Neural Networks 24 (2011) 387–405

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Selective information enhancement learning for creating interpretable representations in competitive learning

Ryotaro Kamimura

IT Education Center, Tokai University, 1117 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan

ARTICLE INFO

Article history:
Received 2 January 2009
Received in revised form 19 October 2010
Accepted 29 December 2010

Keywords:
Selective information enhancement learning
Mutual information
Competitive learning
SOM
Free energy
Variable selection
Attention
Enhancement

ABSTRACT

In this paper, we propose a new information-theoretic method, called "selective information enhancement learning," to explicitly interpret final representations created by learning. More specifically, we aim to make class boundaries obtained by learning as overt as possible by picking up the small number of important variables. The variable selection is performed by information enhancement in which mutual information between input patterns and competitive units is measured, while focusing upon a specific input variable. When this information is larger, the importance of the variable is higher. With selected and important variables, a network is retrained by free energy minimization. With this free energy minimization, we can obtain connection weights by considering the importance of specific variables. We applied the method to an artificial data problem, the Senate problem and the voting attitude problem, all of which are easily obtained for purposes of reproduction. Experimental results for all three problems showed that clear class boundaries could be obtained with a smaller number of variables. In addition, we could observe that a smaller number of input variables tended to have the majority of information on input patterns. This tendency became more explicit when the network size was large.

1. Introduction

1.1. Interpretation and variable selection

One of the most important tasks in neural networks is that of fully interpreting internal representations created by learning (Rumelhart, Hinton, & Williams, 1986). This is not an easy task, because one of the major problems in neural networks is the so-called *black-box* problem. This means that it is difficult to explain how neural networks produce final results, because of greatly distributed representations. With this impossibility of interpretation, it has been very difficult for networks to be applied to practical problems that need an explanation of final representations (Micheli, Sperduti, & Starita, 2001; Nord & Jacobsson, 1998). Thus, many attempts have been made to interpret final representations obtained by learning (Feraud & Clerot, 2002), (Howes & Crook, 1999), (Ishikawa, 1996, 2000), (Setiono, Leow, & Zurada, 2002) and (Towell & Shavlik, 1993), to cite a few. In particular, it has been considered to be very important to examine the meaning and function of input variables for interpretation. Recently, variable selection or feature selection has received more attention, because with huge data being accumulated daily, we must cope with a great number of variables in data analysis. According to Guyon and Elisseeff (2003), several merits in variable selection can be enumerated, such as "facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times and defying the curse of dimensionality". Thus, many methods have been developed to select important variables. For example, Rakotomamonjy (2003) proposed variable selection on a relevance criterion based on the support vector machine. Perkins, Lacker, and Theiler (2003) proposed a method to incorporate feature selection in an integrated criterion optimization scheme. Though many approaches have been proposed, the majority have been based upon supervised learning. Little attention has been paid to the explanation of networks in unsupervised learning, because it has been difficult to identify evaluation functions in unsupervised learning (Guyon & Elisseeff, 2003).

1.2. Selective information enhancement learning

For the explanation of how networks work in unsupervised learning, we propose here selective information enhancement learning. In this learning model, a clear and concrete evaluation function of mutual information is defined with respect to competitive units. We have so far demonstrated that a concept of competition is realized by mutual information maximization. For example, we showed that information maximization could really realize a process of competition and could be used to extract complex rules (Kamimura, 2003a; Kamimura, Kamimura, & Shultz, 2001; Kamimura, Kamimura, & Uchida, 2001). When mutual information between input patterns and competitive units is maximized, conditional entropy between input patterns and

E-mail address: ryo@keyaki.cc.u-tokai.ac.jp.

0893-6080/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2010.12.009

competitive units is minimized. This means that a competitive unit tends to respond to a specific input pattern. In addition, the entropy of competitive units should be as large as possible in mutual information maximization, meaning that all competitive units are fired with equi-probability. This is just a basic idea of competitive learning, presented by Rumelhart and Zipser (1985). As mutual information between input patterns and competitive units is increased, the degree of competition is increased. When mutual information is completely maximized, a pure state of winner-takes-all is realized.

This property of mutual information relative to competitive learning can be used to measure the importance of variables. For example, we have so far introduced a concept of information loss (Kamimura, 2007, 2008a, 2008b) by using the mutual information. The information loss is defined by the difference between mutual information with a full network and its counterpart without a specific element. The methods of information loss have shown that mutual information in competitive learning is really used to detect the importance of specific elements in a network. Profiting from the effectiveness of the information loss, information enhancement (Kamimura, 2008c) is developed to detect the importance of some elements in a network, and it is realized by paying focused attention to specific elements. Suppose that we focus upon an input variable, and then mutual information is significantly increased. This is a case where the input variable has much information or much enhanced information on realizing competitive processes.

