Andrzej PIEGAT [1], **Marek LANDOWSKI** [1,2]
[1] WEST POMERANIAN UNIVERSITY OF TECHNOLOGY
[2] SZCZECIN MARITIME UNIVERSITY

Surmounting Information Gaps Using Average Probability Density Function

Prof. dr hab. inż. Andrzej PIEGAT

He received his PhD degree in 1979 in modeling and control of production systems from Technical University of Szczecin, the DSc degree in control of uderwater vehicles from Rostock University in 1998, and the professor title in 2001. At present he is professor at ZUT. His current research is focused on uncertainty theory, fuzzy logic, computing with words and info-gap theory.

e-mail: apiegat@wi.zut.edu.pl

PhD student Marek LANDOWSKI

He received the MSc degree in Mathematics from the Szczecin University in 2002 and the MSc eng. degree in Computer Science from the West Pomeranian University of Technology in 2004. Currently he is PhD student at the West Pomeranian University of Technology and assistant professor at the Szczecin Maritime University. At present his research interests are focused on identification of probabilistic models of human perception and info-gap theory.

e-mail: m.landowski@am.szczecin.pl

Abstract

In many problems we come across the lack of complete data. The information gap causes that the task seems to be unsolvable. In many cases where the Bayes' networks or Bayes' rule are used, we come across the information gap which is the lack of a priori distribution. The article presents the methods of identifying the average probability density distribution when we know the range of variable and we have some quality knowledge on the distribution. The obtained average probability density distribution minimizes medium squared error. According to the authors' knowledge the average probability density distribution is the novelty in the word literature.

Keywords: Bayes' networks, information gaps, principle of indifference, uncertainty theory, artificial intelligence, probability theory.

Pokonywanie luk informacyjnych za pomocą przeciętnej funkcji gęstości prawdopodobieństwa

Streszczenie

W wielu rzeczywistych problemach często spotykamy się z brakiem danych koniecznych do ich rozwiązania. Dotyczy to zwłaszcza zadań projektowania nowych systemów technicznych, ale i też ekonomicznych, medycznych, agrarnych i innych. Istnienie luk w problemie powoduje, że zadanie wydaje się nierozwiązywalne. W takiej sytuacji, aby w ogóle rozwiązać postawiony problem konieczne jest zaangażowanie ekspertów, którzy są często w stanie podać przybliżone oszacowanie danej brakującej do rozwiązania problemu. Niestety, oszacowania eksperckie zwykle nie są precyzyjnymi liczbami, lecz przedziałami możliwych wartości zmiennej lub też probabilistycznymi rozkładami możliwej wartości brakującej zmiennej. Zatem, aby rozwiązać dany problem konieczne jest wykonywanie operacji na rozkładach gęstości prawdopodobieństwa. Jednym z narzędzi służących do tego celu jest reguła Bayesa. Jest ona np. podstawą do przetwarzania informacji w sieciach wnioskowania probabilistycznego zwanych skrótowo sieciami Bayesa. Zwykle luką informacyjną w tych sieciach jest brak rozkładu a priori zmiennej koniecznego do obliczenia rozkładu a posteriori. W takiej sytuacji, jako rozkład a priori stosowany jest zwykle rozkład równomierny reprezentujący kompletną niewiedzę dotyczącą jakościowych cech rozkładu. Jednak taką wiedzę często posiada ekspert problemu. Artykuł prezentuje metodę identyfikacji przeciętnego rozkładu gęstości prawdopodobieństwa zmiennej dla przypadku, gdy ekspert zna nie tylko zakres możliwych wartości zmiennej, ale także posiada pewną wiedzę o jakościowych cechach rozkładu. Otrzymany z użyciem wiedzy eksperta przeciętny rozkład gęstości prawdopodobieństwa zmniejsza znacznie ryzyko popełnienia katastrofalnie dużych błędów w rozwiązywaniu problemów z lukami informacyjnymi. Według wiedzy autorów koncepcja przeciętnego rozkładu gęstości prawdopodobieństwa jest nowością w literaturze światowej.

Słowa kluczowe: sieci Bayesa, luki informacyjne, zasada nierozróżnialności, teoria niepewności, sztuczna inteligencja, teoria prawdopodobieństwa.

1. Introduction

Solving problems under uncertainty (partial lack of knowledge) is one of the most difficult aims of artificial intelligence (AI). People can solve such problems. To make AI comparable with the human intelligence it has to be also able to solve problems under uncertainty. Problems of information gaps are being intensively investigated at present [1]. An information gap means lack of knowledge about values of variables, about distributions of their probability or possibility, about variability intervals, etc. The problem of information gaps is a common one in Bayesian networks where the prior distributions are necessary but they frequently are unknown.

