UDC 004.8

M.I. SCHLESINGER
International Research and Training Center for Information Technologies and Systems of the NAS of Ukraine and MES of Ukraine, Kyiv, Ukraine, e-mail: *schles@irtc.org.ua.*
E.V. VODOLAZSKIY
International Research and Training Center for Information Technologies and Systems of the NAS of Ukraine and MES of Ukraine, Kyiv, Ukraine, e-mail: *waterlaz@gmail.com.*

MINIMAX DEVIATION STRATEGIES FOR MACHINE LEARNING AND RECOGNITION WITH SHORT LEARNING SAMPLES

Abstract. The article analyses risk-oriented formulation of pattern recognition and machine learning problems. Based on arguments from multicriteria optimization, a class of improper strategies is defined that are dominated by some other strategy. A general form of strategies that are not improper is derived. It is shown that some widely used approaches are improper in the defined sense, including the maximum likelihood estimation approach. This drawback is especially apparent when dealing with short learning samples of fixed length. A unified formulation of pattern recognition and machine learning problems is presented that embraces the whole range of sizes of the learning sample, including zero size. It is proven that solutions to problems in the presented formulation are not improper. The concept of minimax deviation recognition and learning is formulated, several examples of its implementation are presented and compared with the widely used methods based on the maximal likelihood estimation.

Keywords: pattern recognition, machine learning, short learning sample.

INTRODUCTION

The short learning sample problem has been around in machine learning under different names during its whole life. The learning sample is used to compensate for the lack of knowledge about the recognized object when its statistical model is not completely known. Naturally, the longer the learning sample, the better the subsequent recognition. However, when the learning sample becomes too small (2, 3, 5 elements) the effect of small samples becomes evident. In spite of the fact that any learning sample (even a very small one) provides some additional information about the object, it may be better to ignore the learning sample than to utilize it with the commonly used methods.

Example 1. Let us consider an object that can exist in one of two random states $y=1$ and $y=2$ with equal probabilities. In each state the object generates two independent Gaussian random signals x_1 and x_2 with variances equal 1. Mean values of signals depend on the state as it is shown in Fig. 1. In the first state, the mean value is $(2, 0)$. In the second state, the mean value depends on an unknown parameter θ and is $(0, \theta)$. Even when no learning sample is given a minimax strategy can be used to make a decision about the state y. The minimax strategy ignores the second signal and makes decision $y^* = 1$ when $x_1 > 1$ and decision $y^* = 2$ when $x_1 \le 1$.

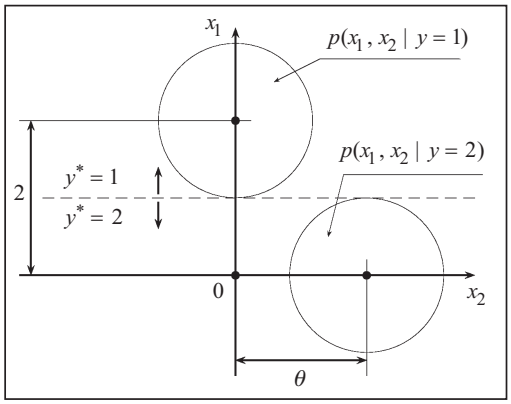

Fig. 1. Example 1: $(x_1, x_2) \in \mathbb{R}^2$ — signal, $y \in \{1, 2\}$ — state

Now let us assume that there is a sample of signals generated by

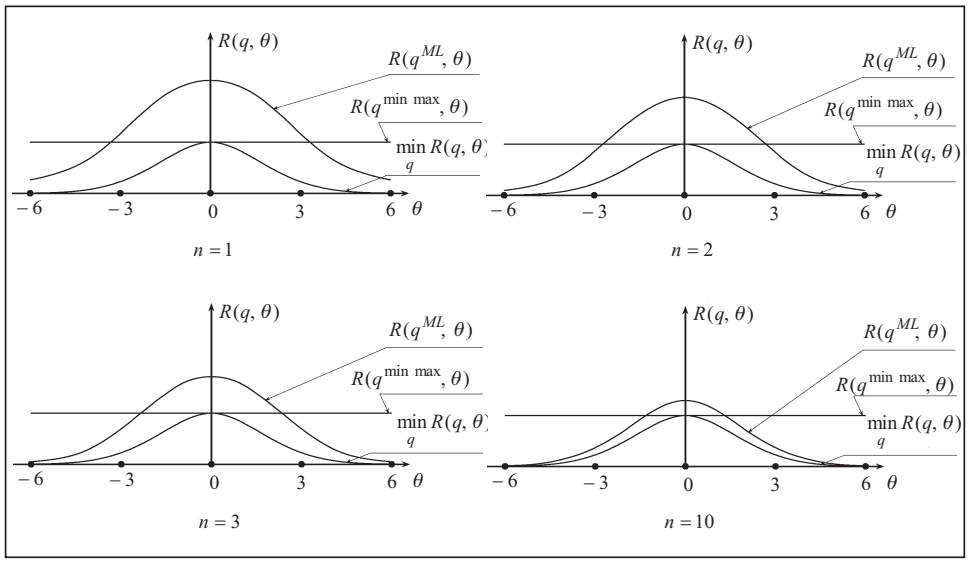

Fig. 2. Probability of a wrong decision (risk) for different sizes n of the learning sample

an object in the second state but with higher variance 16. A maximum likelihood strategy estimates the unknown parameter θ and then makes a decision about y as if the estimated value of the parameter is its true value. Figure 2 shows how the probability of a wrong decision (called the risk) depends on parameter θ for different sizes of the learning sample. In Fig. 2, as well as in all subsequent figures, the curve $R(q^{ML}, \theta)$ is the risk of the maximum likelihood strategy, the curve $R(q^{minmax}, \theta)$ is the risk of the minimax strategy and the curve $\min\limits_{q} R(q, \theta)$ is the minimum possible risk for each model. If the learning sample is sufficiently long, the risk of the maximum likelihood strategy may become arbitrarily close to the minimum possible risk. Naturally, when the length of the sample decreases the risk becomes worse. Furthermore, when it becomes as small as 3 or 2 elements the risk of the maximum likelihood strategy becomes worse than the risk of the minimax strategy that uses neither the learning sample nor the signal x_2 at all. Hence, it is better to ignore available additional data about the recognized object than to try to make use of it in a conventional way. This demonstrates a serious theoretical flaw of commonly used methods, and definitely not that short samples are useless. Any learning sample, no matter how long or short it is, provides some, maybe not a lot information about the recognized object and a reasonable method has to use it. **End of Example.**

Example 2. This is a simple example that has been used by H.Robbins in his seminal article [1] where he initiated the empirical Bayesian approach and explained its main idea. An object can be in one of two possible states $y = 1$ and $y = 2$. In each state, the object generates a univariate Gaussian signal x with variance 1. The mean value of the generated signal depends on the state y so that

$$p(x \mid y = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x+1)^2}{2} \right), \quad p(x \mid y = 2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-1)^2}{2} \right).$$

Only a priori probabilities of states are unknown and θ is the probability of the first state so that $p(y = 1) = \theta$ and $p(y = 2) = 1 - \theta$. Figure 3 illustrates these data.

