NEURAL NETWORK BASED FUZZY MEMBERSHIP FUNCTION ESTIMATION. APPLICATION TO UNCERTAIN TIME-VARYING SYSTEMS SUPERVISION

R. Lengellé, C. Richard

Laboratoire de Modélisation et Sûreté des Systèmes
Université de Technologie de Troyes
B.P. 2060 10010 Troyes cédex, France
Phone: +33-3-25-715681, Fax: +33-3-25-715699
Email: regis.lengelle@univ-troyes.fr

*S. Millemann**

Kurtosis Ingénierie
27 Avenue de la Constellation
B.P. 8295 95802 Cergy Pontoise cédex, France
Phone: +33-1-34-249259, Fax: +33-1-30-384937
Email: infos@kurtosis.fr

This work was supported by Kurtosis Ingénierie in collaboration with Renault.

ABSTRACT

Within the framework of fuzzy logic for pattern recognition, we propose a neural network based membership function procedure. The estimate is obtained as the output of a multilayer network trained to minimize a fuzziness measure of the concept to be learnt. This method is shown to be able to deal with probabilistic, resolutional and fuzzy uncertainty. Fuzziness estimation is also made possible in a better way than using predefined membership functions. We have successfully applied this approach to car behavior evaluation, where fuzzy classes are built on the basis of numerical information in relation with subjective expert evaluations.

1. INTRODUCTION

In system supervision applications, one has to make decisions on the basis of measurements. When a probabilistic model of observations is available, classical statistical decision methods can be applied. On the contrary, a solution can consist in elaborating the decision system from a data set labeled by experts using, for example, statistical pattern recognition methods. Within these approaches, it is often assumed that the expert evaluation is performed with certainty so that the problem is to estimate a regression of the decision (given by the expert) as a function of the measurements (or features extracted from the measurements). However, in various applications, resolutional and/or fuzzy uncertainty can be encountered. As an example, this phenomena appear for slowly time varying systems supervision, where locations of boundaries between stationary states may depend on experts judgements. Examples of such applications can be found in physiological signal analysis (event detection, automatic sleep staging), for which experts conflict or vagueness is often observed. We propose in this paper a neural network based membership function estimation procedure. The universal approximator property of multilayer networks, coupled with the minimization of a fuzziness measure, is used to determine the network parameters. The fuzziness of classes can be automatically determined via analysis of the network output. Furthermore, we show that the neural network also performs gradual fusion of each expert decision from the disjunctive mode (disagreement between experts) to the conjunctive mode (agreement between experts). The method is successfully applied to car behavior estimation.

**S. Millemann is now with PSA Peugeot Citroën*

2. SUBJECTIVE AND FUZZY INFORMATION

Uncertain information can be partitioned into three major categories. The next paragraph give some description of these categories, according to [1].

2.1 Kinds of uncertainty

Probabilistic uncertainty is related to randomness of observations. It is often associated to measurement noise or random fluctuation of the observed system.

Resolutional uncertainty corresponds to the limitations which make the observation not exactly perceptible. This kind of uncertainty is, up to some extent, responsible for the possible conflict between experts.

Fuzziness is the part of uncertainty which comes from the information coding scheme. As an example, in Section 4, fuzziness is associated with the language used by experts to characterize observations. Examples of fuzziness in an expert judgement are: "*possible* apparition of …", "*high* level of …", "*weak* presence of…".

We now remind some basic elements of fuzzy logic.

2.2 Fuzzy logic principles

Let X be the reference universe. First, consider the classical set (crisp set) theory. In this case, every subset $A \subset X$ is uniquely defined by its membership function $\mu_A(x)$ defined as:

$$\mu_A(x) = \begin{cases} 1 & \text{if and only if} \quad x \in A \\ 0 & \text{if and only if} \quad x \notin A. \end{cases} \tag{1}$$

$\mu_A(x)$ is a map from X into $\{0,1\}$. When the membership function $\mu_A(x)$ is a map into $[0,1]$, A is called fuzzy subset of X. References on fuzzy sets theory can be easily found in the literature, see for example [2], [3]. We only remind here the main definitions and properties used in this paper.

