IWANN 2009, Part I, LNCS 5517, pp. 165–173, 2009.

ensembles is presented and included in this comparison. This algorithm aims to obtain a more trustful representation of the datasets by combining the best features of several trained maps. The summarization algorithms are tested for the SOM model in combination with two ensemble techniques such as the Bagging [2] and the AdaBoost [3]. The purpose of this comparison is to verify if the performance of these unsupervised connectionist models can be improved by means of these ensemble meta-algorithms. AdaBoost is applied for the first time in this paper in combination with the WeVoS algorithm.

2 Self-Organizing Maps

The Self-Organizing Map (SOM) algorithm [4] is based on a type of unsupervised learning called competitive learning; an adaptive process in which the neurons in a neural network gradually become sensitive to different input categories, sets of samples in a specific domain of the input space [5]. Its aim is to produce a low dimensional representation of the training samples while preserving the topological properties of the input space.

The main feature of the SOM algorithm is that the neighbours on the lattice are also allowed to learn – i.e. to adapt their characteristics to the input - as well as the winning neuron. Thus, the neighbouring neurons gradually come to represent similar inputs, and their representations become ordered on the map lattice.

This updating of neighbourhood neurons in SOM can be expressed as:

$$w_k(t+1) = w_k(t) + \alpha(t)\eta(v,k,t)\big(x(t) - w_v(t)\big) \tag{1}$$

where, w_k is the weight vector associated with neuron *k*; $\alpha(t)$ is the learning rate of the algorithm; $\eta(v,k,t)$ is the neighbourhood function (usually, the Gaussian function or a difference of Gaussians), in which *v* represents the position of the winning neuron in the lattice, or the best matching unit (BMU); *k*, the positions of neighbouring neurons and *x* , the network input.

3 Quality Measures

Several quality measures have been proposed in literature to study the reliability of the results displayed by topology preserving models in representing the dataset that have been trained with [6, 7]. There is not a global and unified one, but rather a set of complementary ones, as each of them asses a specific characteristic of the performance of the map in different visual representation areas. The three used in this study are briefly presented in the following paragraphs.

Classification Error [8]**.** Topology preserving models can be easily adapted for classification of new samples using a semi-supervised procedure. A high value in the classification accuracy rate implies that the units of the map are reacting in a more consistent way to the classes of the samples that are presented. As a consequence, the map should represent the data distribution more precisely.