**CONTINUED TABLE OF CONTENTS**

**Electronic Noses And Their Applications**  
P. Keller, L. Kangas, L. Liden, S. Hashem, R. Kouzes  

**Target Load For Reducing Acidification Using Genetic Algorithm Approach**  
I. Wong, D.C.L. Lam  

**Neural Networks For Seismic Principal Components Analysis**  
K. Huang  

**Patent Protection For Neural Networks According To The European Patent Convention**  
Y. Skulikaris

**MATHEMATICAL FOUNDATIONS Oral Presentations**

**Adaptive Prototype Formation**  
H. Wechsler, V. Cherkassky, N. Vasilas  

**FLN: A Fuzzy Lattice Neurocomputing Scheme For Clustering**  
V. Petridis, V. Kaburlasos  

**A Maximum Balanced Mapping Certainty Principle For Pattern Recognition And Associative Mapping**  
L. Xu  

**Gradient Learning In Structured Parameter Spaces: Adaptive Blind Separation Of Signal Sources**  
S. Amari  

**Comparison Of VC-Method With Classical Methods For Model Selection**  
V. Cherkassky, F. Mulier, V. Vapnik  

**Cycle-Free Dynamics Of A Cluster-Competitive Net**  
A. Jagota, X. Wang  

**Qualitative Analysis Of Noisy Recurrent Neural Networks**  
O. Oluotomi, S. Das  

**Efficient Numerical Inversion Using Multilayer Feedforward Neural Networks**  
G. Lendaris, K. Mathia  

**Combinatorial Geometry And Vapnik-Chervonenkis Dimension**  
M. Oxley, M. Carter  

**On Completeness Of The Class Of Functions Computable By Neural Networks**  
S.E. Gilev, A.N. Gorban  

**An Almost Analytical Design Of Incremental Discrete Functions Approximation By One-Hidden-Layer Neural Networks**  
B. Beliczynski
A Maximum Balanced Mapping Certainty Principle for Pattern Recognition and Associative Mapping

Lei Xu

1. Dept. of Computer Science, The Chinese University of Hong Kong
   Shatin, Hong Kong (the correspondence address)
2. Information Science Center, Peking University, Beijing, China

Abstract A general principle, called Maximum Balanced Mapping Certainty (Max-BMC), is proposed for pattern recognition and associative mapping. Three unsupervised special cases have been investigated. One is equivalent to maximum mutual information or informax, with a close relation to a special case of the YING-YANG machine. One provides an extension of a recently proposed minimum uncertainty-unbalance PCA-type learning for a single neuron. The other gives a new model particularly suitable for pattern recognition purpose. Furthermore, the counterparts of the three cases for supervised learning have also been studied. In addition, the exclusive, factorial, and separable representations have been discussed with a new type of asymptotic separability for pattern recognition.

1. Introduction

We may come to an age of searching a unified scheme for many different unsupervised and supervised learning models that have been developed for pattern recognition and associative mapping in the literature. Recently, a Bayesian-Kullback scheme, called the YING-YANG Machine, has been proposed as such an effort (Xu, 1995a&96). Its one special case reduces to the EM algorithm (Dempster et al, 1977; Hathaway, 1986; Neal & Hinton, 1993) and the closely related Information Geometry theory and the em algorithm (Amari, 1995a&b), to MDL autoencoder with a “bits-back” argument by Hinton & Zemel (1994) and its alternative equivalent form that minimizes the bits of uncoded residual errors and the unused bits in the transmission channel’s capacity (Xu, 1995d), as well as to Multi-set model learning (Xu, 1995e)—a unified learning framework for clustering, PCA-type learnings and self-organizing map. Its other special case reduces to maximum information preservation (Linker, 1989; Atick & Redlich, 1990; Bell & Sejnowski, 1995). More interestingly its another special case reduces to Helmholtz machine (Dayan et al, 1995; Hinton, 1995) with new understandings. The YING-YANG machine includes also maximum likelihood or least square learning. Furthermore, the YING-YANG Machine has also been extended to temporal patterns with a number of new models for signal modeling, with the Hidden Markov Model (HMM), AMAR and AR models (Xu, 1995b) and other existing models included and extended. In addition, it has also been shown in Xu(1995a&c, 1996b) that one special case of the YING-YANG machine can provide us three variants for clustering or VQ, particularly with criteria and an automatic procedure developed for solving how to select the number of clusters in clustering analysis or Gaussian mixtures—a classical problem that remains open for decades.