Definitions:

$$A \subset B \quad \text{iff} \quad \forall x, \mu_A(x) \le \mu_B(x), \tag{2}$$

$$A = B \quad \text{iff} \quad \forall x, \mu_A(x) = \mu_B(x), \tag{3}$$

$$\mu_{\overline{A}}(x) = 1 - \mu_A(x), \tag{4}$$

$$\mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\}, \tag{5}$$

$$\mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\}. \tag{6}$$

Other functions than min and max have been proposed to define conjunction and disjunction operators. However, these two functions are the most commonly used because of their desirable properties.

Properties:

The properties of conjunction, disjunction and negation operators are generally maintained (commutativity, associativity, distributivity, Morgan's laws, …). However, if min and max are used as $\cap$ and $\cup$, respectively, the properties $A \cap \overline{A} = \varnothing$ and $A \cup \overline{A} = X$ do not still hold. They must be replaced by $\mu_{A \cup \overline{A}}(x) \ge 1/2$ and $\mu_{A \cap \overline{A}}(x) \le 1/2$.

Fuzziness measures H using only membership functions were proposed ([4], [2]), and Ebanks [5] listed some desirable properties for such measures. For example, Kaufmann [2] proposed:

$$H(A,q) = \frac{2}{N^{1/q}} \left[\sum_{x \in X} \left| \mu_A(x) - \mu_{A_{1/2}}(x) \right|^q \right]^{1/q}, \tag{7}$$

where N is the cardinality of X, $q \in [1, +\infty[$ and $A_{1/2}$ is the A $\alpha-cut$ of level ½ defined by:

$$A_{1/2} = \{x \in X : \mu_A(x) \ge 1/2\}, \tag{8}$$

$$\mu_{A_{1/2}}(x) = \begin{cases} 1 & \text{if} \quad \mu_A(x) \ge 1/2 \\ 0 & \text{elsewhere}. \end{cases} \tag{8'}$$

This fuzziness measure is a distance between $\mu_A(x)$ and a crisp membership function of A. Due to the similarity between $H(A,2)$ and the usual mean squared error minimized during the training process of a neural network, the fuzziness measure $H(A,2)$ was used in this work. This will be explained later.

2.3 Classification, learning and neural networks

Let $\{(\underline{x}_n, t_n)\}_{\text{n}}$ denotes a training set of labeled data, where $\underline{x}_n$ is the feature vector associated to the target t_n. Within the framework of statistical pattern recognition, classification can be achieved by estimating the target $t(\underline{x})$ for every $\underline{x} \in X$. When the joint probability function $p(\underline{x}, t)$ is ignored, one must optimize performance measured by a criterion c which depends only on the training set $\{(\underline{x}_n, t_n)\}_{\text{n}}$. One must answer several questions.

1- Which model can be used for the decision function ?

2- Which criterion one must optimize ?

3- Is the resulting performance close to the minimal achievable probability of error P_{Bayes} given by the Bayes decision rule.

4- Is the solution of the optimization problem achievable in a reasonable time ?

Before answering these questions, consider a decision rule $t(\underline{x})$ which is a member of a family T. Let t_{opt} be the decision rule resulting from the optimization of any given criterion c which depends only on $\{(\underline{x}_n, t_n)\}_{\text{n}}$. Let $P_e(t_{opt})$ denote the error probability of the decision rule t_{opt}, and let $\varepsilon(t_{opt})$ be defined as:

$$\varepsilon(t_{opt}) = P_e(t_{opt}) - P_{Bayes}, \tag{9}$$

which can be rewritten as

$$\varepsilon(t_{opt}) = \left(P_e(t_{opt}) - \inf_{t\in T}(P_e(t))\right) \\ - \left(\inf_{t\in T}(P_e(t)) - P_{Bayes}\right) \tag{10}$$

In Eq. 10, one observes that the first term depends only on the criterion c and on its optimization process. This will be called *estimation error*. The second term, usually called *approximation error,* depends only on T . In [6], Vapnik and Chervonenkis have shown that the minimization of $\varepsilon(t_{opt})$ leads to a compromise between these terms: an increase in the *size* (formally defined as the Vapnik-Chervonenkis dimension) of T yields an increment of the estimation error and a decrement of the approximation error, and conversely.

