Statistical Aspects of Neural Networks

By

C.H. Hesse
Department of Statistics
University of California
Berkeley

Technical Report No. 140
December 1987
(under revision)

Department of Statistics
University of California
Berkeley, California

Statistical Aspects of Neural Networks

By

C.H. Hesse

Department of Statistics

University of California at Berkeley

Abstract

We demonstrate how the computational abilities of interconnected 0-1 networks (often referred to in the literature as neural networks) may be employed for the task of efficient combination of statistical evidence from several sources via the Dempster-Shafer theory of upper and lower probability systems. For a certain structure of the evidence (which allows it to be incomplete, inaccurate and partly contradictory) an algorithm is given which has linear time complexity. The work has applications to the handling of uncertainty in knowledge-based expert systems and quantitative knowledge-integration systems.

AMS(MOS) Subject Classification: 62P99, 90A19, 90B15

Key words: upper and lower probability systems, neural networks, expert system, knowledge-integration, parallel processing.

1. Introduction.

Np-complete is one of the new words of the computer age. It is used to characterize problems with respect to their computational complexity. It means, roughly, that a problem is so hard that no polynomial time algorithm that solves it is known. In other words, the time required to solve a problem of a certain size or input length on a given computer grows exponentially with the input length. The number of problems to be classified under this heading grows almost daily. Some of the more famous representatives of this class are the Traveling Salesman problem, the Hamiltonian Path problem, and the Steiner Tree problem (see Garey and Johnson (1979)).

Another problem that is known to be np complete is the combination of basic probability assignments in the Dempster-Shafer theory of upper and lower probability systems (henceforth called DS-theory). The theory was first introduced by Dempster in his 1967 paper and was subsequently extended and enriched by Shafer (1976). It may be regarded as a generalization of Bayesian inference. While Bayesian inference requires a global prior probability distribution for all relevant variables, and observations on some of these variables, no complete prior probability law is needed for DS-modelling. There is, of course, a price to be paid: no exact probabilities can be obtained from such inference, but only upper and lower probability systems (see Dempster (1968)).

The Dempster-Shafer theory is currently widely used among the Artificial Intelligentsia especially in knowledge-based expert systems (e.g. Garey et. al. (1981), Shafer (1984), Buchanan and Shortliffe (1984)) since it lends itself to the representation and combination of evidence from several independent sources and also allows to handle situations where knowledge is both incomplete, inaccurate and partly contradictory.

Section 2 gives a condensed description of the basic structures of DS-theory. For this we claim neither originality nor completeness. The reader should consult Shafer (1976) for an exposition of formal theory. In Section 2 we also give a new interpretation of upper and lower probability systems and DS theory in terms of positive definite functions on semigroups with involution.

Section 3 addresses the problem of representing and combining statistical evidence from several sources. In the DS formalism the main device for knowledge integration is Dempster's rule for the combination of basic probability assignments. We point out a defect of Dempster's rule and suggest a modification (for conflicting evidence only) that eliminates these shortcomings. For a certain structure of the evidence, frequently encountered in practice, an algorithm is then given which has linear time complexity, as measured in arithmetic operations. The innovation is to first map the problem onto a neural network and formulate it in terms of desired maxima of a functional defined over the state space of the network in such a way that the asynchronous parallel processing capabilities of these networks are exploited in a fashion similar to Hopfield

(1982).

2. The Dempster - Shafer Theory.

For many years there has been a need within the Artificial Intelligentsia for a consistent (non ad hoc) theory which can deal with incomplete knowledge, ignorance, uncertainty, and partial contradiction. The Dempster-Shafer theory is such a general purpose tool.

Standard probability models are characterized by a sample space and a probability measure over the sample space. The Dempster-Shafer theory requires only that probability be attached to certain margins (determined by the available evidence) of the sample space (which in this formalism is referred to as frame of discernment) and it operates with basic probability assignments. Formally these have the same structure as probability measures over the *power set* of the sample space (rather than the sample space itself). For a complete description of the formalism the reader is directed to Shafer (1976). Here we define some basic structures of this theory that we will need in the sequel. We confine ourselves to finite frames $\Theta = \{\theta_1, \ldots, \theta_n\}$.