A number of possible new models have also been suggested by this unified YING-YANG Machine scheme (Xu, 1996a). However, there do exist some learning models for pattern recognition and associative mapping that can not be unified under this unified scheme. In this paper, we propose another general learning principle, called Maximum Balanced Mapping Certainty (Max-BMC), as a complement scheme. In Section 2, we present this principle for unsupervised learnings. We show that one special case is equivalent to maximum mutual information or informax (Linker, 1989; Atick & Redlich, 1990; Bell & Sejnowski, 1995), with a close relation to a special case of the YING-YANG machine (Xu, 1996a). One other special case has extended a recently proposed minimum uncertainty-unbalance unsupervised learning for a single neuron, with a close relation to Nonlinear Maximum Variance (NMV) for PCA-type tasks (Xu, 1995f). Particularly, another special case gives us a new model which is more suitable for pattern recognition purpose. Section 3 further introduces the counterparts of the three cases for supervised learnings. In Section 4, based on this general learning principle, we discuss the exclusive, factorial, and separable representations, with a new type of asymptotic separability suggested.

1 This project was supported by the HK RGC Earmarked Grant CUHK484/95E, and by Ho Sin-Hang Education Endowment Fund for Project HSH 95/02.
2. The Max-BMC Principle for Unsupervised Pattern Recognition

All the tasks of pattern recognition and associative mapping can be summarized by setting up a mapping from an input pattern space X to a class or representation space Y. Probabilistically, such a mapping can be modeled by a conditional density $P_M(y|x)$ such that this model $M$ recognizes/classifies/maps an input $x$ into a class or a representation $y$ with probability $P_M(y|x)$.

To obtain this model $M$, we first need to specify a family $\mathcal{M}$ from which our model comes from. This family is usually of two types. For the first type, $M$ is given with $P_M(y|x)$ being a regression model, e.g., $P_M(y|x) = N(f(x), \Sigma)$—a normal density with mean $f(x)$ and variance $\Sigma$ and $f(x)$ implemented by a feedforward network or a radial basis network. Also, $P_M(y|x)$ can be given by normalizing the outputs $\{f_1(x), f_2(x), \ldots, f_k(x)\}$ of a feedforward network into probabilities, e.g., by softmax $e^{f_j(x)}/\sum_{j=1}^{k} e^{f_j(x)}$.

The second type is called generative model with

$$P_M(y|x) = \frac{P_M(x|y)P_M(y)}{\int P_M(x|y)P_M(y)dy}, \quad \text{or} \quad P_M(y|x) = \frac{P_M(x|y)P_M(y)}{\sum_{y=1}^{N} P_M(x|y)P_M(y)}.$$  \hspace{1cm} (1)

with $P_M(x|y)$ given by a backward network, e.g., $P_M(x|y) = N(g(y), \Sigma_y)$—a normal density with mean $g(y)$ and variance $\Sigma_y$ and $g(y)$ implemented by a backward network. In the simplest case, we even let $g(y) = m_y$ being a point, i.e., a usual normal density $P_M(x|y) = N(m_y, \Sigma_y)$.

Next, we need a principle to choose an appropriate one $M \in \mathcal{M}$, based on a given data set.