Another point of view of the approximation/estimation error dilemma can be found in the statistical literature as the bias-variance tradeoff. Assuming that the joint probability density function $p(\underline{x},t)$ is known, the criterion c to be minimized is usually chosen as the mean squared error between the target t and the decision rule $t(\underline{x})$:

$$c = E\left(\left| t - t(\underline{x}) \right|^2 \right). \tag{11}$$

Eq. 11 can be rewritten as:

$$c = E\left(\left| t - E(t/\underline{x}) \right|^2\right) + E\left(\left| t(\underline{x}) - E(t/\underline{x}) \right|^2\right). \tag{12}$$

The first term is the smallest reachable squared error. It depends only on the problem encountered. The second one, which depends on $t(\underline{x})$ estimation accuracy, has to be minimized. It can be rewritten as follows:

$$E\left(\left|t(\underline{x}) - E(t/\underline{x})\right|^2\right) = \left|E(t(\underline{x})) - E(t/\underline{x})\right|^2 \\ + E\left(\left(t(\underline{x}) - E(t(x))\right)^2\right) \tag{13}$$

The first term is recognized as a bias term, and the second as a variance one. Again, if we consider that the decision rule $t(\underline{x})$ is a member of a given family T , an increase in the *size* of T induces a decrement of the bias term and an increment of the variance term, and conversely. Hence, minimization of c results in making a compromise between these antagonistic terms.

Optimization of the squared error was shown to be of major interest, in particular when decision functions are implemented via multilayer networks. Let t be a M-dimensional vector, with M the number of classes, and consider the *one bit encoding* defined as follows:

$$\left(t^i = 1 \quad \text{and} \quad t^j = 0 \quad \forall j \neq i\right) \quad \text{iff} \quad \left(x \in \omega_i\right), \tag{14}$$

where t^i denotes the i$^{\text{th}}$ component of *t.* Then,

$$E(t^i/\underline{x}) = 1.p(t^i = 1/\underline{x}) + 0.p(t^i = 0/\underline{x}) \\ = p(\omega_i/\underline{x}). \tag{15}$$

According to universal properties of multilayer neural networks (see for example [7], [8]), convergence of $t(\underline{x})$ to $E(t/\underline{x})$ can be guaranteed when the cardinality N of the learning set tends to infinity. In our case, the i$^{\text{th}}$ component of the network output tends to $p(\omega_i/\underline{x})$. Therefore, decision that consists in classifying the observation $\underline{x}$ into ω_i if t^i is the largest network output is recognized as the maximum a posteriori probability rule (also known as the Bayes decision rule). This decision rule minimizes the probability of error. Hence, when N is finite, such a decision rule is an estimate of the Bayes one.

In order to deal with the bias variance tradeoff, the network architecture has to be chosen to optimize an estimate of performance. This is usually done with an independent test set (cross validation) or via resampling methods (Jackknife, Bootstrap, for example).

3. APPLICATION TO FUZZY MEMBERSHIP ESTIMATION

3.1 Subjective probabilities and fuzzy membership

In Section 2, it was assumed that labels t_n available in the training set were unambiguously known. However, it often arises that expert judgements are expressed as fuzzy valuations. As an example, this often occurs when expertise is subject to resolutional uncertainty because experts express their valuation using vague concepts. It may also appear disagreement between different experts. Such problems are frequently encountered in many areas including non stationary complex systems monitoring, biomedical signal analysis and recognition, etc. An example is given in Section 4.

In order to perform estimation of fuzzy membership functions from numerical data using a multilayer network, one has to define a coding scheme from the fuzzy valuation made by experts to a numerical value. Remember that multilayer networks perform a posteriori probability estimation (see Eq. 15). Probability of an event can be defined as (the limit when the number of experiments tends to infinity of) the relative frequency of its occurrence. Hence, probability estimation requires that event occurrences are unambiguously known. In the application considered in this paper, experts disagreement may be present. This can result from a different perception of phenomena by different experts, and also because expert may not agree on the generating process of the events considered. Within this context, it becomes impossible to evaluate the relative frequencies (they even may have no sense). We must only deal with the perception made by experts, which was called *subjective probabilities* in the literature (see for example [9], [10]). The following example illustrates this notion: in application described in Section 4, if 3 experts out of 4 say that a given car has the “pumping defect”, we shall conclude that the subjective probability for this car to be a member of the “pumping defect” class equals 0,75. This should not be confused with the (classical) probability that a given car has this defect. Subjective probabilities have been related to fuzzy membership functions. In [11] and [12], for example, authors suggested that the number of positives answers to the question “does x belong to A?” must be (linearly) related to $\mu_A(x)$.