Definition: (a) A non-negative function $m: 2^\Theta \to [0,1]$ from the power set 2^Θ of Θ to the unit interval is a *basic probability assignment* if

$$(2.1) \qquad \begin{aligned} m(\phi) &= 0 \\ \sum_{A \in 2^\Theta} m(A) &= 1 \end{aligned}$$

If $m(A) > 0$, then A is called a focal element.

(b) The *lower probability* $\underline{P}$: $2^{\Theta} \rightarrow [0,1]$ is defined as

$$(2.2) \qquad \underline{P}(A) = \sum_{B \subseteq A} m(B).$$

The simple set function with $\underline{P}(\Theta) = 1$ and $\underline{P}(A) = 0$ for all $A \neq \Theta$ is called the vacuous lower probability function.

(c) The *upper probability* $\overline{P}$: $2^{\Theta} \rightarrow [0,1]$ is defined as

$$(2.3) \qquad \overline{P}(A) = \sum_{B \cap A \neq 0} m(B)$$

(d) Let Θ, Ω be two frames of discernment. If there exists a mapping $w: 2^{\Theta} \rightarrow 2^{\Omega}$ with

$$(2.4) \qquad \begin{array}{ll} 1. & w(\{\theta_i\}) \neq \phi \text{ for all } \theta_i \in \Theta \\ 2. & w(\{\theta_i\}) \cap w(\{\theta_j\}) = \phi \text{ if } \theta_i \neq \theta_j \\ 3. & \bigcup_{\theta_i \in \Theta} w(\{\theta_i\}) = \Omega \end{array}$$

then Ω is a *refinement* of Θ and Θ is a *coarsening* of Ω. The mapping w is a *refining*. (When working with basic probability assignments in practice it will always be assumed that any frame Θ under consideration is embedded in a family of frames **F** consisting of refinements of Θ and coarsenings of the refinements of Θ in such a way, that every pair of elements in **F** has a common refinement in **F**, i.e. if $\Theta_1, \Theta_2 \in \mathbf{F}$, then there exists $w_1: 2^{\Theta_1} \rightarrow 2^{\Omega_1}$ and $w_2: 2^{\Theta_2} \rightarrow 2^{\Omega_2}$ such that $\Omega_1 = \Omega_2$.)

(e) Let $\Theta_1, \Theta_2 \in \mathbf{F}$ be two frames $w: 2^{\Theta_1} \rightarrow 2^{\Theta_2}$ be a refining and let $\underline{P}_2$ be a lower probability over Θ_2. Then $\underline{P}_1$ defined as

$$(2.5) \qquad \underline{P}_1(A) = \underline{P}_2(w(A)) \text{ for all } A \in 2^{\Theta_1},$$

is the *restriction* of $\underline{P}_2$ to Θ_1.

If instead $\underline{P}_1$ is a lower probability over Θ_1, then $\underline{P}_2$ defined by

$$(2.6) \qquad \underline{P}_2(A) = \max_{\substack{B \subseteq \Theta_1 \\ w(B) \subseteq A}} \underline{P}_1(B) \text{ for all } A \in 2^{\Theta_2}$$

is the *minimal extension* of $\underline{P}_1$ to Θ_2.

(It can be shown that $\underline{P}_1$ is the restriction of its minimal extension $\underline{P}_2$ to Θ_1.)

Dempster and Shafer interpret these entities in the following way: m(A) is a piece of probability which can move freely among all the elements θ_i of the set A and $\underline{P}(A)$ is the minimum total probability committed to A in the sense that it cannot move outside A. $\overline{P}(A)$ is the largest amount of probability that can move to a given $\theta_i \in A$. If all the focal elements of m are singleton sets than m reduces to a standard probability distribution over the sample space Θ and $\underline{P}(A)$ is simply the probability attached to the subset A of Θ. Since focal elements are not restricted to singleton elements, basic probability assignments are more general than standard probability distributions over sample spaces: They are probability distributions over the <u>power</u> <u>set</u> of the sample space Θ. This generalization greatly enhances their applicability and makes them much more useful in situations (encountered frequently in knowledge-based expert systems) where exact prior probability distributions are difficult to come by due to limited knowledge, ignorance, uncertainty, or partial contradiction.

