Int. J. Appl. Math. Comput. Sci., 2019, Vol. 29, No. 4, 783–796
DOI: 10.2478/amcs-2019-0058

amcs

REVISITING THE OPTIMAL PROBABILITY ESTIMATOR FROM SMALL SAMPLES FOR DATA MINING

BOJAN CESTNIK [a,b]

[a]Department of Knowledge Technologies
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

[b]Temida d.o.o., Dunajska cesta 51, 1000 Ljubljana, Slovenia
e-mail: bojan.cestnik@temida.si

Estimation of probabilities from empirical data samples has drawn close attention in the scientific community and has been identified as a crucial phase in many machine learning and knowledge discovery research projects and applications. In addition to trivial and straightforward estimation with relative frequency, more elaborated probability estimation methods from small samples were proposed and applied in practice (e.g., Laplace's rule, the m-estimate). Piegat and Landowski (2012) proposed a novel probability estimation method from small samples $Ep_{h\sqrt{2}}$ that is optimal according to the mean absolute error of the estimation result. In this paper we show that, even though the articulation of Piegat's formula seems different, it is in fact a special case of the m-estimate, where $p_a = 1/2$ and $m = \sqrt{2}$. In the context of an experimental framework, we present an in-depth analysis of several probability estimation methods with respect to their mean absolute errors and demonstrate their potential advantages and disadvantages. We extend the analysis from single instance samples to samples with a moderate number of instances. We define small samples for the purpose of estimating probabilities as samples containing either less than four successes or less than four failures and justify the definition by analysing probability estimation errors on various sample sizes.

Keywords: probability estimation, small samples, minimal error, m-estimate.

1. Introduction

When dealing with uncertainty in various situations, probabilities often come into play (Starbird, 2006). Probabilities are used to assign meaningful numerical values (from the interval between 0 and 1) to express the likelihoods that certain random events will occur. Probabilities are the pillars of many scientific and business disciplines, including statistics (DeGroot and Schervish, 2012), game theory (Webb, 2007), quantum physics (Gudder, 1988), strategic and financial decision-making (Grover, 2012), as well as knowledge discovery and data mining (DasGupta, 2011; Flach, 2012).

Conceptually, there seems to be a controversial dispute in the scientific community about the existence and interpretation of various kinds of probabilities (Good, 1966). In particular, there is a substantial conceptual difference between the so-called frequentists and Bayesians. For frequentists, on the one hand, probability represents a frequency of occurrence of a particular event (Rudas, 2008). It can be estimated from a data sample as a single value by the maximum likelihood estimate. On the other hand, a community of Bayesians share the view that probabilities themselves can be random variables distributed according to probability density functions and can, therefore, be used to express also more or less uncertain degrees of (subjective) beliefs that some events will occur (Berger, 1985).

In essence, the problem of probability estimation from data samples can be formulated as follows (Good, 1965). Suppose that we have conducted (observed) and recorded a sample of n independent experiments (instances), out of which there are s successes and f failures ($s + f = n$). How do we estimate the probability p of success in the domain under observation? Or, analogously, based on the collected experiments, how do we estimate the probability p that the next experiment will be a success?

In this paper we deal with probability estimation

methods from small samples, which is a difficult and intricate task (Good, 1965; 1966; Chandra and Gupta, 2011; Chan and Kroese, 2011). The subject is not to be confused with the problem of detecting small samples and determining the optimal sample size for statistical analyses (Good and Hardin, 2012), which is beyond the scope of this paper. To define the concept of "small sample" for the purpose of estimating probabilities, we start with the definition of its opposite: "sufficiently large sample." Probability theory states that if sample size n is large enough, differences between diverse probability estimation methods are negligible (Good, 1965). Therefore, we are more interested in determining which sample size reduces the differences between various probability estimation methods than in determining which sample size reduces the estimation errors close to zero. For the study reported in this paper we presume, as asserted by Good (1965) in his book, the following.

Definition 1. (*Sufficiently large sample*) For the purpose of estimating probabilities, a sample is large enough if both s and f are greater than 3.

As explained and demonstrated in Section 6, on such large samples we can, for all practical purposes, more or less safely use any mathematically sound method for probability estimation, since the differences between average mean absolute errors of different estimation methods tend to drop below the threshold of 0.01.

In addition to Definition 1, we define the concept of small sample.

Definition 2. (*Small sample*) For the purpose of estimating probabilities, a sample is small if either s or f is less than 4.

Note that Definition 2 does not imply that samples containing more than 8 instances are not small samples. In fact, the expected size of a sufficiently large sample depends on the actual probability of success in the sample. The results of our analysis concerning the shortest sufficiently large sample size show that if we are sampling from a population where the actual probability of success is 0.5, the average expected shortest sufficiently large sample size is 10.2 with a median of 10 and a standard deviation of 2.3. If the actual probability of success is 0.2 (or its antonym 0.8), the average expected shortest sufficiently sample size is 20.1 with a median of 19 and a standard deviation of 8.8. More detailed results concerning the analysis of small sample sizes are presented in Section 6. Note, however that, according to Definitions 1 and 2, every sample containing less than or equal to 7 instances is a small sample, regardless of the actual probability of success in the sample.

2. Historical background and related work

According to the classical definition, the probability of obtaining a success in the next experiment can be estimated with *relative frequency* (Feller, 1968)

$$\hat{p}_{\text{relfr}} = \frac{s}{s+f} = \frac{s}{n}. \tag{1}$$

Probability estimation with relative frequency fits nicely into the frequentist paradigm and is perfect in the limit case, when the sample size n approaches infinity (or, is acceptable, if the sample is at least large enough). However, when the sample size is small, there are three major shortcomings concerning such probability estimations. First, the estimates may have extreme values of 0 and 1 which might be intricate in further numerical calculations. Second, the estimation errors might be considerably high. And third, the estimates with relative frequency do not stabilize fast enough with increasing the sample size (Piegat and Landowski, 2013).

To curtail the apparent shortcomings of relative frequency, Laplace (1814) proposed the formula

$$\hat{p}_{\text{Laplace}} = \frac{s+1}{s+f+2} = \frac{s+1}{n+2}, \tag{2}$$

which is also called *Laplace's rule of succession* (e.g., Feller, 1968; Good, 1965). Laplace's formula (2) fits into the Bayesian probability estimation paradigm since it assumes a uniform prior probability distribution by adding one fictitious success and one fictitious failure to the data sample used for the estimation.

Note that (2) solves the first problem of relative frequency (1), since $\hat{p}_{\text{Laplace}}$ can never be equal to 0 or 1. In addition, Laplace's rule of succession typically alleviates the second and third problems of relative frequency by reducing the average errors of the estimations (Feller, 1968; Good, 1965).

Inspired by the work of Good (1965), Cestnik (1990) proposed a more general Bayesian method for probability estimation called the m-estimate:

$$\hat{p}_{\text{Cestnik}}(p_a, m) = \frac{s + p_a m}{s+f+m} = \frac{s + p_a m}{n+m}. \tag{3}$$

The m-estimate has two parameters: p_a and m; p_a is a prior probability and m is a weight assigned to the prior probability. The m-estimate was first used in the context of machine learning and empowered several learning algorithms to significantly improve the classification accuracy of generated models (Cestnik, 1990; Cestnik and Bratko, 1991; Džeroski *et al.*, 1993).

A typical data set used for classification tasks in machine learning (Breiman *et al.*, 1984; Flach, 2012) consists of several learning instances, described with attributes and their values. Each learning instance belongs to a designated class. The task of machine learning

is to construct a model for predicting the value of the class given the values of the attributes. The m-estimate was intentionally designed for estimating conditional probabilities of a class given the values of selected attributes (Cestnik, 1990). The unconditional probability of the class, estimated from the whole sample, is taken as a prior $\hat{p}_a$. Since the whole data set is usually large enough for reliable probability estimation with any probability estimation method, such unconditional estimates are typically considered sufficiently accurate. When estimating the conditional probability of a class given the values of the selected attributes in a rule or a decision tree, the corresponding data set is filtered according to the attributes' values. Such filtered data sets often qualify as small samples, containing just a small subset of the original set of learning instances. In this context, the m-estimate regularly demonstrated its superiority over relative frequency and Laplace's rule of succession (Cestnik, 1990; Cestnik and Bratko, 1991; Mitchell, 1997; Domingos and Pazzani, 1997; Fürnkranz and Flach, 2005; Sulzmann and Fürnkranz, 2009).