A minimax strategy for such an incomplete statistical model makes decision y^* based on the sign of the observed signal and ensures the probability of correct recognition 0.84 independently of a priori probabilities of states.

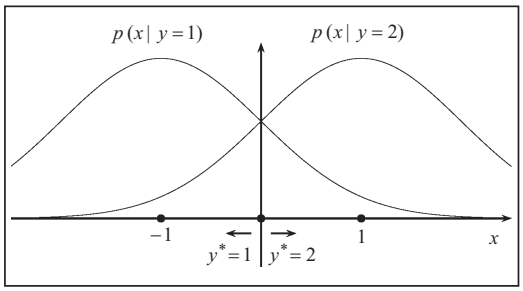

Fig. 3. Example 2: $x \in \mathbb{R}$ — signal, $y \in \{1, 2\}$ — state

Let not only a single object, but a collection of mutually independent objects be available for recognition. Each object is in its own hidden state and is presented with its own signal. Let us also assume that the decision about each object's state does not have to be made immediately when the object is observed and can be postponed until the whole collection is observed. In this case, maximum likelihood estimations of a priori probabilities of states can be computed and then each object of the collection is recognized as if the estimated values of probabilities were the true values. When the presented collection is sufficiently long the probability of a wrong decision can be made as close to the minimum as possible (Fig. 4). However, when the collection is too short, the probability of a wrong decision can be much worse than that of the minimax strategy. **End of Example.**

The considered examples lead to a difficult and so far an unanswered question. What should be done when a fixed sample of 2–3 elements is given and no additional elements can be obtained? Is it really the best way to just ignore these data or is it possible to make use of them? We want to fill up this gap between maximum likelihood and minimax strategies and develop a strategy that covers the whole range of learning samples lengths including zero length. However, this gap, and it is infact a gap, shows a theoretical imperfection of the commonly used learning procedures, namely, of maximum likelihood learning [2, 3]. The short sample problem in whole follows from the fact that maximum likelihood learning as well as many other learning procedures have not been deduced from any explicit risk-oriented requirement to the quality of post-learning recognition. We will formulate such risk-oriented requirements and will see what type of learning procedures follow.

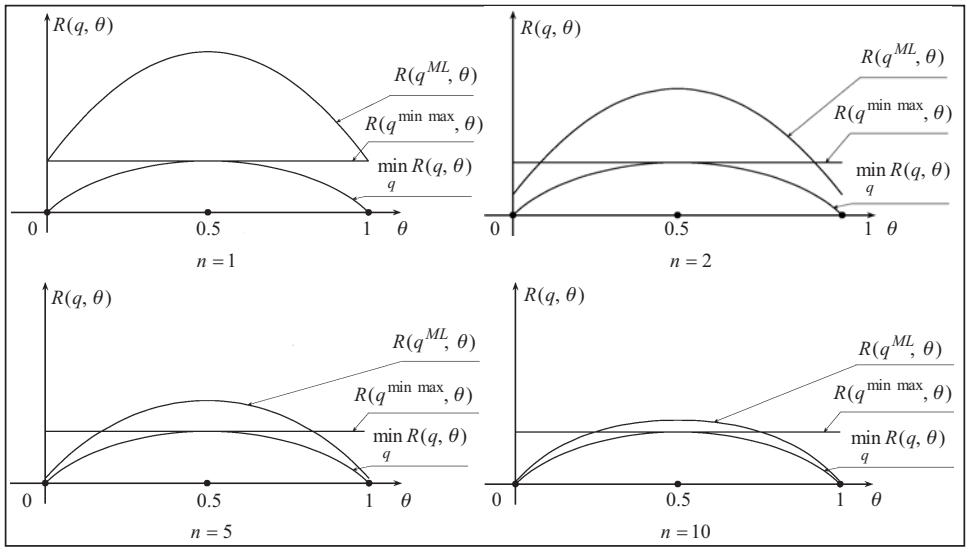

Fig. 4. Probability of a wrong decision (risk) for different sizes n of the learning sample

1. BASIC DEFINITIONS

Definition 1. An object is represented with a tuple

$$\langle X, Y, \Theta, p_{XY} : X \times Y \times \Theta \to \mathbb{R} \rangle$$

where X is a finite set of signal values $x \in X$; Y is a finite set of states $y \in Y$; Θ is a finite set of models $\theta \in \Theta$; $p_{XY}(x, y; \theta)$ is a probability of a pair $(x \in X, y \in Y)$ for a model $\theta \in \Theta$.

A signal x is an observable parameter of recognized object whereas a state y is its hidden parameter. A pair (x, y) is random and for each pair $(x \in X, y \in Y)$ its probability $p_{XY}(x, y; \theta)$ exists. However, this probability is not known because it depends on an unknown model θ. As for the model θ, it is not random, it takes a fixed but unknown value. Only the set Θ is known that the value θ belongs to.

Let z be some random data that depend on a model θ and take values from a finite set Z. The data is specified with a tuple $\langle Z, p_Z : Z \times \Theta \to \mathbb{R} \rangle$, where $p_Z(z; \theta)$ is a probability of data $z \in Z$ for model $\theta \in \Theta$.

Definition 2. Random data $\langle Z, p_Z : Z \times \Theta \to \mathbb{R} \rangle$ that depends on a model is called learning data for an object $\langle X, Y, \Theta, p_{XY} : X \times Y \times \Theta \to \mathbb{R} \rangle$ if $p_{XYZ}(x, y, z; \theta) = p_{XY}(x, y; \theta) \cdot p_Z(z; \theta)$ for all $x \in X$, $y \in Y$, $z \in Z$, $\theta \in \Theta$.