The Dempster-Shafer theory retains the same basic structure of conditioning on observed data as the Bayesian theory but introduces Dempster's rule of combination (see part (f) of the definition) as an updating scheme for upper and lower probability systems:

Definition (continued): (f) If m_1, m_2 are basic probability assignments for $\underline{P}_1$ and $\underline{P}_2$ over the same frame Θ, and if $A_1, \ldots, A_k$ and $B_1, \ldots, B_l$, respectively, are their focal elements then the function m: $2^{\Theta} \rightarrow [0, 1]$ defined as

$$(2.4) \qquad m(A) = \frac{\sum\limits_{A_i \cap B_j = A} m_1(A_i) \cdot m_2(B_j)}{1 - \sum\limits_{A_i \cap B_j = \phi} m_1(A_i) \cdot m_2(B_j)}, \quad A \in 2^{\Theta} \text{ for } \sum\limits_{A_i \cap B_j = \phi} m_1(A_i) \cdot m_2(B_j) < 1$$

is the *orthogonal sum of* m_1 *and* m_2 and is denoted by $m_1 \oplus m_2$.

(It can be shown that $m_1 \oplus m_2$ is a basic probability assignment over Θ. The corresponding lower probability is denoted by $P_1 \oplus P_2$).

If one interprets $\underline{P}_1$ and $\underline{P}_2$ as being derived from independent sources of statistical evidence, then $\underline{P}_1 \oplus \underline{P}_2$ is the lower probability function representing the combined evidence.

Dempster's rule is the fundamental updating mechanism in DS-theory. It describes how incoming new evidence changes upper and lower probabilities: Initial probability bounds are formalized by $\underline{P}_1$ over Θ, new evidence is represented by $\underline{P}_2$ over Θ and if the requirements for the application of Dempster's rule are satisfied, then our new probabilities are represented by $\underline{P}_1 \oplus \underline{P}_2$.

The Dempster-Shafer theory is especially suited to deal with situations where standard Bayesian inference does not succeed because prior probabilities for the entire space are difficult to come by. DS-modelling allows to exactly represent the available evidence, however limited and contradictory it might be, without having to come up with prior probabilities for every possible outcome. For example, in an inference problem that involves integers the structure of apriori knowledge might be such that it allows to specify prior probabilities only for the sets of pairs $\{1,2\}$, $\{3,4\}$,$\{5,6\}$, etc., rather than for each integer individually.

For this and other reasons DS-modelling has frequently been used by the AI community (e.g. Barnett (1981), Garey et. al. (1981), Shafer (1984)) in knowledge-based expert systems. The medical diagnostic system MYCIN (Buchanan and Shortliffe (1984)), for example, operates with DS-theory.

The history of upper and lower probability systems goes back further than indicated by the above references and an extensive literature exists. The following is a partial list only: Boole (1854), Good (1950), Smith (1961, 1965), West (1971) and Beran (1971a,b). There is also a connection to Choquet (1953); in this sense upper probabilities $\overline{P}$ are alternating Choquet capacities of order ∞.

Before going on to Section 3 we mention here that upper and lower probability systems and the Dempster rule fit naturally into the theory of positive definite functions on semigroups with involution.

A semigroup $(S, \circ)$ is a non-empty set S equipped with an associative composition $\circ$ and a neutral element. A semigroup $(S, \circ)$ together with a mapping $* : S \rightarrow S$ that satisfies

$$\text{(i)} \quad (s \circ t)^* = t^* \circ s^* \quad \text{for } s, t \in S$$
$$\text{(ii)} \quad (s^*)^* = s \qquad \text{for } s \in S$$

is a semigroup with involution. A function $\psi : S \rightarrow \mathbb{C}$ is positive definite if $(s, t) \mapsto \psi(s^* \circ t)$ satisfies

$$\sum_{j,k=1}^{n} c_j \bar{c}_k \psi(s_j^* \circ s_k) \geq 0 \text{ for all } n \in \mathbb{N}, \{s_1, \ldots, s_n\} \subseteq S \text{ and } \{c_1, \ldots, c_n\} \subset \mathbb{C}.$$

