A Modular Classification Scheme with Elastic Net Models for Handwritten Digit Recognition

Bai-ling Zhang[1], Min-yue Fu[1] and Hong Yan [2]
'Department of Electrical and Computer Engineering
University of Newcastle, NSW 2308, Australia
[2]Department of Electrical Engineering
University of Sydney, NSW, 2006, Australia

Abstract

This paper describes a modular *classification system for* handwritten *digit* recognition based on *the elastic net model. We use ten* separate elastic *nets to* capture different features in *the ten classes of* handwritten *digits* and represent an *input* sample from *the activations in each net by population decoding. Compared with traditional neural networks based* discriminant *classifiers,* our *scheme features fast training and high recognition* accuracy.

1 Introduction

Self-organizing map (SOM) [1] has been successfully applied in many areas. However, SOM has no cost function and corresponding proof of convergence. In recent years, some alternative paradigms for introducing topology-preserving maps have attracted attention. One example is the elastic net model [2], which is based on a global optimization criterion. Elastic net was suggested for cortical maps via a dimensional reduction framework, i.e., a high dimensional parameter space is mapped smoothly onto a two dimensional cortical surface. Motivated by the elastic net, a generalized deformable model was proposed in [3].

In this paper, we apply the elastic net to handwritten digit recognition. Ten nets are trained parallelly on examples of each digit class and an input digit is recognized by determining which net gives the best reconstruction. The key point is a straightforward decoding of each topographic map by treating it as a kind of population code for the input.

2 Elastic Net for Self-organization

Elastic net algorithm originates from a global optimization criterion, which includes two parts. The first part can be considered as a free energy. Given a data set of N samples, the best representation of them by M prototypes can be proven minimizing the following criterion at a parameter T

$$\begin{aligned} F &= -\frac{1}{\beta}\log Z \\ &= -\frac{1}{\beta}\sum_{x}\log\Big(\sum_{m=1}^{M}\exp(-\beta\|\mathbf{x}-\mathbf{w}(\mathrm{m})\|^2)\Big) \end{aligned} \quad (1)$$

where $\beta = \frac{1}{2T^2}$ is called "inverse temperature". Objective (1) was also proposed in [5] as a maximum-entropy clustering. The second part of the elastic net optimization criterion is a smoothness constraint for the mapping from input to representation. For a 2-dimensional topological map, denote w(m) by its location (k,l) in a c_1 x c_2 rectangular lattice as $\mathbf{w}_{k,l}$, m = $(k-1)c_1 + 1$, then a smoothness measure could be expressed as

$$G = \sum_{k=1}^{c_1-1}\sum_{l=1}^{c_2-1}(\|\mathbf{w}_{k+1,l}-\mathbf{w}_{k,l}\|^2 + \|\mathbf{w}_{k,l+1}-\mathbf{w}_{k,l}\|^2) \quad (2)$$

Boundary conditions could be added as $\mathbf{w}_{c_1,l} = \mathbf{w}_{1,l}$ and $\mathbf{w}_{k,c_2} = \mathbf{w}_{k,1}$. A topology preserving map can be approximately set up based on the constrained optimization criterion

$$\begin{aligned} J &= F + \nu G \\ &= -\frac{1}{\beta}\sum_{x} log\Big(\sum_{m=1}^{M} exp(-\beta\|\mathbf{x}-\mathbf{w}(\mathrm{m})\|^2) + \nu \mathrm{G}\end{aligned} \quad (3)$$

where ν is a trade-off parameter for the smoothness constraint. From objective (3), we can directly apply

the EM algorithm or perform gradient descent algorithm to find w. The weight updating rule by applying gradient descent of J is

$$\begin{aligned}\Delta \mathbf{w}_{k,l} &= \mu p_m(\mathbf{x} - \mathrm{w(m)}) \\ &+ \nu(4\mathbf{w}_{k,l} - \mathbf{w}_{k,l-1} - \mathbf{w}_{k,l+1} - \mathbf{w}_{k-1,l} - \mathbf{w}_{k+1,l}) \\ k &= 1, \cdots, c_1,\ l = 1, \cdots, c_2,\ \ m = (k-1)c_1 + l \end{aligned} \quad (4)$$

where

$$p_m = \frac{\exp(-\beta \|\mathbf{x} - \mathbf{w}(\mathrm{m})\|^2)}{\sum_{k=1}^{M} \exp(-\beta \|\mathbf{x} - \mathbf{w}(\mathrm{k})\|^2)} \quad (5)$$

is a probability for assigning a data point x to reference vector w(m). In algorithm (4), we have ommitted time scale for brevity and alternatively denote w(m) and $\mathbf{w}_{k,l}$.

In the above learning algorithm, a deterministic annealing [5] can be applied, which starts from a small β and all the units equally share their responsibilities in representing an input. Progressively, as β increases, responsibilities are gradually localized. At high β, the function achieves global minimum.

3 Handwritten Digit Recognition

The common practice of using a neural network for digit recognition is to train the model to produce one of the ten labels. A reasonable alternative is to fit a separate model to each class and choose the class of the model that assigns the highest fidelity to a test input. We use elastic nets to establish a modular recognition system as shown in Fig.1, in which an individual network is trained on examples of each digit class and then a digit is classified by deciding which net offers the best reconstruction.

We exploited 20000 digits from the handwritten digit database produced by the U.S. National Institute of Standards and Technology (NIST). 10,000 samples were used for training and the remaining samples for testing. The binary images are scaled to a 25 x20 pixel grid. For each digit class, we train an elastic net with A4 units by examples from the respective class. Learning is proceeded in three cycles with learning parameter μ in eqn (4) changed from 1 to 0.1. The parameter T is initially set to 2 and then decreased to 0.2. Fig. 2 illustrates the converged prototypes of the ten models without smoothness constraint, i.e., $\nu = 0$ in (3). In Fig. 3, converged weight vectors of ten elastic nets ($\nu = 5 \times 10^{-6}$) were demonstrated, from which neighborhood relationships are obvious.