A learning sample $((x_i, y_i) \mid i = 1, 2, \ldots, n)$ used for supervised learning is a special case of learning data when

$$Z = (X \times Y)^n \text{ and } p_Z(z; \theta) = \prod_{i=1}^{n} p_{XY}(x_i, y_i; \theta).$$

A learning sample $(x_i \mid i = 1,2, \ldots, n)$ for unsupervised learning is another special case of learning data when

$$Z = X^n \text{ and } p_Z(z; \theta) = \prod_{i=1}^{n} \sum_{y \in Y} p_{XY}(x_i, y; \theta).$$

Any expert knowledge about the true model is also learning data. One can even consider the case when $|Z| = 1$ and therefore $p_Z(z; \theta) = 1$, which is equivalent to the absence of any learning data at all. We do not restrict learning data in any way except that for any fixed model the learning data z depend neither on the current signal x nor on the current state y so that

$$p_{XYZ}(x, y, z; \theta) = p_{XY}(x, y; \theta) \cdot p_Z(z; \theta) \text{ for all } x \in X, \ y \in Y, \ z \in Z, \ \theta \in \Theta.$$

Definition 3. A non-negative function $q : X \times Y \times Z \to \mathbb{R}$ is called a strategy if $\sum_{y \in Y} q(y \mid x, z) = 1$ for all $x \in X$, $z \in Z$.

A value $q(y \mid x, z)$ of a strategy $q : X \times Y \times Z \to \mathbb{R}$ is a probability of a randomized decision that the current state of an object is y, given the current observed signal x and the available learning data z. The set of all strategies $q : X \times Y \times Z \to \mathbb{R}$ is denoted Q.

Let $\omega : Y \times Y$ be a loss function whose value $\omega(y, y')$ is the loss of a decision y' when the true state is y.

Definition 4. The risk $R(q, \theta)$ of a strategy q on a model θ is the expected loss

$$R(q, \theta) = \sum_{z \in Z} \sum_{x \in X} \sum_{y \in Y} p_{XY}(x, y; \theta) p_Z(z; \theta) \sum_{y' \in Y} q(y' \mid x, z) \omega(y, y').$$

Recall that throughout the paper the sets X, Y, Z, and Θ are assumed to be finite. This allows a much more transparent formulation of the main results. Allowing some

of the sets to be infinite would require finer mathematical tools and the results might be obscured by unnecessary technical details.

2. IMPROPER AND BAYESIAN STRATEGIES

One can see that the risk of a strategy depends not only on the strategy itself but also on the model that the strategy is applied to. Therefore, in a general case, it is not possible to prefer some strategy q_1 to another strategy q_2. The risk of q_1 may be better than the risk of q_2 on some models and worse on others. However, it is possible to prefer strategy q_2 to strategy q_1 if the risk of q_1 is greater than the risk of q_2 on all models. In this case, we will say that q_2 dominates q_1 or q_1 is dominated by q_2.

Definition 5. A strategy q^0 is called improper if a strategy q^* exists such that $R(q^0, \theta) > R(q^*, \theta)$ for all $\theta \in \Theta$.

We want to exclude all improper strategies from consideration and derive a common form of all the rest. Let T denote the set of all non-negative functions $\tau : \Theta \to \mathbb{R}$ such that $\sum_{\theta \in \Theta} \tau(\theta) = 1$. Functions of such type will be referred to as weight functions.

Definition 6. A strategy q^* is called Bayesian if there exists a weight function $\tau \in T$ such that

$$q^* = \operatorname*{argmin}_{q \in Q} \sum_{\theta \in \Theta} \tau(\theta) R(q, \theta).$$

Theorem 1. Each strategy $q^0 \in Q$ is either Bayesian or improper, but never both.

Proof. For a given strategy q^0 let us define a function $F : T \times Q \to \mathbb{R}$,

$$F(\tau, q) = \sum_{\theta \in \Theta} \tau(\theta)[R(q, \theta) - R(q^0, \theta)].$$

According to Definition 4, for any fixed θ the risk $R(q, \theta)$ is a linear function of probabilities $q(y \mid x, z)$. Consequently, for any fixed τ, the function F is a linear function of probabilities $q(y \mid x, z)$ as well. Similarly, the function F is a linear function of weights $\tau(\theta)$ for any fixed strategy q. The set Q of strategies and the set T of weight functions are both closed convex sets. Consequently, due to the known duality theorem [4–6] function F has a saddle point $(\tau^* \in T, q^* \in Q)$ such that

$$\max_{\tau \in T} \min_{q \in Q} F(\tau, q) = F(\tau^*, q^*) = \min_{q \in Q} \max_{\tau \in T} F(\tau, q),$$

where

$$q^* = \operatorname*{argmin}_{q \in Q} \max_{\tau \in T} F(\tau, q), \quad \tau^* = \operatorname*{argmax}_{\tau \in T} \min_{q \in Q} F(\tau, q).$$

It is obvious that $F(\tau, q^0) = 0$ for any $\tau \in T$. Therefore, the inequality $\min_{q \in Q} F(\tau, q) \le 0$ holds for every $\tau \in T$ and, consequently,

$$\max_{\tau \in T} \min_{q \in Q} F(\tau, q) = F(\tau^*, q^*) \le 0.$$

Therefore, there are two mutually exclusive cases: either $F(\tau^*, q^*) < 0$ or $F(\tau^*, q^*) = 0$. In such way, the proof of the theorem is reduced to proving the following four propositions.

Proposition 1. If the strategy q^0 is Bayesian then $F(\tau^*, q^*) = 0$.

Proposition 2. If $F(\tau^*, q^*) = 0$ then the strategy q^0 is Bayesian.

Proposition 3. If the strategy q^0 is improper then $F(\tau^*, q^*) < 0$.

Proposition 4. If $F(\tau^*, q^*) < 0$ then the strategy q^0 is improper.

Proof of Proposition 1. If the strategy q^0 is Bayesian then according to Definition 6 a weight function τ^0 exists such that inequality

$$\sum_{\theta\in\Theta}\tau^0(\theta)R(q,\theta) \geq \sum_{\theta\in\Theta}\tau^0(\theta)R(q^0,\theta)$$

is valid for all $q \in Q$. Consequently, for all $q \in Q$ the chain

$$0 \leq \sum_{\theta\in\Theta}\tau^0(\theta)[R(q,\theta) - R(q^0,\theta)] = F(\tau^0, q) \leq \max_{\tau\in T} F(\tau, q)$$

is also valid. Since all numbers $\max_{\tau\in T} F(\tau, q)$, $q \in Q$, are not negative the least of them is also not negative and

$$\min_{q\in Q}\max_{\tau\in T} F(\tau, q) = F(\tau^*, q^*) \geq 0.$$

It follows from this inequality that $F(\tau^*, q^*) = 0$ because a case $F(\tau^*, q^*) > 0$ is impossible.

