Annals of University of Craiova, Math. Comp. Sci. Ser.
Volume 30(2), 2003, Pages 28–33
ISSN: 1223-6934

Stochastic properties of a learning algorithm based on an aggregation operator

Petrică Badea

Abstract. One of the most used aggregation operators in artificial intelligence is the "probabilistic OR", or POR. If a certain fact has a double estimated confidence through real positive and subunit numbers a and b, then, the overall confidence of that fact is $POR(a,b) = a+b-ab$. POR is an associative function and three or more confidence factors can be aggregated sequentially into an overall confidence factor. When the number of estimations grows up indefinitely one can ask about limit properties of POR. This operator is not a learning algorithm and the paper proposes a learning algorithm based on it. A sequence of random variables using POR like the reward side of a learning and another operator like punish side is defined and it's properties are studied. The proposed algorithm is showed to converge to some simple probability provided certain simple conditions are accomplished. Finally, a comparison between the proposed algorithm and the classic estimator is given.

2000 Mathematics Subject Classification. 03B48, 03B52.
Key words and phrases. learning, aggregation, random variables.

1. Introduction

Artificial intelligence is very often dealing with uncertainty. Handling this concept, different authors have made very different proposals which encapsulate it. One of the most used aggregation operators in artificial intelligence is the "probabilistic OR", or POR. If a certain fact from reality is uncertain and two or more experts try to estimate it's degree of certainty, then that fact has two or more estimated confidences. These are estimations through real positive and subunit numbers a, b, c, and so on. If we have exactly two experts, then, the overall confidence of that fact is $POR(a,b) = a+b-ab$. POR is an associative function and three or more confidence factors can be aggregated sequentially into an overall confidence factor. When the number of estimations grows up indefinitely one can ask about limit properties of POR. It can be interpreted a little bit different writing it's formula as $POR(a,b) = a + b(1-a)$. This means that if a certain expert gives the estimation a, the opinion of the second expert is increasing our confidence a, about the fact, by a fraction of the difference to unity of a. The fraction of the difference is given by the second estimation, b. More specifically, the result is the first estimation a, increased by the fraction b of $(1-a)$. Thinking like this, in learning terms, all can be trivial because the composition of an indefinitely increasing number of estimations, can be convergent when simple conditions are accomplished, but the "reward-punish" paradigm is not encapsulated. We can think a little different: what if a certain expert says the fact is NOT true with a certainty of say, $c \in [0,1]$? More precisely, what is the certainty of a fact, if an expert says YES with the certainty of a, and the other says NOT, with the degree of certainty b? A simple way of answering this question is to put $CON(a,b) = a - b(0-a)$. This formula looks closely related

Received: 19 November 2003.

to the POR fromula, except for the substracting operation instead of summation and the replacement of the number 1 (meaning: TRUE), with 0 (meaning: FALSE). So in a long run problem, the way of acting is like this: at every step, the actual value of confidence is increased using POR formula, if an expert say YES, and is decreased using CON formula if the expert says NO. However, in practice, this problem is out of subject since usually only a small number of experts can estimate a certain fact in the same time. Furthermore, it is rather unlikely for some researcher to combine such a way two or more confidence factors, even if a contradiction ocures. The problem of learning from data arises the following question: when n events are in favor of a certain fact, out of m, and their probability is rather equal, what certainty factor should we assign to that fact? The fraction $\frac{n}{m}$ is the most used and the more rational estimation of the certainty factor. What if the m events come out sequentially? How we should take into account every occurency of such an event? A simple way to deal with this question is to use the "little expert" paradigm: every event is increasing a learning factor if it is in favor of the studied fact and decreases this learning factor if the opposite is true. More precisely: this way of thinking is represented by a sequence of independent variables $(\xi_k)_{k\geq 0}$, defined through sequences of positive real numbers falling into the unit interval $(\rho_k)_{k\geq 0}$ and $(\eta_k)_{k\geq 0}$ by relations:

$$\xi_{k+1} = \begin{cases} \xi_k + (1-\xi_k)\rho_k\,, & \text{if } z_k = 1\,, \\ \xi_k - \eta_k\xi_k\,, & \text{if } z_k = 0\,, \end{cases}$$

where

$$z_{k+1} = \begin{cases} 1\,, & \text{with probability } p\,, \\ 0\,, & \text{with probability } 1-p\,, \end{cases}$$

for fixed values $c, p \in [0,1]$, $\xi_0 = c$. The expectation and the variance are calculated and their properties are studied.

