Control and Cybernetics

vol. **43** (2014) No. 1

Specialized, MSE-optimal m-estimators of the rule probability especially suitable for machine learning*

by

Andrzej Piegat[1] **and Marek Landowski**[2]

[1] Faculty of Computer Science and Information Systems,
West Pomeranian University of Technology,
Zolnierska 49, 71-210 Szczecin, Poland, apiegat@wi.zut.edu.pl
[2] Maritime University of Szczecin,
Waly Chrobrego 1-2, 70-500 Szczecin, Poland,
m.landowski@am.szczecin.pl

Abstract: The paper presents an improved sample based rule-probability estimation that is an important indicator of the rule quality and credibility in systems of machine learning. It concerns rules obtained, e.g., with the use of decision trees and rough set theory. Particular rules are frequently supported only by a small or very small number of data pieces. The rule probability is mostly investigated with the use of global estimators such as the frequency-, the Laplace-, or the m-estimator constructed for the full probability interval [0,1]. The paper shows that precision of the rule probability estimation can be considerably increased by the use of m-estimators which are specialized for the interval $[p_{h\,\min}, p_{h\,\max}]$ given by the problem expert. The paper also presents a new interpretation of the m-estimator parameters that can be optimized in the estimators.

Keywords: machine learning, rule probability, probability estimation, m-estimators, decision trees, rough set theory

1. Introduction

Probability is widely used in artificial intelligence for, e.g., quality evaluation of decision rules in machine learning and data mining (Cestnik,1990, 1991; Chawla and Cieslak 2006; Cichosz, 2000; Cussens,1993; Mozina et al., 2006; Starzyk and Wang, 2004; Sulzmann and Furnkranz, 2009,2010; Witten and Frank, 2005; Zadrozny and Elkan, 2001; Zhang, 1995), in the probabilistic version of rough and fuzzy set theory and clusterization (Polkowski, 2002; Ziarko, 1999), etc. Some of the evaluation methods are based (sometimes not explicitly) on the assumption of a large number of sample pieces. However, in real problems this assumption is frequently not satisfied. Even in the case when we possess an apparently large number of sample pieces and the input space is partitioned into

*Submitted: November 2011; Accepted: January 2014

many influence subspaces of particular rules, the number of sample pieces belonging in a single rule subspace frequently becomes very small. Fig. 1 presents an example of the input space partition typical for the rough set theory or decision trees. The problem of decision rules detected with the method of decision

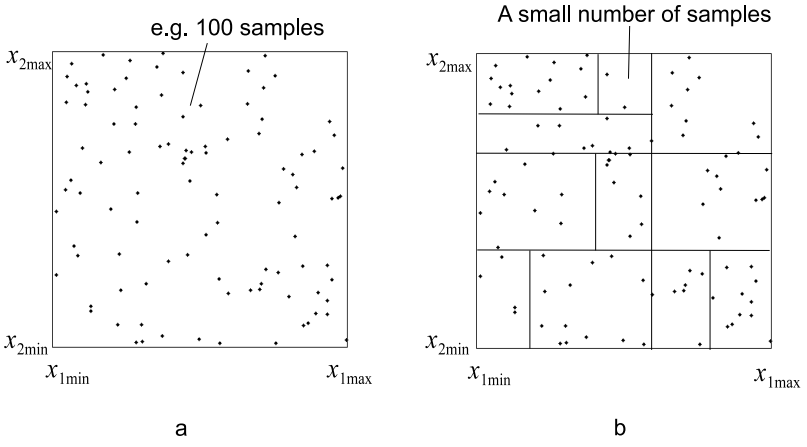


Figure 1. Illustration of small number of sample items occurring in influence subspaces of particular rules detected with the method of decision trees. Fig. 1a – sample pieces in the whole input space. Fig. 1b – sample pieces in subspaces of single rules (non-regular input space partition)

trees was described in Cestnik (1990, 1991), Chawla and Cieslak (2006), Rokach and Maimon (2008), Sulzmann and Furnkranz (2009, 2010). In this case the influence subspaces of particular rules do not involve the regular input-space partition as in the case of rough sets. The problem of a small number of sample pieces in the influence subspaces of particular rules occurs frequently not only in rough set models or decision tree models, but also in classification, clusterization, machine learning, data mining and classic modeling problems. The problem of a small number of sample pieces supporting a rule can be very severe. Sometimes a rule is supported by only one instance, sometimes even by no instance. In general, the domain of a rule can contain both the sample items which support the rule and the samples which negate it.

Now, let us consider one of the possible rules concerning the average gasoline consumption of passenger cars.

$$\begin{array}{l} IF \\ (mp \in [100, 140])\, AND (cw \in [1000, 1200]) \\ THEN \\ P(afc \in [8.0, 9.0]) = p_h \\ OR \\ P(afc \notin [8.0, 9.0]) = p_{\bar{h}} = 1 - p_h \end{array} \tag{1}$$

where: mp - motor power [HP], cw – car weight [kg], afc – average fuel consumption [l/100km].

In the example rule (1) the first conclusion ($afc \in [8.0, 9.0]$) can be called hypothesis h and the second conclusion - its negation or anti-hypothesis $\bar{h}$. Data obtained from the car owners can confirm the hypothesis h or negate it. Thus, the rule conclusion consists of two hypotheses, h and $\bar{h}$, whose probabilities should be determined. Higher probability p_h of the conclusion hypothesis (c-hypothesis) increases its strength and credibility in the set of all k possible or proposed conclusions.

In the first part of the paper the binary case will be analyzed, i.e. the case when the conclusion consists only of the hypothesis h and its negation $\bar{h} = NOTh$. However, in the general case a rule conclusion may consist of k hypotheses $\{h1, h2, ..., hk\}$. An example of such conclusion is given by (2).

$$\begin{array}{l} IF \\ (mp \in [100, 140])\, AND (cw \in [1000, 1200]) \\ THEN \\ P(afc < 7.0) = p_{h1} \\ OR \\ P(afc \in [7.0, 8.0]) = p_{h2} \\ OR \\ P(afc \in [8.0, 9.0]) = p_{h3} \\ OR \\ P(afc > 9.0) = p_{h4} \end{array} \tag{2}$$

where $\sum_{i=1}^{4} p_{hi} = 1$.

Probability of the binomial c-hypotheses h and $\bar{h}$ of the rules is estimated with various estimators. In case of machine learning, the most popular estimators are: the frequency estimator fr_h, the Laplace estimator Ep_{hL}, and the m-estimator Ep_{hM} (Cestnik, 1990, 1991; Chawla and Cieslak, 2006; Cichosz, 2000; Cussens, 1993; Sulzmann and Furnkranz, 2009, 2010). The authors of the present paper purposefully use notations fr_h, Ep_{hL}, and Ep_{hM} instead of p or Pr (which are frequently used in the literature) in order to prevent false suggestion that the estimation results are true probabilities. Results of estimation are, of course, only approximate values (estimates) of true probabilities. The frequency estimate $fr_h = n_h/n$ of c-hypothesis probability is referred to as the naive one (Sulzmann and Furnkranz, 2009, 2010), because it is characterized by many shortcomings. Perhaps its greatest shortcoming is that a single evidence sample piece produces drastic and non-acceptable conclusions concerning probability. If a single sample piece ($n = 1$) is the sample piece of type (1_h), which means that the sample piece confirms the c-hypothesis h, then the frequency estimate is $fr_h = n_h/n = 1/1 = 1$. This suggests the full truth of the hypothesis h and the full falsity of its negation, i.e. of the anti-hypothesis $\bar{h}$. If a single sample piece is of type ($1_{\bar{h}}$), which means that it confirms the anti-hypothesis $\bar{h}$, then the situation turns to reverse. Generally, the frequency estimator has

great average errors of probability estimation for the small number n of sample pieces.

Because of many shortcomings of the frequency estimator, Cestnik (1990, 1991) proposed application of the Laplace estimator Ep_{hL} (3) and m-estimator Ep_{hM} (4).

$$\begin{aligned} Ep_{hL} &= \tfrac{n_h+1}{n+a} \\ Ep_{\bar{h}L} &= 1 - Ep_{hL} = \tfrac{n_{\bar{h}}+a-1}{n+a}. \end{aligned} \tag{3}$$

Coefficient a can be interpreted in different ways. According to Furnkranz (2005); Sulzmann and Furnkranz (2009, 2010), Laplace modified the frequency estimator fr_h by adding one sample to each possible c-hypothesis. Thus, a means, according to this interpretation, the sample number and should be an integer. Next, because the binomial rule conclusion has two hypotheses, h and $\bar{h}$, the coefficient a takes value of 2.

The estimator Ep_{hL} can be viewed as a trade-off between the naive frequency estimator fr_h and the uniform distribution of a priori probability for conclusion hypotheses, namely that h and $\bar{h}$ equals $1/2$ in this case. The m-estimator Ep_{hM} is a generalization of the Laplace-estimator and it enables taking into account differentiated a priori knowledge about probabilities p_h and $p_{\bar{h}}$.

$$\begin{aligned} Ep_{hM} &= \tfrac{n_h+aEp_h(0)}{n+a} = fr_h \tfrac{n}{n+a} + Ep_h(0)\tfrac{a}{n+a} \\ Ep_{\bar{h}M} &= 1 - Ep_{hM}. \end{aligned} \tag{4}$$

The parameter $Ep_h(0)$ is interpreted by Cestnik as the prior estimate of probability p_h based on expert knowledge. The trade-off parameter a should be adapted to the prior $Ep_h(0)$ to allow for the optimal probability estimation. An interesting fact is that Cestnik (1990, 1991), Furnkranz (2005) and Sulzmann nad Furnkranz (2009, 2010) did not suggest that the coefficient a should be an integer. Cestnik only stated that a is a parameter which can be used to manage the trade-off between the prior and the posterior probability and that higher values should be used for noisy data (where the prior should be weighed higher) and lower values should be used for clean data (where the frequency fr_h should be weighed higher). In the case of the binomial conclusion the parameter a has to be equal in Ep_{hM} and $Ep_{\bar{h}M}$. Thus, at complete lack of sample pieces ($n = n_h = n_{\bar{h}} = 0$) the estimates (4) take the form of (5):

$$\begin{aligned} Ep_{hM} &= \tfrac{0+aEp_h(0)}{0+a} = Ep_h(0) \\ Ep_{\bar{h}M} &= 1 - Ep_{hM} = 1 - Ep_h(0). \end{aligned} \tag{5}$$

As it will be shown further on, a different interpretation of parameters $Ep_h(0)$ and a is also possible. It should be noticed that the m-estimate has certain shortcomings.