Proof of Proposition 2. Let $F(\tau^*, q^*) = 0$ then

$$0 = F(\tau^*, q^*) = \max_{\tau\in T}\min_{q\in Q} F(\tau, q) = \min_{q\in Q} F(\tau^*, q) =$$

$$= \min_{q\in Q}\sum_{\theta\in\Theta}\tau^*(\theta)[R(q,\theta) - R(q^0,\theta)] =$$

$$= \min_{q\in Q}\left[\sum_{\theta\in\Theta}\tau^*(\theta)R(q,\theta)\right] - \sum_{\theta\in\Theta}\tau^*(\theta)R(q^0,\theta).$$

It implies the equality

$$\min_{q\in Q}\sum_{\theta\in\Theta}\tau^*(\theta)R(q,\theta) = \sum_{\theta\in\Theta}\tau^*(\theta)R(q^0,\theta)$$

and therefore, the equality

$$q^0 = \operatorname*{argmin}_{q\in Q}\sum_{\theta\in\Theta}\tau^*(\theta)R(q,\theta),$$

which means that q^0 is Bayesian according to Definition 6.

Proof of Proposition 3. If the strategy q^0 is improper then according to Definition 5 a strategy q^1 exists such that inequalities $R(q^1,\theta) < R(q^0,\theta)$ hold for all θ. The set of models is finite and therefore, a value $\varepsilon < 0$ exists such that for any θ inequality $R(q^1,\theta) - R(q^0,\theta) \leq \varepsilon$ holds and the chain

$$0 > \varepsilon \geq \sum_{\theta\in\Theta}\tau(\theta)[R(q^1,\theta) - R(q^0,\theta)] = F(\tau, q^1) \geq \min_{q\in Q} F(\tau, q)$$

is valid for any $\tau \in T$. Since all numbers $\min_{q\in Q} F(\tau, q)$, $\tau \in T$, are not greater than ε the greatest of them is also not greater than ε and

$$\max_{\tau\in T}\min_{q\in Q} F(\tau, q) = F(\tau^*, q^*) \leq \varepsilon < 0.$$

Proof of Proposition 4. Let $F(\tau^*, q^*) < 0$. In this case,

$$F(\tau^*, q^*) = \min_{q \in Q} \max_{\tau \in T} F(\tau, q) = \max_{\tau \in T} F(\tau, q^*) =$$

$$= \max_{\tau \in T} \sum_{\theta \in \Theta} \tau(\theta)[R(q^*, \theta) - R(q^0, \theta)] = \max_{\theta \in \Theta} [R(q^*, \theta) - R(q^0, \theta)]$$

and therefore

$$\max_{\theta \in \Theta} [R(q^*, \theta) - R(q^0, \theta)] < 0.$$

Consequently, the inequality $R(q^*, \theta) < R(q^0, \theta)$ holds for all models $\theta \in \Theta$ and q^0 is improper according to Definition 5. □

The theorem gives good reasons to reappraise a lot of well-known methods that are commonly used as something self-evident. Let us illustrate this criticism with two simple examples. The first example considers a certain method of recognition without learning and the second relates to maximum likelihood learning. In both examples, the loss function is

$$\omega(y, y') = \begin{cases} 0, & \text{if } y = y', \\ 1, & \text{if } y \neq y'. \end{cases}$$

Example 3. Let x be an image of a letter, y be its name and θ be a position of the letter in a field of vision. Let the function $p_{XY}: X \times Y \times \Theta \to \mathbb{R}$ be constructively defined so that probability $p_{XY}(x, y; \theta)$ can be calculated for each triplet x, y, θ. In this case, when an image x with an unknown position θ is observed the decision $y^*(x)$ about the name of the letter has to be of the form

$$y^*(x) = \operatorname*{argmax}_{y \in Y} \sum_{\theta \in \Theta} \tau(\theta) p_{XY}(x, y; \theta). \tag{1}$$

Theorem 1 reveals a certain weakness of the commonly used form

$$y^*(x) = \operatorname*{argmax}_{y \in Y} \max_{\theta \in \Theta} p_{XY}(x, y; \theta). \tag{2}$$

The strategy (2) could be represented in the form (1) if the weights $\tau(\theta)$ in (1) could be chosen individually for each observation $x \in X$. However, each Bayesian strategy is specified with its own weight function $\tau: \Theta \to \mathbb{R}$ so that weights are assigned to elements of the set Θ, not of the set $\Theta \times X$. As a rule, the strategy (2) cannot be represented in the form (1) with fixed weights $\tau(\theta)$ that do not depend on x. It means that the strategy (2) is not Bayesian and is dominated by some other strategy that for each position of the letter recognizes its name better than strategy (2). **End of Example.**

Example 4. Let the sets X, Y and Θ be specified for the recognized object as well as a function $p_{XY}: X \times Y \times \Theta \to \mathbb{R}$. Let the learning information be a random learning sample $z = ((x_i, y_i) \mid i = 1,2, \ldots, n)$ such that $p_Z(z; \theta) = \prod_{i=1}^{n} p_{XY}(x_i, y_i; \theta)$.

