Neurocomputing 74 (2011) 1116–1134

Contents lists available at ScienceDirect

Neurocomputing

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

Supposed maximum information for comprehensible representations in SOM

Ryotaro Kamimura *

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

ARTICLE INFO

Article history:
Received 1 July 2010
Received in revised form
15 November 2010
Accepted 1 December 2010
Communicated by B. Hammer
Available online 22 December 2010

Keywords:
Mutual information
Competitive earning
SOM
Comprehensibility

ABSTRACT

In this paper, we propose a new information-theoretic method to simplify the computation of information and to unify several methods in one framework. The new method is called "supposed maximum information," used to produce humanly comprehensible representations in competitive learning by taking into account the importance of input units. In the new learning method, by supposing the maximum information of input units, the actual information of input units is estimated. Then, the competitive network is trained with the estimated information in input units. The method is applied not to pure competitive learning, but to self-organizing maps, because it is easy to demonstrate visually how well the new method can produce more interpretable representations. We applied the method to three well-known sets of data, namely, the Kohonen animal data, the SPECT heart data and the voting data from the machine learning database. With these data, we succeeded in producing more explicit class boundaries on the U-matrices than did the conventional SOM. In addition, for all the data, quantization and topographic errors produced by our method were lower than those by the conventional SOM.

1. Introduction

We propose a new type of information-theoretic method to produce humanly interpretable representations by maximizing information in components such as input and competitive units in a network. The main characteristics of the method are represented in terms of simplification and unification. In the simplification part, we aim to simplify the procedures of information maximization by supposing that information in components is already maximized. This is called "supposed information," because a maximum information state is not actually attained. In the unification part, we consider several learning methods in the same framework of supposed maximum information. For example, different approaches such as competitive learning and feature selection can be unified in the name of supposed maximum information.

1.1. Simplification

First, a method is proposed to simplify the computation of information maximization. Supposing that information is already maximized before learning, we try to examine what kinds of configuration changes can be observed. The information-theoretic approach has been widely used in neural network because of its possibility to deal with higher order statistics and non-linear transforms [1] and to provide neural networks with optimizing principles. One of the most well-known organizing principles ever proposed is Linsker's information maximization principle [2,3], stating that "the network connections develop in such a way as to maximize the amount of information that is preserved when signals are transformed at each processing stage." Barlow [4,5] proposed minimum entropy coding to generate factorial codes to reduce redundancy among features. These principled approaches have engendered a number of information-theoretic variants [6–10], and the possibility of dealing with complex data and the existence of organizing principles for learning seem to be promising in any aspect of neural networks. However, one of the major shortcomings of the information-theoretic methods lies in their computational complexity. To overcome this shortcoming, a large number of computational methods have already been proposed in many aspects of learning. For example, Linsker has already proposed local learning rules for his information maximization principle [11,12]. Battiti [13] applied mutual information to the feature selection by MIFS (mutual information based feature selection), and a "computationally impossible" computation was substituted for a simple and feasible one. Principe [14,15] used the Renyi entropy and Parzen density estimation for efficient computation and implementation. Morejon and Principe [16] introduced advanced parameter search algorithms in information-theoretic learning. Torkkola [17] simplified the computation of information with the quadratic divergence measure for efficient non-parametric implementation. Kamimura [18,19] borrowed the free energy-like functions from statistical mechanics to simplify the computation of mutual information. These are a few examples. Though much effort has been made to solve the problem, we can

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

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.12.002

say that this difficulty in computing information still remains, which has prevented us from applying information-theoretic methods to many practical problems. In this context, we aim to make the procedure of information maximization as simple as possible by supposing thatinformation is already maximized.

1.2. Unification

After simplification, we can treat several computational methods in a more unified way, because some methods seem to use the supposed maximum information, though implicitly. We take three examples of unification, namely, competitive learning, feature selection and information-theoretic methods. Let us begin with competitive learning [20], where one of the main jobs is to determine a winner by computing distances between input patterns and connection weights. Then, connection weights to the winner with the minimum distance are updated in competitive learning, or the winner with its neighbors must be updated in self-organizing maps [21]. This winner-take-all algorithm is considered to be one realization of supposed maximum information in the context of this paper, because in one of the possible maximum information states, only one competitive unit fires, while all the other units cease to do so.