This paper is showing that provided some conditions are accomplished, the random variable ξ_k, tends to the probability of an event to be in favor of the studied fact. In fact, in certain conditions, ξ_k is an unbiased estimation of this probability. This probability may be estimated classically, simply by taking the ratio of the number of events in favor of the fact, out of the total number of events which occurs. This paper demonstrates that the learned ξ_k often tends to this probability.

2. Main results

Theorem 2.1. *Let $(\rho_k)_{k\geq 0}$, $(\eta_k)_{k\geq 0}$ two sequences of real positive numbers and $\rho_0 < 1$, $\eta_0 < 1$. Let $(\xi_k)_{k\geq 0}$, a sequence of independent random variables defined by:*

$$\xi_{k+1} = \begin{cases} \xi_k + (1-\xi_k)\rho_k\,, & \textit{if } z_k = 1\,, \\ \xi_k - \eta_k\xi_k\,, & \textit{if } z_k = 0\,, \end{cases} \tag{1}$$

$$z_{k+1} = \begin{cases} 1\,, & \textit{with probability } p\,, \\ 0\,, & \textit{with probability } 1-p\,, \end{cases}$$

and let $\xi_0 = c \in [0,1]$ and $q = 1-p$.

Then, one can calculate iteratively the expectation and the variance of ξ_k as:

$$E(\xi_{k+1}) = E(\xi_k)(p(1-\rho_k) + q(1-\eta_k)) + p\rho_k \tag{2}$$

$$\begin{aligned} D^2(\xi_{k+1}) &= D^2(\xi_k)(p(1-\rho_k)^2 + q(1-\eta_k)^2) + pq\rho_k^2 + \\ &\quad +(\eta_k - \rho_k)pqE(\xi_k)(E(\xi_k)(\eta_k - \rho_k) - 2\rho_k), \end{aligned} \tag{3}$$

for $k > 0$.

Proof. Following the definition of ξ_k, the demonstration of the first relation is straightforward:

$$E(\xi_{k+1}) = pE(\xi_k + (1-\xi_k)\rho_k) + qE(\xi_k - \eta_k\xi_k) =$$
$$= pE(\xi_k)(1-\rho_k) + p\rho_k + qE(\xi_k)(1-\eta_k) =$$
$$= E(\xi_k)(p(1-\rho_k) + q(1-\eta_k)) + p\rho_k$$

In order to demonstrate the second relation, let the square of ξ_{k+1} to be:

$$\xi_{k+1}^2 = \begin{cases} \xi_k^2(1-\rho_k)^2 + 2\xi_k\rho_k(1-\rho_k) + \rho_k^2\,, & \text{if } z_k = 1\,, \\ \xi_k^2(1-\eta_k)^2\,, & \text{if } z_k = 0\,, \end{cases}$$

obtained simply, by squaring the relation (1). The first step is the calculus of $E(\xi_{k+1}^2)$:

$$E(\xi_{k+1}^2) = E(\xi_k^2)(p(1-\rho_k)^2 + q(1-\rho_k)^2) + 2E(\xi_k)p\rho_k(1-\rho_k) + p\rho_k^2$$

On the other hand, the square of the relation (2) gives:

$$E^2(\xi_{k+1}) = E^2(\xi_k)(p(1-\rho_k) + q(1-\rho_k))^2 + 2E(\xi_k)p\rho_k(p(1-\rho_k) + q(1-\eta_k)) + p^2\rho_k^2$$