For a frame Θ the power set 2^Θ equipped with the usual intersection of sets, $(2^\Theta, \cap)$ is a commutative, idempotent semigroup. If $m_1, m_2 \in \mathbb{R}^\Theta$ (where $\mathbb{R}^\Theta$ is the set of functions from 2^Θ to $\mathbb{R}$) then

$$m_1 \bullet m_2(A) = \sum_{A_1 \cap A_2 = A} m_1(A_1) \cdot m_2(A_2)$$

is a convolution and $(\mathbb{R}^\Theta, \bullet)$ is again a semigroup. If one requires normalization in $\mathbb{R}^\Theta$, specifically

$$\sum_{A \in 2^\Theta} m(A) = 1, \; m(\phi) = 0, \; m(A) \geq 0 \text{ for all } A \in 2^\Theta$$

then one is in the context of basic probability assignments. Then obviously

$$m_1 \oplus m_2(A) = \frac{m_1 \bullet m_2(A)}{1 - m_1 \bullet m_2(\phi)}$$

is the Dempster combination of basic probability assignments in terms of the convolution $\bullet$. For the basic definitions, see Berg et al. (1984). Harmonic analysis on the semigroup $(\mathbb{R}^{\Theta}, \bullet)$ amounts to a change of basis in $\mathbb{R}^{\Theta}$. The new basis is chosen such that the convolution can be expressed as a simple multiplication which further simplifies both the structure and computations.

Furthermore, if $(2^{\Theta}, \cap, \subseteq)$ is a semigroup $(2^{\Theta}, \cap)$ equipped with the partial order $\subseteq$ set inclusion then for this partial order the Riemann Zeta function ξ is such that for two subsets A, B of Θ, $\xi(A,B) = 1$ if $A \subseteq B$ and $\xi(A,B) = 0$ otherwise. The corresponding Möbius function μ is given by $\mu(A,B) = (-1)^{|B-A|}$ for $A \subseteq B$ and $\mu(A,B) = 0$ otherwise. Here $|A|$ denotes the cardinality of the set A.

Since $\underline{P}(B) = \sum_{A \subseteq B} m(A) = \sum_{A \in 2^{\Theta}} \xi(A,B)\, m(A)$, lower probabilities are Möbius transforms of basic probability assignments and therefore, one may obtain the function m from $\underline{P}$ via Möbius inversion:

$$m(A) = \sum_{B \subseteq 2^{\Theta}} \mu(B,A) \cdot \underline{P}(B) = \sum_{B \subseteq A} (-1)^{|A-B|} \underline{P}(B).$$

In this setting many of the results of harmonic analysis on semigroups become available in the theory of basic probability assignments over frames of discernment.

3. Combining Statistical Evidence.

Let $X_1, \ldots, X_N$ be a set of nodes and C_{ij}; $i,j \in \{1, \ldots, n\}$ a set of edges or connectivities between pairs of nodes. In many applications nodes will be representative of a set of variables, propositions, attributes etc. and connectivities may represent pairwise interactions in the form of joint distributions, upper and lower probability systems, classical or non-monotonic logical relations etc. These network-based knowledge structures appear frequently in the construction and design of expert systems, quantitative knowledge-integration systems and in the theory of reliability (e.g. Buchanan and Shortliffe (1984), Barlow et. al. (1975)).

Consider now, in particular, a set of propositions each one of which has the possible truth values "true" (state 1) and "false" (state 0) and the context where the connectivities are given by basic probability assignments over bivariate frames. In addition, propositions may interact with each other in one of the following ways

(3.1) Proposition j being true gives some support to proposition i being true also.

(3.2) Proposition k being true gives some support to proposition l being false.

In the language of DS-theory, (3.1) and (3.2) translate into the basic probability assignments m_{ij} and m_{lk}, respectively, with focal elements

(3.3) $m_{ij}(\{(1,1), (1,0), (0,0)\}) = r_{ij}, m_{ij}(\{(0,1) \times (0,1)\}) = 1 - r_{ij}$ where in $(\cdot,\cdot)$ the first coordinate represents proposition i and the second coordinate represents proposition j.

(3.4) $m^*_{lk}(\{(0,1), (0,0), (1,0)\}) = s_{lk}$, $m^*_{lk}(\{(0,1) \times (0,1)\}) = 1 - s_{lk}$ where the first coordinate in $(\cdot,\cdot)$ represents proposition l and the second coordinate represents proposition k.