In unsupervised cases, we have only an input data set $\{x_i\}_{i=1}^{N}$ to base on. Hence, any model $M$, $P_M(y|x)$ describe the certainty that we make a decision of classifying or mapping $x$ into $y$. In the ideal case, we hope that this decision should be correct and fully confident. Since we have no supervisor here, it is difficult to check whether it is correct. However, we can maximize $P_M(y|x)$ or $f(P_M(y|x))$ to let our decision more confident, where $f(.)$ is a strictly monotonic increasing function on $[0, 1]$. Considering the whole distribution of $x$ and $y$, we maximize

$$E[f(P_M(y|x))|x] = \int_{y=1}^{N} f(y|x)f(P_M(y|x))dy, \quad \text{f(.) is strictly monotonic increasing on [0,1].}$$  \hspace{1cm} (2)

We further approximate $P(x,y) = P_M(y|x)P_0(x)$ with the empirical estimate $P_0(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x-x_i)$ for the input density. Putting this into eq.(2), we get

$$J_c = \frac{1}{N} \sum_{i=1}^{N} P_M(y|x_i)f(P_M(y|x_i))dy, \quad \text{or} \quad J_c = \frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{N} P_M(y|x_i)f(P_M(y|x_i))$$  \hspace{1cm} (3)

The larger is the $J_c$, the more confident we are on our mapping, that is, the mapping certainty is maximized. However, this maximization alone may result in a trivial solution that all the $\{x_i\}_{i=1}^{N}$ are mapped into a single $y_0$ with $P_M(y_0|x_i) = \delta(y-y_0)$ or $P_M(y_0|x_i) = 1$. In contrary, we hope that the mapped results can well distribute on the full range of $Y$ space. To achieve this target, we minimize $E[f(P_M(y|x))|x]$, i.e.,

$$J_b = \int_{x} P_M(y|x)f(P_M(y|x))dy = \int y=1 f(P_M(y|x))dy, \quad \text{J_b = } \int y=1 P_M(y|x)f(P_M(y|x))dy = \sum_{y=1}^{N} \sum_{y=1}^{N} P_M(y|x_i)$$  \hspace{1cm} (4)

The minimization of $J_b$ will let input data more balance mapped onto the full range of space $Y$.

As a whole, the trade-off between maxim J_c and min J_b gives us a principle as follows:

$$\max_{J_b} J_c - \min J_b \text{ with J_c, J_b given by eq.(3) and eq.(4).} \hspace{1cm} (5)$$

We call it by Maximum Balanced Mapping Certainty (Max-BMC) since it maximizes the mapping certainty and keep the largest balance on the distribution in the space $Y$.

In sequel, we examine three special cases for the function $f(.)$: (1) $f(x) = x$, in this case, eq.(5) becomes

$$J_b = \frac{1}{N} \sum_{i=1}^{N} \int_{x} P_M(y|x_i)dy \quad \text{or} \quad J_b = \frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{N} P_M(y|x_i)dy \quad J_b = \int_{x} P_M(y|x)dy, \quad \text{or} \quad J_b = \sum_{y=1}^{N} P_M(y|x)$$  \hspace{1cm} (6)

This is actually a further development of a unsupervised learning rule for a single neuron in Xu(1995f), called Minimum uncertainty-unbalance, which minimizes

$$J_{ub} = \frac{1}{N} \sum_{i=1}^{N} P_y(y=1|x_i)\sum_{y=1}^{N} p_y(y=0|x_i) - \left( \frac{1}{N} \sum_{i=1}^{N} P_y(y=1|x_i) \right)^2 - \left( \frac{1}{N} \sum_{i=1}^{N} P_y(y=0|x_i) \right)^2$$  \hspace{1cm} (7)

where $p_y(y=1|x_i) = \frac{1+\sigma(y|x_i)^2}{2}$ and $\sigma(\cdot)$ being sigmoid function with $s(0) = 0, s(-\infty) = -1, s(\infty) = 1$. Noticing $p_y(y=0|x_i) = 1 - p_y(y=1|x_i)$, the above equation can be rewritten into $J_{ub} = \sum_{i=1}^{N} P_y(y = 1|x_i) - \sigma(y=0) = \frac{1}{N} \sum_{i=1}^{N} p_y(y = 1|x_i)$. This $J_{ub}$ is just the special case of eq.(5) and eq.(6) for only one neuron. As shown in Xu(1995f), it is closely related to the Nonlinear Maximum Variance (NMV) learning rule

$$\max E[s(w^T x_i + c) - E(s(w^T x_i + c))]^2 = \max \{s^2(w^T x_i + c) - E(s(w^T x_i + c))^2\}$$