Now, one can calculate the variance of ξ_{k+1} as:

$$D^2(\xi_{k+1}^2) = E(\xi_{k+1}^2) - E^2(\xi_{k+1}^2) = E(\xi_k^2)(p(1-\rho_k)^2 + q(1-\eta_k)^2) -$$
$$-E^2(\xi_k)(p^2(1-\rho_k)^2 + q^2(1-\eta_k)^2 + 2pq(1-\rho_k)(1-\eta_k)) +$$
$$+2E(\xi_k)p\rho_k((1-\rho_k) - p(1-\rho_k) - q(1-\eta_k)) + p\rho_k^2 - p^2\rho_k^2 =$$
$$= p(1-\rho_k)^2(E(\xi_k^2) - pE^2(\xi_k)) + q(1-\eta_k)^2(E(\xi_k^2) - qE^2(\xi_k)) -$$
$$-2pqE^2(\xi_k)(1-\rho_k)(1-\eta_k) + 2E(\xi_k)p\rho_k q(\eta_k - \rho_k) + pq\rho_k^2 =$$
$$= p(1-\rho_k)^2(D^2(\xi_k) + qE^2(\xi_k)) + q(1-\eta_k)^2(D^2(\xi_k) + pE^2(\xi_k)) -$$
$$-2pqE^2(\xi_k)(1-\rho_k)(1-\xi_k) + 2pq\rho_k E(\xi_k)(\eta_k - \rho_k) + pq\rho_k^2 =$$
$$= D^2(\xi_k)(p(1-\rho)k)^2 + q(1-\eta_k)^2) + E^2(\xi_k)pq((1-\rho_k)^2 + (1-\eta_k)^2)) -$$
$$-2pqE^2(\xi_k)(1-\rho_k)(1-\eta_k) + 2pq\rho_k E(\xi_k)(\eta_k - \rho_k) + pq\rho_k^2 =$$
$$= D^2(\xi_k)(p(1-\rho)k)^2 + q(1-\eta_k)^2) + pq\rho_k^2 + Q$$

Where:

$$Q = E^2(\xi_k)pq((1-\rho_k)^2 + (1-\eta_k)^2)) - 2pqE^2(\xi_k)(1-\rho_k)(1-\eta_k) + 2pq\rho_k E(\xi_k)(\eta_k - \rho_k)$$

If one can show that:

$$Q = (\eta_k - \rho_k)pqE(\xi_k)(E(\xi_k)(\eta_k - \rho_k) + 2\rho_k)$$

the demonstration is done. In fact, taking the expression of Q, it follows that:

$$Q = E^2(\xi_k)pq((1-\rho_k)^2 + (1-\eta_k)^2)) - 2pqE^2(\xi_k)(1-\rho_k)(1-\eta_k) + 2pq\rho_k E(\xi_k)(\eta_k - \rho_k) =$$
$$= E^2(\xi_k)pq((1-\rho_k) - (1-\eta_k))^2 + 2pq\rho_k E(\rho_k)(\eta_k - \rho_k) =$$
$$= E^2(\xi_k)pq(\eta_k - \rho_k)^2 + 2pq\rho_k E(\xi_k)(\eta_k - \rho_k) =$$
$$= (\eta_k - \rho_k)pqE(\xi_k)(E(\xi_k)(\eta_k - \rho_k) + 2\rho_k)$$

The last expression equals Q and demonstrate the theorem. □

Corollary 2.1. *If the sequences $(\rho_k)_{k>0}$ and $(\eta_k)_{k>0}$ are both convergent to real positive and subunit numbers ρ and, respectively, η, then the limit of expected value and of the variance of ξ_k are as follows:*

$$E(\xi_k) \longrightarrow \frac{p\rho}{p\rho + q\eta} \tag{4}$$

$$D^2(\xi_k) \longrightarrow \frac{pq(\frac{\rho\eta}{p\rho+q\eta})^2}{1-(p(1-\rho)^2+q(1-\eta)^2)} \tag{5}$$

