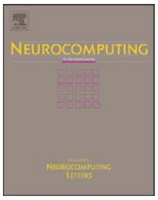
Neurocomputing 72 (2009) 2693–2704

Contents lists available at ScienceDirect

Neurocomputing

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



An information-theoretic approach to feature extraction in competitive learning

Ryotaro Kamimura *

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

ARTICLE INFO

Article history:
Received 21 February 2008
Received in revised form
21 September 2008
Accepted 29 September 2008
Communicated by D. Xu
Available online 1 November 2008

Keywords:
Mutual information
Competitive learning
Information loss
Conditional information

ABSTRACT

In this paper, we propose a new information-theoretic approach to competitive learning and self-organizing maps. We use several information-theoretic measures, such as conditional information and information losses, to extract main features in input patterns. For each competitive unit, conditional information content is used to show how much information on input patterns is contained. In addition, for detecting the importance of each variable, information losses are introduced. The information loss is defined as the difference between information with all input units and information without an input unit. We applied the information loss to conventional competitive learning to show that distinctive features could be extracted by the information loss. Then, we analyzed the self-organizing maps by the conditional information and the information loss. Experimental results showed that main features in input patterns were more clearly detected.

1. Introduction

Several information-theoretic principles, such as information maximization and minimization, have been proposed to describe neural information processing. Linsker sated in his infomax principle [29–31] that connection weights were modified so as to maximize mutual information between input and output layers. On the other hand, Barlow's minimum redundancy coding or principle [5,6] was proposed to produce factorial codes for efficient representations. These information-theoretic principles have had much influence on many attempts to apply information-theoretic methods to neural networks. For example, Deco and Obradovic [10] tried to reduce statistically correlated combination of inputs, and he also tried to apply this minimum redundancy principle to feature extraction with stochastic neural networks [11]. Atick and Redlich [3] showed that spatial coherent features were detected by mutual information. Becker and Hinton [8,7] also used mutual information maximization for coherent features. Levia-Murillo and Atres-Rodriguez [28] introduced mutual information maximization in linear feature extraction. Jeon et al. [19] proposed an information-theoretic learning using unlabeled data for improved performance in supervised classification. Deco et al. [9] also minimized mutual information to eliminate overfitting for improved generalization performance.

Of particular concern to this paper, we found a similarity between competitive learning and mutual information. Mutual information between input patterns and competitive units can be used to realize competitive processes [20–24,27]. In addition, the method can solve the fundamental problems of competitive learning, such as a dead neuron problem [12,2,39,32,18]. Because mutual information in neural networks is a good indicator to show a degree of organization in neural networks, it can surely be used to extract some features in input patterns by maximizing mutual information. On the other hand, by minimizing mutual information, statistically independent features can be constructed. Thus, these recent research developments seem to suggest that information-theoretic methods can be applied to many feature detection problems.

Though information-theoretic methods seem to be promising for improved performance in neural network, at least, two problems for the use of mutual information in neural networks have been solved for real applications of the methods, that is, computational complexity and inability to extract detailed features in input patterns. In the first place, we can say that one of the main problems of mutual information is that it usually needs heavy computation and its computational complexity become serious for complex and practical problems [15]. Recent developments in the use of mutual information have made it possible to use less expensive and more efficient computational methods for mutual information. For example, mutual information was restricted to class labels and discrete labels in the inexpensive feature selection of maximum output information

* Tel.: +81 463 58 1211.
E-mail address: ryo@cc.u-tokai.ac.jp

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2008.09.013

[35]. A new information discriminant analysis was proposed that was treatable analytically and computationally feasible [34]. Quadratic divergence measures such as information potentials and information forces instead of the conventional Kullback–Leibler divergence have been introduced for efficient computation [36]. Morejon and Principe [33] proposed an efficient computational procedure with Kernel estimators in information-theoretic learning. In addition, we have found that the computation of entropy is greatly simplified by introducing free energy. Instead of direct computation, we have only to compute the partition functions, which are much simpler than entropy functions. In addition, this free energy approach is closely related to mixture models [17,4,37,38,26,25].