The selective information enhancement learning proposed here uses this information procedure to produce internal representations by considering the importance of input variables. With the use of this information enhancement, several important input variables are initially selected. Then, using detected important input variables, a network is trained to modify connection weights that are expected to include more information on important input variables. First, with the most important input variable enhanced, a network is trained; then, with the two most important input variables enhanced, the network is trained, and so on. In this way, we can successively obtain connection weights, taking into account the effect of several important input variables.

Because the combination of information enhancement and selective information enhancement learning has been based upon competitive learning, it is most suited for demonstrating its performance when applied to competitive learning. However, instead of competitive learning, we use the conventional SOM for learning, because it is easy to demonstrate the change caused by reducing the number of variables. The self-organizing maps by Kohonen (1988, 1995) have been used for clustering, feature detection and, particularly, for visualizing complex data. To obtain clearer maps, many visualization techniques have been developed (Kaski, Nikkila, & Kohonen, 1998; Mao & Jain, 1995; Ultsch & Siemon, 1990; Vesanto, 1999). However, even with these sophisticated visualization techniques, it has been difficult to extract important features from the ambiguous representations generated by the conventional methods. Our method of selective information enhancement can significantly modify connection weights so as to take into account the effect of important variables and present clearer class boundaries. Thus, the visualization performance of SOM can be improved, and explicit class boundaries can be generated.

1.3. Outline of the paper

In Section 2, we first explain the concept of selective information enhancement learning and introduce how to compute update rules for the conventional SOM. Then, we compute mutual information by focusing upon a specific input unit and compute

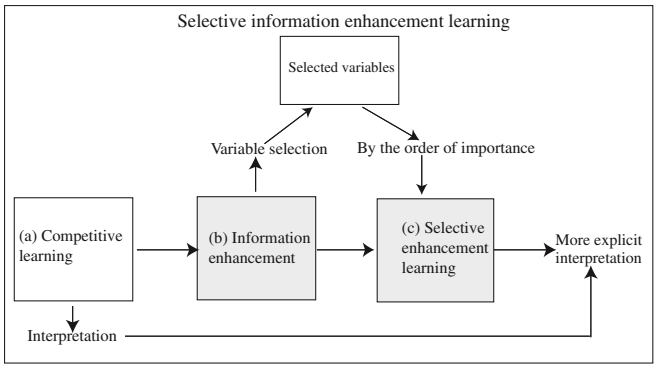

Fig. 1. A schematic diagram of selective information enhancement learning.

enhanced information. This enhanced information is used to evaluate the importance of input variables. With selected and important input variables, a network is retrained to minimize the free energy. The free energy is introduced to simplify the computation of mutual information between competitive units and input patterns. Then, we introduce relative information to show how much information is stored in a variable with respect to total information content. In Section 3, we present the results of three experiments: classification of artificial data, the Senate problem and the voting attitude problem. In the experiments, we demonstrate that the importance of input variables is easily interpreted and that final U-matrices show clearer boundaries than those obtained by the conventional SOM when the number of input variables is small. Then, we show that a large quantity of information is stored in a small number of important input variables. This means that, for these well-known data, a few variables only are enough to produce explicit feature maps. In addition, we try to show that the explicit class structure can be obtained even when the network size is larger, while the conventional SOM fails to produce an explicit one. The extraction of the explicit class structure means that we can extract major and minor class boundaries and the relations between them. Finally, we try to show in the voting attitude data that input patterns are classified into two groups with inverse responses to input patterns and a peripheral group responding indifferently to input patterns.

2. Theory and computational methods

2.1. Concept of selective information enhancement learning

Selective information enhancement learning is a method to train connection weights with a limited number of input variables. Fig. 1 shows a schematic diagram of selective information enhancement learning. The selective information enhancement learning is composed of all procedures in the diagram. In the first place, we use competitive learning or SOM to obtain connection weights, as shown in Fig. 1(a). If simple competitive learning is used, this component is built into the learning procedure. However, if we need lateral interactions between neurons, we need methods outside the learning procedure, such as the conventional SOM. To interpret connection weights more explicitly, we select important variables by information enhancement, as shown in Fig. 1(b). The information enhancement procedure is used to select important input variables by the enhancement of competitive units, focusing upon specific input variables. Then, with the selected variables, we retrain networks so as to reflect the importance of the chosen input variables. This procedure is the selective enhancement learning shown in Fig. 1(c). Learning consists in minimizing the cross entropy between ordinary firing probabilities of competitive units and those of competitive units