- The a priori estimate $Ep_h(0)$ is understood as a point value, e.g. 0.15. However, because we have no precise knowledge about the true value of probability p_h, it would be more reasonable to assume certain interval $[p_{h\,\min}, p_{h\,\max}]$ in which, according to a problem expert, the estimated probability is contained.

- Both the Laplace estimator and the m-estimator are global estimators, i.e. they are based on the assumption that the estimated probability can take any value in interval [0,1]. However, in many real problems such assumption is not reasonable. Let us consider the task of estimating the probability that a certain political party (party P) will win in the future elections. If we know that the party has at present the popularity of about 5% in the public opinion polls, then in the investigation the global estimators for $p_h \in [0,1]$ should not be used. It is more useful to apply an m-estimator that is specialized in probability estimation in the interval $p_h \in [0, 0.1]$ and produces more precise predictions within it.
- The parameter $Ep_h(0)$ is in case of the m-estimate interpreted as the suspected value of probability p_h. As it will be shown, this interpretation cannot always be used.

The general organization of the paper is as follows: in Section 2 the optimal estimate $Ep_h(1_h)$ of hypothesis h from one data piece (1_h) and the global probability estimator $Ep_h(n)$ for the case of no expert knowledge will be derived. In Section 3 specialized, local m-estimator of probability in limited intervals $p_h \in [p_{h\,\min}, p_{h\,\max}]$ resulting from expert knowledge for the binary case will be derived. In Section 4 specialized m-estimator for non-binary rule conclusions with k-hypothesis (k-nary case) are presented.

2. The optimal probability estimate $Ep_h(1_h)$ of hypothesis h from one data piece (1_h) and the global probability estimator $Ep_h(n)$ for the case of no expert knowledge

In the section, the derivation of the optimal m-estimate of the global character ($p_h \in [0,1]$) will be presented. The result of this derivation is not new. However, this time the derivation will be oriented on the 'single case' problem, i.e. on concluding about probability from only one, single sample item. This problem has often been discussed in the literature of the subject and is in principle not solved. Some specialists are of the opinion that probability from single evidence pieces has no sense (Hajek, 2010; Mises, 1957). We disagree. In the presented derivation, it will be shown that such a probability estimate is useful for improving probability estimates based on small number of sample items. The optimal value $Ep_h(1_h)$ of the probability estimate from one sample item will be derived. Such estimate has very small credibility because it is based on only one piece of evidence. This optimal estimate $Ep_h(1_h)$ is the basis for probability estimation for larger number of sample items. It will also be very important for determination of the parameters $Ep_h(0)$ and a in the formula of the m-estimate, both in the global and in the specialized case.

Let us assume a rule with the binomial conclusion $c = \{h, \bar{h}\}$, where p_h means probability of the conclusion hypothesis h and $p_{\bar{h}}$ probability of its negation $\bar{h} = NOTh$, and $p_h + p_{\bar{h}} = 1$. Let us next assume that only one sample piece (1) concerning the considered rule ($n = 1$) is at our disposal. If the sample item confirms the c-hypothesis h ($n_h = 1$, $n = 1$) then it is denoted by (1_h), if

it negates the hypothesis h and confirms the anti-hypothesis $\bar{h}$ ($n_{\bar{h}} = n = 1$), it will be denoted by $(1_{\bar{h}})$. Now, let us analyze the case of the sample piece (1_h) confirming the c-hypothesis. Which value $Ep_h(1_h)$ should we infer from this sample piece about probability of the c-hypothesis? Let us notice that the naive frequency estimator infers a drastic estimate from the confirming sample piece:

$$\begin{aligned} &fr_h(1_h) = n_h/n = 1/1 = 1 \\ &fr_h(1_{\bar{h}}) = n_{\bar{h}}/n = 0/1 = 0 \\ &fr_h(1_h) + fr_h(1_{\bar{h}}) = 1. \end{aligned} \tag{6}$$

Such estimate value seems unacceptable. Thus, the question arises: which estimate value $Ep_h(1_h)$ or $Ep_h(1_{\bar{h}})$ from one sample piece would be acceptable? Both estimates should satisfy the condition (7):

$$Ep_h(1_h) + Ep_h(1_{\bar{h}}) = 1. \tag{7}$$

To determine the optimal value that would hopefully be acceptable, an optimality criterion has to be chosen. In this case the MSE criterion (mean-square-error) was chosen. The probability estimate $Ep_h(1_h)$ inferred from one sample piece can not in the general case be precise and it will have the MSE error expressed by (8):

$$\Delta^{sqr}(1_h) = [p_h - Ep_h(1_h)]^2. \tag{8}$$

In a similar way the m-estimate of anti-hypothesis $\bar{h}$ will also have the MS error expressed by (9):

$$\Delta^{sqr}(1_{\bar{h}}) = [p_h - Ep_h(1_{\bar{h}})]^2 = [p_h - [1 - Ep_h(1_h)]]^2. \tag{9}$$

Usually, for probability estimation we have N sample pieces ($N > 1$) and N_h of them confirm the c-hypothesis of the rule and $N_{\bar{h}}$ confirm the anti-hypothesis $\bar{h}$ ($N_h + N_{\bar{h}} = N$).

The sum of the estimation errors $\Delta^{sqr}(N_h)$ from N_h sample pieces (1_h) confirming the c-hypothesis h is expressed by formula (10) and the error sum $\Delta^{sqr}(N_{\bar{h}})$ from all $N_{\bar{h}}$ sample pieces (1_h) confirming the anti-hypothesis h is expressed by formula (11):

$$\Delta^{sqr}(N_h) = [p_h - Ep_h(1_h)]^2 N_h \tag{10}$$

$$\Delta^{sqr}(N_{\bar{h}}) = [p_h - Ep_h(1_{\bar{h}})]^2 N_{\bar{h}} = [p_h - [1 - Ep_h(1_h)]]^2 N_{\bar{h}}. \tag{11}$$

The error sum $\Delta^{sqr}(N)$ of probability estimations from all N sample items, both from N_h sample items confirming the c-hypothesis h and $N_{\bar{h}}$ sample items confirming the anti-hypothesis $\bar{h}$ is determined by formula (12):

$$\Delta^{sqr}(N) = [p_h - Ep_h(1_h)]^2 N_h + [p_h - [1 - Ep_h(1_h)]]^2 N_{\bar{h}}. \tag{12}$$

Let us assume that the number of sample pieces at disposal, N, approaches infinity. Then, the number N_h of sample items confirming the hypothesis h,

according to the probability definition, approaches $N_h = N \cdot p_h$ and the number $N_{\bar{h}}$ of sample items negating the hypothesis $\bar{h}$ approaches $N_{\bar{h}} = N \cdot p_{\bar{h}} = N(1-p_h)$. Thus, for $N \to \infty$, equation (12) takes the form of (13):

$$\Delta^{sqr}_{N\to\infty}(N) = [p_h - Ep_h(1_h)]^2 N p_h + [p_h - [1 - Ep_h(1_h)]]^2 N(1-p_h). \quad (13)$$

Thus, the average MSE error, denoted by $\Delta^{sqr}_{aver}(1)$, of one sample item, independently of the fact whether the sample piece confirms (1_h) or negates $(1_{\bar{h}})$, the hypothesis, has a finite value:

$$\Delta^{sqr}_{aver}(1) = \frac{1}{N}\Delta^{sqr}_{N\to\infty}(N) = [p_h - Ep_h(1_h)]^2 p_h + [p_h - [1 - Ep_h(1_h)]]^2 (1-p_h). \quad (14)$$

Fig. 2 shows the functional surface of the average square error resulting from sample items (1_h) confirming the hypothesis h.

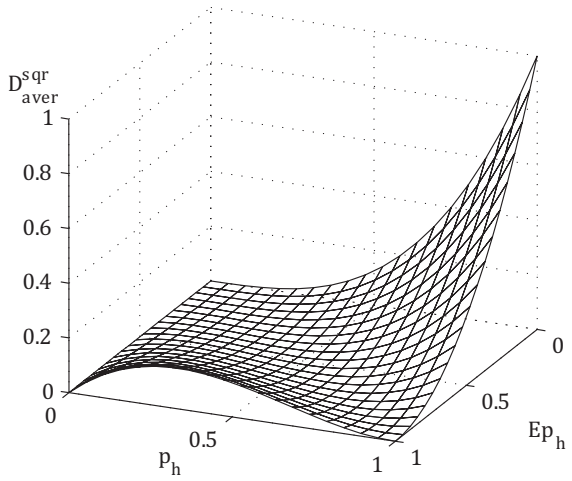


Figure 2. The functional surface of the first component $\Delta^{sqr}_{aver}(1_h) = [p_h - Ep_h(1_h)]^2 \cdot p_h$ of the average error $\Delta^{sqr}_{aver}(1)$, resulting from the sample items confirming the hypothesis h

The functional surfaces $\Delta^{sqr}_{aver}(1_h)$ and $\Delta^{sqr}_{aver}(1_{\bar{h}})$ are mutually symmetrical. Fig. 3 shows the complete functional surface of the full error $\Delta^{sqr}_{aver}(1)$.