947
which is an extension of linear PCA learning to the nonlinear case.

\[ f(x) = \ln(x), \]  

in this case, eq.(5) becomes

\[ J_c = \frac{1}{N} \sum_{i=1}^{N} \int p_M(y|x_i) \ln p_M(y|x_i) \, dy, \quad \text{or} \quad J_c = \frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{k} p_M(y|x_i) \ln p_M(y|x_i), \]

\[ J_s = \int p_M(y) \ln p_M(y) \, dy, \quad \text{or} \quad J_s = \sum_{y=1}^{k} p_M(y) \ln p_M(y). \]  

(7)

That is, this Max-BMC is equivalent to maximum mutual information or informax (Linsker, 1989; Atick & Redlich, 1990; Bell & Sejnowski, 1995), and also a special case of the YING-YANG machine (Xu, 1996a).

(3) \[ f(x) = s(x) \] and \( s(x) \) is a sigmoid function with \( s(0.5) = 0.5 \) and is convex on \([0, 0.5]\), e.g., \( s(x) = \tanh(x - 0.5) + 0.5 \). In this case, eq.(5) becomes

\[ J_c = \frac{1}{N} \sum_{i=1}^{N} \int p_M(y|x_i) s(p_M(y|x_i)) \, dy, \quad \text{or} \quad J_c = \frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{k} p_M(y|x_i) s(p_M(y|x_i)) \, dy, \]

\[ J_s = \int p_M(y) s(p_M(y)) \, dy, \quad \text{or} \quad J_s = \sum_{y=1}^{k} p_M(y) s(p_M(y)) \]  

(8)

Comparing \( \ln(x) \) with \( z \), we find that \( J_c - J_s \) in eq.(7) attempts to discount those values of probabilities near 1 (which is helpful to increase the decision's certainty) and to over-count those near 0 (which is actually not desirable). In contrast, \( s(x) \) can also discount those values of probabilities near 0. As a whole, it pays more attention on deciding those data which are difficult to make a decision (which is surely desirable to increase decision certainty). Thus, eq.(8) should be more suitable for those tasks with binary outputs, such clustering or classification. This is a new model that deserves further theoretical and experimental explorations.

3. The Max-BMC Principle for Supervised Pattern Recognition

In supervised cases, we have a data set \( \{ x_i, y_i \}_{i=1}^{N} \) to base. Now the distribution of \( y \) is controlled by the data \( \{ y_i \}_{i=1}^{N} \), we no longer need the term \( J_s \) for this purpose. Thus, we can maximize only the following simplified \( J_c \) to choose \( M \in M \):

\[ J_c = \frac{1}{N} \sum_{i=1}^{N} p_M(y|x_i) f(p_M(y|x_i)) \]

\[ J_c = \frac{1}{N} \sum_{i=1}^{N} p_M(y|x_i) f(p_M(y|x_i)), \quad \text{for } f(x) = x, \]

\[ J_c = \frac{1}{N} \sum_{i=1}^{N} p_M(y|x_i) \ln p_M(y|x_i), \quad \text{for } f(x) = \ln(x) \]

(9)

The case for \( f(x) = x \) is the same as a new supervised learning model obtained from a special case of the YING-YANG machine (Xu, 1996a). All the others in eq.(9) are also new models that deserve further theoretical and experimental explorations.

4. Exclusive, Factorial, and Separable Representations

For \( J_c \) in eq.(3), in the case that \( f(\cdot) \) is convex, we have

\[ J_c \leq f(\frac{1}{N} \sum_{i=1}^{N} \int p_M(y|x_i) \, dy) \leq f(1), \quad \text{or} \quad J_c \leq f(\frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{k} p_M(y|x_i)) \leq f(1) \]  

(10)

That is, it is up-bounded by \( f(1) \). For \( f(x) = x, f(1) = 1 \) is achieved when \( \frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{k} p_M(y|x_i) \) = 1, which holds only when \( p_M(y|x_i) = 1 \) for a single \( y \) and \( p_M(y|x_i) = 0 \) for all the other \( y \)'s. In other words, each \( x_i \) is classified into one class with probability 1, or each \( x_i \) is exclusively represented by a single \( y \). For \( f(x) = \ln(x) \), similarly \( f(1) = 0 \) is also achieved when each \( x_i \) is exclusively represented by a single \( y \). So, we see that max \( J_c \) is to ensure an exclusive representation for each input \( x_i \).