In their paper, Piegat and Landowski (2012) proposed a novel formula for the probability estimation:

$$\hat{p}_{\text{Piegat}}(a) = \frac{1}{2} + \frac{s-f}{2(s+f+a)} = \frac{1}{2} + \frac{s-f}{2(n+a)}. \quad (4)$$

Formula (4) is in their paper denoted by Ep_{ha} and has one parameter a. The theoretical optimization of the mean absolute error (MAE) with the proposed formula (4) yielded the optimal value of $a = \sqrt{2}$ (Piegat and Landowski, 2012). After the replacement with the optimized value of a, the following formula, denoted by $Ep_{h\sqrt{2}}$ in their paper, was obtained:

$$\hat{p}_{\text{Piegat}}(\sqrt{2}) = \frac{1}{2} + \frac{s-f}{2(n+\sqrt{2})}. \quad (5)$$

Piegat and Landowski (2012) demonstrated, both theoretically and experimentally, that on small samples (sample size $<$ 25) the accuracy achieved with their formula (5) was considerably better than the accuracy of relative frequency (1) and closely comparable to the accuracy of Laplace's rule (2) and the m-estimate (3) used with parameters $p_a = 1/2$ and $m = 2$.

The probability estimation methods studied in this paper are the following: relative frequency (1), Laplace's rule of succession (2), Piegat's formula $Ep_{h\sqrt{2}}$ (5), and Cestnik's m-estimate (3). In calculations and graphical presentations, as well as for shorter annotations in text, the methods are denoted by shorter names *"relfr"*, *"Laplace"*, *"Piegat"*, and *"Cestnik"*, respectively. For denoting the actual probability of a sample we use the symbol p, while the estimated probability is denoted as $\hat{p}$. The prior probability used in the m-estimate is denoted as p_a.

As an error measure of a probability estimation method we use the *mean absolute error* (abbreviated as MAE in the paper) for easier comparison with the findings reported in Piegat and Landowski (2012). Also, the preliminary experiments with another measure of error, *root mean squared error* (RMSE), revealed that the general observations and conclusions remain the same regardless of the error measure used. However, in their subsequent papers, Piegat and Landowski (2013; 2014) showed that after using RMSE to optimize errors of their probability estimation formula (4) they obtained the optimal value $a = 2$, which made their formula equivalent to Laplace's rule.

Piegat and Landowski (2012) argue that having no knowledge about prior probabilities is often the case. However, in several situations it is possible to acquire background (prior) knowledge that can be effectively used in calculating posterior probability estimates after observing a series of experiments. In such situations, the more elaborated m-estimate reveals its advantages. We can specify both parameters: p_a as a prior probability, and m as the strength of our confidence in prior. Even if our knowledge about the prior probability is partially limited (e.g., interval, distribution), it can still help reducing the estimation errors.

Piegat and Landowski (2012) made a comparison of the results achieved using their formula (5) with the m-estimate; they set the parameters of the m-estimate such that $p_a = 1/2$ and $m = 2$, which effectively transformed the m-estimate into Laplace's rule. They argue that the comparison with such an uninformed version of the m-estimate is fair, since the relative frequency, Laplace's rule and Piegat's formula do not use any additional information about the priors.

In this paper we compare four probability estimation methods labelled as "relfr", "Laplace", "Piegat" and "Cestnik", with respect to their mean absolute errors in the context of a carefully designed experimental framework for conducting experiments and comparisons. The aim is to study the impact of availability/absence of prior information on the final estimation errors, as well as the effect of various levels of noise in the training samples on the estimation errors of the four probability estimation methods. Our prediction is that, especially when dealing with the estimations from small samples, information about prior probabilities can help reducing the estimation errors.

The paper is organized as follows. In the next section we provide additional theoretical background for the four observed probability estimation methods and demonstrate that Piegat's formula is in fact a special case of Cestnik's m-estimate. In Section 4 we describe the design and implementation of the experimental framework used in our study and clarify some additional methodological issues. We demonstrate the utility of the experimental

framework in Section 5 by presenting experimental results of the four studied probability estimation methods. In Section 6 we present the analysis of the sample size (small sample, sufficiently large sample) impact on the estimation errors of the studied probability estimation methods. In conclusions we summarize the most important contributions of this research.

3. Theoretical background

In this section we revisit the probability estimation methods (relative frequency, Laplace's rule of succession, Piegat's formula, and the m-estimate) and further elaborate their representations to deepen the understanding. We present methods' parameters (if required by the method) and discuss their role in the estimation. Finally, we demonstrate that Piegat's formula is a special case of the m-estimate.

3.1. Methods for probability estimation. As stated in Section 1, the relative frequency (1) is a standard method for probability estimation. It works well if the sample is large enough. As can be observed throughout this subsection, all other probability estimation methods incorporate the ratio s/n, either explicitly or implicitly, in their calculation.

Laplace's formula (2) can be rewritten as a weighted average of relative frequency s/n and $1/2$,

$$\hat{p}_{\text{Laplace}} = \frac{s}{n} \times \frac{n}{n+2} + \frac{1}{2} \times \frac{2}{n+2}. \tag{6}$$

Technically, (2) assumes ignorance of the prior probability by adding two fictitious experiment outcomes to the collected sample: one success and one failure. In the Bayesian sense, it introduces a uniform prior distribution with the expected value of $\hat{p}_a = 1/2$ (Good, 1965).

Piegat's formula (5) can also be rewritten as a weighted average of relative frequency s/n and $1/2$,

$$\hat{p}_{\text{Piegat}}(\sqrt{2}) = \frac{s}{n} \times \frac{n}{n+\sqrt{2}} + \frac{1}{2} \times \frac{\sqrt{2}}{n+\sqrt{2}}. \tag{7}$$

At first sight, the equivalence of formulas (5) and (7) is not obvious. A more detailed explanation of the derivation is given in Section 3.2. At this point, however, observe that we can simplify both (5) and (7) to the following expression:

$$\hat{p}_{\text{Piegat}}(\sqrt{2}) = \frac{s + \frac{\sqrt{2}}{2}}{n+\sqrt{2}}. \tag{8}$$

In consequence, the only difference between Piegat's formula (7) and Laplace's rule (6) is the number of added fictitious instances to the sample. While in Laplace's rule there are two fictitious instances added (one success and one failure), Piegat's formula adds $\sqrt{2}$ instances ($\sqrt{2}/2$ successes and $\sqrt{2}/2$ failures).

For the purpose of explaining the role of parameters p_a and m, we can transform Cestnik's formula (3) to the form of a weighted average of relative frequency s/n and p_a (Cestnik and Bratko, 1991):

$$\hat{p}_{\text{Cestnik}}(p_a, m) = \frac{s}{n} \times \frac{n}{n+m} + p_a \times \frac{m}{n+m}. \tag{9}$$

In (9) we introduce

$$\alpha = \frac{m}{n+m}. \tag{10}$$

To express the factor $n/(n+m)$ in (9) in terms of α, observe that

$$\frac{n}{n+m} = \frac{n+m-m}{n+m} = 1 - \alpha. \tag{11}$$

Since both m and n are supposedly non-negative numbers, α can take values from the interval $[0, 1]$. Using α, we can rewrite (9) as

$$\hat{p}_{\text{Cestnik}}(p_a, m) = \frac{s}{n} \times (1-\alpha) + p_a \times \alpha. \tag{12}$$

Formulas (9) and (12) show that the estimate $\hat{p}_{\text{Cestnik}}$ is a linear combination of relative frequency s/n and prior probability p_a. In the extreme case, when α is equal to 0 ($m = 0$), the result of formula (12) is equal to relative frequency s/n. On the other hand, if α is equal to 1 ($n = 0$; or, if $n > 0$: $m \to \infty$), then the result of formula (12) converges to prior probability p_a.