Fig. 3 allows for interesting observations. The true probability value p_h is not known. It is to be estimated. But we can choose such a value $Ep_h(1)$ of the probability estimate for one confirming sample item (1_h) that will be optimal in the sense of the MSE-criterion. The optimal estimate will minimize the risk of making large errors of probability estimation. Fig. 4 presents the section of the functional surface of the square error $\Delta^{sqr}_{aver}(1)$ for the estimate value

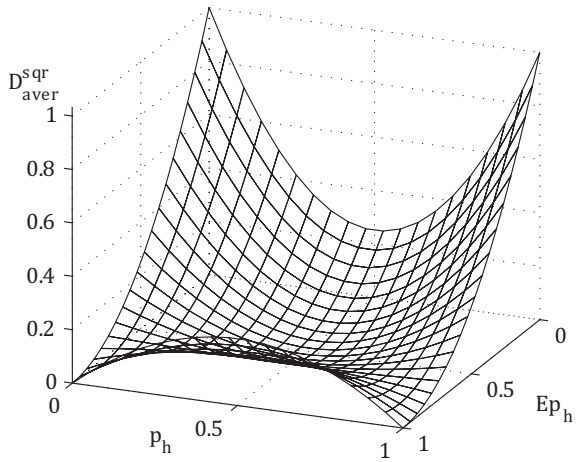


Figure 3. The functional surface $\Delta^{sqr}_{aver}(1) = \Delta^{sqr}_{aver}(1_h) + \Delta^{sqr}_{aver}(1_{\bar{h}})$ of the average error $\Delta^{sqr}_{aver}(1)$, resulting from the sample items both confirming and negating the hypothesis h, formula (14)

$Ep_h(1_h) = 1$. This estimate value corresponds to the probabilistic conclusion drawn from one sample item (1_h) by the naive frequency estimator fr_h.

The visual analysis of Fig. 3 shows that there are many values of the estimate $Ep_h(1_h)$, which generate smaller values of the square error than the value $Ep_h(1_h) = 1$. It means that assigning the radical confirmation strength equal to 1 by the universally used frequency estimator $fr_h = n_h/n$ to the single sample item 1_h is not the best idea. Thus, the optimal value of the one-sample item estimate $Ep_h(1_h)$ that minimizes the cross section area A of the one-sample item square-error function $\Delta^{sqr}_{aver}(1)$ should be determined.

The square-error area A that should be minimized is expressed by (15):

$$A = \int_{p_{h\,\min}}^{p_{h\,\max}} \Delta^{sqr}_{aver}(1)dp_h = \int_{0}^{1} \Delta^{sqr}_{aver}(1)dp_h. \tag{15}$$

In formula (15) the interval $p_h \in [p_{h\,\min}, p_{h\,\max}]$ is the probability interval in which, according to the problem expert, lies the true probability p_h of the hypothesis h contained in the rule conclusion. Let us first analyze the situation, in which the problem expert has no knowledge about the true probability p_h. Therefore, the expert assumes the full probability interval $p_h \in [0, 1]$ as admissible.

If the full probability interval $p_h \in [0, 1]$ were assumed, then the MS-error area A for this interval, given by formula (15), should be minimized.

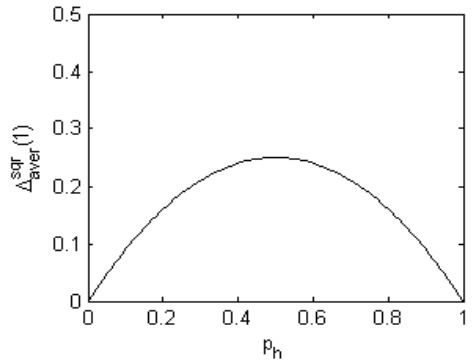


Figure 4. Cross section of the average single sample item square error $\Delta^{sqr}_{aver}(1)$ from Fig. 3 for the probability estimate $Ep_h(1_h) = 1$ that corresponds to the estimate calculated by the frequency estimator $fr_h = n_h/n$

After integrating formula (15), the formula (16) for the error area A is obtained:

$$A = Ep_h^2(1_h) - \frac{4}{3}Ep(1_h) + \frac{1}{2}. \tag{16}$$

The derivative of A equated to 0 is given by (17):

$$\frac{\partial A}{\partial Ep_h(1_h)} = 2Ep(1_h) - \frac{4}{3} = 0. \tag{17}$$

The solution of equation (17) delivers the optimal value of the one-sample item estimate:

$$Ep_h^{opt}(1_h) = \frac{2}{3} \tag{18}$$

where $p_h \in [0, 1]$.

After inserting the optimal value 2/3 in (16), the minimal value of the square-error area $A = 1/18$ is obtained.

For visual comparison, Fig. 5 presents the cross section of the error-area function for the value $Ep_h(1_h) = 1$ and for $Ep_h^{opt}(1_h) = 2/3$.

The error area $A = 1/6$ of the frequency estimator fr_h is three times larger than the minimal area $A = 1/18$ that can be obtained with the optimum estimator. An interesting fact should be noted: from one, single sample item (1_h) confirming the c-hypothesis h, the frequency estimator draws the drastic conclusion $fr_h = n_h/n = 1/1 = 1$ meaning the complete certainty of the c-hypothesis. It also draws a similar, drastic conclusion about the complete falsehood of the c-anti-hypothesis $\bar{h}$ ($fr_h = n_{\bar{h}}/n = 0/1 = 0$). Instead, the optimal conclusion drawn from one sample item should be different: $Ep_h(1_h) = 2/3$ and $Ep_h(1_{\bar{h}}) = 1/3$. Such inference is more rational.

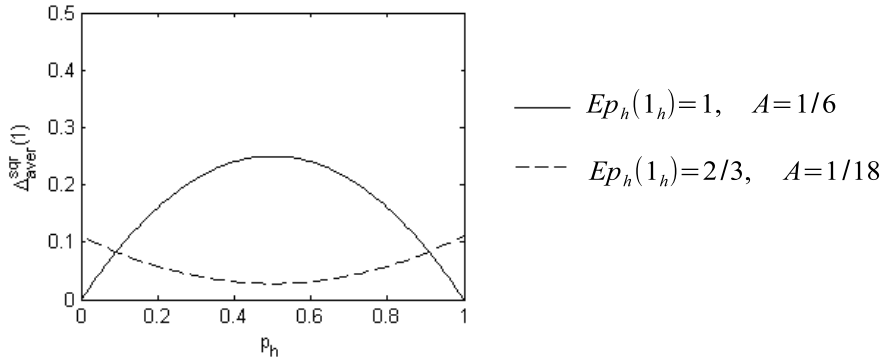


Figure 5. Comparison of the square error function $\Delta^{sqr}_{aver}(1)$ for the value $Ep_h(1) = 1$ corresponding to the frequency estimator $fr_h = n_h/n$ and of the function for the optimal estimate $Ep_h(1) = 2/3$

In the literature of the subject, various probability estimators have been proposed. The one-sample item estimate value $Ep_h(1_h)$, calculated by the considered estimator can be a good and simple test of its optimality in the sense of the criterion (14), minimizing the expected square-error.

M-estimator (4) contains the coefficient a which enables the existence of the prior estimate $Ep_h(0)$ in the case of complete lack of sample items ($n = 0$). However, the task of this coefficient, according to Cestnik, was to enable a trade-off between the prior estimate and the relative frequency estimate based on evidence. Cestnik (1990, 1991) proposed using higher values of a for low-quality experimental data (noisy data) and lower values for high-quality experimental data. It should be noted that the choice of a concrete value of a has to be made by the problem expert. The problem expert has to choose both the prior $Ep_h(0)$ and the coefficient a. Based on (4) we have:

$$Ep_h(n) = w_1 Ep_h(0) + w_2 fr_h(n) \tag{19}$$

where $w_1 = \frac{a}{n+a}$ and $w_2 = \frac{n}{n+a}$, $w_1 + w_2 = 1$.

Only for a sufficiently high number of data pieces the influence of the expert-opinion decreases and the estimate becomes more and more objective. This statement constitutes the basis for a new interpretation of the coefficient a. The coefficient w_1 in formula (19) of the m-estimate is a relative value of the expert-knowledge expressed in the form of the prior $Ep_h(0)$. The coefficient w_2 expresses the relative value of the experimental knowledge $fr_h(n)$. With the increase of the number of data pieces, n, the value of $w_1(n)$ decreases to zero (20) and the value of the experimental knowledge $w_2(n)$ increases to 1:

$$\lim_{n\to\infty} w_1(n) = \lim_{n\to\infty} \frac{a}{n+a} = 0 \tag{20}$$

$$\lim_{n\to\infty} w_2(n) = \lim_{n\to\infty} \frac{n}{n+a} = 1. \tag{21}$$

It can be easily concluded from these two facts that for a certain number of sample items, n, values of both weight coefficients $w_1(n)$ and $w_2(n)$ of the prior- and of the experimental knowledge must be equal and will have the value of 1/2:

$$w_1(n) = \frac{a}{n+a} = w_2(n) = \frac{n}{n+a} = \frac{1}{2}. \tag{22}$$

Solution of equation (22) yields the value of $n = a$. Thus, the coefficient a can be interpreted as the sample item number n at which the knowledge relating to probability obtained from the experiment in the form of frequency $fr_h = n_h/n$ is equally important with the prior $Ep_h(0)$. In real problems, the experimental knowledge can be of high quality (clean) or of low quality (noisy). This quality must be evaluated by experts.