Evidence with the structure (3.1), (3.3) or (3.2), (3.4) is by no means artificial. An example for (3.3) is the percolation of water through a system of sites and valves:

The links (edges) denote valves which allow water to flow from site to site. Each valve is open independently of other valves with probability p (otherwise closed) in the direction indicated by the arrow. Hence, if site (3,5), say is wet, then site (4,5) is wet with probability p. However, if (3,5) is dry (4,5) may still be wet due to a connection with (4,4). The relevant basic probability assignment has the same focal elements as in (3.3).

It gives further insight to note that in the above sense basic probability assignments bridge the gap between probabilistic and logical relations. Probabilistic relations between random variables are represented either by joint distributions or conditional distributions. Logical relations, on the other hand, correspond to subsets of the joint outcome space: For example, if X and Y are two Boolean variables then the implication "if X = 1 then Y = 1" may be represented by the subset $\{(1,1),(1,0),(0,0)\}$ of the product space $\{(0,1) \times (0,1)\}$. The basic probability assignment (3.3) attaches a probability to this subset and hence allows to randomize logical relations.

We now address the following rather general problem: For a large and possibly highly interconnected network of nodes and connectivities of type (3.1) and (3.2) inferences about the states of nodes both locally (i.e. for an individual node) and globally (i.e. for the set of nodes as a whole) are desired given the states of certain nodes.

(3.5)

In (3.5) ● denotes a node with known state. The basic building block of the network (3.5) is the unit

(3.6)

where the links (the respective probability assignments on the edges) and the states of X_1 and X_2 are given and these interact with X_3 through (3.1) or (3.2).

Consider first the case where both interactions are of type (3.1), i.e.

$$m_{31}(\{(1,1),\ (1,0),\ (0,0)\}) = r_{31},\ m_{31}(\{(0,1) \times (0,1)\}) = 1 - r_{31} \tag{3.7}$$

$$m_{32}(\{(1,1),\ (1,0),\ (0,0)\}) = r_{32},\ m_{32}(\{(0,1) \times (0,1)\}) = 1 - r_{32}. \tag{3.8}$$

Since inferences about the state of X_3 are desired one needs to first minimally extend (Definition (e)) the corresponding probability assignments (3.7), (3.8) to the entire space $(0,1) \times (0,1) \times (0,1)$, combine them via Dempster's rule (Definition (f))over this space, and marginalize with respect to X_3 conditional on states of X_1 and X_2, respectively again via Dempster's rule.

Here and below in $(\cdot,\cdot,\cdot)$ the sequence of states refers to (X_3, X_2, X_1). Minimal extension of (3.7), (3.8) generates

$$m_{31}(\{(1,1,1),\ (1,0,1),\ (1,1,0),\ (1,0,0),\ (0,1,0),\ (0,0,0)\}) = r_{31}$$

$$m_{31}(\{(0,1) \times (0,1) \times (0,1)\}) = 1 - r_{31}$$

$$m_{32}(\{(1,1,1),\ (1,1,0),\ (1,0,1),\ (1,0,0),\ (0,0,1),\ (0,0,0)\}) = r_{32}$$

$$m_{32}(\{(0,1) \times (0,1) \times (0,1)\}) = 1 - r_{32}.$$

Dempster-combination gives rise to

$$m_{31} \oplus m_{32}(\{(1,1,1), (1,0,1), (1,1,0), (1,0,0), (0,0,0)\}) = r_{31} \cdot r_{32}$$
$$m_{31} \oplus m_{32}(\{(1,1,1), (1,1,0), (1,0,1), (1,0,0), (0,0,1), (0,0,0)\}) = r_{32}(1 - r_{31})$$
$$m_{31} \oplus m_{32}(\{(1,1,1), (1,0,1), (1,1,0), (1,0,0), (0,1,0), (0,0,0)\}) = r_{31}(1 - r_{32})$$
$$m_{31} \oplus m_{32}(\{(0,1) \times (0,1) \times (0,1)\}) = (1 - r_{31})(1 - r_{32})$$

and conditioning on $X_1 = 1$, $X_2 = 1$ leads to

$$m_{31} \oplus m_{32}(\{(1,1,1)\}) = r_{31} + r_{32} - r_{31} \cdot r_{32}$$
$$m_{31} \oplus m_{32}(\{(1,1,1), (0,1,1)\}) = 1 - r_{31} - r_{32} + r_{31} \cdot r_{32},$$