Formula (12) can be rearranged as a sum of relative frequency and a residual,

$$\hat{p}_{\text{Cestnik}}(p_a, m) = \frac{s}{n} + \left(p_a - \frac{s}{n}\right) \times \alpha. \tag{13}$$

The role of the residual is to push the estimation away from relative frequency s/n towards prior probability p_a. The strength of the push is proportional to the value of α.

The process of Bayesian probability estimation is sometimes called smoothing or flattening (Flach, 2012; Fienberg and Holland, 1972; Bouguila, 2013). The main role of smoothing is to push the estimated probability away from the extreme value obtained with relative frequency (1) towards some sort of average value ($1/2$ for Laplace's and Piegat's formulae, p_a for the m-estimate). Contrary to formula (13), formula (12) can be represented as the sum of prior probability and the residual:

$$\hat{p}_{\text{Cestnik}}(p_a, m) = p_a + \left(\frac{s}{n} - p_a\right) \times (1-\alpha). \tag{14}$$

In (14) the role of the residual is to push the estimate from prior probability p_a towards relative frequency s/n. The strength of the push is proportional to the value of $(1-\alpha)$. The expressions of the m-estimate used in (13) and (14) are suitable to better understand the role of parameter α in error analyses reported in the subsequent sections.

3.2. Piegat's formula is a special case of the m-estimate.

Theorem 1. *Piegat's formula (5) is equivalent to Cestnik's m-estimate (3) with parameters $m = \sqrt{2}$ and $p_a = 1/2$.*

Proof. We start with Piegat's formula (Piegat and Landowski, 2012) for estimating probabilities from small samples (note that in his original formula Piegat used n_h for the number of confirmations of hypothesis h and $n_{\bar{h}}$ for the number of confirmations of anti-hypothesis $\bar{h}$):

$$\hat{p}_{\text{Piegat}}(\sqrt{2}) = \frac{1}{2} + \frac{s-f}{2(n+\sqrt{2})}. \tag{15}$$

At first sight, the m-estimate

$$\hat{p}_{\text{Cestnik}}(p_a, m) = \frac{s + p_a m}{n+m}. \tag{16}$$

looks very different from Piegat's formula (15). In (15) we substitute n with $s+f$ and rearrange it as a single fraction:

$$\hat{p}_{\text{Piegat}}(\sqrt{2}) = \frac{s+f+\sqrt{2}+s-f}{2(s+f+\sqrt{2})}. \tag{17}$$

After simplifying the numerator, we obtain

$$\hat{p}_{\text{Piegat}}(\sqrt{2}) = \frac{2s+\sqrt{2}}{2(s+f+\sqrt{2})}. \tag{18}$$

Next, we divide both the numerator and denominator by 2 and replace $s+f$ with n:

$$\hat{p}_{\text{Piegat}}(\sqrt{2}) = \frac{s+\frac{1}{2}\sqrt{2}}{s+f+\sqrt{2}} = \frac{s+\frac{1}{2}\sqrt{2}}{n+\sqrt{2}}. \tag{19}$$

The comparison of (19) and (16) demonstrates that Piegat's formula is a special case of Cestnik's m-estimate; if in (16) we set parameters $m = \sqrt{2}$ and $p_a = 1/2$, we obtain (19), which is equivalent to (15). ■

Note that the m-estimate (16) is more general than the other three estimation methods, since it subsumes all the others. We can obtain all the other estimation methods by setting the m-estimate parameters m and p_a. The appropriate parameter setting of m and p_a for obtaining the other probability estimation methods is shown in Table 1.

Table 1. Setting the parameters of the m-estimate for obtaining the other probability estimation methods.

method / parameters of the m-estimate	p_a	m
relative frequency	–	0
Laplace's rule of succession	0.5	2
Piegat's $Ep_{h\sqrt{2}}$	0.5	$\sqrt{2}$

4. Methodology and the experimental framework

In this section we describe the implementation of the experimental framework that was used for the evaluation of the presented methods with respect to their mean absolute errors (MAE). The framework is conceptually similar to the framework used by Piegat and Landowski (2012) for their practical experiments; additionally, it introduces several new features and improvements. The most important requirements and decisions for the experimental framework were as follows:

- Generate samples with large numbers of instances distributed according to the binomial distribution with the probability of success p. From each generated sample we would like to be able to extract a substantial number of sub-samples for probability estimation tests, so that the statistical validity of the estimation errors is achieved. The decision was to perform 100000 experiments. Therefore we set the generated sample size to 100500. Note that Piegat and Landowski (2012) used the sample size of 10000.

- Probabilities of successes in the sample should be evenly represented and the interval (0, 1) should be reasonably well covered. We decided to use 21 different probabilities with the interval of 0.05 between two adjacent probabilities $(0.01, 0.05, 0.10, \ldots, 0.95, 0.99)$; the lowest and the highest probability were set deliberately to avoid 0.0 and 1.0. Note that Piegat and Landowski (2012) used 11 selected probabilities in their experiments.

- Probability estimation methods should be implemented in a unified and comparable way. Access to data samples should be unified. The framework should provide an additional functionality to introduce noise to the artificial training samples to mimic more realistic practical situations.

- The experimental framework implemented in R (R Core Team, 2018) should be publicly available for download and experimentation. It is available on GitHub (Cestnik, 2018) under the MIT license.

4.1. Sample generation in the experimental framework.

The instances that comprise artificial samples are drawn from the binomial distribution. Artificial samples (Larose, 2010) are often generated to mimic the samples obtained from the real world, which are generally more costly and sometimes also more difficult to obtain.

However, there is one major difference between the two kinds of samples. In samples collected from real world we often do not know the actual probability p; so, we have to estimate $\hat{p}$ from the sampled data. Since the actual p is unknown, the absolute error $|\hat{p} - p|$. One of the benefits of artificial samples is that, due to the known p, we can reliably calculate the absolute error $|\hat{p} - p|$ of the probability estimation.

Let us first generate a random sample S_p of n binary instances, where each instance can be either a success with probability p or failure with probability $1 - p$. A success is encoded as 1 and a failure as 0. From sample S_p we select a sequence of l consecutive instances, starting with instance j; the selected sub-sample is denoted as $S_p^{[j,l]}$. We can use sub-sample $S_p^{[j,l]}$ to estimate the probability with a given method:

$$\hat{p}_{\text{method}}(S_p^{[j,l]}) \tag{20}$$

and compare the estimate with the known probability p that was used to generate the sample S_p. We can calculate the absolute error (AE) of the estimation method:

$$\text{AE}_{\text{method}}(S_p^{[j,l]}) = |\hat{p}_{\text{method}}(S_p^{[j,l]}) - p|. \tag{21}$$

The mean absolute error (MAE) of the probability estimate with a given method on 100000 sub-samples of size l is the following:

$$\text{MAE}_{\text{method}}(S_p^{[*,l]}) = \frac{\sum_{j=1}^{100000} \text{AE}_{\text{method}}(S_p^{[j,l]})}{100000}. \tag{22}$$

In (22) index j runs from 1 to 100000, so that each sub-sample $S_p^{[j,l]}$ is obtained from sample S_p as a sequence of l instances starting at instance j.

To obtain the average of MAEs (AMAE) for all 21 samples $S_{p_j}, j \in \{1, \ldots, 21\}$ we used the following formula:

$$\text{AMAE}_{\text{method}}(l) = \frac{\sum_{j=1}^{21} \text{MAE}_{\text{method}}(S_{p_j}^{[*,l]})}{21}. \tag{23}$$

$\text{AMAE}_{\text{method}}(l)$ calculates the average error of MAEs of the $method$ on sub-samples of size l from all the samples generated with the preselected probabilities within the experimental framework. AMAE can be used as a single value estimate of the absolute error of the method. Note that even though AMAE is calculated as an average of averages (MAEs), it yields an accurate average error estimate from all AEs, since each MAE is calculated as an average error from the same number of sub-samples.

