Function approximation in the framework of evidence theory: A connectionist approach *

Thierry Denœux

U.M.R CNRS 6599 Heudiasyc
Université de Technologie de Compiègne
BP 20529 - F-60205 Compiègne cedex - France
email: Thierry.Denoeux@utc.fr

Abstract

We propose a novel approach to functional regression based on the Transferable Belief Model, a variant of the Dempster-Shafer theory of evidence. This method uses reference vectors for computing a belief structure that quantifies the uncertainty attached to the prediction of the target data, given the input data. The procedure may be implemented in a neural network with specific architecture and adaptive weights. It allows to compute an imprecise assessment of the target data in the form of lower and upper expectations. The width of this interval reflects the partial indeterminacy of the prediction resulting from the relative scarcity of training data.

1. Introduction

Neural network models such as multilayer perceptrons and radial basis function networks are powerful tools for non linear function approximation. If each pair of input and output vectors is assumed to be drawn from a certain joint probability distribution $f(\mathbf{x}, \mathbf{y})$, it is well known that minimizing the sum-of-squares criterion results asymptotically in finding the best approximation (in the least-squares sense) to the regression function, i.e. in estimating the conditional average of the target data, given the input data. However, this result holds only when the number of training vectors goes to infinity. For finite sample size, the accuracy of the prediction made for some input $\mathbf{x}$ depends on the local density of training vectors around $\mathbf{x}$. In some applications, the probability distribution of the data may even not be rigorously the same in the training and test sets, yielding very poor predictions for those input vectors situated far from training samples. For these reasons, it is usually very important to provide, together with the predictions, an assessment of their reliability, for example in the form of lower and upper bounds. Such an approach has been investigated in the context of radial basis function networks by Leonard et al. [5], who have proposed a heuristic method for determining confidence limits of the predictions, based on estimated residuals and local data density. However, their approach still assumes the input vector to be situated in the region of influence of a learnt prototype vector, a condition that may not be verified, especially in high dimensional spaces.

In this paper, we propose a new approach to this problem, based on the Dempster-Shafer (D-S) theory of belief functions [6, 7]. The main idea behind this approach is to introduce principles whereby the uncertainty pertaining to a prediction of the target data, given the input data, may be assessed and quantified. The formalism of *belief functions* introduced by Shafer [6] and justified axiomatically by Smets [7] is used for uncertainty representation. This formalism allows to introduce the concepts of *lower* and *upper* expectations, which can be used to describe the confidence that may be attached to the prediction, given the presence or absence of training data in a given region of the input space. An implementation of this method in a neural network architecture with adaptive weights is proposed and is demonstrated on simulated data.

The paper is organized as follows. The main definitions related to the D-S theory of evidence are first recalled in Section 2. We then show how this approach may be applied to classification and regression problems (Section 3). The connectionist implementation is then described in Section 4, and demonstrated in Section 5.

*To appear in *Proceedings of the 1997 International Conference on Neural Networks (ICNN'97)*, Houston, June 1997. IEEE.

2. The D-S theory of evidence

The idea of using belief functions for representing someone's feeling of uncertainty was first explored by Shafer [6], following the seminal work of Dempster [1] about upper and lower probabilities induced by multi-valued mappings. The use of belief functions as an alternative to subjective probabilities for representing uncertainty was later justified axiomatically by Smets [7], who introduced the Transferable Belief Model (TBM), providing a clear and coherent interpretation of the various concepts underlying the theory. A general scheme for applying this theoretical framework to pattern classification has been proposed by Denœux [3, 2]. In the following the main definitions pertaining to the theory of belief functions (according to the interpretations provided by the TBM) are briefly recalled.