Now, we show that using a crisp coding of the experts valuation enables a neural network to estimate subjective probabilities and, with this coding scheme, that a multilayer network performs gradual fusion from the conjunctive mode (agreement between experts) to the disjunctive mode.

3.2 Membership function estimation and gradual fusion

The criterion usually minimized by the output of a multilayer network is the squared error Q defined as:

$$Q = \frac{1}{N}\sum_{n=1}^{N}\left\| t_n - g(\underline{x}_n,\theta)\right\|^2, \tag{16}$$

where $g(\underline{x}_n,\theta)$ and θ are the output and the weights of the network, respectively. If a pool of experts is used, this error can be rewritten as:

$$Q = \frac{1}{N}\frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N}\left\| t_{n,k} - g(\underline{x}_n,\theta)\right\|^2, \tag{17}$$

where K is the number of experts and $t_{n,k}$ the label of the n^{th} observation for the k^{th} expert. Note that this function is minimized when

$$g(\underline{x}_n,\theta) = \frac{1}{K}\sum_{k=1}^{K} t_{n,k}\ . \tag{18}$$

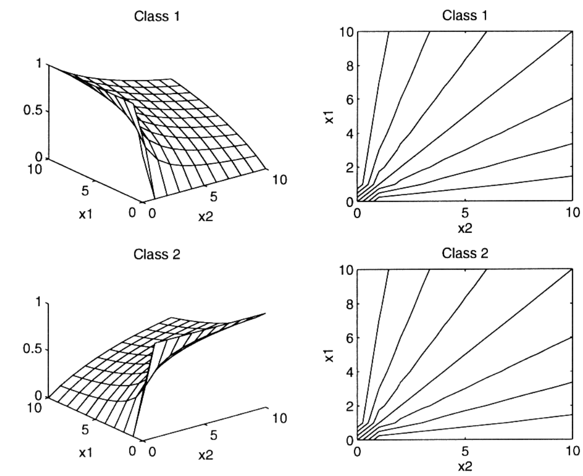
Hence, a crisp (0,1) coding of experts judgements and the universal approximator property of multilayer networks allows to estimate the subjective probabilities (membership functions) via the minimization of Eq. 17. Note that Eq. 16 (so does Eq. 17) is of the form of Eq. 7 in which the $\mu_{A_{1/2}}(x)$ crisp coding has been replaced by a binary coding of experts judgements. Therefore, minimizing Eq. 16 (or Eq. 17) is equivalent to minimizing a fuzziness measure of the concept to be learnt. The gradual fusion property is now illustrated. In the following experiment, observations (x_1,x_2) are assumed to be members of the set $\{0,1,\ldots,10\}\times\{0,1,\ldots,10\}$. For each observation (x_1,x_2), one supposes that x_1 experts vote in favor of class 1 and x_2 in favor of class 2. Each observation is presented to the network x_1 times as an observation from class 1 and x_2 times as an observation from class 2. Obviously, according to Eq. 18, the network output is an estimation of the membership functions defined by:

$$\mu_1(x_1,x_2) = \frac{x_1}{x_1+x_2} \text{ and } \mu_2(x_1,x_2) = \frac{x_2}{x_1+x_2}, \tag{19}$$

with

$$\mu_1(0,0) = \mu_1(0,0) = 1/2\,. \tag{19'}$$

Figure 1 shows interpolated theoretical and estimated membership functions obtained with a one hidden layer (10 neurons) network.

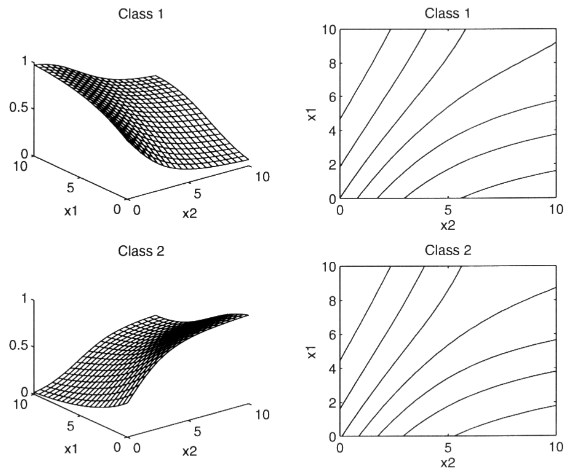


(a)

Figure 1 (a). Theoretical membership functions.