To measure how different probability estimation methods m_1 and m_2 are, we can compare the AMAEs (23) of the two methods. However, since the AMAE is the average of MAEs, it can be similar even if the two MAE arrays are quite different in various probability intervals. For this reason, a more precise measure of the similarity of two methods $\text{ADIFF}_{\text{m1, m2}}(l)$ was used:

$$\text{ADIFF}_{\text{m1, m2}}(l) = \frac{\sum_{j=1}^{21} |\text{DF}_{\text{m1, m2}}(j,l))|}{21}, \tag{24}$$

where

$$\text{DF}_{\text{m1, m2}}(j,l) = \text{MAE}_{\text{m1}}(S_{p_j}^{[*,l]}) - \text{MAE}_{\text{m2}}(S_{p_j}^{[*,l]}). \tag{25}$$

4.2. Data structures, functions and procedures in the experimental framework. The proposed experimental framework consists of the generated instance samples, the functions for probability estimation and calculation of estimation errors, and a procedure that implements various testing scenarios.

The samples S_p are generated for all the preselected probabilities p_i ($i \in [1, \ldots, 21$). Each sample consists of 100500 instances. 100000 sub-samples of various sizes (from 1 to 500) starting at consecutive positions can be generated from each sample S_p.

The probability estimation functions are defined for each of the studied methods: relative frequency (1), Laplace's rule (2), Piegat's formula (5), and the m-estimate (3). The framework also includes a function that calculates MAEs (22) of the selected probability estimation method.

Algorithm 1 defines the basic evaluation procedure of a probability estimation method. Two main parameters of the algorithm are the $method$ for the estimation of probabilities and the number of instances (sample size) from which the probability is estimated. In addition, we have to provide probability estimation parameters if required by the $method$. Such parameters include prior probability for the estimation and the value of m for the m-estimate. To simulate more realistic settings for conducting experiments, we also included distortion of prior probability for the m-estimate, and distortion of the index of the training sample that can be used to introduce noise in the sample.

5. Evaluation of probability estimation methods in the experimental framework

The mathematical probability theory (Feller, 1968) states that, in an idealized situation when the sample size approaches infinity, the probability of success p is equal to the relative frequency:

$$p = \lim_{n \to \infty} \frac{r}{n} \tag{26}$$

In reality, however, obtaining samples of an infinite size n is not physically feasible. Therefore, we have to take into account the fact that probability estimation errors are bound to occur. Among several factors that influence

Algorithm 1. Evaluation of a probability estimation method.

Step 1. Select a $method$ for probability estimation and set the parameters required by the $method$.

Step 2. Select the level of noise in the sample: distortion p_d of p_a for the m-estimate and distortion d_s of the index of the training sample.

Step 3. Select estimation sample size l (e.g., 1, 2, 3, ...).

Step 4. Use formula (22) on each of the 21 samples $S_{p_i}(i \in 1, \ldots, 21)$ to calculate the MAEs of the selected $method$ and store the results in an array, where each element of the array represents the MAE of one sample.

Step 5. Use the array of stored results for displaying graphs and computing the AMAE with formula (23) of the $method$.

the validity of statistical analyses as well as probability estimation errors (Good and Hardin, 2012), in our study we focused on the following two: small samples and biased prior assumptions. Small samples, even if obtained without any other bias, are themselves the cause of probability estimation errors (Good, 1965). And second, biased prior assumptions, if incorporated in probability estimation methods, also affect the probability estimation errors.

In this section we present the results of our experimental work. There were three tasks that we wanted to accomplish with our experiments. The first task (1) was to compare all four probability estimation methods on sub-samples of the same size and to examine the impact of sub-sample sizes 1 to 7 (small samples) on the performance of each probability estimation method. In the second experiment (2) we evaluated the impact of a random distortion of prior probability p_a (biased prior assumptions) used in the m-estimate and compared such a distorted m-estimate with the other estimation methods. The third task (3) was to compare the performances of all the estimation methods with the metrics proposed in Demšar (2006) for comparing the performances of data mining algorithms.

Three probability estimation methods (relative frequency (1), Laplace's rule (2) and Piegat's formula (5) have no additional parameters except n and s. However, the m-estimate (3) requires two parameters: p_a and m. If not explicitly stated otherwise, in the experiments with the m-estimate the actual probability p of sample S_p was taken as parameter p_a and, the parameter m was set to 2, which was a sort of standard value in many studies that used the m-estimate (e.g., Cestnik, 1990; Flach, 2012). Note that taking the probability p of S_p as the parameter p_a, which means that the actual probability of the sample was taken as a prior in probability estimation, is idealized and seems to offer an unfair advantage to

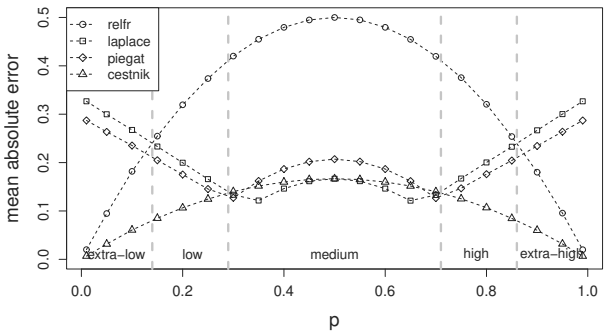

Fig. 1. Mean absolute errors of the relative frequency, Laplace's rule, Piegat's formula and Cestnik's m-estimate when the estimation is performed on samples of size 1. Vertical lines indicate boundaries between loosely defined probability intervals named extra-low, low, medium, high and extra-high.

the m-estimate; therefore, the obtained results have to be considered as the minimal possible absolute errors achieved with the m-estimate. To compensate for such an ideal setting and to obtain a more realistic comparison of the studied probability estimation methods, we introduced p_d distortions of p_a in the experiments described in later subsections.

5.1. Comparison of probability estimation methods on sub-samples of the same size. The first task was to compare the four estimation methods according to their mean absolute errors (MAEs) on small sub-samples of the same size. We started with sub-samples of size 1 (single instance samples). The comparison of MAEs of the four estimation methods is shown in Fig. 1. The relative frequency produced the highest MAEs and the m-estimate the lowest MAEs. The behaviour of the MAEs of "Piegat" is quite similar to the MAEs of "Laplace"; the only difference between the two methods is in their smoothing constant: while "Laplace" uses the value 2, "Piegat" uses smaller $\sqrt{2}$ and, therefore, enforces a sightly weaker push of the probability estimate towards prior probability $1/2$. Note that the value of α (as in (12)) on a single instance sample is $2/3 \approx 0.6667$ for "Laplace" and $\sqrt{2}/(1+\sqrt{2}) \approx 0.5858$ for "Piegat".

The obtained average mean absolute errors (AMAEs) are shown in Table 2. The relative frequency produced the highest AMAE (0.3185) and the m-estimate the lowest (0.1062). Laplace's rule and Piegat's formula resulted in mutually comparable AMAEs (0.2037 and 0.1993) that lied between the former two. Relative frequency estimation is the most different from all the other estimation methods, while Laplace's rule and Piegat's formula are the most similar.

Table 2. AMAEs of the four methods on sample sizes 1 to 7.

Sample size	"relfr"	"Laplace"	"Piegat"	"Cestnik"
1	0.3185	0.2037	0.1993	0.1062
2	0.2192	0.1709	0.1670	0.1096
3	0.1776	0.1499	0.1469	0.1065
4	0.1525	0.1358	0.1325	0.1017
5	0.1359	0.1246	0.1217	0.0971
6	0.1245	0.1154	0.1127	0.0934
7	0.1152	0.1083	0.1059	0.0896

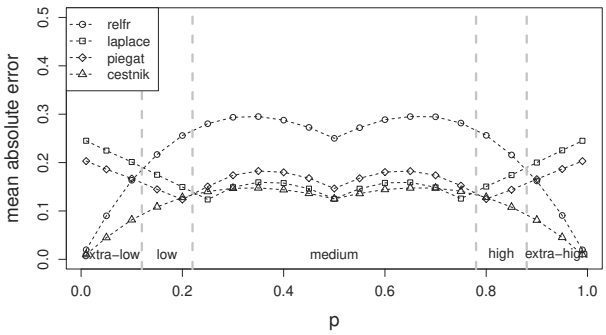

Fig. 2. Mean absolute errors of the relative frequency, Laplace's rule, Piegat's formula and Cestnik's m-estimate when the estimation is performed on samples of size 2. Vertical lines indicate boundaries between loosely defined probability intervals that are named *extra-low*, *low*, *medium*, *high* and *extra-high*.