Then the decision y^* about the current state y_0 based on the current signal x_0 and available learning sample z has to be of the form

$$y^* = \operatorname*{argmax}_{y_0 \in Y} \sum_{\theta \in \Theta} \tau(\theta) \prod_{i=0}^{n} p(x_i, y_i; \theta) \tag{3}$$

for some fixed τ that does not depend on z. One can see that the commonly used

maximum likelihood strategy

$$y^* = \arg\max_{y_0 \in Y} p(x_0, y_0; \theta^{ML}(z)), \tag{4}$$

$$\theta^{ML}(z) = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p(x_i, y_i; \theta)$$

can almost never be represented in the form (3) with constant weights and therefore is not Bayesian. It means that some other strategy exists that makes a decision about the current state based both on current signal and learning information and for each model makes it better than strategy (4). **End of Example.**

3. A GAP BETWEEN MAXIMUM LIKELIHOOD AND MINIMAX STRATEGIES

We consider maximum likelihood and minimax strategies and specify a gap between them. Let us define a strategy $q^{\text{opt}}(\theta) = \arg\min_{q \in Q} R(q, \theta)$ for each $\theta \in \Theta$ that assigns a probability $q^{\text{opt}}(y \mid x, z; \theta)$ for each triplet (x, y, z). The strategy $q^{\text{opt}}(\theta)$ is the best possible strategy that should be used if a true model was known. Since the model is known no learning data are needed. For any fixed model θ a strategy $q(\theta): X \times Y \times Z \to \mathbb{R}$ can be replaced with a strategy $q_X(\theta): X \times Y \to \mathbb{R}$ with the same risk. Probabilities $q(y \mid x, z; \theta)$ have to be transformed into probabilities $q_X(y \mid x; \theta)$ according to expression

$$q_X(y \mid x; \theta) = \sum_{z \in Z} p_Z(z; \theta) q(y \mid x, z; \theta)$$

and so the chain

$$R(q, \theta) = \sum_{z \in Z} \sum_{x \in X} \sum_{y \in Y} p_{XY}(x, y; \theta) p_Z(z; \theta) \sum_{y' \in Y} q(y' \mid x, z; \theta) \omega(y, y') =$$

$$= \sum_{x \in X} \sum_{y \in Y} p_{XY}(x, y; \theta) \sum_{y' \in Y} \omega(y, y') \sum_{z \in Z} p_Z(z; \theta) q(y' \mid x, z; \theta) =$$

$$= \sum_{x \in X} \sum_{y \in Y} p_{XY}(x, y; \theta) \sum_{y' \in Y} q_X(y' \mid x; \theta) \omega(y, y') = R(q_X, \theta)$$

is valid for each model θ. Consequently, the equality

$$\min_{q \in Q} R(q, \theta) = \min_{q_X \in Q_X} R(q_X, \theta) \tag{5}$$

is valid for each θ. The symbol Q_X in (5) designates the set of all strategies of the form $q_X: X \times Y \to \mathbb{R}$ that do not use the learning data.

Definition 7. A strategy $q^{ML}: X \times Y \times Z \to \mathbb{R}$ is called a maximum likelihood strategy if for each triplet (x, y, z) it specifies a probability

$$q^{ML}(y \mid x, z) = q_X^{\text{opt}}(x \mid y; \theta^{ML}(z)),$$

where $q_X^{\text{opt}}(\theta) = \arg\min_{q_X \in Q_X} R(q_X, \theta)$ and $\theta^{ML}(z) = \arg\max_{\theta \in \Theta} p_Z(z; \theta)$.

In other words, maximum likelihood strategies use the learning data z to estimate a model θ and make a decision that minimizes the expected loss with an assumption that the estimated model is the true model.

As it has been quoted for Examples 3 and 4, as a rule, maximum likelihood strategies cannot be represented in the form of a Bayesian strategy

$$q^B = \arg\min_{q \in Q} \sum_{\theta \in \Theta} \tau(\theta) R(q, \theta)$$

with fixed weights $\tau(\theta)$ that do not depend on the learning data. In such cases, the maximum likelihood strategy q^{ML} may be dominated by another strategy of the form $X \times Y \times Z \to \mathbb{R}$. The so-called minimax strategies, however, are free of this flaw.

Definition 8. Strategy $\operatorname{argmin}_{q \in Q} \max_{\theta \in \Theta} R(q, \theta)$ is called a minimax strategy.

Theorem 2. No minimax strategy is improper.

Proof. Let us prove an equivalent statement that any improper strategy q^0 is not minimax. Indeed, as far as q^0 is improper another strategy q^1 exists such that $R(q^1, \theta) < R(q^0, \theta)$ for all θ. Therefore, $\max_{\theta} R(q^1, \theta) < \max_{\theta} R(q^0, \theta)$ and $\min_{q} \max_{\theta} R(q, \theta) < \max_{\theta} R(q^0, \theta)$ and q^0 is not $\operatorname{argmin}_{q} \max_{\theta} R(q, \theta)$. □

Though the maximum likelihood strategy may be improper whereas the minimax strategy is never improper the first one has an essential advantage over the second. There is a rather wide class of learning data such that the maximum likelihood strategy is in a sense consistent for any recognized object whereas there is a rather wide class of recognized objects such that the minimax strategy is not consistent for any learning data. Let us exactly formulate these statements and prove them.

Let $z \in Z$ be a random variable that depends on model θ and let for each $z \in Z$ and $\theta \in \Theta$ a probability $p_Z(z; \theta)$ be given. We will say that this dependence is essential if for each two different models $\theta_1 \neq \theta_2$ a value z^* exists such that $p_Z(z^*; \theta_1) \neq p_Z(z^*; \theta_2)$. Let $z^n = (z_i \mid i = 1, 2, \ldots, n) \in Z^n$ be a learning sample, $p_{Z^n}(z^n; \theta^*) = \prod_{i=1}^{n} p_Z(z_i; \theta^*)$ be a probability of the sample and $\theta^{ML}(z^n) =$ $= \operatorname{argmax}_{\theta} p_{Z^n}(z^n; \theta)$ be a maximum likelihood estimation of the model.

Consistency is a generally known property of maximum likelihood estimate. In the considered case this property can be formulated in a simple way that the probability of inequality $\theta^{ML}(z^n) \neq \theta^*$ converges to zero when n increases or, formally,

$$\lim_{n \to \infty} \sum_{z^n \in Z^n_{err}} \prod_{i=1}^{n} p_Z(z_i; \theta^*) = 0, \tag{6}$$

where

$$Z^n_{err} = \{z^n \in Z^n \mid \theta^{ML}(z^n) \neq \theta^*\}. \tag{7}$$

Consistency of maximum likelihood estimations is the base for the proof of the following theorem about consistency of the maximum likelihood strategy.

Theorem 3. Let z be a random variable that takes values from a set Z according to probability distribution $p_Z(z; \theta)$ that essentially depends on θ;

let n be a positive integer and $z^n = (z_i \mid i = 1, 2, \ldots, n) \in Z^n$ be a random learning sample with probability distribution $p_{Z^n}(z^n; \theta) = \prod_{i=1}^{n} p_Z(z_i; \theta)$;

let $q_n^{ML}: X \times Y \times Z^n \to \mathbb{R}$ be a maximum likelihood strategy for an object $\langle X, Y, \Theta, p_{XY}: X \times Y \times \Theta \to \mathbb{R} \rangle$ and learning data $\langle Z^n, p_{Z^n}: Z^n \times \Theta \to \mathbb{R} \rangle$.