Prior distributions are also necessary in all other problems where Bayes' rule [2] is applied. Let A and B be two events. The conditional probability $p(A|B)$ of event A is not known but necessary for a problem solving. If the inverse conditional probability $p(B|A)$ and the prior probability $p(A)$ are known then the unknown probability $p(A|B)$ can be calculated with *Bayes' theorem* [2],

$$p(A|B)= \alpha\, p(B|A)\, p(A), \tag{1}$$

where: α - normalizing coefficient, $p(B|A)$ - likelihood function.

Bayes' theorem is used in probabilistic, automated reasoning. Unfortunately, it requires knowledge of the prior probability $p\,(A)$ that frequently is not known. Also in many other problems, which have nothing to do with Bayes' rule, some data frequently are unknown.

Sometimes to solve a problem we need numerical value x^* of variable x. But its value is unknown (information gap). Experts can give us then the interval $[x_{\min}, x_{\max}]$ in which the value is contained.

Bayes' theorem is very valuable tool because it tells us how to update or revised beliefs, expressed in form of probabilities, in light of new evidence a posteriori. However, in practical applications it is sometimes difficult to determine the prior probability distribution. Example of a simple problem with an information gap can be the famous Bertrand's problem [3].

"The train leaves at noon to travel a distance of 300 km. It travels at a speed of between 100 km/h and 300 km/h. What is the probability that it arrives before 2 p.m.?"

To give an answer to the above question knowledge of the pdf of the train velocity is necessary. However, this distribution is not given - we have to do here with an information gap. But if the problem solution is, for certain reasons, necessary for us, we can calculate an approximate and quite credible solution with the use of the principle of indifference (for short: PI) formulated by Laplace [2]. This principle gives us the advice to assume equal probability density for all possible values of x contained in the interval $[x_{\min}, x_{\max}]$ if any of the x-values can not be distinguished

in respect of the probability density. It means, the uniform pdf(x) can be assumed. The uniform distribution can according to the PI be assumed for variables about which nothing, absolutely nothing is known. However, sometimes in the considered problems we have certain, general, qualitative knowledge concerning the character (shape) of the distribution. In the next chapter the situation will be analyzed, when we know that the distribution is unimodal one with boundaries values run to zero. According to the authors' knowledge the average probability density distribution is the novelty in the word literature.

2. The Average, Unimodal Distribution of Probability Density with Boundary Values Aiming at Zero (UBlimit0)

Let us assume that our knowledge concerning the real value x^* of variable x is as follows:

I. The value x^* is contained in the interval $[x_{\min}, x_{\max}]$. For simplicity let us assume the normalized interval [0,1].
II. The pdf(x) is unimodal one. However, we do not know whether it is symmetric or asymmetric one. Therefore, we have to allow both the left asymmetry, the symmetry, and the right asymmetry of the pdf(x).
III. We know that the boundary values aiming at zero, so maximum lies between the boundaries.

Fig. 1 presents a few examples of unimodal distributions, which satisfy conditions I-III.

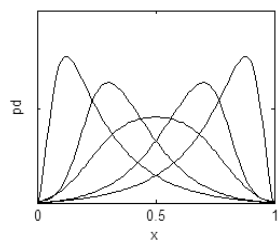

Fig. 1. A few examples of unimodal distributions from the infinite number of distributions that satisfy conditions I, II, III.
Rys. 1. Kilka przykładów rozkładów unimodalnych z nieskończonej liczby rozkładów, które spełniają warunki I, II, III.

The number of possible unimodal distributions, where boundaries values run to zero (UBlimit0) is infinitely large. The real pdf(x) distribution in the considered problem can be any one of them. Is it, in general, possible to determine the average distribution of the infinite number of distributions?

It seemingly seems impossible. However, there exists a certain possibility. The average distribution can be determined with *the method of decreasing granulation of elementary events* that was proposed by A. Piegat in [4]. According to the condition I the variable x can take values only in the normalized interval [0.1]. This interval can be partitioned in n subintervals Δx_i of width $1/n$. Let i be the number of a subinterval, $1 \le i \le n$. As *elementary event* will be understood the event $x^* \in \Delta x_i$. By *granulation* of the elementary event will be understood the width $1/n$ of the subinterval Δx_i of the event. Granulated will not only be the variable x but also probability p of elementary events. In this case the granulation of probability $1/n$ will mean that probability can take only $(n+1)$ discrete values. E.g. for $n = 3$ probability can only take values $p \in \{0, 1/3, 2/3, 1\}$. Now, let us consider the question how many distributions are possible for granulation 1/3. This granulation is assumed both for the variable x and for probability p of elementary events. All possible UBlimit0-distributions (histograms) for this case are shown in Fig. 2.