The second problem of mutual information is directly related to this paper; that is, mutual information is ineffective in detecting detailed features. Mutual information represents a degree of organization in a system. When mutual information is larger, the system is considered to be more organized. Thus, mutual information between input patterns and competitive units could be used to extract important features in neural networks. However, mutual information is the average of information over all competitive units. We cannot see how much information is contained in each competitive unit. Thus, we introduce conditional mutual information to examine information contained in each competitive unit. Though conditional information becomes negative, it can be an indicator of the organization of each competitive unit, with due attention to the negative property. A concept of conditional mutual information is not new, and Abramson [1] used the conditional mutual information in a discussion of channel capacity. Gurney [14] proposed a new concept of information spectrum, which is based upon condition information in describing information processing in dendrites.

Now, for detecting the importance of each variable, we introduce information losses. As mentioned, mutual information is an indicator of organization. As information is increased, more organized activations can be seen in competitive units. Then, it is important to see the change in organization when an element is deleted. If this deletion considerably influences the degree of organization, the element is very important to maintain organization. Thus, the information loss, or the loss of organization, is used to describe the importance of elements. Information losses are defined by the difference between information content with a full network and that without an element. As the importance of the element becomes higher, the information loss becomes higher. When the element is considered to be a feature, we can estimate to what extent this feature contributes to organization. When we apply the information loss to competitive learning, we can expect the detection of some important features that cannot be detected only by examining connection weights.

In Section 2, we briefly present the concept of mutual information and two activation functions of the inverse of Euclidean distances and Gaussian functions. By applying them to actual networks, we present how to compute mutual information. Then, we introduce conditional mutual information by skipping some average operations in mutual information so as to obtain some information depending upon specific elements. Thus, conditional information can be used to extract information specific to the elements. Then, we introduce a concept of information loss where mutual information is computed by deleting some elements in a network. When the information loss is larger, the corresponding elements have more information on the elements and surely play more important roles. In Section 3, we apply the method to competitive learning and present experimental results in which we try to show that distinctive features can be extracted by information losses. In Section 4, we apply the information measures to self-organizing maps to shows that more explicit boundaries and features can be seen. In addition, a small number of important features can be detected in all the problems.

2. Theory and computation methods

2.1. Conditional and mutual information

We consider information content stored in competitive unit activation patterns. For this purpose, let us define information to be stored in a neural system. Information stored in a system is represented by decrease in uncertainty [13]. Uncertainty decrease, that is, information I, is defined by

$$I = \sum_{\forall s} \sum_{\forall j} p(s)p(j|s) \log \frac{p(j|s)}{p(j)}, \tag{1}$$

where $p(j)$, $p(s)$ and $p(j|s)$ denote the probability of firing of the jth unit, the probability of the sth input pattern and the conditional probability of the jth unit, given the sth input pattern, respectively. In addition, we can define conditional mutual information. An example of conditional mutual information is defined by

$$I_j = \sum_{\forall s} p(s)p(j|s) \log \frac{p(j|s)}{p(j)}. \tag{2}$$

It should be noted that conditional mutual information can be negative.

Let us define information in neural networks. As shown in Fig. 1, a network is composed of input units x_k^s and competitive units v_j^s. An output from the jth competitive unit can be computed by

$$v_j^s = \frac{1}{\sum_{k=1}^{L} (x_k^s - w_{jk})^2}, \tag{3}$$

where L is the number of input units and w_{jk} denotes connections from the kth input unit to the jth competitive unit. The output is increased as connection weights come closer to input patterns. In addition, we can compute the activity by using the Gaussian function

$$v_j^s = \exp\left(- \frac{\sum_{k=1}^{L} (x_k^s - w_{jk})^2}{2\sigma^2} \right), \tag{4}$$

where σ denotes the Gaussian width. We use the inverse of Euclidean distance and Gaussian functions for the activation functions, depending upon objectives. When connection weights should be as faithful as possible to input patterns, the inverse function is appropriate. On the other hand, when we must flexibly control information, Gaussian functions are needed, because

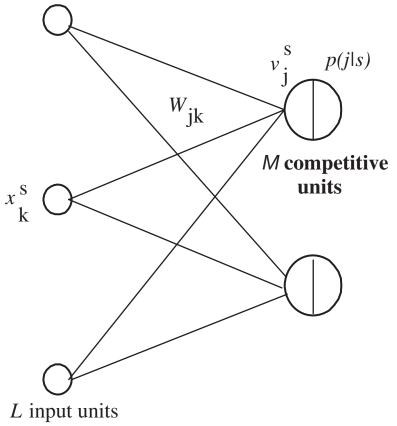


Fig. 1. A network architecture for competitive learning.