To interpret the behaviour of MAEs of the four probability estimation methods in Fig. 1 we grouped the probabilities p on the x axis into loosely defined intervals. In particular, in Fig. 1 we observed two compelling interval boundary values: the intersection point between MAE of "relfr" and MAE of "Laplace" (or MAE of "Piegat"), and the intersection point between MAE of "Cestnik" and MAE of "Laplace" (or MAE of "Piegat"). Note that the MAEs of "Laplace" and "Piegat" are noticeably similar, so we decided to use MAEs of "Laplace" in calculations of the interval boundary values. Since we were not interested in the exact determination of the interval boundary values, they were determined experimentally within the experimental framework. The decision was to use 0.01 tolerance: if two values differed less than the tolerance 0.01, they were considered equal to each other. The first interval boundary was set at 0.14, which was the lowest probability where MAE of "relfr" rose equal to MAE of "Laplace" (0.01 tolerance). The second boundary value was determined accordingly and was set to 0.29.

The loosely defined probability intervals in Fig. 1 were named as *extra-low*, *low*, *medium*, *high*, and *extra-high*. The interval boundaries were named accordingly: boundary *extra* between the extra-low and low intervals, and boundary *medium* between the low and medium intervals. Boundaries extra and medium were conversely used as $(1 - \text{boundary})$ for determining the antonym boundaries between medium and high and high and extra-high intervals. Accordingly, in Fig. 1 there are five loosely defined intervals of actual probabilities p: extra-low (0.00, 0.14], low (0.14, 0.29), medium [0.29, 0.71], high (0.71, 0.86), and extra-high [0.86, 1.00).

On the samples with actual probabilities of success from extra-low and extra-high relative frequency "relfr" performed much better than "Laplace" and "Piegat". On the other three intervals (low, medium, high) "Laplace" and "Piegat" performed considerably better than "relfr". The m-estimate evidently performed better than all other methods on all the intervals except on the medium interval, when the performances of "Cestnik", "Laplace" and "Piegat" were roughly comparable. Note, however, that the results for "Cestnik" might be too optimistic to generalize, since the actual p was taken as hypothetical $\hat{p}_a$, $m = 2$.

The comparison of the errors of the four methods on sub-samples of size 2 is shown in Table 2 and in Fig. 2. The interval boundaries were recomputed using the same procedure as for single instance samples. Note that boundary *extra* decreased to 0.12 and boundary *medium* decreased to 0.22. Average mean absolute errors (AMAEs) are shown in Table 2. The relative frequency still produced the highest AMAE (0.2192) and the m-estimate the lowest (0.1069). Laplace's rule and Piegat's formula resulted in comparable AMAEs (0.1709 and 0.1670) that lie between the former two.

By adding an additional instance to a single instance sample AMAE decreased substantially for "relfr", decreased moderately for "Laplace" and "Piegat", while for "Cestnik", contrary to our expectations, AMAE slightly increased. A more detailed analysis of such unexpected behaviour revealed that for probabilities 0.35 to 0.65 $\text{MAE}_{\text{Cestnik}}$ slightly decreased on samples of size 2 with respect to single instance samples; however, for probabilities up to 0.30 and over 0.70 $\text{MAE}_{\text{Cestnik}}$ slightly increased by adding an additional instance to a single instance sample. The increase was especially notable for probabilities between 0.1 and 0.2, as well as 0.8 and 0.9. Those probabilities were for the m-estimate harder to estimate correctly from two instance samples than from single instance samples. The main reason for such a slight increase in the m-estimate's AMAE was found by observing parameter α in formula (12): $\alpha = 2/3$ for single instance samples, while $\alpha = 1/2$ for two instance samples, which corresponded to smaller weight attached to prior probability p_a in the latter case. Since the actual

prior of the sample was taken as p_a, the higher α enlarged the impact of p_a to the final probability estimation and, therefore, induced the lower estimation error.

The comparison of the errors of the four probability estimation methods on sub-samples of size 3 is shown in Fig. 3. Note that boundary *extra* further dropped to 0.11 and boundary *medium* dropped to 0.18. Average mean absolute errors (AMAEs) are shown in Table 2. AMAEs of all four methods were reduced with the introduction of a new instance to the sample; however, the order of the estimation methods from the highest to the lowest AMAE was preserved.

Figures 4 and 5 show the calculated MAEs on samples with 10 and 100 instances, respectively. Note that the values on the y axis in both figures are rescaled to show the differences between the methods in more detail. The AMAEs of the four probability estimation methods on the samples of 10 and 100 instances are shown in Table 2. While Fig. 4 still exhibits subtle differences between the MAEs of distinctive methods, the differences between the methods in Figure 5 are practically negligible. Note that both boundaries *extra* and *medium* in Fig. 5 further dropped to 0.06 and 0.08, respectively. Since the differences between the methods in Fig. 5 on samples of size 100 were almost negligible, the boundaries *extra* and *medium* became irrelevant.

In summary, AMAEs of the four probability estimation methods on sample sizes 1 to 7 are shown in Table 2. The interval boundaries changed substantially with increasing the number of instances in the samples, as shown in Table 3. The size of the medium interval increased with increasing the sample size and slowly prevailed on account of diminishing the other four intervals. For further analyses and interpretations we generalized the boundaries of loosely defined intervals from Table 3 by computing the average of the first five sample sizes (from 1 to 5). Generalized boundary extra was set to 0.10 and generalized boundary medium to 0.20.

Table 3. Interval boundaries *extra* and *medium* depending on the sample size.

Sample size	Boundary *extra*	Boundary *medium*
1	0.14	0.29
2	0.12	0.22
3	0.11	0.18
4	0.10	0.16
5	0.09	0.14
6	0.08	0.12
7	0.08	0.11
10	0.06	0.08
100	0.01	0.01

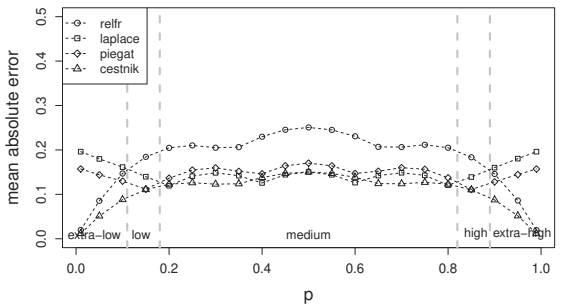

Fig. 3. Mean absolute errors of relative frequency, Laplace's rule, Piegat's formula and Cestnik's m-estimate when the estimation is performed on samples of size 3. Vertical lines indicate boundaries between loosely defined probability intervals that are named *extra-low*, *low*, *medium*, *high* and *extra-high*.

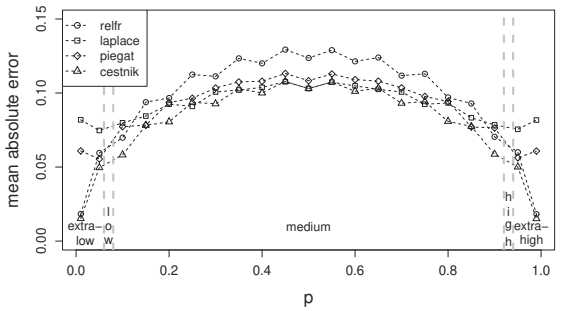

Fig. 4. Mean absolute errors of relative frequency, Laplace's rule, Piegat's formula and Cestnik's m-estimate when the estimations are performed on samples of size 10. Vertical lines indicate boundaries between loosely defined probability intervals that are named *extra-low*, *low*, *medium*, *high* and *extra-high*.