Now, let us consider probability estimation for the case when the problem expert has completely no knowledge about the value of the estimated probability p_h and about the probability interval $p_h \in [p_{h\,\min}, p_{h\,\max}]$. Then, the full, global interval of potential values of probability p_h should be assumed. It is rational to assume as the initial value of the probability estimate $Ep_h(0) = 0.5$, because this value minimizes the criterion Cr^{abs} of the maximal, possible absolute-error to 0.5:

$$Cr^{abs} = \min\left[\max\left|p_h - Ep_h(0)\right|\right]. \tag{23}$$

Assumption of $Ep_h(0) = 0.5$ results in the mathematical form of the estimator given by (24):

$$Ep_h(n) = \frac{n_h + 0.5a}{n+a}. \tag{24}$$

Now, the optimal value of the coefficient a is to be determined. Therefore, the optimal value of the estimate $Ep_h(1)$, given by (18) for $n = n_h = 1$, can be used:

$$Ep_h^{opt}(n) = \frac{2}{3} = \frac{n_h + 0.5a}{n+a} = \frac{1 + 0.5a}{1+a}. \tag{25}$$

The solution of equation (25) gives the optimal value $a_{opt} = 2$ and the estimator $Ep_h(n)$ takes its final form:

$$Ep_h(n) = \frac{n_h + 1}{n+2}. \tag{26}$$

It is the Laplace estimator given by (3). Thus, Laplace estimator is the optimal (in the sense of the MSE-criterion) and global ($p_h \in [0, 1]$) estimator for the situation when the problem expert is not able to even approximately evaluate the probability p_h.

An interesting information resulting from the Laplace estimator is the one concerning the trade-off coefficient value $a = 2$. It means that in the considered binomial problem the value w_1 of the experimental knowledge, expressed by the

frequency fr_h, becomes the knowledge expressed by the prior $Ep_h(0) = 0.5$ at only two sample items. Such prior value seems low. It also should be noticed that this value was obtained as a result of application of the square-error (MSE)-criterion (14). It can be shown that application of the absolute error criterion in derivation of the optimal value of the estimate $Ep_h(n)$ results in a different value of $a = \sqrt{2} \approx 1.41$, not being an integer (Piegat and Landowski, 2012). It also shows that the probability estimate depends on the criterion chosen. The decision which error-criterion (incompatibility criterion) is to be chosen depends on the problem expert.

Now, let us consider the Cestnik m-estimator in its shortened form:

$$Ep_h(n) = \frac{n_h + aEp_h(0)}{n + a}. \tag{27}$$

The main aim of introduction of the m-estimator by Cestnik was to enable the use of a priori expert knowledge in probability estimation. As it will be shown, the coefficient $Ep_h(0)$ can be interpreted and used as a priori expert knowledge about probability. However, this is not always true and must be used with certain limitations (in a limited probability interval). Moreover, a new interpretation of this coefficient, that seems to be more general, will be presented. Cestnik proposed the following way of determining the coefficient: in the first step the prior value $Ep_h(0)$ should be given by the problem expert, and next the value of the trade-off coefficient a should be selected. According to this proposal, the prior value $Ep_h(0)$ is independent and it should not be correlated with the value of a. It will be shown that such advice will sometimes lead to a considerable decrease of the estimation accuracy. This is caused by the fact that not all $Ep_h(0)$-values are allowed (from the optimality point of view) and that two-sided correlation of both coefficients is necessary. Correlating $Ep_h(0)$ with the value of the coefficient a is realized on the basis of the optimal estimate-value for the "single case problem" $n = n_h = 1$. The optimal value in the global case is $Ep_h^{opt}(1_h) = 2/3$, formula (18). From (18) and (27), the formula (28) is obtained:

$$Ep_h^{opt}(1_h) = \frac{n_h + aEp_h(0)}{n + a} = \frac{1 + Ep_h(0)a}{1 + a} = \frac{2}{3}. \tag{28}$$

Solving equation (28) gives

$$a = \frac{1 - Ep_h^{opt}(1_h)}{Ep_h^{opt}(1_h) - Ep_h(0)} = \frac{1}{2 - 3Ep_h(0)}. \tag{29}$$

The value of the initial estimate $Ep_h(0)$ should be assumed in such a way that the correlated coefficient a will be positive. Negative a-values result in negative probability estimates. Therefore, conditions (30) have to be satisfied:

$$\begin{aligned} &Ep_h^{opt}(1_h) - Ep_h(0) > 0 \\ &Ep_h(0) < Ep_h^{opt}(1_h) \\ &Ep_h(0) < \tfrac{2}{3}\ . \end{aligned} \tag{30}$$

It results from (30) that the prior $Ep_h(0)$ should not take greater values than the optimal, one-sample item estimate $Ep_h(1_h)$. Is the above conclusion correct? To check this statement let us investigate what happens if we apply the prior-value exactly equal to the limit value of 2/3. Then, according to (27), the optimal value of the trade-off coefficient a approaches infinity:

$$a^{opt} = \lim_{Ep_h(0)\to 2/3^-} \frac{1}{2 - 3Ep_h(0)} = \infty. \tag{31}$$

What does this value mean in practice? To answer this question let us analyze formula (32) for the m-estimate in the not-shortened form:

$$\begin{array}{l} Ep_h(n) = w_1 Ep_h(0) + w_2 fr_h \\ = \frac{a}{n+a} Ep_h(0) + \frac{n}{n+a} fr_h \ . \end{array} \tag{32}$$

It can be easily checked that if the trade-off coefficient a approaches infinity then the relative weight w_1 of the prior knowledge increases to 1 and the relative value w_2 of the experimental knowledge decreases to zero, which means that the prior $Ep_h(0)$ would be of the highest value, because w_2 is multiplied in (32) by fr_h:

$$\begin{array}{l} \lim\limits_{a\to\infty} w_1 = \lim\limits_{a\to\infty} \frac{a}{n+a} = 1 \\ \lim\limits_{a\to\infty} w_2 = \lim\limits_{a\to\infty} \frac{n}{n+a} = 0. \end{array} \tag{33}$$

The coefficient value $w_2 = 0$ would also mean that the experimental results expressed by the frequency $fr_h = n_h/n$ could not introduce any correction to the prior $Ep_h(0)$, no matter how imprecise would be the prior and how high would be the number of data pieces n. Such situation is not acceptable, since the expert knowledge contained in the prior-estimate $Ep_h(0)$ is only of approximate character, it is virtually never precise and has to be corrected and improved by experiments. Thus, very high values of the trade-off coefficient a should not be used. Fig. 6 shows the dependence between the optimal values of this coefficient and the assumed prior value $Ep_h(0)$.

As it is shown in Fig. 6, from the point of view of estimation optimality the allowed interval of the prior $Ep_h(0)$ is the interval $0 \le Ep_h(0) < 2/3$ and the allowed interval of the coefficient a is $1/2 \le a < \infty$. Even though also for values $Ep_h(0) > 2/3$ the "optimal" values of the trade-off coefficient a "theoretically" can be calculated from formula (29), these values will be negative and therefore they have no substantial sense. Their application results in negative values of calculated probabilities, in probabilities greater than one and in infinitely large probabilities, as it can easily be checked by simple calculations. However, at this point the following question can be stated: "What happens if we use a forbidden value of the prior $Ep_h(0)$ together with a positive, allowed value of the coefficient a?" If we apply such an option, then the coefficient pair $\{Ep_h(0), a\}$ will not be optimal according to the MSE-criterion (14) and the estimation error will be unnecessarily higher than in case when we use the optimal, correlated pair $\{Ep_h^{opt}(0), a^{opt}\}$.

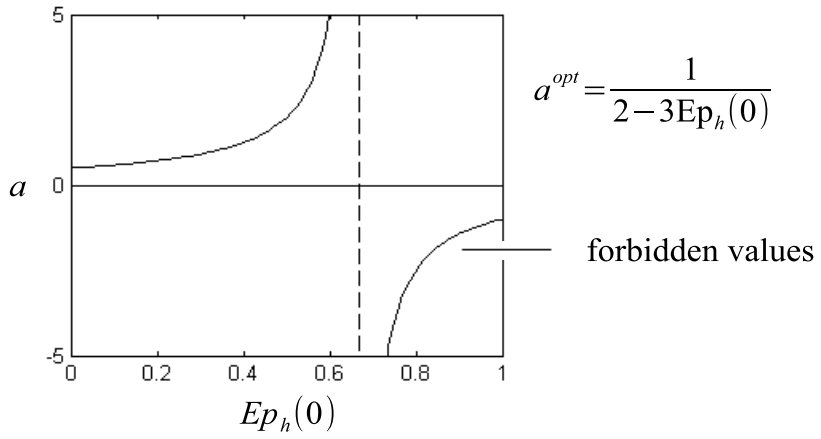


Figure 6. Dependence between the optimal value of the trade-off coefficient a and the assumed prior-value $Ep_h(0)$ of the estimated probability p_h

In the sequel, an example will be shown, demonstrating how much the estimation error can increase when inappropriate values of $Ep_h(0)$, which are not correlated with the coefficient a, are assumed. In this example let us assume that the true probability $p_h = 0.75$ is known and that it was correctly evaluated by the problem expert as being about 0.75. Now, let us assume, according to the proposal from Cestnik, the same "prior" value $Ep_h(0) = 0.75$. Let us notice that $Ep_h(0) = 0.75$ is greater than the allowed value 2/3 of $Ep_h^{opt}(1_h)$. For $a = 2$, according to (27), the m-estimate will take the form of:

$$Ep_h(n) = \frac{n_h + aEp_h(0)}{n + a} = \frac{n_h + 1.5}{n + 2}. \tag{34}$$

For $n = n_h = 1$, the estimate value $Ep_h(1_h) = 5/6$ is obtained. This value is not the optimal value $Ep_h^{opt}(1_h) = 2/3$. Now, it will be shown how large is the MSE-error (14) of the estimate for single sample items generated with exact probability 0.75. The results are given by (35):

$$\begin{aligned} &\Delta_{aver}^{sqr}(1) = [p_h - Ep_h(1_h)]^2 p_h + [p_h - (1 - Ep_h(1_h))]^2 (1 - p_h) \\ &= \left(\tfrac{3}{4} - \tfrac{5}{6}\right)^2 \tfrac{3}{4} + \left[\tfrac{3}{4} - \left(1 - \tfrac{5}{6}\right)\right]^2 \left(1 - \tfrac{3}{4}\right) = \tfrac{52}{576} \ . \end{aligned} \tag{35}$$