Let ω be some variable of interest taking on values in a finite set Ω called the *frame of discernment*. Let us assume that an agent entertains beliefs concerning the value of ω, given a certain evidential corpus. We postulate that these beliefs may be represented by a *belief structure* (or *belief assignment*), i.e. a function from 2^Ω to $[0, 1]$ verifying $\sum_{A\subseteq\Omega} m(A) = 1$ and $m(\emptyset) = 0$. For all $A \subseteq \Omega$, the quantity $m(A)$ represents the mass of belief allocated to the proposition "$\omega \in A$", and that cannot be allocated to any strict sub-proposition because of lack of evidence. The subsets A of Ω such that $m(A) > 0$ are called the *focal elements* of m. The information contained in a belief structure may be equivalently represented as a *belief function* bel, or as a *plausibility function* pl, defined respectively as $\text{bel}(A) = \sum_{B\subseteq A} m(B)$ and $\text{pl}(A) = \sum_{B\cap A\neq\emptyset} m(B)$. The quantity $\text{bel}(A)$, called the *credibility* of A, is interpreted as the total degree of belief in A (i.e. in the proposition "$\omega \in A$"), whereas $\text{pl}(A)$ denotes the amount of belief that could potentially be transfered to A, taking into account the evidence that does not contradict that hypothesis.

Let us now assume that two distinct pieces of evidence induce two belief structures m_1 and m_2. The *orthogonal sum* of m_1 and m_2, denoted as $m = m_1 \oplus m_2$ is defined as:

$$m(A) = K^{-1} \sum_{B\cap C=A} m_1(B)m_2(C) \qquad (1)$$

with $K = \sum_{B\cap C\neq\emptyset} m_1(B)m_2(C)$, for $A \neq \emptyset$, and $m(\emptyset) = 0$. The orthogonal sum (also called *Dempster's rule of combination*) is commutative and associative. It plays a fundamental operation for combining different evidential sources in evidence theory.

Decision making is an important issue in any theory of uncertainty. In the TBM, a distinction is made between two levels of uncertainty representation: a *credal* level at which beliefs are entertained and represented using the formalism of belief functions, and a *decision* level at which belief functions are converted to probability distributions to allow coherent betting behaviors [7]. Given a belief structure m, the Generalized Insufficient Reason Principle [7] leads to the definition of the *pignistic* probability distribution BetP as $\text{BetP}(\omega) = \sum_{A\ni\omega} \frac{m(A)}{|A|}$, where $|A|$ denotes the cardinality of A. According to the TBM, pignistic probabilities should be used for computing expected utilities or losses in a decision context. Another approach to decision making in D-S theory consists in considering the set $\mathcal{C}$ of probability distributions P *compatible* with belief function bel, i.e., verifying

$$\text{bel}(A) \leq P(A) \leq \text{pl}(A)$$

$\forall A \subseteq \Omega$. The lower and upper expectations of a function $f : \Omega \mapsto \mathbb{R}$ are then defined respectively as:

$$\begin{aligned} \mathbb{E}_*(f) &= \min_{P\in\mathcal{C}} \mathbb{E}_P(f) && (2)\\ &= \sum_{A\subseteq\Omega} m(A) \min_{\omega\in A} f(\omega) && (3)\\ \mathbb{E}^*(f) &= \max_{P\in\mathcal{C}} \mathbb{E}_P(f) && (4)\\ &= \sum_{A\subseteq\Omega} m(A) \max_{\omega\in A} f(\omega) && (5) \end{aligned}$$

where $\mathbb{E}_P(\cdot)$ denotes expectation relative to P. These definitions lead to different possible decision strategies, based on minimization of the lower or upper expected loss.

3. Application to classification and regression

3.1 Classification

We consider the task of classifying a feature vector $\mathbf{x}$ into one of M predefined groups or categories [3, 2]. The variable of interest is then the class ω of that pattern, and the frame of discernment is the set $\Omega = \{\omega_1, \ldots, \omega_M\}$ of classes. The available evidence is supposed to consist in a set of n representative patterns or *prototypes* $\mathbf{p}^1, \ldots, \mathbf{p}^n$. We do not for the moment consider the way in which this set of patterns has been obtained from training data. Each prototype $\mathbf{p}^i$ is assumed to have some known degree of membership u^i_j to each class ω_j, with $\sum_{j=1}^M u^i_j = 1$. The membership vector of prototype i is denoted as $\mathbf{u}^i = (u^i_1, \ldots, u^i_M)$.