conditioning on $X_1 = 1$, $X_2 = 0$

$$m_{31} \oplus m_{32}(\{(1,0,1)\}) = r_{31}$$
$$m_{32} \oplus m_{32}(\{(1,0,1), (0,0,1)\}) = 1 - r_{31},$$

hence producing the expected results in this simple case. Similar relations are obtained when both basic probability assignments are of type (3.2). Consider now the case of conflicting evidence. Specifically, let

$$m_{31}(\{(1,1), (1,0), (0,0)\}) = r_{31}$$
$$m_{31}(\{(1,1) \times (0,1)\}) = 1 - r_{31}$$
$$m_{32}^*(\{(0,1), (0,0), (1,0)\}) = s_{32}$$
$$m_{32}^*(\{(0,1) \times (0,1)\}) = 1 - s_{32}$$

then after minimal extension, Dempster combination and

conditioning on $X_1 = 1$, $X_2 = 1$:

$$m_{31} \oplus m_{32}^*(\{(1,1,1)\}) = \frac{r_{31}(1 - s_{32})}{1 - s_{32} r_{31}}$$

$$m_{31} \oplus m_{32}^{*}(\{(0,1,1)\}) = \frac{s_{32}(1 - r_{31})}{1 - s_{32} r_{31}}$$

conditioning on $X_1 = 1, X_2 = 0$:

$$m_{31} \oplus m_{32}^{*}(\{(1,0,1)\}) = r_{31}$$
$$m_{31} \oplus m_{32}^{*}(\{(0,0,1)\}) = 0$$
$$m_{31} \oplus m_{32}^{*}(\{(0,0,1), (1,0,1)\}) = 1 - r_{31}$$

conditioning on $X_1 = 0, X_2 = 1$:

$$m_{31} \oplus m_{32}^{*}(\{(0,1,0)\}) = s_{32}$$
$$m_{31} \oplus m_{32}^{*}(\{(1,1,0)\}) = 0$$
$$m_{31} \oplus m_{32}^{*}(\{(0,1,0), (1,1,0)\}) = 1 - s_{32}.$$

These are the required computations for one basic building block only. Nodes may have more than two incoming links requiring repeated pairwise Dempster combination. It is clear that for a large and highly interconnected network the number of operations explodes.

Our goal is therefore to use the computational abilities (especially their ability for parallel processing) of neural networks for the combination and propagation of probability assignments. Towards this end, we require a transformation which accomplishes combination of basic probability assignments on an additive scale, i.e. we require an Abelian group which is isomorphic to the additive group of real numbers on $(-\infty, +\infty)$. A candidate for such a transformation is $-\log(1 - \cdot)$, Shafer's (1976) weight of evidence function:

Adjusted to our context write $S(r_{31})$ for the weight of evidence for $X_3 = 1$ (due to a probability assignment of type (3.1)) conditional on $X_1 = 1$:

$$S(r_{31}) = -\log(1 - r_{31}).$$

If there also is $S(r_{32})$ due to a probability assignment on the second edge in (3.6), then the total (conditional) support for $X_3 = 1$ is

$$S(r_{31}) + S(r_{32}) = -\log(1 - r_{31} - r_{32} + r_{31} \cdot r_{32})$$

and since $m_{31} \oplus m_{32}(\{(1,1,1)\}) = r_{31} + r_{32} - r_{31} \cdot r_{32}$ clearly

$$S(r_{31}) + S(r_{32}) = S(r_{31} + r_{32} - r_{31} \cdot r_{32}).$$

Hence in this case transformation to the weight of evidence scale and Dempster combination are exchangeable operations. This is true also for two probability assignments of type (3.2). Unfortunately, this commutativity collapses in the case of conflicting evidence. When two basic probability assignments of type (3.1) and (3.2), respectively, are combined, then the second probability assignment erodes part of the support for $X_3 = 1$. If Dempster's rule and Shafer's weights of evidence scale jointly are to be consistent then the support for $X_3 = 1$ conditional on both $X_1 = 1$ and $X_2 = 1$ should be $S(r_{31}) - S(s_{32})$, if $r_{31} \geq s_{32}$ and, in particular, this difference should be zero if $r_{31} = s_{32}$. Instead, the combined support is equal to $S(r_{31}) - S(r_{31} \cdot s_{32})$ and nonzero for $r_{32} = s_{32}$.