Now, let us take into account the knowledge that $Ep_h(0)$ has to be correlated with the coefficient a according to formula (29) and also correlated with the optimal value $Ep_h(1_h) = 2/3$ ($Ep_h(0) < Ep_h(1_h) = 2/3$). In the example, the value $Ep_h(0) = 0.6$ was assumed. Then, on the basis of (29), the optimal and correlated value of the coefficient a can be calculated:

$$a^{opt} = \frac{1}{2 - 3Ep_h(0)} = \frac{1}{2 - 3 \cdot 0.6} = 5 \ . \tag{36}$$

In this way, formula (37) is obtained, which differs from formula (34), for the m-estimate:

$$Ep_h(n) = \frac{n_h + aEp_h(0)}{n+a} = \frac{n_h + 5 \cdot 0.6}{n+5} = \frac{n_h + 3}{n+5} \ . \tag{37}$$

For $n = n_h = 1$ the value $Ep_h(1_h) = Ep_h^{opt}(1_h) = 2/3$ is obtained. Now, let us calculate the average MSE-error for estimating probability $p_h = 3/4$ on the basis of single sample items, formulas (14) and (38):

$$\begin{aligned} &\Delta_{aver}^{sqr}(1) = \left[p_h - Ep_h(1_h)\right]^2 p_h + \left[p_h - (1 - Ep_h(1_h))\right]^2 (1 - p_h) \\ &= \left(\tfrac{3}{4} - \tfrac{2}{3}\right)^2 \tfrac{3}{4} + \left[\tfrac{3}{4} - \left(1 - \tfrac{2}{3}\right)\right]^2 \left(1 - \tfrac{3}{4}\right) = \tfrac{28}{576} \ . \end{aligned} \tag{38}$$

The comparison of formulas (35) and (38) shows that assuming the "a priori" estimate according to Cestnik's suggestion, which is precisely equal to 3/4, gives the average error $\Delta_{aver}^{sqr}(1) = 52/576$. However, assuming the value $Ep_h(0) = 0.6$, which is smaller than the true probability value of 0.75, gives a considerably smaller error $\Delta_{aver}^{sqr}(1) = 28/576$. These results show that Cestnik's opinion suggesting interpretation of $Ep_h(0)$ as "a priori", expected value of estimated p_h-probability, is in the general case not correct. Fig. 7 shows a comparison of the expected MSE-error $\Delta_{aver}^{sqr}(1)$ generated by single pieces and the error calculated with formula (14) for interval $p_h \in [0.5, 1]$, in which the true probability value 0.75 lies. In the first case the prior value $Ep_h(0) = 0.75$, according to Cestnik's suggestion, was assumed and in the second case the value $Ep_h(0) = 0.6$, which is compatible with the optimal form of the estimator $Ep_h(n) = (n_h + 3)/(n + 5)$, formula (37).

Resulting from the analysis of the above example the following question arises: if $Ep_h(0)$ does not have in the general case the meaning of the prior probability then how should it be interpreted?

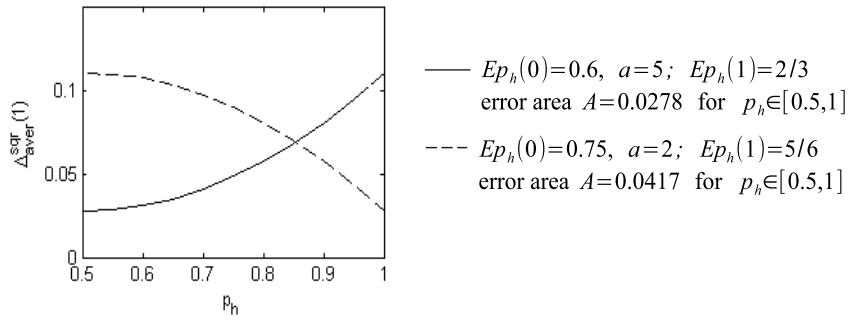
Let us assume that we have an information source delivering statistical data of binary character. An example of such a generator can be an oncology institute which provides data about consecutive patients with suspected lung cancer. On the basis of the data provided we want to determine the probability of the conclusion of following rule:

IF (a person smokes over 20 cigarettes a day) AND (smoking period is over 30 years long)

THEN [(the person has the lung cancer with probability p_h)

OR (the person has no lung cancer with probability $(1 - p_h)$]

The first part of the conclusion "the person has the lung cancer with probability p_h" is the c-hypothesis h and the second part of the conclusion is the anti-hypothesis $\bar{h}$. Let us assume that the true probability value p_h equals 0.75. However, as $Ep_h(0)$ the value of 0.6 was assumed. Then, in the result of obtaining successive information pieces about patients we will be able to calculate increasingly precisely the estimation $Ep_h(n)$ of the probability p_h of lung cancer. If the number of patients with lung cancer is greater than those without it, then the diagram of the probability estimation process will be approximately as in Fig. 8.



p_h	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1
$\Delta^{sqr}_{aver}(1)$ ——	0.0278	0.0286	0.0311	0.0353	0.0411	0.0486	0.0578	0.0686	0.0811	0.0953	0.1111
$\Delta^{sqr}_{aver}(1)$ ---	0.1111	0.1103	0.1078	0.1036	0.0978	0.0903	0.0811	0.0703	0.0578	0.0436	0.0278

Figure 7. Comparison of the MSE-errors of estimation of the probability $p_h = 0.75$ with the use of $Ep_h(0) = 0.75$ $(a = 2)$ according to Cestnik's suggestions and with the use of $Ep_h(0) = 0.6$ $(a = 5)$ according to the optimal form of the m-estimator $Ep_h(n) = [n_h + aEp_h(0)]/(n + a) = (n_h + 3)/(n + 5)$

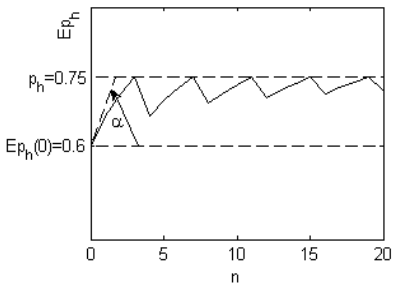
The meaning of the coefficients $Ep_h(0)$ and a can be explained on the basis of the initial derivative of the estimation process. The formula of the estimator $Ep_h(n)$ is given by (27). The general formula for its derivative is given by (39), for $n_h = n$:

$$\frac{\partial Ep_h(n)}{\partial n} = \frac{a(1 - Ep_h(0))}{(n + a)^2}. \tag{39}$$

In particular, for $n = 0$, formula (39) takes the form:

$$\frac{\partial Ep_h(n)}{\partial n}(n = 0) = \frac{1 - Ep_h(0)}{a} = \tan\alpha. \tag{40}$$

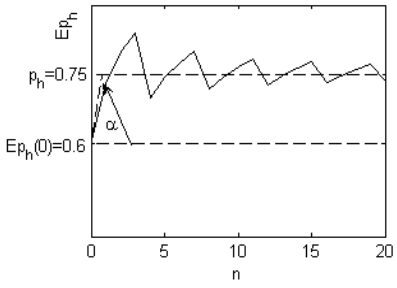
As shown in Fig. 7, the coefficient $Ep_h(0)$ that is interpreted by Cestnik as the "a priori" value of the estimated probability (as the value of this probability evaluated by the problem expert: in the example it takes the value of 0.75) in the general case does not have this meaning. The general interpretation of $Ep_h(0)$ is as follows: $Ep_h(0)$ is the initial value of the estimate, from which the identification process of probability begins when successive information sample items come. As formula (40) shows, assuming the initial value $Ep_h(0)$ near 1 makes the value of $\tan\alpha$ (Fig. 8) decrease. The derivative $\tan\alpha$ determines the speed of probability estimation (in technical control systems $\tan\alpha$ determines the time-constant of the control process, Lutz and Wendt, 1998). The estimation speed should be high, however, it should not be excessively high, because this may bring about oscillations in the estimation process. One can see from



$$\tan\alpha = \frac{\partial Ep_h}{\partial n}(n=0) = \frac{1 - Ep_h(0)}{a}$$

Figure 8. Example of the estimation (identification) process of the probability p_h of the c-hypothesis h in the course of obtaining successive sample items confirming or negating the c-hypothesis in case of domination of confirming sample items (small oscillations below p_h)

formula (40) that too big values of $Ep_h(0)$ in connection with too small values of the coefficient a cause high estimation speed, which results in oscillating estimation of probability in the course of appearance of the successive sample items, Fig. 9.



$$\tan\alpha = \frac{1 - Ep_h(0)}{a}$$

Figure 9. Example of the oscillatory process of the probability estimation in the course of appearance of successive information sample items as a result of too large speed of probability estimation (greater oscillations around p_h)

It should also be added that one of estimators with an excessively large estimation speed is just the naive frequency estimator $fr_h = n_h/n$. Estimation processes with its use are frequently oscillatory; examples can be found e.g. in Larose (2010).

Summarizing this section, we provide the advice relating to the practical application of the global m-estimator ($p_h \in [0, 1]$) of the hypothesis probability in the case when no expert knowledge is at disposal.

Step 1. Assume prior approximate value $Ep_h(0)$ of the estimated hypothesis probability p_h. This value should be less than 2/3, e.g. 0.5.