As can be seen, the estimated membership functions closely resembles the theoretical ones. The diagonal line represent the most disjunctive case when equal number of experts have chosen class 1 and class 2. As expected, membership estimates are very close to 0,5 on this line. The most conjunctive case can be seen for observations

(0,10) or (10,0).



(b)

Figure 1 (b). Estimated membership functions.

As can be seen, the estimated membership functions closely resemble the theoretical ones. The diagonal line represents the most disjunctive case when equal number of experts have chosen class 1 and class 2. As expected, membership estimates are very close to 0,5 on this line. The most conjunctive case can be seen for observations (0,10) or (10,0). For example, the observation (0,10) for which no expert chose class 1 and 10 experts chose class 2 has estimated membership degrees very close to 0 for class 1 and 1 for class 2, respectively.

4. APPLICATION TO CAR BEHAVIOR EVALUATION

The objective is to perform an automated diagnosis of 6 possible behavior defects of a car:

1- engine or wheel vibration;

2- on seat compression;

3- pumping;

4- transversal movements (high roll);

5- copying (low roll in relation with the road transversal profile);

6- on seat movements.

During an on-road experiment, experts give on-line impressions. The training data set is composed of time intervals during which experts evaluate their impressions using usual vocabulary (linguistic variables). The absence of synchronicity between phenomena and expert judgements implied a manual synchronization of signals. For the characterization of defects, several accelerometers were set up into the car. The resulting signals were judiciously combined in order to compress information. Then representative features, mainly related to the time-frequency power content, were extracted from these synthetic signals. Experts valuations were crisply coded as indicated previously.

A major difference between usual classification and our application is that more than one defect can be present at any given time instant. Therefore, the following usual relation

$$\sum_{i=1}^{Nclass} \mu_i(\underline{x}) = 1$$

does not still hold. This is the reason why a network was dedicated to each defect. The networks structures and weights were determined according to the internal representation optimization algorithm proposed in [13]. A cross validation procedure was used in order to cope with the bias-variance dilemma. The resulting networks were constituted of a single hidden layer of 9, 5, 7, 2, 8, 5 neurons for the defects listed previously, respectively.

After the optimization process, the fuzziness measures were estimated for each defect. The results obtained on the cross validation set are shown in figure 2.

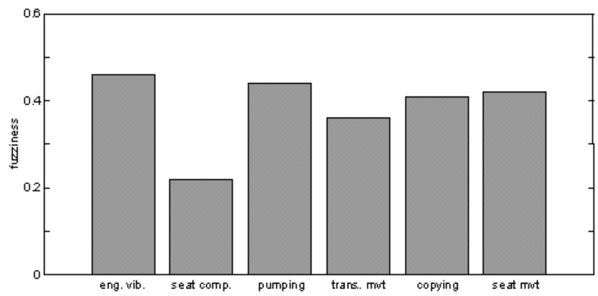


Figure 2. Fuzziness measures.

One observes that fuzziness is much smaller for the “on seat compression” defect than for others. This was confirmed by experts who agreed on the feeling that on seat compression can be less ambiguously detected than other defects. Furthermore, networks output time representations have shown that the transitions between absence and presence of on seat compression are much sharper than for every other defect.

Figure 3 represents an example of the networks outputs as a function of time during a time interval where on seat movements were detected twice by an expert (bold line). It also appears “engine or wheel vibrations” during the first period labeled "on seat movements". This can be explained by the similar frequency bands occupied by these phenomena. One can also observe that "on seat movements" decrease when a lower frequency phenomenon appears (pumping), as usually experienced by experts.