On this deeper level the DS-theory has the same weakness as that ascribed to the Bayesian theory by Shafer (1976 p.22) on a less profound level i.e. equal support for both sides of a dichotomy should combine to no support for either, but, when using Shafer's weights of evidence scale in combination with Dempster's rule, they do not.

We therefore chose to modify Dempster's rule introducing a slightly modified updating mechanism (for conflicting evidence only) eliminating this defect. To make the difference clear also notationally, we called this the concept of *degrees of confirmation.*

Definition: For the basic probability assignments of type (3.1) and (3.2), respectively:

Given $X_1 = 1$, we say that $X_3 = 1$ has degree of confirmation r_{31}.

Given $X_2 = 1$, we say that $X_3 = 1$ has degree of confirmation $-s_{32}$.

For non-conflicting evidence we want the combination rule (for confirmation numbers) to produce the same result as Dempster's rule applied to two basic probability assignments of type (3.1) or (3.2). For conflicting evidence the above-mentioned defect should be eliminated. We propose

Definition: (Combination Rule for Degrees of Confirmation). Let X be confirmed independently to degrees a and b. Then the combined degree of confirmation is $a \oplus b$ where

$$a \oplus b = \begin{cases} a + b - ab & \text{if } a, b \geq 0 \\ a + b + ab & \text{if } a, b \leq 0 \\ a + b/(1 - \min\{|a|, |b|\}) & \text{if sign } a \neq \text{sign } b. \end{cases}$$

In addition, define the modified transformation

$$S^*(a) = \begin{cases} -\log(1-a), & a \geq 0 \\ \log(1+a), & a < 0. \end{cases}$$

Then the transformation S^* is compatible with the combination rule for degrees of confirmation in the sense that one may first transform both degrees of confirmation and then combine them, or first combine degrees of confirmation and then transform them. And this holds for both nonconflicting and conflicting evidence:

$$S^*(a \oplus b) = S^*(a) + S^*(b) \text{ for all } a,b \in (-1,1).$$

Now everything is in place for an efficient computational handling of the network (3.5) and the desired task. We will allow evolution of the global state of the network (represented as a binary word of length N) over time and write $X_i(t) = 0$ or $X_i(t) = 1$ for $i = 1,2, \ldots, N$ depending on whether the i-th node is in state 0 or state 1 at time t. (A mechanism for state changes will be introduced momentarily.)

The connectivities C_{ij} for $i,j = 1,2, \ldots, N$ will be taken to be constants, in particular, they are derived from the transformation S^*:

$$C_{ij} = -\log(1 - r_{ij})$$

for a probability assignment between propositions (nodes) i and j of type (3.1) and

$$C_{ij} = +\log(1 - s_{ij})$$

if the probability assignment is of type (3.2). Hence the connectivities C_{ij} are the conditional degrees of confirmation given $X_i = 1$.

Define also the *confirmation function*

$$C(t) = \sum_{\substack{i \ j \\ i \neq j}}\sum C_{ij} X_i(t) X_j(t) - \sum_i \gamma_i X_i(t)$$

and introduce the following mechanism for state changes governing the evolution of the system of nodes

$$(3.9) \quad \begin{array}{ll} \text{at time } T_i^k \text{ node i changes its state from 0 to 1} & \\ & \text{if } \sum\limits_{j \neq i} C_{ij} X_i(T_i^k) \\ \text{at time } T_i^k \text{ node i changes its state from 1 to 0} & \end{array} \begin{array}{l} > \gamma_i \\ \\ \leq \gamma_i \end{array}$$

Here γ_i is the individual threshold of the i-th neuron, and for given $i = 1, \ldots, N$ the random variables T_i^k, $k = 1,2,\ldots$ are the times at which changes occur in a Poisson process with parameter λ. Individual nodes therefore wake up at random times and decide whether of not to change their state. The T_i^k and T_j^l for $i \neq j$ are independent random variables for all k, l. The time evolution of the state of the system of nodes is based on asynchronous parallel processing.

