Stochastic Neural Network Classifiers

A

Eitan Gross
University of Arkansas, USA

INTRODUCTION

Traditional back propagation algorithms have been employed to train neural networks to compute the difference between the network's prediction and the true value of the input signal. This error function is then used to adjust the values of the synaptic weights during training, *via* an optimization algorithm that searches for a global minimum in the error function's parameter space. This approach ignores however the stochastic nature of the neuronal code and is therefore not suitable for classification tasks involving uncertainty in the input variables. Such uncertainties may arise for instance during mapping of text to phonemes, or in using diagnostic test results for prognosis of critically ill patients. In this article we address this issue and discuss the use of the Kullback-Leibler distance (relative entropy) between the probability distributions of the input and output vectors as a cost function to be optimized during training of supervised multilayer perceptron (MLP) classifiers. Using data from the Commission on Cancer's breast and colorectal carcinoma Patient Care Evaluation and the National Cancer Institute's Surveillance, Epidemiology and End Results breast carcinoma data sets, we compared predictions for 5-year and 10-year survival made by our MLP to that of the TNM system. In all categories studied, our MLP outperformed the TNM system. The improvement in prognostic ability offered by artificial neural networks over regression methods, has clinical importance for therapy, clinical trials, patient information, and quality assurance.

BACKGROUND

Neural networks with back propagation algorithms have been used as classifiers in numerous applications including face (Lawrence, Giles, Tsoi, & Back, 1997) and pattern (Kwak et al., 2002) recognition, analysis of electroencephalograms (Khorasani & Weng, 1996) for medical diagnostics, analysis of the finite impulse response in optoelectronic processors (Silveira, Pati, & Wagner, 2002) and financial forcasting (Kaastra & Boyd, 1996; Wun, Hua, Jen, Ying, & Soushan, 2006).

The back propagation algorithm computes the diference between the network's output (prediction) and a target (or true) value of the input signal of a training set (Rumelhart, Hinton, & Williams, 1986). This error function is then used during training to adjust the values of the synaptic weights via the steepest-descend or other optimization algorithms that search for a global minimum in the error function's parameter space. In many applications however, such as mapping text to phonemes (Sejnowski & Rosenberg, 1987), or initial symptom-based diagnosis of illness (Yan, Jiang, Zheng, Peng, & Li, 2006) where a patient in a doctor's office is asked to grade the severity of his symptoms on a scale of 1 to 10, a "probabilistic" learning algorithm is more approperiate in which both the input and output vectors are expressed in probability terms. Moreover, the deterministic learning algoithm ignores the stochastic nature of the nuronal response (Faisal, Selen, & Wolpert, 2008) *in vivo*. For instance, it has been suggested that neuronal variability provides a ''probabilistic population code''(Ma, Beck, Latham, & Pouget, 2006), which allows the brain to represent probability distributions, and perform Bayesian inference (Knill & Pouget, 2004). The stochastic nature of neuronal activity suggests that a given stimulus is coded by a *distribution* of values, representing either the spike frequency (rate code) or interspike interval (temporal code) (Huxter, Burgess, & O'Keefe, 2003).

MAIN FOCUS OF THE ARTICLE

For an ideal information processing system, the information-theoretic distance between the output response d_Y to two stimuli, α_0, α_1, must be smaller then the cor-

DOI: 10.4018/978-1-4666-5888-2.ch026

responding distance of the input response d_X to these stimuli, i.e.: $d_Y(\alpha_o, \alpha_1) \leq d_X(\alpha_o, \alpha_1)$. The choice of which distance measure to use is not always trivial and will normally depend on the type of stimulus (continuos or discrete) and the manner in which the distance chosen scales with the size (number of neurons in our case) of the information-processing system.

SOLUTIONS AND RECOMMENDATIONS

To address the stochastic nature of the neuronal response we propose, in this article, the Kullback-Leibler distance (Kullback & Leibler, 1951) between the probability distributions of the input and output vectors as the distance of choice. Here, we will use the Kullback-Leibler distance as a cost function to be optimized during training of a supervised stochastic multi-layer perceptron classifier. The Kullback-Liebler distance has been used in information-theoretic studies as well as in various combinatorial optimization problems including the travelling sales person (Wu & Hsu, 2011), the knapsack problem (Caserta, Quinonez Rico, & Marquez Uribe, 2008) and the max-cut problem (Klein & Lu, 1996).

As we will show below by treating the Kullback-Leibler distance (D) between a probability *p* and its approximate solution *q* as a cost function, a learning rule (Δw) can be derived which maximizes the likelihood that a proposition *k* is true, given a stimulus α at the input. Givan a database consisting a number of medical cases, for instance, our learning rule can potentially be used to help diagnose new patients based on a given set of symptoms they have.

D is an Ali-Sivey class distance measure (Ali & Sivley, 1966). This class of distances have the general form: $d_X(\alpha_o, \alpha_1) = f(\varepsilon_0[c(\Lambda(X)])$, where $\Lambda(\cdot)$ is the likelihood ratio: $p_X(\cdot; \alpha_1)/p_X(\cdot; \alpha_0)$, and $c(\cdot)$ is a convex function of its argument. $\varepsilon_0[\cdot]$ is the expected value with respect to the probability function specified by the parameter α_0 and $f(\cdot)$ is a non-decreasing function. D is formally given by (Kullback & Leibler, 1951):

$$D_X(\alpha_1 \,||\, \alpha_0) = \int p_X(x;\alpha_1) log \frac{\int p_X(x;\alpha_1)}{\int p_X(x;\alpha_0)}$$

For which $c(x)=x\ logx$, $f(x)=x$ and $f(c(1))=0$. It should be noticed that while the Kullback-Leibler distance is not symmetrical it is an additive quantity even when the random variables are not identically distributed.

MODEL

In our proposed back propagation model (Figure 1), minimization of D between *p* and its approximate solution *q* is carried out iteratively using a multi-layer perceptron (MLP) with one input layer, one hidden layer and one output layer (Callan, 1999). The induced local field of neuron *j* in the hidden layer is given by:

$$v_{j|\,\alpha} = \sum\nolimits_i w_{ji} x_{i|\,\alpha} \tag{1}$$

where w_{ji} is the synaptic weight of hidden neuron *j* connected to source node *i* in the input layer and xi|α is the i^{th} component of the input vector **X**, given stimulus α. A "stimulus" here may represent for instance an object or a pattern to be recognized by the network, or a patient with a set of symptoms to be diagnosed. Accordingly, the input vector **X** may represent the feature space for the object or pattern recognition classifier, or in medical diagnostic application it may represent the patient's set of symptoms or diagnostic results. The output of hidden neuron *j* for stimulus α is then given by:

$$yj|\alpha = \varphi vj|\alpha \tag{2}$$

where $\varphi(\cdot)$ is the logistic function, given by:

$$\varphi\left(v\right) = \frac{1}{1+e^{-v}} \tag{3}$$

Let the induced local field of output neuron *k*, representing a component in the stimulus' feature space, be given by:

$$v_k = \sum_j w_{kj} y_j \tag{4}$$