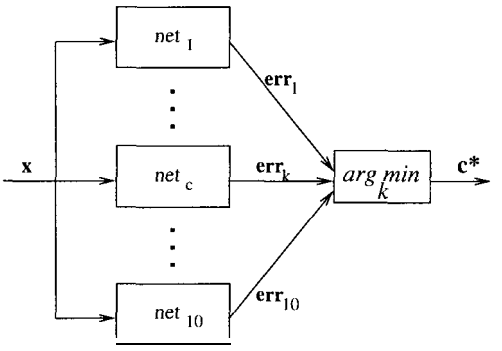

Figure 1. Our proposed modular classification system based on elastic nets.

The trained elastic nets can then be directly used as classifiers. Topographic map provides a population coding of representations and a simple decoding [4] can be proceeded as follows. In an elastic net, we regard mth prototype w(m) as a partial description of input x and a complete representation $\hat{\mathbf{x}}$ is the center of gravity of w(k), $k = 1, \cdots, M$. Specifically, the representation vector $\hat{\mathbf{x}}$ for each net can be expressed as

$$\hat{\mathbf{x}} = \frac{\sum_k a_k \mathbf{w}(\mathrm{k})}{\sum_k a_k} \quad (6)$$

where a_k is an interpolating function. A simple form is the following Gaussian kernel function, i.e.,

$$a_k = \exp(-\frac{\|\mathbf{x} - \mathbf{w}(\mathrm{k})\|^2}{2\rho_k^2}) \quad (7)$$

where ρ_k is the width of kth unit's receptive field, which should be chosen relatively sharp [4]. When a test sample x is presented to all the ten networks, reconstruction errors $err_l, l = 1, \cdots, 10$, are obtained,

$$err_l = \|\mathbf{x} - \hat{\mathbf{x}}^{(l)}\|^2,\ l = 1, \cdots, 10 \quad (8)$$

where l is the number of a network. We build a classifier by using a decision module which compare the distance or reconstruction errors (8) between the reconstructed vectors and presented pattern. A minimum operator is the most simple form for associating the class of a model with the smallest error, i.e., we assign x to the class c^* where

$$c^* = \underset{c}{\mathrm{argmin}}\ err, \quad (9)$$

For comparison purpose, we experimented with different comprise parameter ν. The main results are shown in Table 1. In the experiment, we took $\rho_k = \sqrt{(2)}/2$. We also compared different sizes of the elastic

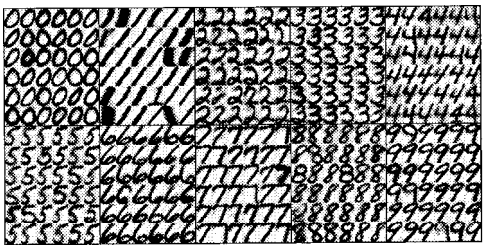

Figure 2. The converged reference vectors of ten mixture Gaussian models, i.e., $\nu = 0$.

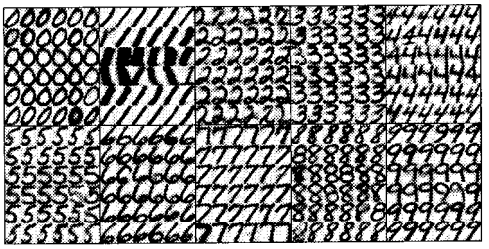

Figure 3. The converged reference vectors of ten elastic net model with parameter $\nu = 5 \times 10^{-e}$.

net. As can be expected, the larger the network, the more accurate the recognition result. However, as the number of nodes increases in each network, the learning will slow down and the improvement over the recognition rate is not obvious.

Table 1. Recognition accuracy for the elastic net models with different comprise parameter ν. Each network has 36 units. Both training and testing data set has 10,000 samples.

ν	0	5×10^{-6}	5×10^{-5}	5×10^{-4}	1×10^{-4}
training	**95.84%**	**95.97%**	**96.63%**	**96.91%**	**96.89%**
testing	**94.50%**	**94.74%**	**94.99%**	**95.32%**	**95.70%**

4 Conclusion

In this paper, we proposed an application of elastic net in handwritten digit recognition, based on a modular recognition scheme which incorporate population decoding concept. Our experiment results showed a fast training and high recognition accuracy. Our classification scheme can be considered as an extension of the well-known Learning Vector Quantization (LVQ) proposed by Kohonen [1]. LVQ finds a set of cluster centers for each class and classification is performed by finding the closest center by assigning the associated class. The online learning algorithms for LVQ are similar to the k-means clustering. In other words, LVQ extends clustering to classification problems. In a similar way, our recognition scheme generalizes mixture density estimation (with smoothness constraints) to classification problems. Such an extension is meaningful in both theory and practice.

References

[1] T.Kohonen, Self-organization and associative memory, Springer-Verlag, Berlin, 1989

[2] R.Durbin, G.Mitchison, A dimension reduction framework for understanding cortical maps, *Nature, 343, 644-647,* 1990

[3] A.L.Yulle, etc, Dimension reduction, generalized deformable models and the development of ocularity and orientation, Neural *Networks, 9,* 309-319, 1996.

[4] H.P.Snippe, Parameter extraction from population codes: a critical assessment, Neural *Comput., 8,* 511-529, 1996.

[5] K.Rose, etc. Statistical mechanics and phase transitions in clustering, *Phys. Rev. Lett., 65, 945-948,* 1990.