Each data item $(\mathbf{p}^i, \mathbf{u}^i)$ constitutes a piece of evidence regarding the class of feature vector $\mathbf{x}$. This evidence may be assumed to induce a belief structure $m^i(\cdot|\mathbf{x})$ on Ω with focal elements $\{\omega_j\}$ for $j \in$

$\{1, \ldots, M\}$ and Ω. The mass of belief assigned to $\{\omega_j\}$ is postulated to be proportional to the degree of membership of $\mathbf{p}^i$ to class ω_j, and a decreasing function of the distance (according to some relevant metric δ) between vectors $\mathbf{x}$ and $\mathbf{p}^i$:

$$\begin{aligned} m^i(\{\omega_j\}|\mathbf{x}) &= u_j^i \phi^i(\delta(\mathbf{x}, \mathbf{p}^i)) \quad 1 \leq j \leq M \\ m^i(\Omega|\mathbf{x}) &= 1 - \phi^i(\delta(\mathbf{x}, \mathbf{p}^i)) \end{aligned}$$

where ϕ is a decreasing function verifying $\phi^i(0) \leq 1$ and $\lim_{d \to \infty} \phi^i(d) = 0$. In [2], δ was chosen to be the Euclidean distance, and ϕ^i was assumed to be of the form $\phi^i(d) = \alpha^i \exp(-\gamma^i d^2)$ with $0 \leq \alpha^i \leq 1$ and $\gamma^i > 0$. However, any other metric and family of functions could be considered as well.

As a result of the consideration of the n prototypes, n belief structures may be defined. These structures are obtained from distinct sources of information and may therefore be combined using Dempster's rule to yield a global belief structure $m(\cdot|\mathbf{x}) = m^1(\cdot|\mathbf{x}) \oplus \ldots \oplus m^n(\cdot|\mathbf{x})$ summarizing our final belief concerning the class of $\mathbf{x}$. The focal elements of m are the singletons of Ω, and Ω itself. In the decision phase, pattern $\mathbf{x}$ may be assigned to the class of maximum credibility, which is also in this case the one with greatest plausibility. More sophisticated decision strategies are discussed in [4].

3.2 Regression

The above procedure may be transposed to regression problems in the following way. We consider the task of predicting the value of a continuous variable y based on an input vector $\mathbf{x}$. The data generating process is unknown, but it may be assumed to be properly described by a joint probability distribution $f(\mathbf{x}, y)$. Let $\mathcal{Y}$ denote the set of values taken by variable y. As an approximation to y, we consider a discrete variable $\tilde{y}$ obtained by partitioning $\mathcal{Y}$ into M disjoint subsets $\omega_1, \ldots, \omega_M$, and associating to each ω_i a representative value y_i.

Let us now denote by Ω the set $\{\omega_1, \ldots, \omega_M\}$, and let us assume that we have gathered some evidence regarding the value of y associated to some input vector $\mathbf{x}$. Then, according to the TBM, the belief induced by this evidence may be represented by a belief structure. Ideally, the frame of discernment for that belief structure should be $\mathcal{Y}$, the domain of variation of variable y. However, the theory of evidence is much more complex when considering continuous frames, so we prefer to define a belief structure on the discretized version Ω of $\mathcal{Y}$.

Let $m(\cdot|\mathbf{x})$ denote that structure, assumed to be obtained by comparing $\mathbf{x}$ to n prototypes as explained in the previous section. The focal elements of $m(\cdot|\mathbf{x})$ are thus $\{\omega_i\}$ for $i = 1, \ldots, M$ and Ω. Then, the upper and lower expectations of $\tilde{y}$ are defined respectively as:

$$\begin{aligned} \mathbb{E}_*(\tilde{y}|\mathbf{x}) &= \sum_{i=1}^{M} m(\{\omega_i\}|\mathbf{x}) y_i + m(\Omega|\mathbf{x}) \min_i y_i \\ \mathbb{E}^*(\tilde{y}|\mathbf{x}) &= \sum_{i=1}^{M} m(\{\omega_i\}|\mathbf{x}) y_i + m(\Omega|\mathbf{x}) \max_i y_i \end{aligned}$$