If $X_i(T_i^k+)$ and $C(T_i^k+)$ denote the state of the ith node and the confirmation function, respectively, immediately after the i-th node has evaluated its field $\sum\limits_{j \neq i} C_{ij} X_i(T_i^k)$ for the kth time, then

$$(3.10) \quad X_i(T_i^k+) - X_i(T_i^k) = \delta \Rightarrow C(T_i^k+) - C(T_i^k) = -\delta(\sum_{j \neq 1} C_{ij} X_j(T_i^k) - \gamma_i)$$

for $\delta = -1, 0, +1$.

In view of our definition of degree of confirmation and with respect to the envisioned task of the network $\gamma_i = 0$ for all i should be chosen. It is clear from (3.9) and (3.10) that the above scheme for state changes causes $C(t)$ to monotonically increase in time and the evolution of the network is towards a (local) maximum of the confirmation function. The system is started up with random states except for the nodes whose states are part of the input information.

Interconnected systems of 0-1 nodes (sometimes referred to as neural networks) of similar structure have been employed by Hopffield (1982, 1984) for the task of solving large scale optimization problems with constraints. Among other things his findings were:

(a) If $C_{ij} = C_{ji}$ for $i \neq j$ then the system has stable limit points.

(b) Stable limit points persist when $C_{ij} \neq C_{ji}$ but additional noise is introduced into the system.

(c) Convergence is rapid. After only a few multiples of the stochastic mean processing time the system settles into limiting behaviors, the most common of these being a stable state.

(d) When (apart from the prescribed input states of the network) the dynamics were started from randomly assigned configurations convergence usually was towards a small number of stable limits only for repeated runs with different (random) starting configurations. Hence a small number of stable states (the local maxima of the confirmation function) usually collect the system flow from initial configurations.

We expect similar behavior for our network in the context of the specific task but have not done simulation studies yet. This clearly needs to be done.

References

Barlow, R.E., Fussell, J.B. and Singpurwalla, N.D. (1975). Reliability and Fault Tree Analysis. SIAM, Philadelphia.

Barnett, J.A. (1981). Computational methods for a mathematical theory of evidence. *In Proceedings of the 7th International Joint Confernece on Artifical Intelligence*, Vancouver, 868-875.

Beran, R.J. (1971a). On distribution-free statistical inference with upper and lower probabilities. *Ann. Math. Statist.* **42**, 157-168.

Beran, R.J. (1971b). A note on distribution-free statistical inference with upper and lower probabilities. *Ann. Math. Statist.* **42**, 1943-1948.

Berg, C., Christensen, J.P.R. and Ressel, P. (1984). Harmonic Analysis on Semigroups. Springer, New York.

Boole, G. (1854). An Investigation of the Laws of Thought. London, reprinted in 1958 by Dover, New York.

Buchanan, B.G. and Shortliffe, E.H. (1984). Rule-based Expert Systems: the MYCIN Experiment of the Stanford Heuristic Programming Project. Reading, Addison-Wesley.

Choquet, G. (1953). Theory of capacities. *Annales de l'Institut Fourier* **V**, 131-295.

Dempster, A.P. (1967). Upper and lower probabilities induced by a multi-valued mapping. *Ann. Math. Statist.* **38**, 325-339.

Dempster, A.P. (1968). A generalization of Bayesian inference. *J. Roy. Statist. Soc., Series B*, **30**, 205-247.

Garey, M.R. and Johnson, D.S. (1979). Computers and Intractability. New York: W.H. Freemann.

Garey, T.D. et.al. (1981). An inference technique for integrating knowledge from disparate sources. *Proceedings of the 7th International Conference on Artificial Intelligence*, 319-325.

Good, I.J. (1950). Probability and the Weighing of Evidence. Hafner.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **79**, 2554-2558.

Hopfield, J.J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci. USA* **81**, 3088-3092.

Shafer, G. (1976). A Mathematical Theory of Evidence, Princeton: University Press.

Shafer, G. (1984). Probability judgement in artificial intelligence and expert systems. Working Paper No. 165, School of Business, University of Kansas.

Smith, C.A.B. (1961). Consistency in statistical inference and decision (with discussion). *J. Roy Statist. Soc. B*, **23**, 1-25.

Smith, C.A.B. (1965). Personal probability and statistical analysis (with discussion). *J. Roy. Statist. Soc., A*, **128**, 469-499.

West, S. (1971). Upper and lower probability inference for the logistic function. Ph.D. thesis, Department of Statistics, Harvard University.

Department of Statistics
University of California
Berkeley, CA 94720