http://dx.doi.org/10.1109/SIPNN.1994.344853

1994 International Symposium on Speech, Image Processing and Neural Networks, 13-16 April 1994, Hong Kong

Risk Assessment of Axillary Lymph Node Metastases in Early Breast Cancer Patients using the Maximum Entropy Network

Poh Lian Choong[‡], Christopher J.S. deSilva[‡], Hugh J.S. Dawkins[†],
Peter Robbins [†], Jennet M. Harvey[†], Gregory F. Sterrett[†],
John Papadimitriou [†] and Yianni Attikiouzel[‡]

Abstract

This paper describes an Artificial Neural Network (ANN) architecture for constructing Maximum Entropy (MaxEnt) models based on discrete distributions. Entropy is maximized by a partition function method involving the use of Lagrange multipliers which is capable of implementation by an ANN architecture. The *Maximum Entropy Network* (MaxEN), consists of a training module and a testing module of interconnected processing elements. The practical use of the MaxEN network is illustrated with an application in the clinical management of early breast cancer patients.

1 Introduction

This paper outlines a non-parametric method for estimating probability distributions from small sample sets. One major drawback of conventional non-parametric methods such as the Probabilistic Neural Network [6] or the Parzen windowing method [5] is the extensive amount of sample data required. In this paper, we describe the use of Maximum Entropy Estimation (MEE) for constructing multinomial distributions.

The Maximum Entropy Network (MaxEN) approach incorporates concepts from neural network theory, information theory, optimization and statistical inference. The most unbiased estimate of a probability distribution given only partial data (incomplete information) is obtained by maximizing Shannon's entropy measure subject to constraints derived from a set of moments [3]. The MEE allows us to construct a probability distribution that satisfies the specified constraints but no other conditions (implicit or explicit) relating to the data and whose entropy is greater than that of any other distributions that satisfies the same constraints.

Cheeseman [1] has proposed algorithm for constructing expert systems based on MaxEnt models. Our aim and approach are essentially the same as those of Cheeseman, in that we wish to derive MaxEnt distributions based on small data sets, which can be used to make various types of decisions. Our main application is in the field of medicine, in particular, the problems of diagnosis and prognosis: can we automate the process of making inferences about the causes and possible future course of a cancer based on a description of the measurement of various tumour parameters (risk factors)?

2 The Maximum Entropy Formalism

Let $S = \{0, 1, 2, \ldots, N\}$ be a finite set and let $\{p_0, p_1, p_2, \ldots, p_N\}$ be a probability distribution on S, where p_j is the probability of occurrence of j. In order that this should be a probability distribution, we must have $p_j > 0$ for all j, and $\sum_{j=0}^{N} p_j = 1$. It will be be convenient to regard the probability distribution as a vector

$$\mathbf{p} = (p_0, p_1, \ldots, p_N)^T \in [0,1]^{N+1} \subset \mathbb{R}^{N+1}. \qquad (1)$$

The entropy of the probability distribution $\mathbf{p}$ is defined to be

$$H(\mathbf{p}) = -\sum_{j=0}^{N} p_j \log(p_j), \qquad (2)$$

where log denotes the natural logarithm. It is easy to show that H has a unique maximum on the set of probability distributions when $p_j = 1/(N+1)$ for all j.

Let $f_k, k = 1, \ldots, C$ be functions defined on S. The

[0] This research is supported by the Cancer foundation of Western Australia.

[‡] Center for Intelligent Information Processing Systems
Department of Electrical & Electronic Engineering
University of Western Australia
Nedlands 6009
Western Australia

[†] Department of Pathology and Hospital and University Pathology Services
Queen Elizabeth II Medical Center
Sir Charles Gairdner Hospital
Nedlands 6009
Western Australia

0-7803-1865-X/94/$3.00 © 1994 IEEE

average values of these functions are given by

$$\bar{f}_k = \sum_{j=0}^{N} p_j f_k(j) \tag{3}$$

The MEE process is a means of finding the probability distribution which satisfies the constraints imposed by the equations above whose entropy is greater than any other distribution which satisfies the same constraints. This is a constrained optimization problem, and may be solved in the standard way by the use of Lagrange multipliers, $\lambda_0, \lambda_1, \ldots, \lambda_C$.