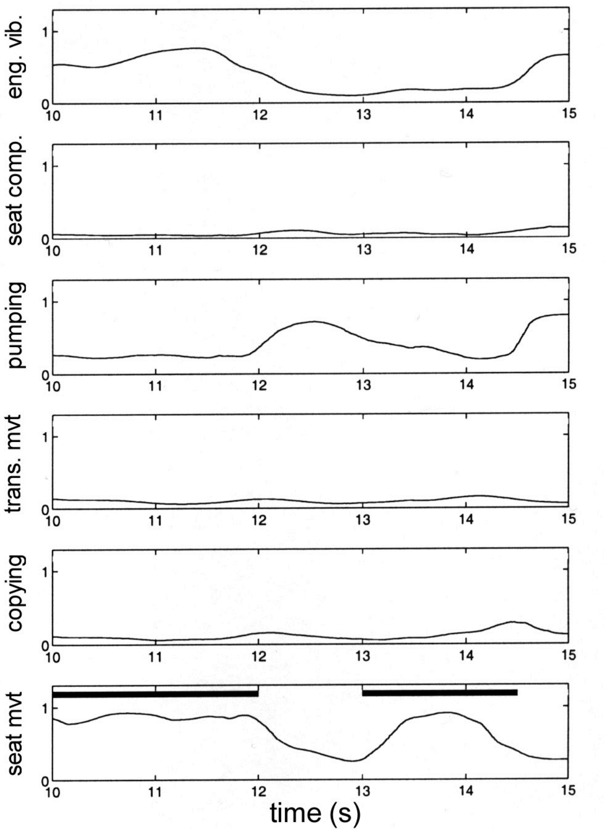


Figure 3. Example of membership functions estimation

5. CONCLUSIONS

In this paper, we have shown that fuzziness can be automatically extracted from the concept to be learnt. The proposed method relies on the universal approximator property of multilayer networks. Using a crisp coding of the experts evaluations, minimization of a fuzziness measure allows to make the least fuzzy decision according to the data. Furthermore, this method takes into account every kind of uncertainty (probabilistic, resolutional, fuzziness), the separate contribution of which, however, cannot be extracted. We have also shown that the decision rule automatically performs gradual fusion from the conjunctive mode (experts agreement) to the disjunctive mode (experts disagreement).

We have successfully applied this method to car behavior estimation where input data are vibration measurements and desired output (experts evaluations) are expressed in terms of linguistic variables. The results were correlated with the experts feeling. The *intelligent detectors* obtained after the networks training were implemented in real time on a car and successfully validated.

REFERENCES

[1] N. R. Pal and J.C. Bezdek: "Measuring fuzzy uncertainty", IEEE trans. On Fuzzy Systems, Vol. 2, pp.107-118, 1994.

[2] A. Kaufmann: "Introduction à la théorie des sous-ensembles flous", Masson, Paris, 1973.

[3] D. Dubois and H. Prade: "Fuzzy sets and systems: theory and applications", Academic Press, London, 1980.

[4] A. Deluca and S. Termini: "A definition of non probabilistic entropy in the setting of fuzzy sets theory", Information and Control, Vol. 20, pp. 301-312, 1972.

[5] B. R. Ebanks: "On measures of fuzziness and their representations", J. Math. Anal. And Appl., Vol. 94, pp. 24-37, 1983.

[6] V. N. Vapnik: "Estimation of dependences based on empirical data", Spronger-Verlag, New York, 1982.

[7] G. Cybenko: "Approximation by superpositions of a sigmoidal function", Mathematics of Control, Signal and Systems, Vol. 2, nr 4, pp. 303-314, 1989.

[8] K. Hornik: "Approximation capabilities of multilayer feedforward networks", Neural Networks, Vol. 4, pp. 251-257, 1991.

[9] G. Shafer: "A mathematical theory of evidence", Princeton University Press, Princeton, New Jersey, 1976.

[10] L. A. Zadeh: "Fuzzy sets and common sense knowledge", Cognitive Science Report nr 21, University of California, Berkeley, CA, 1984.

[11] M. Nowakowska: "Fuzzy concepts in the social sciences", Behav. Sci., Vol. 22, pp. 107-115, 1977.

[12] H. M. Hersh and A. Caramazza: "A fuzzy set approach to modifiers and vagueness in natural languages", J. Exp. Psychol.: Gen., Vol. 105, pp. 254-276, 1976.

[13] R. Lengelle and T. Denoeux: "Training multilayer perceptrons layer by layer using an objective function for internal representations", Neural Networks, Vol. 9, nr 1, pp. 83-98, 1996.