Then $\lim_{n \to \infty} \max_{\theta \in \Theta} \left[R(q_n^{ML}, \theta) - \min_{q \in Q} R(q, \theta) \right] = 0.$

Proof. As far as a set Θ is finite the proof of the theorem is reduced to proof of the equality

$$\lim_{n\to\infty}\left[R(q_n^{ML},\theta)-\min_{q\in Q}R(q,\theta)\right]=0 \tag{8}$$

for any θ. The subsequent proof is based on equality (5), on equalities (6) and (7) that express consistency of maximum likelihood estimates and on equality

$$R(q_n^{ML},\theta)=\sum_{z^n\in Z^n}p_{Z^n}(z^n;\theta)\min_{q_X\in Q_X}R(q_X,\theta^{ML}(z^n)),$$

$$\text{where } \theta^{ML}(z^n)=\operatorname*{argmax}_{\theta\in\Theta}p_{Z^n}(z^n;\theta),$$

that follows from Definition 7. The following chain is valid:

$$\lim_{n\to\infty}\left[R(q_n^{ML},\theta)-\min_{q\in Q}R(q,\theta)\right]=\lim_{n\to\infty}\left[R(q_n^{ML},\theta)-\min_{q_X\in Q_X}R(q_X,\theta)\right]=$$

$$=\lim_{n\to\infty}\left[\sum_{z_n\in Z^n}p_{Z^n}(z^n;\theta)\min_{q_X\in Q_X}R(q_X,\theta^{ML}(z^n))-\min_{q_X\in Q_X}R(q_X,\theta)\right]=$$

$$=\lim_{n\to\infty}\sum_{z_n\in Z^n}p_{Z^n}(z^n;\theta)\left[\min_{q_X\in Q_X}R(q_X,\theta^{ML}(z^n))-\min_{q_X\in Q_X}R(q_X,\theta)\right]=$$

$$=\lim_{n\to\infty}\sum_{z^n\in Z^n_{err}}p_{Z^n}(z^n;\theta)\left[\min_{q_X\in Q_X}R(q_X,\theta^{ML}(z^n))-\min_{q_X\in Q_X}R_X(q_X,\theta)\right]\le$$

$$\le\lim_{n\to\infty}\sum_{z^n\in Z^n_{err}}p_{Z^n}(z^n;\theta)\left[\max_{y\in Y}\max_{y'\in Y}w(y,y')-\min_{y\in Y}\min_{y'\in Y}w(y,y')\right]=$$

$$=\lim_{n\to\infty}\left\{\left[\max_{y\in Y}\max_{y'\in Y}w(y,y')-\min_{y\in Y}\min_{y'\in Y}w(y,y')\right]\sum_{z^n\in Z^n_{err}}p_{Z^n}(z^n;\theta)\right\}=$$

$$=\left[\max_{y\in Y}\max_{y'\in Y}w(y,y')-\min_{y\in Y}\min_{y'\in Y}w(y,y')\right]\lim_{n\to\infty}\sum_{z^n\in Z^n_{err}}p_{Z^n}(z^n;\theta)=0.$$

It follows from the chain that for any θ the inequality

$$\lim_{n\to\infty}\left[R(q_n^{ML},\theta)-\min_{q\in Q}R(q,\theta)\right]\le 0$$

holds. The difference $R(q_n^{ML},\theta)-\min_{q\in Q}R(q,\theta)$ is never negative and so (8) is proved. □

So, with the increasing length of the learning sample the risk function of the maximum likelihood strategy becomes arbitrarily close to the minimum possible risk function. Minimax strategy does not have this nice property. Moreover, for a certain class of objects, minimax strategies simply ignore the learning sample, no matter how long it is.

Theorem 4. Let for an object $\langle X, Y, \Theta, p_{XY}: X\times Y\times\Theta\to\mathbb{R}\rangle$ a pair (θ^*, q_X^*) exists such that

$$q_X^*=\operatorname*{argmin}_{q_X\in Q_X}R(q_X,\theta^*),\ \theta^*=\operatorname*{argmax}_{\theta\in\Theta}R(q_X^*,\theta).$$

Then the inequality

$$\max_{\theta\in\Theta} R(q,\theta) \geq \max_{\theta\in\Theta} R(q^*_X,\theta) \tag{9}$$

is valid for any learning data $\langle Z, p_Z : Z\times\Theta\to\mathbb{R}\rangle$ and any strategy $q : X\times Y\times Z\to\mathbb{R}$.

Proof. For any strategy $q\in Q$ we have the chain

$$\max_{\theta\in\Theta} R(q,\theta) \geq R(q,\theta^*) \geq \min_{q\in Q} R(q,\theta^*) =$$

$$= \min_{q_X\in Q_X} R(q_X,\theta^*) = R(q^*_X,\theta^*) = \max_{\theta\in\Theta} R(q^*_X,\theta). \qquad \square$$

The theorem shows that for some objects the minimax approach is particularly inappropriate because it enforces to ignore any learning data. There is nothing unusual in conditions of Theorem 4. Examples 1 and 2 in Introduction show just the cases when these conditions are satisfied.

So, there is a following gap between maximum likelihood and minimax strategies. The maximum likelihood strategy may be dominated with another strategy. In this case, it can be improved and, consequently, it is not optimal from any point of view. However, maximum likelihood strategies are consistent for a wide class of learning data and so this shortage does not become apparent when a learning sample of arbitrary size may be obtained. Cases of learning samples of fixed sizes, especially, short ones form an area of improper application of maximum likelihood strategies. This area is not covered with minimax strategies. Though minimax strategies are dominated with no strategy, for a rather wide class of objects minimax requirement enforces to ignore any learning sample, no matter how long it is.

4. MINIMAX DEVIATION STRATEGIES

This section is aimed at developing a consistent Bayesian strategy that has to fill the previously mentioned gap between maximum likelihood and minimax strategies.

Definition 9. A strategy $\operatorname{argmin}_{q\in Q} \max_{\theta\in\Theta} \left[R(q,\theta) - \min_{q'\in Q} R(q',\theta) \right]$ is called a minimax deviation strategy.