Gibbs defined the *partition function*, Z, which is a function of the Lagrange multipliers:

$$Z(\lambda_1, \ldots, \lambda_C) = \sum_{j=0}^{N} \exp\left(-\sum_{k=1}^{C} f_k(j)\lambda_k\right), \tag{4}$$

from which the probabilities may be computed from

$$p_j = \frac{\exp\left(-\sum_{k=1}^{C} f_k(j)\lambda_k\right)}{Z(\lambda_1, \ldots, \lambda_C)}. \tag{5}$$

To find the λ_k, we have to solve the equations

$$\frac{\partial \log(Z)}{\partial \lambda_k} = \bar{f}_k \tag{6}$$

for $k = 1, \ldots, C$. When the values of the $f_k(j)$ are integers, which is the case when the constraints specify the values of moments of the distribution, these equations can be re-written as a set of C simultaneous polynomial equations for the $e^{-\lambda_k}$. (It may be noted that the partition function formulation always produces a set of probabilities that satisfy $\sum_{j=0}^{N} p_j = 1$, so there is no need to apply this constraint explicitly.) Substituting $v_k = e^{-\lambda_k}$ simplifies the optimization process. We shall call the v_k the *Lagrange Coefficients*.

3 The Maximum Entropy Network

The complexity of the MEE process using this method has been addressed by the use of an ANN architecture. The Maximum Entropy Network (MaxEN) consists of two modules of interconnected processing elements, each capable of carrying out simple operations. The network entropy maximization state is characterized by a set of Lagrange multipliers, λ_k, that is obtained by solving a set of C non-linear equations:

$$\mathbf{Y}(\lambda) = \mathbf{0} \tag{7}$$

or

$$\begin{bmatrix} y_1(\lambda) \\ \vdots \\ y_C(\lambda) \end{bmatrix} = \begin{bmatrix} \bar{f}_1 - \frac{\partial \log Z}{\partial \lambda_1} \\ \vdots \\ \bar{f}_C - \frac{\partial \log Z}{\partial \lambda_C} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{8}$$

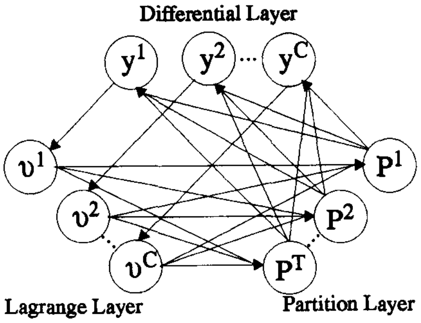

Figure 1: MaxEN Training Architecture

3.1 Training Module

The MaxEN *training module, shown in* Figure 1, employs three layers of processing units. The first layer of units is the *Lagrange Layer*, which has C units, whose states correspond to the values of the Largrange coefficients. The second layer is the *Partition Layer*, which computes the components of the partition function. The number of units in this layer is equal to the product of the numbers of discrete values of each of the input variables. The third layer, the *Constraint Layer*, also has C units, which compute the extent to which the constraints are not satisfied.

The connections between the Lagrange Layer and the Partition Layer have associated weights, which are determined by the constraints.

Training the MaxEN is a two-phase process. The forward propagation step involves determining the magnitude of the objective function after each update to the Lagrange states or coefficients, v_k. The back propagation step then calculates the difference between the actual and target values of the objective function value, $\mathbf{Y}(v)$, and changes the states of the Lagrange units to minimize the error. The process is iterated until the error is less than some tolerance value.

3.1.1 Forward Propagation

To initialize the network, all the Lagrange coefficients, v_k are initially set to unity. Incoming connections to the Partition Unit, j are at the left and originate from Lagrange Units of the layer below. Output values of the Lagrange Units arriving at the Partition Unit, j are calculated,

$$P_j = \prod_{k=1}^{C} v_k^{w^1_{kj}} \tag{9}$$

where

v_k = activation level of *Lagrange Unit, k*;
w^1_{kj} = weight from the *Lagrange Units, k* to *Partition Unit, j*.