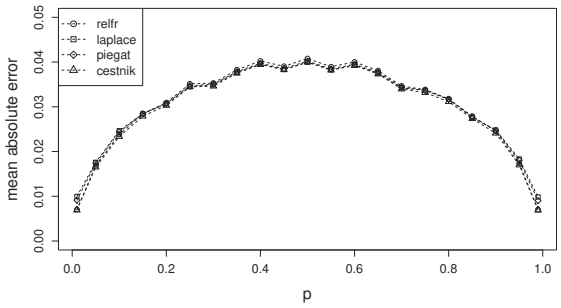

Fig. 5. Mean absolute errors of relative frequency, Laplace's rule, Piegat's formula and Cestnik's m-estimate when the estimation is performed on samples of size 100.

5.2. Introducing biased prior assumptions as distortions p_d of p_a to obtain hypothetical $\hat{p}_a$ in the m-estimate. In the experiments described in the previous subsection we used the actual p of the samples as a prior

probability p_a in the m-estimate. Such a serendipity could rarely happen in a real-world situation; normally, we are bound to start the estimation with an approximate prior probability $\hat{p}_a$. Therefore, to make the estimation set-up more realistic and to compensate for such an "overly informed" m-estimate, we introduced deliberate distortions to prior probabilities; given the maximal distortion p_d, we, in each estimation, randomly selected hypothetical $\hat{p}_a$ from the interval $[p_a - p_d, p_a + p_d]$.

The AMAEs of the m-estimate increased with the growing distortion in $\hat{p}_a$, which can be observed from Table 4. The increase was moderate for distortion $p_d = 0.3$, substantial for distortion $p_d = 0.5$, and extensive for distortion $p_d = 1.0$. By comparing the results from Tables 4 and 2 we observed that while the AMAEs of the m-estimate with $p_d = 0.3$ were still well below the corresponding AMAEs of "Laplace" and "Piegat". The difference between the AMAEs of the m-estimate with $p_d = 0.5$ and those of "Laplace" and "Piegat" were only slightly in favour of the m-estimate. With distortion $p_d = 1.0$ the m-estimate's AMAEs rose close to the AMAEs of relative frequency and was thus far above AMAEs of "Laplace" and "Piegat". Increasing the number of instances in samples decreased AMAEs of the m-estimate in all cases.

Comparisons of the m-estimate using p_d distortions (0.3, 0.5 and 1.0) of p_a with other probability estimation methods on single instance samples are shown in Figs. 6, 7 and 8. With increasing distortions to p_d, the MAEs and AMAEs of the m-estimate increased. While the m-estimate with $p_d = 0.3$ distortion of p_a on the average still performed better than Laplace's rule and Piegat's formula (Fig. 6), the distortion of 0.5 increased the MAEs of the m-estimate slightly over the MAEs of "Laplace" and "Piegat" in the medium interval (Fig. 7), even though the overall AMAE of the m-estimate with $p_d = 0.5$ (0.1773) was still slightly smaller than AMAEs of "Laplace" (0.2037) and "Piegat" (0.1993) on single instance samples. The slight superiority of the m-estimate's AMAEs with $p_d = 0.5$ over AMAEs of "Laplace" and "Piegat" can be observed from Tables 4 and 2. With the distortion of $p_d = 1.0$ the m-estimate performed worse than "Laplace" and "Piegat" on all 21 single size samples from the experimental framework (Fig. 8). Therefore, when $\hat{p}_a$ is for the m-estimate selected randomly (e.g., distortion $p_d = 1.0$), the AMAEs of the m-estimate from Table 4 show the inferiority of the m-estimate to "Laplace" and "Piegat" from Table 2 on all small sample sizes 1 to 7.

Table 4. AMAEs of the m-estimate with $m = 2$ on sample sizes 1 to 7 and distinctive distortions p_d of p_a.

Sample size	p_d distortion in $\hat{p}_a$ 0.0	0.3	0.5	1.0
1	0.1062	0.1351	0.1773	0.2489
2	0.1096	0.1263	0.1523	0.2025
3	0.1065	0.1178	0.1360	0.1737
4	0.1017	0.1103	0.1240	0.1539
5	0.0971	0.1039	0.1148	0.1391
6	0.0934	0.0985	0.1074	0.1278
7	0.0896	0.0937	0.1012	0.1187

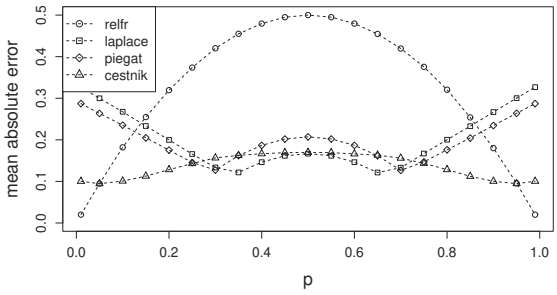

Fig. 6. Comparison of MAEs of the four estimation methods, p_d distortion of p_a for the m-estimate $= 0.3$.

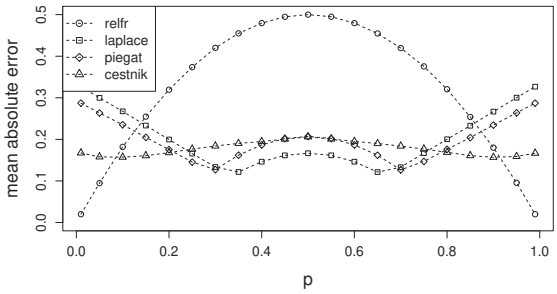

Fig. 7. Comparison of MAEs of the four estimation methods, p_d distortion of p_a for the m-estimate $= 0.5$.

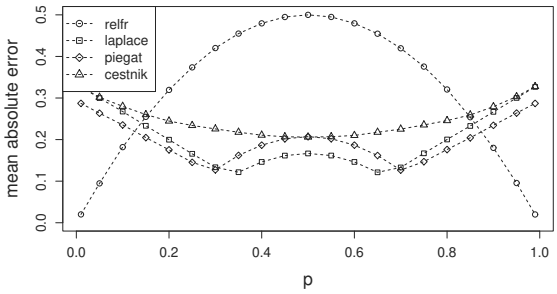

Fig. 8. Comparison of MAEs of the four estimation methods, p_d distortion of p_a for the m-estimate $= 1.0$.

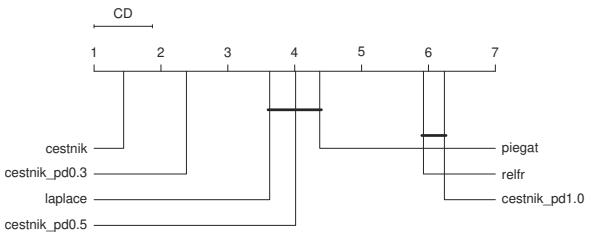

Fig. 9. Ranking the probability estimation methods in a critical difference plot. Calculated MAEs on sub-samples from 1 to 7 instances were used for comparison and ranking. CD is the critical difference obtained with $\alpha = 0.01$.

5.3. Ranking the probability estimation methods. For comparing and ranking the results obtained by the four studied probability estimation methods (relative frequency, Laplace's rule, Piegat's formula, and the m-estimate) we applied the approach proposed by Demšar (2006) and refined by García and Herrera (2008) as well as García *et al.* (2010) for statistically comparing the performance of different machine learning algorithms. For the m-estimate we included also three additional variations of biased prior assumptions by introducing distortions p_d of p_a: 0.3, 0.5 and 1.0. MAEs for each of the seven probability estimation methods on sub-samples of sizes 1 to 7 were calculated on 100000 sub-samples extracted from 21 samples in the experimental framework. As a result, 147 distinct MAE calculations were obtained for each probability estimation method and used for pairwise performance comparison.