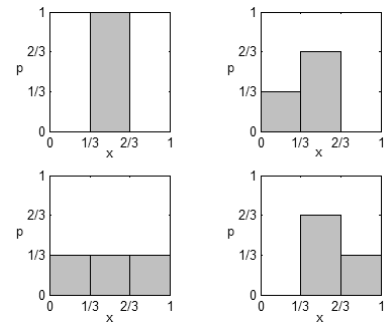

Fig. 2. All 4 possible UBlimit0-distributions of probability for granulation 1/3 of the variable x and of probability p
Rys. 2. Wszystkie 4 rozkłady prawdopodobieństwa Ublimit0 dla granulacji 1/3 zmiennej x i prawdopodobieństwa p

On the basis of all possible distributions from Fig. 2 the average probabilities p_1, p_2, p_3 of particular elementary events $x \in [0,1/3]$, $x \in [1/3, 2/3]$, and $x \in [2/3, 1]$ can be calculated.

$$p_1 = \frac{1}{4}\sum_{j=1}^{4} p_{1j} = \frac{1}{4}(0 + 1/3 + 1/3 + 0) = 1/6$$

$$p_2 = \frac{1}{4}\sum_{j=1}^{4} p_{2j} = \frac{1}{4}(1 + 2/3 + 1/3 + 2/3) = 4/6$$

$$p_3 = \frac{1}{4}\sum_{j=1}^{4} p_{3j} = \frac{1}{4}(0 + 0 + 1/3 + 1/3) = 1/6$$

The average UBlimit0-distrubution for granulation 1/3 is shown on Fig. 3.

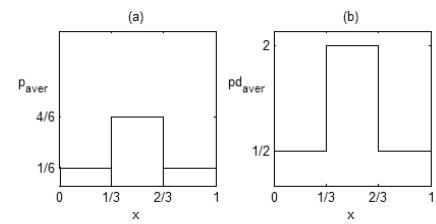

Fig. 3. The average UBlimit0-distribution for granulation 1/3 in the form of a histogram (a) and of pdf(x),(b)
Rys. 3. Przeciętny rozkład Ublimit0 dla granulacji 1/3 w postaci histogramu (a) i pdf(x) (b)

In the second step of the *method of decreasing granulation* the granulation was decreased from 1/3 to 1/4. Decreasing granulation causes an increase of the number of possible UBlimit0-distributions. Fig. 4 shows the comparison of the average pdf(x) for granulation 1/3 and 1/4.

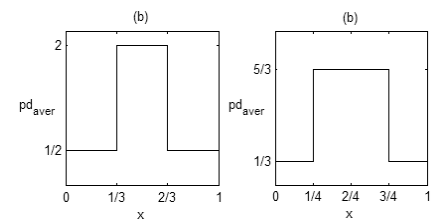

Fig. 4. The average distribution of probability density (pd) of 4 possible distributions from Fig.2 for granulation 1/3 (a) and the average distribution of 12 possible distributions for granulation 1/4 (b)
Rys. 4. Przeciętny rozkład gęstości prawdopodobieństwa (pd) z 4 możliwych rozkładów z rys. 2 dla granulacji 1/3 (a) oraz przeciętny rozkład z 12 możliwych rozkładów dla granulacji 1/4 (b)

If we stepwise decrease the granulation of the variable x and of probability p (1/3, 1/4, 1/5, 1/6/ ... ,1/n), then we will more and more approach a limiting distribution that represents the infinitive number of possible UBlimit0-distributions and which corresponds to the infinitely small granulation 1/n: $n \to \infty$. The increase of the distributions' number is very strong and rapid. So, granulation 1/3 is corresponded by 4 possible distributions, 1/4 by 12 distributions, ... , 1/25 by 899 276, and 1/27 by 1 835 932 distributions. This rapidly increasing number of distributions causes a very large memory burden for computers. However, the investigations made by the authors have shown that succeeding average distributions corresponding to decreasing granulation of elementary events quite quickly approach a certain limiting distribution and that differences between distributions corresponding to small granulation 1/n become negligible. In the practice one can observe this phenomenon already for granulations 1/24, 1/25, 1/26, Therefore the authors stooped generating the distributions for granulation 1/27. The average UBlimit0-distribution for this granulation is shown in Fig. 5.