The expectation of $\tilde{y}$ relative to the pignistic probability distribution is:

$$\mathbb{E}_{bet}(\tilde{y}|\mathbf{x}) = \sum_{i=1}^{M} \mathrm{BetP}(\omega_i|\mathbf{x}) y_i \qquad (6)$$

with $\mathrm{BetP}(\omega_i|\mathbf{x}) = m(\{\omega_i\}|\mathbf{x}) + m(\Omega|\mathbf{x})/M$. The quantity $\mathbb{E}_{bet}(\tilde{y}|\mathbf{x})$ may thus be considered as a point prediction of y given $\mathbf{x}$, while the interval $[\mathbb{E}_*(\tilde{y}|\mathbf{x}), \mathbb{E}^*(\tilde{y}|\mathbf{x})]$ constitutes an *imprecise* assessment of the target variable.

4. Neural network implementation

4.1 Architecture

A possible neural network implementation of the above procedure is represented in Figure 1. The first hidden layer (L_1) is similar to the hidden layer of a RBF network with Gaussian activation function. The activation of neuron i in that layer is

$$\phi^i = \alpha^i \exp(-\gamma^i \|\mathbf{x} - \mathbf{p}^i\|^2).$$

The second hidden layer (L_2) computes the belief structure m^i associated to prototype $\mathbf{p}^i$. It is composed of n modules of $M+1$ units. The outputs from the i-th module is a vector $\mathbf{m}^i$ with

$$m_j^i = m^i(\{\omega_j\}|\mathbf{x}) = u_j^i \phi^i$$

for $1 \leq j \leq M$ and

$$m_{M+1}^i = m^i(\Omega|\mathbf{x}) = 1 - \phi^i$$

The third hidden layer (L_3) performs the combination of the n belief structures from the former layer using the Dempster's rule *without normalization.* It is composed of n *interconnected* modules of $M+1$ sigma-pi units. The outputs $\mathbf{o}^i = (o_1^i, \ldots, o_{M+1}^i)$ from module i correspond to the unnormalized orthogonal sum of $m^1, \ldots, m^i$. They are computed iteratively using the outputs from module i of layer L_2 and module $i-1$ of layer L_3:

$$o_j^i = o_j^{i-1} m_j^i + o_j^{i-1} m_{M+1}^i + o_{M+1}^{i-1} m_j^i \qquad (7)$$

for $1 \leq j \leq M$ and

$$o_{M+1}^i = o_{M+1}^{i-1} m_{M+1}^i \qquad (8)$$

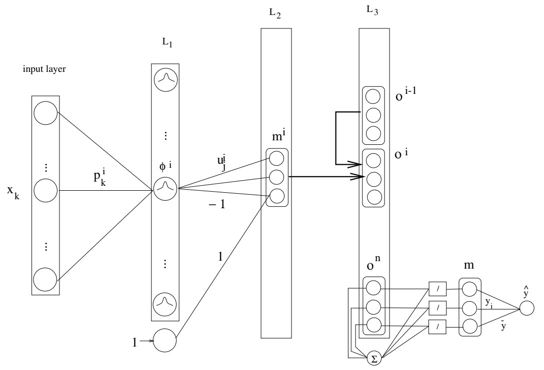

Figure 1. Architecture of the Belief Function Neural Network.

with $o_j^1 = m_j^1$ for $1 \leq j \leq M+1$.

The output from layer L_3 is the vector $\mathbf{o}^n$ of activations in the last module of that layer. Its values are the belief masses corresponding to the unnormalized orthogonal sum of $m^1, \ldots, m^n$. The normalized output vector $\mathbf{m}$ is obtained by normalizing the components of $\mathbf{o}^n$ to unity: $\mathbf{m} = \mathbf{o}^n / K$ with $K = \sum_{j=1}^{M+1} o_j^n$.