In addition to the winner-take-all as a realization of supposed maximum information, we can consider variable selection as another realization of supposed maximum information. Note that the variable selection in the present paper corresponds to the input unit selection. As information in input units becomes larger, the number of important input units becomes smaller. Finally, when the information is maximized, only one important input unit remains. Thus, input unit (variable) selection can be considered in terms of supposed maximum information. Variable selection has received much attention recently, because the number of input variables to be processed has become larger and larger [22]. Variable selection aims to "improve the prediction performance of the predictors, to provide faster and more cost-effective predictors and to provide a better understanding of the underlying process that generated the data" [22]. Thus, many methods have been developed to select important variables. For example, Sung [23] compared three methods for ranking input importance, namely, sensitivity analysis, fuzzy curves and the change of the mean square error. Steppe and Bauer [24] classified saliency measures into a derivative-based and weight-based one. Belue and Bauer [25] proposed a confidence interval around the average for identifying less important features. Egmont-Pertersen et al. [26] showed a mathematical framework in which several measures for input units were given. All these approaches fundamentally aim to reduce the number of input units as much as possible, and they can be considered as realizations of supposed maximum information.

Third, the supposed maximum information can be used to unify the information-theoretic methods for feature selection. As explained above, feature selection plays an important role in learning. However, the majority of those methods have been focused mainly upon supervised learning, because it is easy to find evaluation functions in that learning [22]. In unsupervised learning, explicit evaluation functions have not been established for variable selection [22]. We have introduced variable selection in unsupervised competitive learning by introducing a method of information loss [27–29] or information enhancement [30,31]. In the information loss method, a specific input unit or variable is temporarily deleted, and the change in mutual information between competitive units and input patterns is measured. If the difference between mutual information with and without the input unit is increased, the target input unit certainly plays a very important role. On the other hand, in information enhancement, a specific input unit is used to enhance competitive units or to increase the selectivity of competitive units. If the selectivity measured by mutual information between competitive units and input patterns is large, the target input unit is important in increasing the selectivity. Though the two methods seem to be different from each other, they can be unified in a framework of the supposed maximum information, because in both methods, one component in a network is supposed to play a major role at the initial stage of learning, which corresponds to the supposed maximum information.

In addition, one of the major difficulties of these information-theoretic methods is that of determining how much information should be acquired for measuring the information loss or information enhancement. There is no way to determine the amount of information to be acquired. However, when we can see the information enhancement [30,31] and the information loss [27–29] in a framework of supposed maximum information, we can solve the problem of the amount of information to be acquired. Namely, all we have to do is to increase information on input units as much as possible.

1.3. Outline of the paper

In Section 2, before explaining the details of the new method, we show what the meaning of maximum and minimum information states are in this paper. Then, we present a general framework of the proposed model, in which the two steps in learning are explained conceptually and technically. These two steps of learning follow completely the same computation procedures. Information on input units is supposed to be maximized, and an optimal firing probability is determined, namely, maximum information learning. We use conventional self-organizing maps (SOM) instead of pure competitive learning, because the SOM make it easy to visually demonstrate the better performance of our method. The optimal firing probabilities of input units are estimated by using mutual information between competitive units and input patterns, because we already know that mutual information maximization between competitive units and input patterns corresponds to competitive processes in competitive learning. In Section 3, we apply the method to three well-known data, namely, Kohonen's animal data, the SPECT heart data and the voting attitude data from the machine learning database. With all these data, we try to show that information on input units is increased and reaches its steady state as the spread parameter is increased. By retraining competitive networks by taking into account the importance of the input units by their firing probabilities, clearer class boundaries can be generated on the U-matrices.

2. Theory and computational methods

2.1. A general framework for learning

In this paper, we try to interpret network configurations clearly, and the interpretability is measured by using the information content of input units as well as competitive units. Though we try to maximize information in competitive units as well as input units, our focus is on the information in input units. This is because mutual information between competitive units and input patterns, shown in Fig. 1(b2), is realized in conventional competitive learning [32]. The unification of the two approaches will be further explored and discussed in Section 4.2. One of the simplest ways to interpret input units is to reduce the number of input units as much as possible so as to be able to focus upon a small number of important input units for interpretation. Thus, intuitively, a state with a small number of input units is one with high information