Proof. From the previous theorem, taking the equation (2) we obtain:

$$\lim_{n\to\infty} E(\xi_{k+1}) = \lim_{n\to\infty}(E(\xi_k)(p(1-\rho_k)+q(1-\eta_k))+p\rho_k)$$

$$\lim_{n\to\infty} E(\xi_{k+1}) = \lim_{n\to\infty} E(\xi_k)(p(1-\lim_{n\to\infty}\rho_k)+q(1-\lim_{n\to\infty}\eta_k))+p\lim_{n\to\infty}\rho_k$$

$$E = E(p(1-\rho)+q(1-\eta))+p\rho$$

$$E = \frac{p\rho}{p\rho+q\eta}$$

Using equation (3), it is easy to see that taking the limit, we have the following calculus:

$$D^2 = D^2(p(1-\rho)^2+q(1-\eta)^2)+pq\frac{\eta-\rho}{p\rho+q\eta}(\frac{p\rho(\eta-\rho)}{p\rho+q\eta}+2\rho)+pq\rho^2$$

This is equivalent to:

$$D^2(1-(p(1-\rho)^2)+q(1-\eta)^2) = pq(\frac{(\eta-\rho)^2p^2\rho^2}{(p\rho+q\eta)^2}+2\rho\frac{(\eta-\rho)p\rho}{p\rho+q\eta}+\rho^2)$$

Or, by compacting the terms:

$$D^2(1-(p(1-\rho)^2)+q(1-\eta)^2) = pq(\frac{(\eta-\rho)p\rho}{p\rho+q\eta}+\rho)^2$$

Because it is easy to see that it is true the relation:

$$\frac{(\eta-\rho)p\rho}{p\rho+q\eta}+\rho = \frac{\rho\eta}{p\rho+q\eta},$$

it follows that the limit of the variance is:

$$D^2 = \frac{pq(\frac{\rho\eta}{p\rho+q\eta})^2}{1-(p(1-\rho)^2+q(1-\eta)^2)}.$$

□

Corollary 2.2. *If the sequences in corollary 2.1 are convergent to the same value, denoted by ρ, then the limit of the expected value and of the variance of ξ_k, when k tends to infinity are:*

$$\lim_{k\to\infty} E(\xi_k) = p$$

$$\lim_{k\to\infty} D^2(\xi_k) = \frac{pq\rho}{2-\rho}$$

Proof. The proof is trivial, since it is sufficient to replace in equations (4) and (5) the value of η with ρ. □

Corollary 2.3. *Let $(\rho_k)_{k\geq 0}$, a sequence of real positive numbers and $\rho_k < \frac{1}{k+1}$, for $k = 1, ...n, n \in N$. Let $(\xi_k)_{k\geq 0}$, a sequence of independent random variables defined by:*

$$\xi_{k+1} = \begin{cases} \xi_k + (1-\xi_k)\rho_k\,, & \text{if } z_k = 1\,, \\ \xi_k - \rho_k\xi_k\,, & \text{if } z_k = 0\,, \end{cases} \tag{6}$$

$$z_{k+1} = \begin{cases} 1\,, & \text{with probability } p\,, \\ 0\,, & \text{with probability } 1-p\,, \end{cases}$$

Then, the variance of ξ_k is inferior to the variance of the classic estimator, for $k <= n$.

Proof. Firstly, it has to be observed from the main theorem that for $\rho_k = \eta_k$, the variance of ξ_k follows the formula:

$$D^2(\xi_{k+1}) = D^2(\xi_k)(1-\rho_k)^2 + pq\rho_k^2$$

On the other hand, it is easy to see that $D^2(\xi_0) = 0$ and $D^2(\xi_1) = pq\rho_0$. A trivial induction shows that we have the following formulas:

$$D^2(\xi_{k+1}) = pqf(\rho_0, \rho_1, ...\rho_k) = pqf_{k+1}$$

$$f_{k+1} = f_k(1-\rho_k)^2 + \rho_k^2$$

Since $f_1 = \rho_0 < 1 = \frac{1}{1}$, and suppose that $f_k < \frac{1}{k}$, for any integer between 1 and k, let show that we can choose a value for ρ_k, such that $f_{k+1} < \frac{1}{k+1}$. Indeed, the last inequality means that:

$$f_k(1-\rho_k)^2 + \rho_k^2 < \frac{1}{k+1}$$

The attached second degree equation in ρ_k, has always two distinct roots, because of the fact that it's discriminant is positive:

$$\Delta = \frac{1-kf_k}{k+1} > 0$$

as induction hypothesis says. The two distinct roots of the attached equation are:

$$f_{k1} = \frac{f_k - (\frac{1-kf_k}{k+1})^{\frac{1}{2}}}{f_k + 1}$$

$$f_{k2} = \frac{f_k + (\frac{1-kf_k}{k+1})^{\frac{1}{2}}}{f_k + 1}$$

Because the second root is strictly positive, the choice for ρ_k such that $f_{k+1} < \frac{1}{k+1}$ is straightforward. In fact, similar inequalities are valid for ρ_k: provided $\rho_k < \frac{1}{k+1}$, it is true that:

$$\frac{\rho_k - (\frac{1-k\rho_k}{k+1})^{\frac{1}{2}}}{\rho_k + 1} < 1$$

In fact, ρ_k must be inside the interval $[0, \frac{\rho_k - (\frac{1-k\rho_k}{k+1})^{\frac{1}{2}}}{\rho_k+1}]$. Finally, let take again the inequality:

$$f_k(1-\rho_k)^2 + \rho_k^2 < \frac{1}{k+1}$$

and recall that:

$$D^2(\xi_{k+1}) = pqf(\rho_0, \rho_1, ...\rho_k) = pqf_{k+1}$$

This means that:

$$D^2(\xi_{k+1}) < pq\frac{1}{k+1}$$

This last inequality demonstrates the corollary, because it simply says that the variance of ξ_k is lower that the variance of the classic estimation. □

The last corollary shows that for a finite number of steps, the estimation of the probability of a fact given by ξ_k, is less dispersed than the classic one. This result is not as surprising as it seems to be, because sooner or later, in the learning process, the classic estimation became to be less dispersed. The great importance of this result is that the learning algorithm using ξ_k formula, may be maintained to be more stable for a little number of steps. This mike be very important in a learning process with a great number of parameters to be estimated, when it is well known that the classic probability needs a great number of steps until a certain stability appears.

References

[1] ***, Bayesian Models in Medicine, *The European Conference on Artificial Intelligence in Medicine*, Cascais, Portugal, July, 2001.

[2] C. Berenstein, L. Kanal, D. Lavine, Consensus Rules, In: *Uncertainty in Artificial Intelligence*, Edited by L. Kanal and J.F. Lemmer, Elsevier Science Publishers, 1986.

[3] R.K. Bhatnagar, L.N. Kanal, Handling Uncertain Information: A review of numeric and Non-numeric Methods, In: *Uncertainty in Artificial Intelligence*, Edited by L. Kanal and J.F. Lemmer, Elsevier Science Publishers, 1986.

[4] Th. Eiter, Th. Lukasiewicz, Complexity Results for Structure-Based Causality, *Proceedings of the Sixteen International Joint Conference on Artificial Intelligence*, IJCAI-2001.

[5] P. Lucas, *Certainty factor-like Structures in Bayesian Belief Networks*.

[6] P. Lucas, Computer-based Decision Support in the Management of Primary Gastric non-Hodgkin Lymphoma, *Methods of Information in Medicine*, **37**, 206-219 (1998).

[7] J. Pearl, A Constraint-Propagation Approach to Probabilistic Reasoning, In: *Uncertainty in Artificial Intelligence*, Edited by L. Kanal and J.F. Lemmer, Elsevier Science Publishers, 1986.

[8] D.J. Spiegelhalter, Probabilistic Reasoning in Predictive Expert Systems, In: *Uncertainty in Artificial Intelligence*, Edited by L. Kanal and J.F. Lemmer, Elsevier Science Publishers, 1986.

(Petrică Badea) Department of Biostatistics, University of Medicine and Pharmacy, Craiova RO-200585, Romania
E-mail address: bip@umfcv.ro