Finally, the output layer is composed of a single neuron with activation $\hat{y}$ defined as:

$$\hat{y} = \sum_{i=1}^{M} y_i m_i + \bar{y} m_{M+1} \tag{9}$$

with $\bar{y} = \frac{1}{M} \sum_{j=1}^{M} y_j$. Hence, the network output is exactly equal to the expectation of $\tilde{y}$ relative to the pignistic probability distribution as defined by Eq. 6. Once the input signal has been propagated in the network, the lower and upper expectations of $\tilde{y}$ may be readily obtained as $\mathbb{E}_*(\tilde{y}|\mathbf{x}) = \sum_{i=1}^{M} y_i m_i + m_{M+1} \min_i y_i$ and $\mathbb{E}^*(\tilde{y}|\mathbf{x}) = \sum_{i=1}^{M} y_i m_i + m_{M+1} \max_i y_i$.

4.2 Training

The training of the above network may be performed in two phases: (1) initialization and (2) minimization of an error criterion.

First of all, we start by defining M classes by partitioning the output space $\mathcal{Y}$ so as to have approximately the same number samples in each class (more sophisticated discretization techniques could also be used). Assuming $\mathcal{Y}$ to be of the form $[y_{min}, y_{max}]$, we thus find $M-1$ thresholds $\xi_1, \ldots, \xi_{M-1}$ and define $\omega_1 = [y_{min}, \xi_1[$, $\omega_i = [\xi_{i-1}, \xi_i[$ for $i = 1, \ldots, M-1$, and $\omega_M = [\xi_{M-1}, y_{max}]$. The initial value of y_i may then be set to the middle of interval ω_i, for $i = 1, \ldots, M$.

The next step in the initialization phase consists in the construction of a set of n prototypes using a clustering or vector quantization procedure such as the c-means algorithm. The degree of membership of prototype i to class ω_j may then be initialized as the proportion of training pattern from that class in the region of influence V^i of prototype i. As a heuristic, the initial value of the scale parameter γ^i may be set equal to the inverse of the mean squared distance between prototype $\mathbf{p}^i$ and training patterns in V^i.

Once the network parameters have been initialized, the next step consists in the definition of an error function to be minimized. In our case, since we are performing classification *and* function approximation simultaneously, we shall define the total error for pattern $\mathbf{x}$ as a weighted sum of a classification error E_c and a regression error E_r:

$$E(\mathbf{x}) = \nu E_c(\mathbf{x}) + (1-\nu) E_r(\mathbf{x}) \tag{10}$$

where $0 \leq \nu \leq 1$ is a parameter controlling the trade-off between both types or error. A natural choice for the regression error is $E_r(\mathbf{x}) = (y - \hat{y})^2$, whereas the classification error may be defined as:

$$E_c(\mathbf{x}) = \sum_{j=1}^{M} (\mathrm{BetP}(\omega_j|\mathbf{x}) - t_j)^2 \tag{11}$$

where t_j is a binary variable indicating the true membership of $\mathbf{x}$ to class ω_j ($t_j = 1$ if $\mathbf{x} \in \omega_j$, and $t_j = 0$ otherwise), and $\mathrm{BetP}(\omega_j|\mathbf{x})$ is the pignistic probability of class ω_j computed for pattern $\mathbf{x}$, which is equal to $m(\{\omega_j\}|\mathbf{x}) + m(\Omega|\mathbf{x})/M$.

A mean output error E can be computed by averaging $E(\mathbf{x})$ for all $\mathbf{x}$ in the training set. Since the prototypes vectors should remain in regions of high data density, they are not changed during the optimization process. The free parameters are therefore γ^i, α^i and u_j^i for $i = 1, \ldots, n$ and $j = 1, \ldots, M+1$. The constraints are $\gamma^i > 0$, $0 < \alpha^i < 1$ and $\sum_{j=1}^{M} u_j^i = 1$ for all $1 \leq i \leq n$. These constraints are automatically satisfied by introducing new parameters η^i, ξ^i and β_q^i such that

$$\begin{aligned} \gamma^i &= (\eta^i)^2 \\ \alpha^i &= (1 + \exp(-\xi^i))^{-1} \\ u_q^i &= (\beta_q^i)^2 / \sum_{k=1}^{M} (\beta_k^i)^2 \end{aligned}$$

and minimizing E with respect to these new parameters. Calculation of the whole gradient can be performed in linear time with respect to the input dimension, the number M of classes and the number n of prototypes [2].