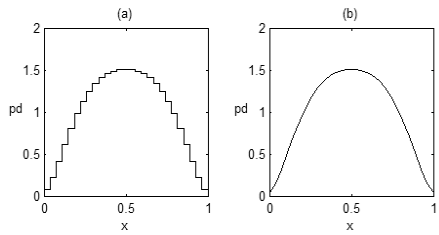

Fig. 5. The average distribution of probability density (pd) representing 1 835 932 possible UBlimit0-distributions for granulation 1/27 (a) and its smoothed approximation (b)
Rys. 5. Przeciętny rozkład gęstości prawdopodobieństwa (pd) z 1 835 932 możliwych rozkładów Ublimit0 dla granulacji 1/27 (a) oraz jego aproksymacja (b)

The polynomial approximation (2) of UBlimit0-distrubution:

$$pd_{śr}(x) = (125.6575\,x^6 - 376.9724\,x^5 + 431.6136\,x^4 - 234.9397\,x^3 + 54.3417\,x^2 + 0.2994\,x + 0.0254)/0.9991 \tag{2}$$

The mean absolute error of the smoothed approximation (2) equals 0,0026.

3. The Average, Unimodal Right-Asymmetric Distribution of Probability Density with Boundary Values Aiming at Zero (UAB-limit0)

Let us assume that our knowledge concerning the real value x^* of variable x is as follows:

IV. The value x^* is contained in the interval $[x_{\min}, x_{\max}]$. For simplicity let us assume the normalized interval [0,1].
V. The pdf(x) is unimodal and right-asymmetric one. The probability of the left side exceeds 0.5.
VI. We know that the boundary values aiming at zero, so maximum lies between the boundaries.

Fig. 6 presents a few examples of UABlimit0 distributions, which satisfy conditions IV-VI.

The average of the 912 084 boundary distributions possible for granulation 1/27 is shown in Fig. 7.

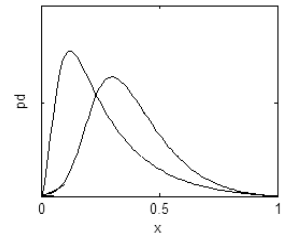

Fig. 6. Examples of UABlimit0 distributions from the infinite number of distributions that satisfy conditions IV, V, V
Rys. 6. Przykłady rozkładów UABlimit0 z nieskończonej liczby rozkładów, które spełniają warunki IV, V, VI

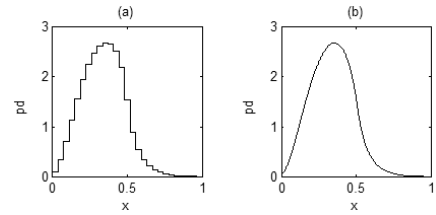

Fig. 7. The average distribution of probability density (pd) representing 912 084 possible UABlimit0-distributions for granulation 1/27 (a) and its smoothed approximation (the mean absolute error of the smoothed approximation equals 0,006) (b)
Rys. 7. Przeciętny rozkład gęstości prawdopodobieństwa (pd) z 912 084 możliwych rozkładów UABlimit0 dla granulacji 1/27 (a) oraz jego aproksymacja (średni błąd bezwzględny aproksymacji wynosi 0,006) (b)

The mean absolute error of the polynomial approximation (3) of UABlimit0 distributions equals 0,006

$$pd_{śr}(x) = \begin{cases} -537.1208\,x^5 + 737.8176\,x^4 - 429.8345\,x^3 + 101.6936\,x^2 + 1.0226\,x + 0.0474 & \text{for } x \in [0;14/27] \\ 1200.4484\,x^6 - 5854.7203\,x^5 + 11861.6852\,x^4 - 12789.5251\,x^3 + 7750.1852\,x^2 - 2507.2873\,x + 339.2145 & \text{for } x \in (14/27;1] \end{cases} \tag{3}$$

4. Conclusion

In the article the two average probability density distributions are presented. The first one is the average unimodal distribution density probability with boundary values aiming at zero whereas when identifying the second average distribution the additional knowledge on right side asymmetry of distribution is assumed. The received the average distributions of probability density minimizes mean squared error and can be used in case of the lack of information about the real a priori distribution of variable. According to the author's knowledge the above results are new in the world scientific literature.

5. References

[1] Yakov B.H., Info-gap decision theory-decisions under severe uncertainty. Second edition, Academic Press, London, 2006.
[2] Russel R., Norwig P., Artificial Intelligence- A Modern Approach. Second edition, Prentice Hall, Upper Saddle River, NJ, 2003.
[3] Magidor O., The classical theory of probability and the principle of indifference. 5th Annual Carnegie Mellon/University of Pittsburgh Graduate Philosophy Conference, pp.1-17, 2003. http://www.andrew.cmu/org/conference/2003
[4] Piegat A., Landowski M., Bayes' Rule, Principle of Indifference, and Safe Distribution, Artificial Intelligence and Soft Computing – ICAISC 2008, Lecture Notes in Computer Science, vol. 5097, Germany, 661-670, 2008.

Artykuł recenzowany