Step 2. Calculate the optimal value of the trade-off coefficient a^{opt} from formula (41):

$$a^{opt} = \frac{1}{2 - 3Ep_h(0)}. \tag{41}$$

Step 3. The optimal, global estimator $Ep_h^{opt}(n)$ of probability is then given by formula (42):

$$Ep_h^{opt}(n) = \frac{aEp_h(0) + n_h}{n + a^{opt}}. \tag{42}$$

3. Specialized, local m-estimators of probability in limited intervals $p_h \in [p_{h\,\min}, p_{h\,\max}]$ resulting from expert knowledge, for binary case

The global estimator of probability that is optimized for the full probability interval $p_h \in [0, 1]$ (Laplace estimator $Ep_{hL} = (n_h + 1)/(n + 2)$ is an example of such estimator) should be used only when the problem expert does not have any knowledge concerning the approximate interval in which the estimated probability p_h of the c-hypothesis lies. However, usually the expert has certain knowledge and is able to give an approximate probability interval $p_h \in [p_{h\,\min}, p_{h\,\max}]$ ($p_{h\,\min} < p_{h\,\max}$) of the c-hypothesis appearing in the rule. The problem expert can, e.g., know that $p_h \in [0, 1/4]$, or that $p_h \in [3/4, 1]$, or that $p_h \in [1/2, 3/4]$. Application of global estimators in these cases can result in considerable decrease of the estimation accuracy determined by, e.g., MSE-error $\Delta_{aver}^{sqr}(1)$, as expressed by formula (14). Instead, application of a local estimator, which was optimized for the interval $[p_{h\,\min}, p_{h\,\max}]$ of the estimated probability, allows for considerable reduction of the average estimation error. It should also be added that even when the interval $[p_{h\,\min}, p_{h\,\max}]$ is evaluated with a considerable excess, we will still achieve smaller estimation errors than with the global, not-specialized estimators. Even in drastic case, if the probability interval $[p_{h\,\min}, p_{h\,\max}]$ is determined fully erroneously and the true probability value lies outside of it, the specialized estimator will correctly identify this probability outside of the interval. However, the estimation MSE-error will not be the smallest one. The value of this error is given by (14) and by (43):

$$\Delta_{aver}^{sqr}(1) = [p_h - Ep_h(1_h)]^2 p_h + [p_h - [1 - Ep_h(1_h)]]^2 (1 - p_h). \tag{43}$$

In the sequel, on the basis of the MSE-error, the formulas for determining the values of the parameters $Ep_h(0)$ and a will be derived. The parameters of the m-estimator $Ep_{hM} = [n_h + aEp_h(0)]/(n + a)$ will be optimal for the interval $p_h \in [p_{h\,\min}, p_{h\,\max}]$ if they secure minimization of the area A of the functional

surface of the MSE-error given by (15) and (44):

$$\begin{aligned} A &= \int_{p_{h\,\min}}^{p_{h\,\max}} \Delta_{aver}^{sqr}(1) dp_h \\ &= \left[\left(1 - \tfrac{4}{3} Ep_h^{opt}(1_h)\right) p_h^3 + \left(2Ep_h^{opt}(1_h) - \tfrac{3}{2}\right) p_h^2 \right. \\ &\left. + [(Ep_h^{opt}(1_h))^2 - 2Ep_h^{opt}(1_h) + 1] p_h \right]_{p_{h\,\min}}^{p_{h\,\max}}. \end{aligned} \tag{44}$$

By minimizing the error area A in relation to $Ep_h(1_h)$, the optimal value of $Ep_h(1_h)$, given by (45), can be derived:

$$Ep_h^{opt}(1_h) = \frac{2}{3}\left(p_{h\,\min}^2 + p_{h\,\min} p_{h\,\max} + p_{h\,\max}^2\right) - (p_{h\,\min} + p_{h\,\max}) + 1 \tag{45}$$

where $p_{h\,\min} < p_{h\,\max}$.

The formula (45) was derived from formula:

$$Ep_h^{opt}(1_h) = 1 + \frac{\frac{2}{3}\left(p_{h\,\max}^3 - p_{h\,\min}^3\right)}{p_{h\,\max} - p_{h\,\min}} - \frac{p_{h\,\max}^2 - p_{h\,\min}^2}{p_{h\,\max} - p_{h\,\min}}.$$

In formula (45) $Ep_h^{opt}(1_h)$ means the optimal estimate of probability concluded from only one, single sample item confirming the c-hypothesis h of the rule conclusion. On the basis of $Ep_h^{opt}(1_h)$, the parameters $Ep_h(0)$ and a (for $n = n_h = 1$) can be chosen from (46):

$$Ep_h^{opt}(1_h) = \frac{1 + aEp_h(0)}{1 + a}. \tag{46}$$

When choosing $Ep_h(0)$ and a one should take into account conditions (35) and (36), given previously, and recalled here in (47):

$$\begin{aligned} &Ep_h(0) < Ep_h^{opt}(1_h) \\ &a = \frac{1 - Ep_h^{opt}(1_h)}{Ep_h^{opt}(1_h) - Ep_h(0)}. \end{aligned} \tag{47}$$

In the sequel, the examples of optimal, specialized m-estimators for various intervals $[p_{h\,\min}, p_{h\,\max}]$ of probability will be determined. First, let us analyze the case of p_h being contained in the interval $[0, 1/4]$. According to (44), the optimal value $Ep_h^{opt}(1_h)$ of the single-sample item estimate that minimizes the error area A should be determined with (48):

$$A = \int_0^{1/4} \Delta_{aver}^{sqr}(1) dp_h. \tag{48}$$

Thus, the ready formula (45) will be used. In the analyzed example it takes the form of (49):

$$Ep_h^{opt}(1_h) = \frac{2}{3}\left(0 + 0 \cdot \frac{1}{4} + \left(\frac{1}{4}\right)^2\right) - \left(0 + \frac{1}{4}\right) + 1 = \frac{19}{24} \approx 0.79167. \tag{49}$$

It should be noted that the calculated value $Ep_h^{opt}(1_h) = 19/24 = 0.79167$, determined for the interval $[0, 1/4]$, differs from the optimal value $Ep_h^{opt}(1_h) = 2/3$, obtained for the global interval $[0, 1]$. The value $2/3$ refers to the Laplace estimator $Ep_{hL}(n) = (n_h + 1)/(n + 2)$. The optimal value $Ep_h^{opt}(1_h) = 19/24$ is the basis for choosing values of the coefficients $Ep_h(0)$ and a using formulas (47):

$$Ep_h(0) < Ep_h^{opt}(1_h) = 19/24$$
$$a = \frac{1 - Ep_h^{opt}(1_h)}{Ep_h^{opt}(1_h) - Ep_h(0)}.$$

In the result of choosing the initial value $Ep_h(0) = 1/8 = 0.125$ as the middle value in the interval $[0, 1/4]$, the following value of the parameter a is obtained: $a^{opt} = 5/16 = 0.3125$. Thus, the optimal m-estimator $Ep_{hM}(n)$ for the interval $[0, 1/4]$ is given by:

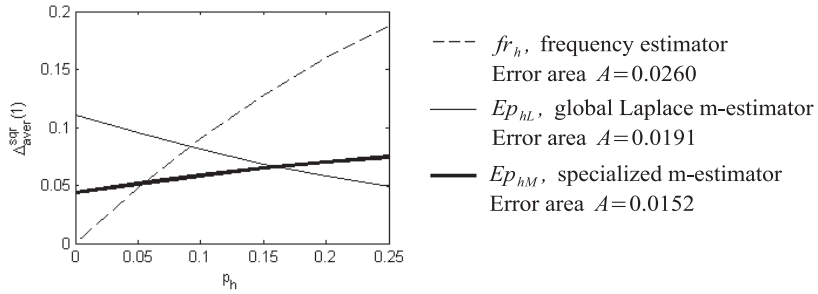
$$Ep_{hM}(n) = \frac{n_h + aEp_h(0)}{n + a} = \frac{n_h + \frac{5}{16} \cdot \frac{1}{8}}{n + \frac{5}{16}} = \frac{n_h + 0.03906}{n + 0.3125}. \tag{50}$$

One should note here that the value of the parameter $a = 5/16$ is not an integer. In the case of the Laplace estimator, which is a special case of the global m-estimator with $Ep_h(0) = 1/2$ and $a = 2$, the estimator form is $Ep_{hL}(n) = (n_h+1)/(n+2)$. Fig. 10 shows the diagrams of MSE-errors $\Delta_{aver}^{sqr}(1)$, given by formula (14), calculated for three various estimators of probability: for the frequency estimator fr_h, for Laplace estimator and for the optimal m-estimator given by formula (50).

Fig. 10 shows the superiority of the m-estimator specialized for the interval $p_h \in [0, 1/4]$ over the non-specialized, global frequency estimator and over the Laplace estimator with the integer value $a = 2$. A similar superiority of specialized m-estimators is achieved for other intervals $[p_{h\,\min}, p_{h\,\max}]$. If (for example) the expert knows that the true probability $p_h \in [1/4, 3/4]$ then the naive frequency estimator fr_h with $Ep_h(1_h) = 1$ has the error area $A = 0.1146$, the Laplace estimator with $Ep_h(1_h) = 2/3$ has $A = 0.0174$, and the specialized m-estimator with $Ep_h(0) = 1/2$, $a = 11$ and $Ep_h(1_h) = 13/24 = 0.5417$ has the error area A equal to only 0.0095, that is: more than two times less than the Laplace estimator. The MSE-error diagrams are presented in Fig. 11.

In the case considered above, when the hypothesis probability p_h belongs to the interval $[1/4, 3/4]$, the initial value of the estimate, $Ep_h(0) = 0.5$, could be chosen as the value which lies in the middle of the uncertainty interval of p_h and thus it is intuitively very acceptable, Fig. 12.

Now let us consider the case when, according to the expert knowledge, the hypothesis probability is contained in the interval $p_h \in [2/3, 1]$. Then, according to formula (45), the optimal, one-sample item estimate has the value $Ep_h(1_h) = 20/27$. In the case of the interval $[2/3, 1] = [4/6, 1]$ it seems intuitively rational to assume the initial estimate $Ep_h(0)$ in the middle of the interval i.e. $Ep_h(0) = 5/6$. However, this value does not satisfy the optimality condition (47) $Ep_h(0) < Ep_h(1_h)$ because $Ep_h(0) = 5/6 = 45/54$ is larger than



p_h		0	0.05	0.1	0.125	0.15	0.2	0.25
$\Delta^{sqr}_{aver}(1)$	---	0	0.0475	0.0900	0.1094	0.1275	0.1600	0.1875
$\Delta^{sqr}_{aver}(1)$	——	0.1111	0.0953	0.0811	0.0746	0.0686	0.0578	0.0486
$\Delta^{sqr}_{aver}(1)$	——	0.0434	0.0513	0.0584	0.0616	0.0647	0.0701	0.0747

Figure 10. Comparison of MSE-errors $\Delta^{sqr}_{aver}(1) = f(p_h)$ for interval $p_h \in [0, 1/4]$ for the naive frequency estimator $fr_h = n_h/n$, Laplace estimator $Ep_{hL} = (n_h + 1)/(n+2)$, and m-estimator $Ep_{hM} = (n_h+5/128)/(n+5/16)$ that was optimized according to the advice given in the paper

$Ep_h(1_h) = 40/54$. Assumption of the initial estimate $Ep_h(1_h) = 5/6 = 45/54$ would increase the expected square error of the estimation. Fig. 13 presents the parameter value distribution of the m-estimator with the initial estimate $Ep_h(0) = 5/6$ chosen in the middle of the interval $p_h \in [2/3, 1]$.