The layer of *Partition Units* computes the component of the *Partition Function, Z*(4).

The activation level of each of the Partition Units is then propagated to the Differential Layer, which computes the objective function, $\mathbf{Y}(v)$ from the weighted state of the Partition Units. The weight matrix defined between the layer of Partition Units and Differential Units is, $w^2_{jk} = w^1_{kj} - f_k$ where w^2_{jk} = weight from the Partition Unit, j to Differential Unit, k. The outputs of the *Differential Units* are:

$$y_k = \sum_{j=1}^{T} w^2_{jk}.P_j \tag{10}$$

where

P_j = activation level of *Partition Unit, j*;
f_k = k^{th} constraint value.

3.1.2 Backward Propagation

The backward propagation state involves adjusting the activation level of the Lagrange Units to minimize the objective function. The minimization algorithm used is the Error Propagation Method. The difference between the desired and predicted value is propagated back from the Differential Units to the Lagrange Units and the values of the Lagrange Coefficients, v_k are adjusted accordingly to reduce the error.

Using the Error Propagation Method, the update algorithm is:

$$v_k^{t+1} = v_k^t + \alpha.(-y_k) + \beta.(v_k^t - v_k^{t-1}) \tag{11}$$

where

α is the gain parameter set to 0.001

β is the momentum parameter set to 0.001

The network is trained successfully when the magnitude of the objective function, $\mathbf{Y}(v)$ is less than the predefined tolerance value.

3.1.3 MENN Testing Module

The Lagrange coefficients, v_k, determined from the training process are used as weights in the testing module. *Figure 2 shows the network architecture* of the testing module for classifying input variables $X = (x^2, \ldots, x^N)$ into $(n_1 + 1)$ categories. The input units are merely distribution units that supply the same input variables to all the Pattern units. The weight matrix, W, between the Input units and the Pattern units is calculated from the determined Lagrange coefficients, v_k as follows:

$$w_{ij} = \begin{cases} j. - \log(v_1) & \text{if } i = 1 \\ -\log(v_i) + j. - \log(v_{(i+(N-1))}) & \text{if } i \neq 1 \end{cases} \tag{12}$$

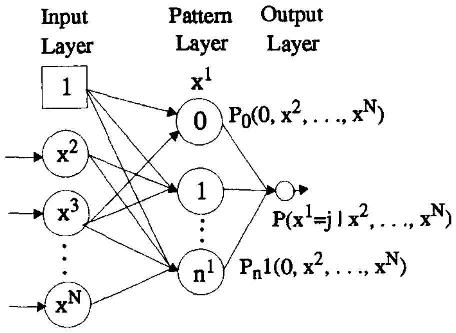

Figure 2: MaxEN Testing Architecture

where $N < C$, $i = 1, \ldots, N$ and $j = 0, \ldots, n^1$.

Each Pattern unit computes

$$S_j = (\sum_{i=2}^{N} x_i.w_{ij}) + w_{1j} \tag{13}$$

where

w_{ij} = weightfrom Input unit i to Pattern unit j.

and then performs a non-linear operation on S_j before outputting the activation level, $P(x^1 = j, x^2, \ldots, x^N)$ to the output units. The non-linear operation used is $f(S_j) = \frac{exp(-S_j)}{Z(\lambda_1,\ldots,\lambda_C)}$ and the resultant activation level of each pattern unit represents the MaxEN probability distribution, $P(x^1 = j, x^2, \ldots, x^N)$. The Output units merely sum the output of the pattern units and calculate the conditional probabilities as follows:

$$P(x^1 = j|x^2, \ldots, x^N) = \frac{P(x^1 = j, x^2, \ldots, x^N)}{\sum_{i=0}^{n^1} P(x^1 = i, x^2, \ldots, x^N)} \tag{14}$$

4 Breast Cancer Prognosis

4.1 Medical Background

Currently, the only certain way to determine the status of the axillary lymph node is to carry out complete axillary clearance. Accurate knowledge of axillary lymph nodes status has an essential role in the management of early breast cancer patients because the axillary node represents the principal site of regional metastases and is a marker of systemic metastases. It would be of value to be able to predict which patients are likely to have metastases on the basis of characteristics of the primary tumour.