A common approach to avoid overfitting is to perform regularization. In our model, a way to moderate the importance of a prototype i is to decrease α^i: for $\alpha^i = 0$, we have $m^i(\Omega|\mathbf{x}) = 1$, and m^i no longer influences the result of the orthogonal sum. To obtain a smoother solution, it has also proved beneficial to avoid large absolute values of the y_i. We therefore add to the error function a regularization term C equal to:

$$C = \sum_{i=1}^{n} \alpha^i + \sum_{j=1}^{M} y_j^2 \tag{12}$$

The new error function to be minimized then becomes $J = E + \mu C$, where μ is a hyper-parameter, the optimal value of which may be determined by cross-validation.

5. Example

As an example, we consider a regression problem where the input $x \in \mathbb{R}$ is taken from a mixture of two Gaussian distributions:

$$f(x) \sim 0.5\mathcal{N}(-2, 0.25) + 0.5\mathcal{N}(2, 0.25).$$

The target variable y is defined as $y = \sin(3x) + x + \varepsilon(x)$, where $\varepsilon(x)$ is a Gaussian white noise with variance 0.01 if $x \leq 0$ and 1 if $x > 0$. A training set of $N = 150$ samples was generated from that distribution.

Figure 2 shows a result of our simulations with $n = 20$, $M = 7$, $\nu = 0.5$ and $\mu = 0.005$. As can be seen in that example, the absence of training data around $x = 0$ is correctly reflected by a large difference between lower and upper expectations. Note that the prediction interval does not become larger in the region of large variance of $\varepsilon(x)$, since the width of this interval essentially indicates uncertainty about the conditional average of the target data resulting from low density of training data. The scatter of the target variable around the mean might be estimated independently, for example by approximating the squared prediction error $(\hat{y} - y)^2$.

6. Conclusions

We have proposed a novel approach to functional regression based on the Transferable Belief Model, a variant of the Dempster-Shafer theory of evidence. This method uses reference vectors for computing a belief structure that quantifies the uncertainty attached to the prediction of the target data, given the input data. The procedure may be implemented in a neural network with specific architecture and adaptive weights. It allows to compute an imprecise assessment of the target data in the form of lower and upper expectations. The width of this interval can be interpreted as reflecting the partial indeterminacy of the prediction resulting from the relative scarcity of training data. Detailed comparison between this method and other approaches based e.g. on confidence intervals in the classical sense such as described in [5] is under way and will be reported in future papers.

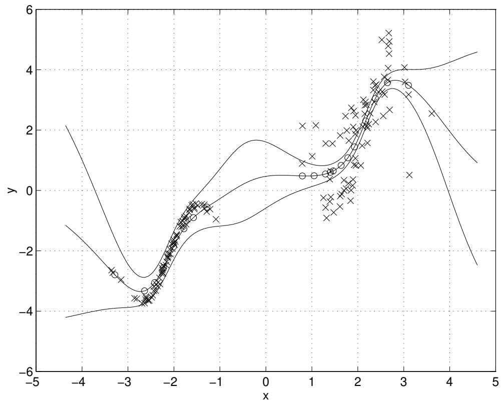

Figure 2. Plot of the lower, pignistic and upper expectations. Training points and prototypes are indicated by x and o, respectively.

References

[1] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, AMS-38:325–339, 1967.

[2] T. Denœux. An evidence-theoretic neural network classifier. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 712–717, Vancouver, October 1995.

[3] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

[4] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition (In press)*, 1997.

[5] J. A. Leonard, M. A. Kramer, and L. H. Ungar. Using radial basis functions to approximate a function and its error bounds. *IEEE Transactions on Neural Networks*, 3(6):624–627, 1992.

[6] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.

[7] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.