The process of analysing the results was implemented in R using package *scmamp* (Calvo and Santafé, 2016). The obtained critical difference plot is shown in Fig. 9. The critical difference calculated with Nemenyi test with parameter $\alpha = 0.01$ was 0.8724. When the difference between two probability estimation methods in the critical difference plot was greater than the critical difference, the performances of the two methods can be regarded as significantly different from each other. In Fig. 9 the probability estimation methods whose performance was not significantly different are connected with a horizontal line.

Not surprisingly, the m-estimate with no distortions in p_a (labeled as "Cestnik" in Fig. 9) was ranked the first. The second method was the m-estimate with distortion $p_d = 0.3$ (labelled as "Cestnik_pd0.3"); when distortions p_d were small enough, the m-estimate still performed significantly better than all the other probability estimation methods. Next came a group of methods including Laplace's rule ("Laplace"), the m-estimate with distortion $p_d = 0.5$ ("Cestnik_pd0.5") and Piegat's formula ("Piegat"); the performances of these three methods were not significantly different. The lowest ranked methods were relative frequency ("relfr") and the m-estimate with distortion $p_d = 1.0$ ("Cestnik_pd1.0"), whose mutual performance differences were also statistically insignificant. The effect of distortions p_d of prior probability p_a in the m-estimate can be summarized as follows: when distortions p_d were relatively small (e.g., smaller than 0.3), the m-estimate performed significantly better than the other probability estimation methods; with distortions p_d around 0.5 the m-estimate performed similarly to Laplace's rule and Piegat's formula, while distortions p_d close to 1.0 degraded the performance of the m-estimate to become even slightly worse than the performance of relative frequency.

6. Experiments for finding the shortest large enough sample

Besides defining the sample size directly by its length, we can specify it indirectly by defining a certain criterion that the sample must satisfy. For example, Definition 1 states that a sample is large enough if it contains more than three successes and more than three failures. Samples containing at least s successes and at least f failures can be of variable length, with their average length depending heavily on the authentic probability of success in the sample. In this subsection we present both theoretical and experimental findings concerning the determination of the shortest large enough sample according to Definition 1.

When sampling from binomial distribution $B(n, p)$, the probability of obtaining exactly s successes in n experiments (number of failures is $f = n - s$) is

$$\Pr(s, n) = \frac{(n)!}{s!\,(n-s)!}\, p^s (1-p)^{n-s}. \tag{27}$$

We used (27) to calculate the exact probability that, according to $B(n, p)$, at least s_{lim} successes and at least f_{lim} failures are obtained in n experiments:

$$\Pr_{s \geq s_{\text{lim}}, f \geq f_{\text{lim}}}(n) = \sum_{i=s_{\text{lim}}}^{i=n-f_{\text{lim}}} \Pr(i, n). \tag{28}$$

In (28) all the probabilities $\Pr(i, n)$ are summed over the index i from s_{lim} to $n - f_{\text{lim}}$. The lower limit of index i guarantees that the number of successes is equal to or greater than s_{lim}, while the upper limit guarantees that the number of failures is equal to or greater than f_{lim} by limiting the number of success to less than or equal to $n - f_{\text{lim}}$. For shorter notation in the subsequent formulas we denote λ the condition $s \geq s_{\text{lim}}, f \geq f_{\text{lim}}$. To estimate the average sample size that contains at least s_{lim} successes and at least f_{lim} failures, we first transformed the cumulative probability distribution $\Pr_\lambda(n)$ to the

Table 5. Statistical moments of sizes of the shortest large enough samples (Definition 1) distributions for all 21 probabilities used in the experimental framework.

ind	probab. of success	mean	median	mode	sd
1, 21	0.01, 0.99	400.0	367	300	199.0
2, 20	0.05, 0.95	80.0	74	60	39.0
3, 19	0.10, 0.90	40.0	37	30	19.0
4, 18	0.15, 0.85	26.7	25	20	12.3
5, 17	0.20, 0.80	20.1	19	15	8.8
6, 16	0.25, 0.75	16.2	15	8	6.7
7, 15	0.30, 0.70	13.7	12	8	5.2
8, 14	0.35, 0.65	12.0	11	8	4.1
9, 13	0.40, 0.60	11.0	10	8	3.2
10, 12	0.45, 0.55	10.4	10	8	2.6
11	0.50	10.2	10	8	2.3

probability density for each n:

$$\mathrm{pr}_\lambda(n) = \begin{cases} \mathrm{Pr}_\lambda(n) & \text{if } n = 1, \\ \mathrm{Pr}_\lambda(n) - \mathrm{Pr}_\lambda(n-1) & \text{if } n > 1. \end{cases} \tag{29}$$

Then, we calculated the expected average of $\mathrm{pr}_\lambda(n)$ for all sample sizes from 1 to ∞:

$$\mu_\lambda = \sum_{n=1}^{\infty} n \times \mathrm{pr}_\lambda(n). \tag{30}$$

The standard deviation of $\mathrm{pr}_\lambda(n)$ was calculated accordingly:

$$\sigma_\lambda = \sqrt{\sum_{n=1}^{\infty} (n-\mu_\lambda)^2 \times \mathrm{pr}_\lambda(n)}. \tag{31}$$

In numerical calculations of formulas (30) and (31) that are presented in Table 5 we used the upper limit of 1500 instead of ∞, since the values of $\mathrm{pr}_\lambda(n)$ tend to drop to 0 for n larger than 1500. The so calculated statistical moments of samples' length distributions are presented in Table 5. The decreasing order of the sample size mean, median and mode indicates that the probability distributions of sample sizes are highly positively skewed with a long right tail. In fact, the calculated skewness (DeGroot and Schervish, 2012) was greater than 1 for all sample size probability distributions shown in Table 5.

In this subsection we present experimental results that additionally justify Definitions 1 and 2 by measuring and comparing AMAEs of the four probability estimation methods on sub-samples described by such indirect sample-size criteria. We compared the methods' errors on the shortest samples containing at least three, four, and five successes and failures.

The results of the four probability estimation methods obtained on such sub-samples from the experimental framework obtained as the average of 100000 experiments in the experimental framework are shown in Table 6. AMAEs of the four probability estimation methods decreased with increasing the descriptive sample sizes. However, note that in these experiments we were not interested that much in minimizing the values of AMAEs; our goal was to observe and minimize the differences between the estimation methods in terms of ADIFFs. On the samples described with $4s$ and $4f$ all ADIFFs dropped below 0.01 (or very close to 0.01 for ADIFF between "relfr" and "Cestnik"). Note that, for example, on the $3s$ and $3f$ samples all ADIFFs between "Cestnik" and the other estimation methods were quite higher than 0.01.

The MAEs obtained by the four probability estimation methods on the shortest large enough sample (as in Definition 1) are shown graphically in Fig. 10. The differences between the used estimation methods are small; in fact, on the average they are smaller than 0.01, as can be observed from Table 6 part $4s$ and $4f$. Since the differences between AMAEs and ADIFFs of parts $4s$ and $4f$, and $3s$ and $3f$ are really small, one might conditionally accept also the samples with $3s$ and $3f$ as large enough, sacrificing a small additional estimation error for the sake of simplicity in selecting relative frequency as the estimation method. However, the decision of choosing the samples of $4s$ and $4f$ as the shortest sufficiently large samples is supported also by the assertion of Good (1966), as already mentioned in the context of Definition 1. Therefore, in spite of the relatively small differences between ADIFFs of $4s$ and $4f$, and $3s$ and $3f$, the proposed threshold of 0.01 seems

Table 6. AMAEs and ADIFFs of the four probability estimation methods on indirectly described sample sizes.