Minimax deviation strategies do not have the drawback of the minimax strategies. The following theorem, which is similar to Theorem 3 for maximum likelihood strategies, is valid for minimax deviation strategies as well.

Theorem 5. Let z be a random variable that takes values from a set Z according to probability distribution $p_Z(z;\theta)$ that essentially depends on θ;

let n be a positive integer and $z^n = (z_i \mid i = 1, 2, \ldots, n) \in Z^n$ be a random learning sample with probability distribution $p_{Z^n}(z^n;\theta) = \prod_{i=1}^{n} p_Z(z_i;\theta)$;

let $q^*_n : X\times Y\times Z^n\to\mathbb{R}$ be a minimax deviation strategy for an object $\langle X, Y, \Theta, p_{XY} : X\times Y\times\Theta\to\mathbb{R}\rangle$ and learning data $\langle Z^n, p_{Z^n} : Z^n\times\Theta\to\mathbb{R}\rangle$.

Then

$$\lim_{n\to\infty} \max_{\theta\in\Theta} \left[R(q^*_n,\theta) - \min_{q\in Q} R(q,\theta) \right] = 0. \tag{10}$$

Proof. The theorem is a straightforward consequence of Definition 9 and Theorem 3. Let q^{ML}_n be a maximum likelihood strategy for an object $\langle X, Y, \Theta, p_{XY} : X\times Y\times\Theta\to\mathbb{R}\rangle$ and learning data $\langle Z^n, p_{Z^n} : Z^n\times\Theta\to\mathbb{R}\rangle$. It follows from Definition 9 that

$$\max_{\theta\in\Theta}\left[R(q_n^*,\theta)-\min_{q\in Q}R(q,\theta)\right]\le\max_{\theta\in\Theta}\left[R(q_n^{ML},\theta)-\min_{q\in Q}R(q,\theta)\right]$$

for any n. It follows from Theorem 3 that

$$\lim_{n\to\infty}\max_{\theta\in\Theta}\left[R(q_n^*,\theta)-\min_{q\in Q}R(q,\theta)\right]\le\lim_{n\to\infty}\max_{\theta\in\Theta}\left[R(q_n^{ML},\theta)-\min_{q\in Q}R(q,\theta)\right]=0.$$

As far as the difference $\left[R(q_n^*,\theta)-\min_{q\in Q}R(q,\theta)\right]$ is non-negative for any model the equality (10) is proved. □

Let us note that the proof of Theorem 10 shows not only consistency of the minimax deviation strategy. It also shows that the minimax deviation strategy converges to the desired result not slower than the maximum likelihood strategy. Similarly, one can show that this advantage of the minimax deviation strategy holds as compared with any consistent strategy and from this point of view it is the best of all consistent strategies. Nevertheless, the following theorem states that minimax deviation strategies are also inappropriate for recognizing objects of a certain type.

Theorem 6. Let for an object $\langle X, Y, \Theta, p: X\times Y\times\Theta\to\mathbb{R}\rangle$ a model θ^* and a strategy q_X^* exist such that

$$q_X^*=\operatorname*{argmin}_{q_X\in Q_X}\left[R_X(q_X,\theta^*)-\min_{q_{X'}\in Q_X}R_X(q_{X'},\theta^*)\right],\tag{11}$$

$$\theta^*=\operatorname*{argmax}_{\theta\in\Theta}\left[R_X(q_X^*,\theta)-\min_{q_{X'}\in Q_X}R_X(q_{X'},\theta)\right].\tag{12}$$

Then the inequality

$$\max_{\theta\in\Theta}\left[R(q,\theta)-\min_{q_X\in Q_X}R(q_X,\theta)\right]\ge\max_{\theta\in\Theta}\left[R(q_X^*,\theta)-\min_{q_X\in Q_X}R(q_X,\theta)\right]$$

holds for any learning data $\langle Z, p_Z: Z\times\Theta\to\mathbb{R}\rangle$ and any strategy $q\in Q$.

Proof. In fact, proof of the theorem does not differ from the proof of the Theorem 4. □

However, the consequences of this theorem for minimax deviation strategies are not so destructive as those of Theorem 4 for minimax strategies. In fact, conditions (11) and (12) imply that a strategy $q_X^*\in Q_X$ exists that does not use learning information and assures minimal possible risk for each model,

$$R(q_X^*,\theta)=\min_{q_X\in Q_X}R(q_X,\theta)\text{ for all }\theta\in\Theta.$$

In this case, any learning data are needless and have to be omitted by any strategy.

Evidently, the minimax deviation strategy is not improper and, consequently, is Bayesian. The following theorem shows how the corresponding weight function has to be obtained.

Theorem 7. Minimax deviation strategy

$$q^*=\operatorname*{argmin}_{q\in Q}\max_{\theta\in\Theta}\left[R(q,\theta)-\min_{q_X\in Q_X}R(q_X,\theta)\right]$$

is a Bayesian strategy $\operatorname*{argmin}_{q\in Q}\sum_{\theta\in\Theta}\tau^*(\theta)R(q,\theta)$ with respect to weight function

$$\tau^*=\operatorname*{argmax}_{\tau\in T}\left[\min_{q\in Q}\sum_{\theta\in\Theta}\tau(\theta)R(q,\theta)-\sum_{\theta\in\Theta}\tau(\theta)\min_{q_X\in Q_X}R(q_X,\theta)\right].\tag{13}$$

Proof. Let us define a function $F: T \times Q \to \mathbb{R}$,

$$F(\tau, q) = \sum_{\theta \in \Theta} \tau(\theta) R(q, \theta) - \sum_{\theta \in \Theta} \tau(\theta) \min_{q_X \in Q_X} R(q_X, \theta)$$

and express q^* and τ^* in terms of F,

$$q^* = \operatorname*{argmin}_{q \in Q} \max_{\theta \in \Theta} \left[R(q, \theta) - \min_{q_X \in Q_X} R(q_X, \theta) \right] =$$

$$= \operatorname*{argmin}_{q \in Q} \max_{\tau \in T} \sum_{\theta \in \Theta} \tau(\theta) \left[R(q, \theta) - \min_{q_X \in Q_X} R(q_X, \theta) \right] = \operatorname*{argmin}_{q \in Q} \max_{\tau \in T} F(\tau, q),$$