4.2 Description of Data

The Breast Cancer data obtained from the Department of Pathology, Hospital and University Pathology Services of the Sir Charles Gairdner Hospital, The University of Western Australia related to 247 patients diagnosed with primary breast cancer in Western Australia from 1990 - 1992. Only a consecutive series of 176 patients, treated by surgical excision for whom complete histological information was available were considered. All the 176 patients were diagnosed with *primary infiltrating carcinoma* [4] and had complete axillary dissection. The extent of axillary lymph node metastases of all the 176 patients were available for the following analysis.

Table 1 lists the factors used in this study. All of the factors evaluated have been previously shown to have value as indicators of prognosis [2].

Table 1: Clinical and histopathological factors

Risk Factor	Values
1. Age	Continuous
2. Mitotic Count	Continuous
3. Tubule	0, 1, 2
4. Nuclear Size	0, 1, 2
5. Nuclear Pleomorphism	0, 1, 2
6. Tumour Grade	0, 1, 2
7. Tumour Size	Continuous
8. Vascular Invasion	0, 1

5 MaxEnt Model Construction

The model that we construct give estimate of the probability of classifying node-positive (LN+) and node-negative (LN−) patients with a given set of risk factors values. For example we might construct a model with the risk factors Nuclear Size and Vascular Invasion. The model would give us probabilities of the forms $P(LN + \ and\ NuclearSize = 1\ and\ VascularInvasion = 1)$ for example. From these it is easy to calculate the conditional probabilities of LN+ and LN− given any set of values of the risk factors in the model, which are the probabilities of interest for prediction. These probabilities are then used to assign patients to either LN− or LN+ group, based on a cut-off point of 0.5.

We constructed models incorporating more than one risk factor. In these models, the constraints imposed were the mean of the outcome, the risk factors and of the products of the outcome and the individual risk factors. Table 2 gives details of the various models.

The overall accuracy of the models needs to be increased if they are to be of practical use. We expect to improve the accuracy by constructing models based on larger training sets with other sets of risk factors.

Table 2: Results of the most accurate models

Risk factors	Sensitivity (%)	Specificity) (%)
Nuclear Pleomorphism Tumour Size Vascular Invasion	83.3	80.0
Tumour Grade Tumour Size Vascular Invasion	81.0	82.0

6 Conclusion

Non-parametric modelling using the MEE provides probabilistic inference for small sample. Apart from this advantage, the multinomial models can be constructed with minimum computation time. Overall MEE proves to be a useful method for pattern recognition in situation whereby the number of sample data is small. In the clinical example illustrated, the overall accuracy of the models needs to be increased if they are to be of practical use. We expect to improve the accuracy by constructing models based on other sets of risk factors.

References

[1] P. Cheeseman, "A Method of Computing Maximum Entropy probability values for Expert Systems ", *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, C. R. Smith and G.J. Erickson (eds.), D. Reidel Pub. Comp., pp. 1042 - 1052, 1989.

[2] C.W. Elston and I.O. Ellis, "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long term follow-up", *Histopathology*, **19**, pp. 403 - 410, 1991.

[3] E.T. Jaynes, "On the rationale of maximum entropy methods", *Proc. IEEE*, **70**, pp.939 - 952, 1982.

[4] D.L. Page and T.J. Anderson, *Diagnostic Histopathology of the Breast*, Logman Group UK Ltd, 1987.

[5] E. Parzen, "On estimation of a probability density function and mode", *Annals of Mathematical Statistics*, **33**, pp.1065 - 1076, 1962.

[6] D.F. Specht, "Probabilistic Neural Network", *Neural Networks*, **3**, pp. 109 - 118, 1990.

[7] C.E. Shannon, "A Mathematical Theory of Communication", *The Bell System Technical Journal*, **27**, pp. 379 - 423, 1948.