Method	AMAE	ADIFF "relfr"	ADIFF "Laplace"	ADIFF "Piegat"
$3s$ and $3f$				
"relfr"	0.0851			
"Laplace"	0.0841	0.0097		
"Piegat"	0.0839	0.0070	0.0027	
"Cestnik"	0.0701	0.0150	0.0140	0.0137
$4s$ and $4f$				
"relfr"	0.0744			
"Laplace"	0.0739	0.0070		
"Piegat"	0.0736	0.0050	0.0020	
"Cestnik"	0.0641	0.0103	0.0098	0.0095
$5s$ and $5f$				
"relfr"	0.0671			
"Laplace"	0.0667	0.0055		
"Piegat"	0.0666	0.0039	0.0016	
"Cestnik"	0.0595	0.0076	0.0071	0.0070

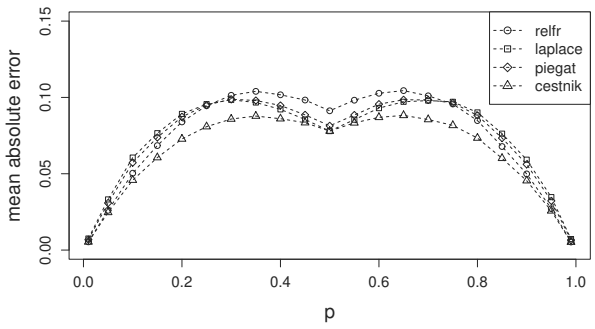

Fig. 10. Comparison of MAEs of "relfr", "Laplace", "Piegat" and "Cestnik" on the shortest sub-samples that contain four or more successes and four or more failures.

to additionally support and justify Definition 1.

7. Conclusions

In this paper we tested and compared four probability estimation methods in the context of small samples: relative frequency, Laplace's rule of succession, Piegat's formula and the m-estimate. The main contributions of the paper can be summarized as follows. We defined the concepts of a sufficiently large sample and a small sample for probability estimation in Definitions 1 and 2. We proved that Piegat's estimate is a special case of the m-estimate. We described an experimental framework in R that was used to conduct our study. The framework is available for download from GitHub (Cestnik, 2018) under the MIT license and provides a playground for future studies to assess the performance characteristics of various probability estimation methods. In an in-depth analysis and comparison of several probability estimation methods we identified their strengths and weaknesses on small samples. We ranked several probability estimation methods on a critical distance plot according to the approach of Demšar (2006).

For Definition 2 of small samples for probability estimates we offered a justification in terms of error analysis. The definition is meaningful (small sample contains either less than four successes or less than four failures) in the sense that each individual sample can be effectively classified as small or large enough. Consequently, appropriate probability estimation methods can be applied. For example, if a sample is classified as large enough, relative frequency can be used without hesitation. However, on a small sample it might be beneficial to invest additional effort in selecting a more appropriate probability estimation method to reduce the estimation error to an acceptable level. On the other hand, we can strive to enlarge the sample to become large enough. However, there are situations in which such enlargements are not viable. In such cases, the only sensible decision is to take what we have and do the best with it.

For further work we plan to investigate the impact of biased samples to the estimation errors of various probability methods. In addition, we would like to develop procedures to determine an optimal m for the m-estimate with respect to a given distortion p_d of the prior probability. Within the experimental framework we plan to develop new estimation methods that combine the high precision of the m-estimate and the simplicity of Laplace's rule or Piegat's formula.

Acknowledgment

The author wishes to thank Marko Bohanec from the Department of Knowledge Technologies at Jožef Stefan Institute for his comments and suggestions on earlier drafts of this paper.

References

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer, New York, NY.

Bouguila, N. (2013). On the smoothing of multinomial estimates using Liouville mixture models and applications, *Pattern Analysis and Applications* **16**(3): 349–363.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont.

Calvo, B. and Santafé, G. (2016). SCMAMP: Statistical comparison of multiple algorithms in multiple problems, *The R Journal* **8**(1): 248–256.

Cestnik, B. (1990). Estimating probabilities: A crucial task in machine learning, *Proceedings of the 9th European Conference on Artificial Intelligence, London, UK*, pp. 147–149.

Cestnik, B. (2018). Experimental framework in R for experimenting with probability estimations from small samples, `https://github.com/BojanCestnik/probability-estimation.R`.

Cestnik, B. and Bratko, I. (1991). On estimating probabilities in tree pruning, *Proceedings of the European Working Session on Learning, Porto, Portugal*, pp. 138–150.

Chan, J.C.C. and Kroese, D.P. (2011). Rare-event probability estimation with conditional Monte Carlo, *Annals of Operations Research* **189**(1): 43–61.

Chandra, B. and Gupta, M. (2011). Robust approach for estimating probabilities in naïve-Bayes classifier for gene expression data, *Expert Systems with Applications* **38**(3): 1293–1298.

DasGupta, A. (2011). *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*, Springer, New York, NY.

DeGroot, M. and Schervish, M. (2012). *Probability and Statistics*, Addison-Wesley, Boston, MA.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets, *Journal of Machine Learning Research* **7**(1): 1–30.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss, *Machine Learning* **29**(2): 103–130.

Džeroski, S., Cestnik, B. and Petrovski, I. (1993). Using the m-estimate in rule induction, *Journal of Computing and Information Technology* **1**(1): 37–46.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Willey, Hoboken, NJ.

Fienberg, S.E. and Holland, P.W. (1972). On the choice of flattening constants for estimating multinomial probabilities, *Journal of Multivariate Analysis* **2**(1): 127–134.

Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, Cambridge University Press, New York, NY.

Fürnkranz, J. and Flach, P.A. (2005). ROC 'n' rule learning—towards a better understanding of covering algorithms, *Machine Learning* **58**(1): 39–77.

García, S., Fernández, A., Luengo, J. and Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, *Information Sciences* **180**(10): 2044–2064.

García, S. and Herrera, F. (2008). An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons, *Journal of Machine Learning Research* **9**(12): 2677–2694.

Good, I.J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, MIT Press, Cambridge, MA.

Good, I.J. (1966). How to estimate probabilities, *IMA Journal of Applied Mathematics* **2**(4): 364–383.

Good, P. and Hardin, J. (2012). *Common Errors in Statistics (and How to Avoid Them)*, Wiley, Hoboken, NJ.

Grover, J. (2012). *Strategic Economic Decision-Making: Using Bayesian Belief Networks to Solve Complex Problems*, Springer New York, NY.

Gudder, S. (1988). *Quantum Probability*, Academic Press, Boston, MA.

Laplace, P.-S. (1814). *Essai philosophique sur les probabilités*, Courcier, Paris.

Larose, D. (2010). *Discovering Statistics*, W.H. Freeman, New York, NY.

Mitchell, T.M. (1997). *Machine Learning*, McGrawHill, Maidenhead.

Piegat, A. and Landowski, M. (2012). Optimal estimator of hypothesis probability for data mining problems with small samples, *International Journal of Applied Mathematics and Computer Science* **22**(3): 629–645, DOI: 10.2478/v10006-012-0048-z.

Piegat, A. and Landowski, M. (2013). Mean square error optimal completeness estimator eph2 of probability, *Journal of Theoretical and Applied Computer Science* **7**(3): 3–20.

Piegat, A. and Landowski, M. (2014). Specialized, MSE-optimal m-estimators of the rule probability especially suitable for machine learning, *Control and Cybernetics* **43**(1): 133–160.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, `https://www.R-project.org/`.

Rudas, T. (2008). *Handbook of Probability: Theory and Applications*, SAGE Publications, Thousand Oaks, CA.

Starbird, M. (2006). *What Are the Chances? Probability Made Clear*, Chantilly, VA.

Sulzmann, J.N. and Fürnkranz, J. (2009). An empirical comparison of probability estimation techniques for probabilistic rules, *in* J. Gama *et al.* (Eds), *Discovery Science*, Springer, Heidelberg, pp. 317–331.

Webb, J. (2007). *Game Theory: Decisions, Interaction and Evolution*, Springer, London.

Bojan Cestnik received his PhD in computer science from the University of Ljubljana, Faculty of Computer and Information Science, in 1991. Presently, he is the managing director at the company of Temida, a researcher at the Department of Knowledge Technologies at Jožef Stefan Institute, and a professor of computer science at the University of Nova Gorica. His professional and research interest include knowledge based information systems and machine learning.

Received: 15 December 2018
Revised: 24 March 2019
Accepted: 23 April 2019