Since for the initial estimate $(Ep_h(0) = 40/54) > (Ep^{opt}_h(1_h) = 40/54)$ there exists no positive, optimal value of the trade-off coefficient a, therefore certain positive, though not optimal, value of this coefficient has to be assumed. It can be the value $a = 2$ applied in the Laplace estimator $Ep_h(n) = (n_h + 1)/(n + 2)$. Then, according to (50), the formula (51) for the m-estimator is obtained:

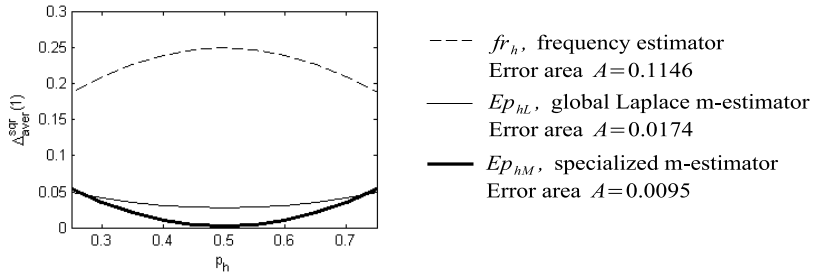
$$Ep_h = \frac{n_h + 1.667}{n + 2}. \tag{51}$$

Instead, if optimality advices (47) are followed, then the initial value $Ep_h(0)$ should be chosen, lower than the one-sample item estimate $(Ep_h(0) < (Ep_h(1_h) = 40/54)$. Let us choose a slightly smaller value of $Ep_h(0) = 38/54$. For this value there exists, (47), the optimal and positive value of the trade-off coefficient $a = 8$. Thus, the optimal specialized m-estimator is given by

$$Ep^{opt}_h = \frac{n_h + 5.63}{n + 8}. \tag{52}$$

The parameter value distribution of this estimate is shown in Fig. 14.

According to formula (44), the area A of the square error for both compared estimators (51) and (52), and the average estimation error Δ_{aver} for interval



p_h	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75
$\Delta^{sqr}_{aver}(1)$ ---	0.1875	0.2100	0.2275	0.2400	0.2475	0.2500	0.2475	0.2400	0.2275	0.2100	0.1875
$\Delta^{sqr}_{aver}(1)$ ——	0.0486	0.0411	0.0353	0.0311	0.0286	0.0278	0.0286	0.0311	0.0353	0.0411	0.0486
$\Delta^{sqr}_{aver}(1)$ ——	0.0538	0.0351	0.0205	0.0101	0.0038	0.0017	0.0038	0.0101	0.0205	0.0351	0.0538

Figure 11. Comparison of MSE-error diagrams for interval $p_h \in [1/4, 3/4]$ of the specialized m-estimator $Ep_{hM} = (n_h + 5.5)/(n + 11)$ and of the global, not-specialized estimators: of the frequency estimator $fr_h(n)$ and of Laplace estimator $Ep_{hL}(n)$

$p_h \in [2/3, 1]$ was calculated as below:

$$Ep_h(n) = \frac{n_h + 1.667}{n + 2}, A = 0.0281, \Delta_{aver} = A/(1/3) = 0.0844 \tag{53}$$

$$Ep_h^{opt}(n) = \frac{n_h + 5.63}{n + 8}, A = 0.0208, \Delta_{aver} = A/(1/3) = 0.0624. \tag{54}$$

As the results show, the estimator (54) with the correctly chosen initial estimate $Ep_h(0)$ has in the interval $p_h \in [2/3, 1]$ a smaller, average estimation error than the estimator with the incorrect initial estimate. Fig. 15 presents the charts of the MSE-error $\Delta^{sqr}_{aver}(1)$ calculated according to formula (14) for four compared estimators.

As shown in Fig. 15, the specialized m-estimator $Ep^{opt}_{hM}(n)$ has the smallest estimation error in the interval $p_h \in [2/3, 1]$. This result is rather clear because it is understandable that estimator, which does not have the optimal parameters, will have lower accuracy than an estimator with optimal parameters. Further on, results of comparison of three classifiers applied to detection of classification rules for a benchmark problem will be presented.

The Balance Scale data set comes from UC Irvine Machine Learning Repository (Siegler, 1994) and was generated to model the results of psychological experiments carried out by Siegler (1976). Using three types of naive Bayes classifiers with different probability estimators (frequency, Laplace and specialized) the examples from data set were classified to one of three classes: tip of the balance scale to the right, tip to the left, or the balance scale is balanced. The

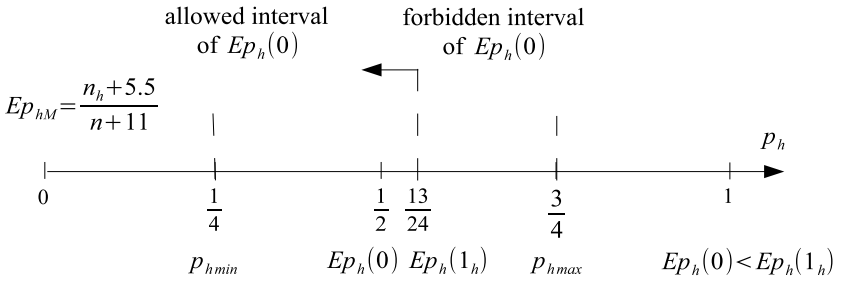


Figure 12. Parameter value distribution of the optimal specialized m-estimator for interval $p_h \in [1/4, 3/4]$

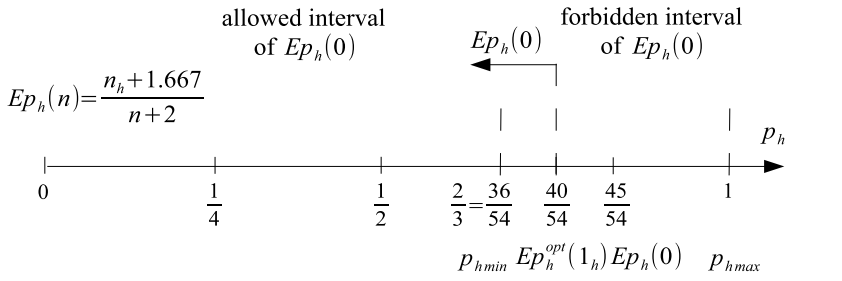


Figure 13. Incorrect choice of the initial estimate $Ep_h(0)$ in the middle ($45/54 = 5/6$) of the estimated probability interval $p_h \in [2/3, 1]$

data set consists of 625 instances and 5 attributes: class name (left, balance, right), left weight (1, 2, 3, 4, 5), left distance (1, 2, 3, 4, 5), right weight (1, 2, 3, 4, 5), and right distance (1, 2, 3, 4, 5). For example, element of the data set "2, 5, 2, 1" (left weight = 2, left distance = 5, right weight = 2, right distance = 1) should be classified to the class "left".

Fig. 16 shows the results of correct classifications by naive Bayes classifiers using three estimators and different number of elements in the learning data set. Results are the mean of 100 experiments in each case, expressed as a percentage of correct classifications of elements from the testing data set. The highest differences in classification were obtained for five examples in the learning data set. In this case the best result is given by the naive Bayes using specialized estimator (59% of correct classifications), the next is naive Bayes using Laplace estimator (54% of correct classifications) and the worst is naive Bayes with frequency estimator (47% of correct classifications). The results for 75 elements in the learning data sets show that the naive Bayes with frequency estimator

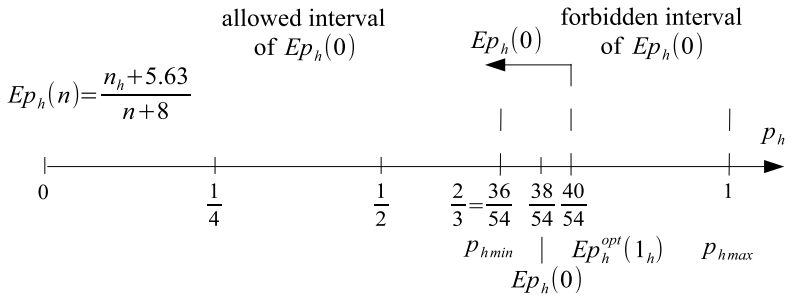


Figure 14. Parameter value distribution of the m-estimator with correctly chosen initial estimate $Ep_h(0) = 38/45$ satisfying the condition $Ep_h(0) < Ep_h^{opt}(1_h)$

gives 3% less of correct classifications than other classifiers, Fig.16.

At the end of this section, the pieces of advice referring to the parameter choice of the specialized m-estimators will be repeated:

- Determine the probability interval $[p_{h\,\min}, p_{h\,\max}]$.
- On the basis of formula (45) calculate the optimal value of the single-sample item estimate $Ep_h^{opt}(1_h)$.
- Taking into account the condition $Ep_h(0) < Ep_h^{opt}(1_h)$ and formula (46) choose value of $Ep_h(0)$. This value should lie possibly near the suspected value of the estimated probability p_h.
- Taking into account the chosen value $Ep_h(0)$ and using formula (46) choose the value of the parameter a that is correlated with $Ep_h(0)$.