On the other hand, \( \max(-J_s) \) is actually the entropy of \( P_M(y) \). When \( y = \{ y_1, \cdots, y_k \} \) with each binary \( y_j = 1 \) or \( y_j = 0 \), \( \max(-J_s) \) is achieved when \( P_M(y) = \prod_{j=1}^{k} P_M(y_j) \), i.e., whether \( y \), taking 1 or 0 is independent of what values \( y_j, j \neq i \) takes. This is called factorial representation, which is different from the above exclusive representation by which there's only one \( y_j = 1 \) with all \( y_i = 0, i \neq j \).

So we see that the Max-BMC Principle is a combination of exclusive representation on \( P(y|x) \) and factorial representation on \( P(y) \).

In sequel, we further study the condition that \( \frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{k} p_M^2(y|x_i) = 1 \). When \( N \) is finite, it holds only when \( p_M(y|x_i) \) is exclusive for each \( x_i \). When \( N \to \infty \), this requirement can be relaxed into

\[ \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{k} p_M^2(y|x_i) = 1, \quad \text{for } f(x) = x \]

\[ \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{k} p_M(y|x_i) \ln p_M(y|x_i) = 0, \quad \text{for } f(x) = \ln(x) \]  

(11)

which may not imply that \( p_M(y|x_i) \) is exclusive for each \( x_i \). In this case, we call those \( p_M(y|x_i) \) asymptotic exclusive. Clearly, exclusive implies asymptotic exclusive, but its inverse is not true.

948
Moreover, from \( \sum_{i=1}^{N} P_M(y|z_i) = 1 \), we have \( 1 = \sum_{i=1}^{N} P_M(y|z_i) + \sum_{i=1}^{N} \sum_{j=1}^{N} P_M(p|z_i)P_M(q|z_i) \). Thus, the condition eq.(11) for \( f(x) = x \) is equivalent to

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} P_M(p|z_i)P_M(q|z_i) = 0
\]

which holds when \( \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} P_M(p|z_i)P_M(q|z_i) = 0 \) for any pair \( p \neq q \). In other words, asymptotically there is no overlap between classes \( p, q \). That is, the data of classes \( p, q \) are asymptotically separable. Here, the meaning of "separable" is more general and relaxed than the conventional deterministic "linear separable" "nonlinear separable" in pattern recognition literature, which does not permit \( P_M(p|z_i)P_M(q|z_i) = 0 \) on any pair \( p \neq q \) even for a single point \( z_i \) and thus is too idealistic in practice. In the special case that \( M \) is given by the generative model eq.(1) with Gaussian density \( P_M(z|y) = N(\mu_y, \Sigma_y) \), the condition eq.(13) is exactly the Cond-EMG—a necessary and sufficient condition for the correct convergence by the EM algorithm on Gaussian mixture (Ma & Xu, 1996).

With the above regards, we can take the conditions eq.(11) or eq.(12) as the definition of asymptotic separable data or distributions, which are more general and practical for pattern recognition purpose. Moreover, from eq.(10) we know that making eq.(12) satisfied only results in the maximized upbound \( \ln 1 = 0 \) for \( J_c \) with \( f(x) = \ln x \), which does not imply the maximization of this \( J_c \) itself.

5. Conclusions

The Maximum Balanced Mapping Certainty (Max-BMC) can serve as a general principle for both unsupervised and supervised pattern recognition. It unifies some existing models and also provides several new models. It is a combination of exclusive representation on \( P(y|x) \) and factorial representation on \( P(x) \). Based on it, we can also obtain more general and practical definitions on separability for pattern recognition.

References


Ma, J.W., & Xu, L. (1996), "A necessary and sufficient condition for the correct convergence by the EM algorithm on Gaussian mixture ", to be submitted.


