Estimates of constrained multi-class a posteriori probabilities in time series problems with neural networks

Juan Ignacio Arribas*, Jesus Cid-Sueiro* , Tulay Adali**, Hongmei Ni**, Bo Wang** and Anibal R. Figueiras-Vidal***

*Dept. of Teoría de la Señal, Comunicaciones e Ing. Telemática, Universidad de Valladolid, Valladolid, Spain

**Dept. of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Baltimore, MD, USA

***Dept. of Tecnologías de las Comunicaciones, Universidad Carlos III de Madrid, Madrid, Spain

email: jarribas@tel.uva.es, jescid@tel.uva.es, adali@engr.umbc.edu, arfv@ing.uc3m.es

Abstract

In time series problems, where time ordering is a crucial issue, the use of Partial Likelihood Estimation (PLE) represents a specially suitable method for the estimation of parameters in the model. We propose a new general supervised neural network algorithm, Joint Network and Data Density Estimation (JNDDE), that employs PLE to approximate conditional probability density functions for multi-class classification problems. The logistic regression analysis is generalized to multiple class problems with softmax regression neural network used to model the a-posteriori probabilities such that they are approximated by the network outputs. Constraints to the network architecture, as well as to the model of data, are imposed, resulting in both a flexible network architecture and distribution modeling. We consider application of JNDDE to channel equalization and present simulation results.

Introduction

Partial likelihood is a recent extension of maximum likelihood introduced by Cox [1]. It provides a particularly suitable formulation for time series problems and a partial likelihood formulation for real-time signal processing is given in [2]. The maximum PLE (MPLE) can be written as follows:

$$PL(\mathbf{w},\mathbf{v})=\prod_{i=1}^{N}\prod_{j=1}^{J} f_{\mathbf{w},\mathbf{v}}(C_{i,j}\mid \mathbf{x}_i)^{T_{i,j}}$$

$$T_{i,j}=\begin{cases}1 & \mathbf{x}_i\in C_{i,j}\\ 0 & else\end{cases} \tag{1}$$

where **w** represents the network parameters, **v** the model of data over-parameters, J the number of classes, N the number of samples, $C_{i,j}$ the class label at time instant i for class j, $\mathbf{x}_i$ the observed history vector at time i. Function $f(.)$ represents the a posteriori probabilities and $T_{i,j}$ the indicator index of class j at time i.

There are several approaches to estimation of the a-posteriori class probabilities. One is based on *Strict Sense Bayessian classifiers* (SSB). Following [3], we call a classifier *Strict Sense Bayesian* (SSB) if its outputs are estimates of the a posteriori probabilities of the classes. In a similar way, for a training viewpoint, we call a cost function is SSB if it is minimized when the classifier is SSB. In [3] and [4], a SSB cost function is defined such that it has a unique minimum when output **y** coincides with a posteriori class probabilities, which we are trying to determine. Necessary and sufficient conditions for a cost to be SSB are also given in the reference.

We propose softmax regression for the a posteriori class probabilities,

$$f_{\mathbf{wv}}(C_{i,j}\mid \mathbf{x}_i,\mathbf{w}_{j,k}) \tag{2}$$

and Gaussian mixture model for observation,

$$f_{\mathbf{vw}}(\tilde{\mathbf{x}}_i\mid \pi_{m,j,k},v_{m,j,k},\sigma_m^2) \tag{3}$$

where the dimensionality of both models does not need to be the same. OPDE-MPL learning rule is obtained for the *Generalized Softmax Perceptron* (GSP) universal classifier architecture, see Fig.1.

0-7803-5529-6/99/$10.00

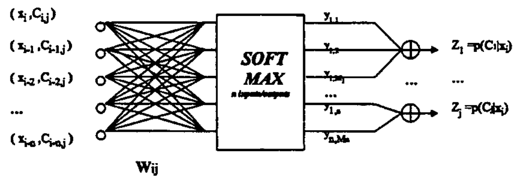

Fig 1- The GSP-PL network

We show that for a Gaussian mixture model, the class conditional probabilities, using Bayes theorem, are:

$$f_{\mathbf{w},\mathbf{o}}(\mathbf{x}_i \,|\, C_{i,j}) = \sum_{k=1}^{M_j}\sum_{m=1}^{L} \pi[j,k,m;\mathbf{w},\mathbf{o},\mathbf{x}_i] \quad (4)$$
$$N[\tilde{\mathbf{x}}_i - v(j,k,m;\mathbf{w},\mathbf{o},\mathbf{x}_i), \sigma_m^2]$$

where $\tilde{\mathbf{x}}_i$ represents the, in general, different dimensionality of the model of data with respect to network model for the observation $\mathbf{x}_i$, k the subclass index for GSP, M_j the number of subclasses in class j, m the mixture component index, L the number of mixture components, and $N(x-v,\sigma^2)$ the uncorrelated Gaussian kernel.

Probability term π and modified mean v, can be properly defined as a function of network weights $\mathbf{w}$ and data model over-parameters $\mathbf{o}$ at each time instant i. It is important to note the uncoupled treatment of network model and data model order in this approach. Details of previous formulation, can be found in [4] and [5].

Results

Under two regularization conditions, it can be shown that PL maximization is equivalent to Accumulated Relative Entropy (ARE) or Kullback-Leibler distance minimization [2]. PL (or the ARE cost) measures relative errors as opposed to the popular Mean Square Error (MSE) cost that uses absolute errors. In [2], it is shown that the PL (or the ARE) cost is well-formed in the sense of Wittner and Denker in that the algorithm will always recover from convergence from the wrong extreme.

For binary coding, we obtain the tracking performance for PL shown in Fig. 2, which highlights the well-formed property of the PL cost.

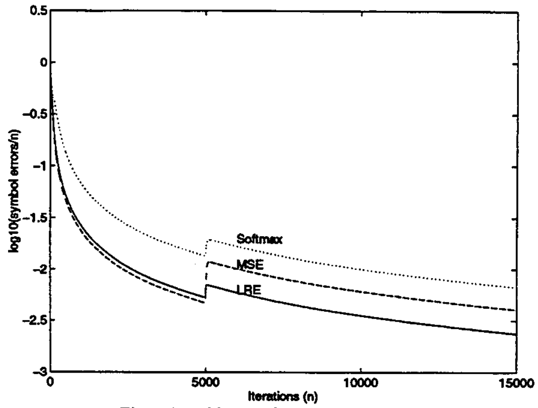

Fig 2- Tracking performance ARE cost

Conclusions

A new parametric density estimation algorithm has been proposed to train neural classifiers in time series regression problems. MPL gradient learning rules for the general OPDE have been derived. The method requires making some hypothesis about data density functions, as well as about the time series structure with softmax based networks, but this can be assumed as general as desired and both effects are uncoupled. Simulations are carried out to study the performance of the algorithm.

Acknowledgements

This work has been partially supported by the Spanish Government under C.I.C.Y.T. grant numbers TIC96-0500-C10-03 and TIC96-0500-C10-09. T. Adali's work was partially supported by the National Science Foundation Career Award, NSFNCR-9703161.

References

[1] D.R.Cox, "Partial likelihood", **Biometrika**, vol. 62, pp.69-72.

[2] T.Adali, X. Liu and M.K. Sönmez, "Conditional distribution learning with neural networks and its application to channel equalization", **IEEE Trans. On Signal Processing**, Vol. 45, No. 4, pp. 1051-1064, April, 1997.

[3] J. Cid-Sueiro, J. I. Arribas, S. Urban-Muñoz, and A. R. Figueiras-Vidal , "Cost Functions to Estimate A Posteriori Probabilities in Multiclass Problems" **IEEE Transactions on Neural Networks**, May 1999, Volume 10, Number 03, p. 645.

[4] J.Ignacio Arribas, Jesus Cid-Suerio, Tülay Adali and Anibal R. Figuerias-Vidal, "Neural architectures for parametric estimation of a posteriori probabilities by constrained conditional density functions", **Proceedings of the IEEE Workshop on Neural Networks**, Madison, Wisconsin, USA, 1999.

[5] Juan Ignacio Arribas, Jesus Cid-Sueiro, Tulay Adali and Anibal R. Figueiras-Vidal, "Neural networks to estimate ML multi-class constrained conditional probability density functions", **International Joint Conference on Neural Networks**, Washington D.C., USA, July 1999.