Table 1 presents an example of investigations on parameters of specialized m-estimator depending on a priori knowledge of interval probability. Results were obtained from formulas (45) and (46).

Table 1. Parameters of specialized m-estimators depending on a priori knowledge

$[p_{h\,\min}, p_{h\,\max}]$	[0,0.1]	[0,0.2]	[0.1,0.3]	[0.2,0.4]	[0.3,0.5]	[0.4,0.6]
$Ep_h^{opt}(1_h)$	0.9067	0.8267	0.6867	0.5867	0.5267	0.5067
$Ep_h^{opt}(0)$	0.05	0.1	0.2	0.3	0.4	0.5
a^{opt}	0.1089	0.2385	0.6438	1.4419	3.7368	74
$[p_{h\,\min}, p_{h\,\max}]$	[0.5,0.7]	[0.6,0.8]	[0.7,0.9]	[0.8,1]	[0.9,1]	
$Ep_h^{opt}(1_h)$	0.5267	0.5867	0.6867	0.8267	0.9067	
$Ep_h^{opt}(0)$	0.52	0.58	0.68	0.82	0.9	
a^{opt}	71	62	47	26	14	

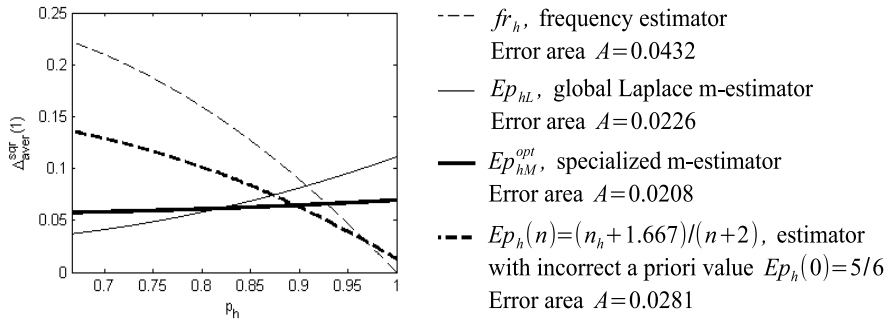


Figure 15. Comparison of MSE-erorrs $\Delta^{sqr}_{aver}(1) = f(p_h)$ for interval $p_h \in [2/3, 1]$ for the global ($p_h \in [0, 1]$) frequency estimator $fr_h = n_h/n$, the global Laplace estimator $Ep_{hL} = (n_h + 1)/(n + 2)$, the specialized m-estimator $Ep_{hM}(n) = (n_h + 1.667)/(n + 2)$ with incorrect initial estimate $Ep_h(0)$, and specialized m-estimator $Ep^{opt}_{hM}(n) = (n_h + 5.630)/(n + 8)$ with correctly chosen initial estimate

4. Specialized m-estimators for non-binary rule-conclusions with k hypotheses (k-nary case)

If a rule conclusion contains k hypotheses $\{h1, h2, \ldots, hk\}$ and on the basis of data pieces k-nary m-estimates $Ep_{hM1}, Ep_{hM2}, , Ep_{hMk}$ of probability are calculated, their sum is mostly not equal to 1,

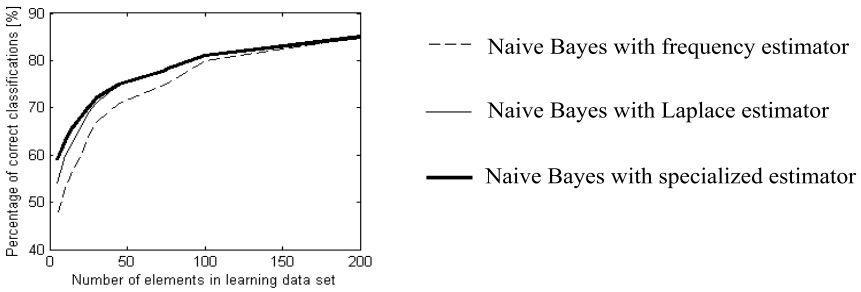
$$\sum_{i=1}^{k} Ep_{hMi} \neq 1.$$

Therefore, the individually calculated binary estimates Ep_{hMi} should be normalized according to formula (55), where Ep^k_{hMi} denotes the k-nary estimate of probability,

$$Ep^k_{hMi} = \frac{Ep_{hMi}}{\sum\limits_{i=1}^{k} Ep_{hMi}}. \tag{55}$$

5. Conclusions

The paper shows that the increase of rule quality in machine learning systems, such as, e.g., decision trees or rough set theory, is possible. At present, the probability of rule conclusions is determined with the frequency estimator, the Laplace estimator, and the m-estimator that are of global character. Because the problem expert frequently knows the approximate interval of probability, therefore probability estimators can be used that are specially tuned for the given probability interval. This allows for considerable increase of accuracy of probability estimation of rule conclusions. Specialized probability estimators



Number of elements in learning data set	5	10	15	20	25	30	45	75	100	200
Naive Bayes with frequency estimator [%]	47	53	57	60	64	67	71	75	80	85
Naive Bayes with Laplace estimator [%]	54	60	63	66	69	71	75	78	81	85
Naive Bayes with specialized estimator [%]	59	63	66	68	70	72	75	78	81	85

Figure 16. Percentage of correct classifications of examples from the testing data set, depending on the number of elements in the learning data set, using naive Bayes classifiers with frequency estimator, Laplace estimator, and specialized estimator

are of great importance for problems with small numbers of data pieces (instances). The paper explains how parameters of specialized m-estimators can be adapted to the probability interval given by the problem expert. The paper also explains why the hitherto existing interpretation of the m-estimator parameters is rather incorrect and gives the correct interpretation. The idea of specialized m-estimators of probability was conceived by Andrzej Piegat.

References

CESTNIK, B. (1990) Estimating probabilities: A crucial task in machine learning. In: L. C. Aiello (Ed.), *ECAI'90.* Pitman, London, 147-149.

CESTNIK, B. (1991) *Estimating probabilities in machine learning.* Ph.D. thesis, University of Ljubljana, Faculty of Computer and Information Science.

CHAWLA, N. V. and CIESLAK, D. A. (2006) Evaluating calibration of probability estimation from decision trees. *AAAI Workshop on the Evaluation Methods in Machine Learning*, The AAAI Press, Boston, July 2006, 18–23.

CICHOSZ, P. (2000) *Systemy uczące się (Learning systems).* Wydawnictwo Naukowo–Techniczne, Warsaw, Poland.

CUSSENS, J. (1993) Bayes and pseudo–bayes estimates of conditional probabilities and their reliabilities. In: *Proceedings of European Conference on Machine Learning*, *ECML-93.* **LNCS 667**, 136–152.

FURNKRANZ, J. & FLACH, P. A. (2005) ROC 'n' rule learning - towards a better understanding of covering algorithms. *Machine Learning*, **58**(1), 39–77.

HAJEK, A. (2010) Website: *Interpretations of probability. The Stanford Encyclopedia of Philosophy* (E.N. Zalta ed.). Available from: http://plato.stanford.edu/entries/probability-interpret/.

LAROSE, D. T. (2010) *Discovering Statistics*. W. H. Freeman and Company, New York.

LUTZ, H. and WENDT, W. (1998) *Taschenbuch der Regelungstechnik*. Verlag Harri Deutsch, Frankfurt am Main.

MOZINA, M., DEMSAR, J., ZABKAR, J. and BRATKO, I. (2006) Why is rule learning optimistic and how to correct it. In: *European Conference on Machine Learning, ECML 2006*. **LNCS 4212**, 330–340.

PIEGAT, A. and LANDOWSKI, M. (2012) Optimal estimator of hypothesis probability for data mining problems with small samples. *Int. J. Appl. Math. Comput. Sci.*, **22**, 3, 629–645.

POLKOWSKI, L. (2002) *Rough Sets*. Physica-Verlag, Heidelberg, New York.

ROKACH, L. and MAIMON, O. (2008) Data mining with decision trees, theory and applications. *Machine Perception and Artificial Intelligence*, **69**. World Scientific Publishing Co. Pte. Ltd, New Jersey, Singapore.

SIEGLER, R. S. (1976) Three Aspects of Cognitive Development. *Cognitive Psychology*, 8, 481–520.

SIEGLER, R. S. (1994) *Balance Scale Weight & Distance Database. UCI Machine Learning Repository*. Available from: http://archive.ics.uci.edu/ml/datasets/Balance+Scale.

STARZYK, A. and WANG, F. (2004) Dynamic probability estimator for machine learning. *IEEE Transactions on Neural Networks*, March **15**(2), 298–308.

SULZMANN, J. N. and FURNKRANZ, J. (2009) An empirical comparison of probability estimation techniques for probabilistic rules. In: J. Gama, V. S. Costa, A. Jorge, P. Brazdil, *Proceedings of the 12th International Conference on Discovery Science (DS-09)*, Porto, Portugal. Springer-Verlag, 317–331.

SULZMANN, J. N. and FURNKRANZ, J. (2010) *Probability estimation and aggregation for rule learning*. Technical Report TUD-KE-201-03, TU Darmstadt, Knowledge Engineering Group.

WITTEN, I. H. and FRANK, E. (2005) *Data Mining*. Second edition, Elsevier, Amsterdam.

ZADROZNY, B. and ELKAN, C. (2001) Learning and decision making when costs and probabilities are both unknown. In: *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*. San Francisco, August 2001. ADM, 204–213.

ZHANG, Z. (1995) *Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting. M-estimators. INRIA*. Available from: http://research.microsoft.com/en-us/um/people/zhang/INRIA/Publis/Tutorial-Estim/Main.html.

ZIARKO, W. (1999) Decision making with probabilistic decision tables. In: N. Zhong, ed., *RSFDGrC'99 Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing*, Yamaguchi, Japan. Springer-Verlag, Berlin, Heidelberg, New York, 463–471.

VON MISES, R. (1957) *Probability, Statistics and the Truth.* Macmillan, Dover, New York.