$$\tau^* = \operatorname*{argmax}_{\tau \in T} \min_{q \in Q} F(\tau, q).$$

The function F is a linear function of q for fixed τ and a linear function of τ for fixed q and is defined on the Cartesian product of two closed convex sets T and Q. In this case a pair (τ^*, q^*) is a saddle point [1, 2, 4],

$$\min_{q \in Q} \max_{\tau \in T} F(\tau, q) = F(\tau^*, q^*) = \max_{\tau \in T} \min_{q \in Q} F(\tau, q),$$

that implies $F(\tau^*, q^*) = \min_{q \in Q} F(\tau^*, q)$ and

$$q^* = \operatorname*{argmin}_{q \in Q} F(\tau^*, q) = \operatorname*{argmin}_{q \in Q} \left[\sum_{\theta \in \Theta} \tau^*(\theta) R(q, \theta) - \sum_{\theta \in \Theta} \tau^*(\theta) \min_{q_X \in Q_X} R(q_X, \theta) \right] =$$

$$= \operatorname*{argmin}_{q \in Q} \sum_{\theta \in \Theta} \tau^*(\theta) R(q, \theta). \qquad \square$$

In such a way, developing a minimax deviation strategy is reduced to calculating weights $\tau(\theta)$ of models that maximize concave function (13). General purpose methods of non-smooth optimization [7] were used to calculate $\tau(\theta)$ in the following experiments.

5. EXPERIMENTS

Minimax deviation strategies have been built for objects considered in Introduction in Examples 1 and 2.

Minimax deviation strategies have been compared with maximum likelihood and minimax strategies. Results are presented in Figures 5 and 6 that show risk $R(q, \theta)$ of the strategies as a function of a model for several learning sample sizes. Figure 5 relates to Example 1 and Figure 6 to Example 2.

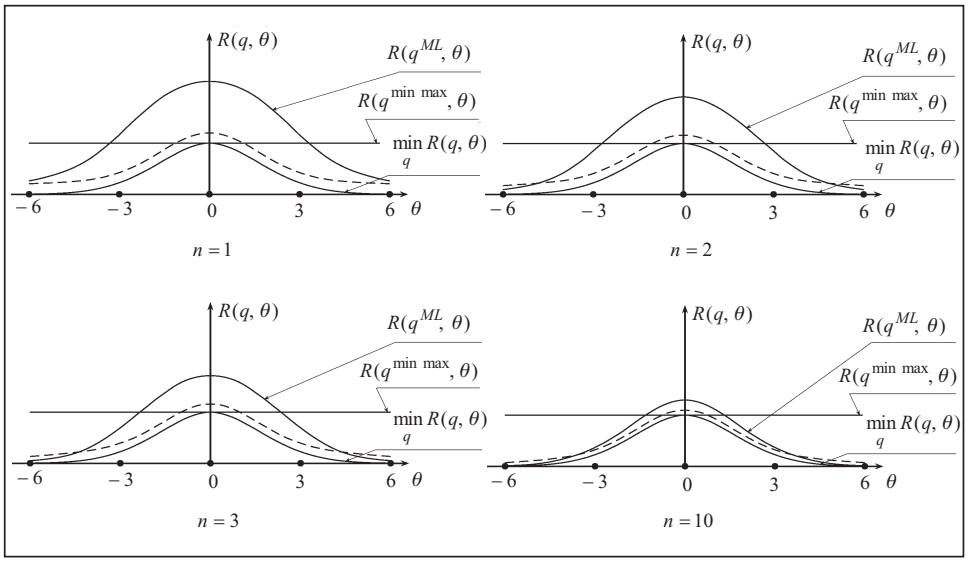

Fig. 5. Example 1. Probability of making a wrong decision for different sizes n of the learning sample. The dashed line shows the risk of the minimax deviation strategy

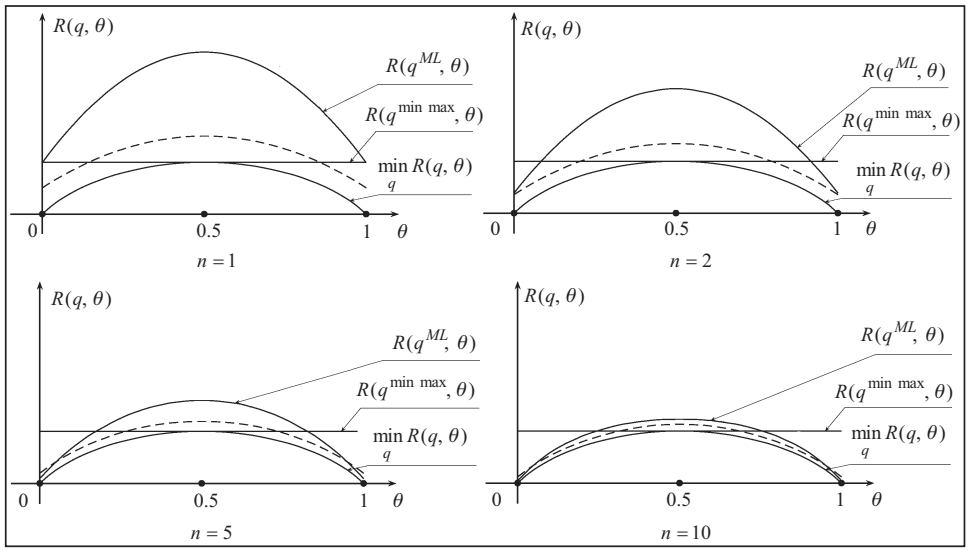

Fig. 6. Example 2. Probability of making a wrong decision for different sizes n of the learning sample. The dashed line shows the risk of the minimax deviation strategy

CONCLUSIONS

The paper analyzes the problem when for given object

$$\langle X, Y, \Theta, p_{XY}: X \times Y \times \Theta \to \mathbb{R} \rangle,$$

loss function $w: Y \times Y \to \mathbb{R}$, learning data $\langle Z, p_Z: Z \times \Theta \to \mathbb{R} \rangle$, observed current signal x and available learning data z a decision y^* about the current hidden state y has to be made. Many commonly used strategies make decisions of the form

$$y^* = \operatorname*{argmin}_{y' \in Y} \sum_{y \in Y} p_{XY}(x, y; \theta^{est}(z)) w(y, y'), \tag{14}$$

where $\theta^{est}: Z \to \Theta$ is a reasonable estimation of model θ based on learning data z. It means that the learning data are used to choose a single best model and the objects are recognized as if this best model equals the true model. The approach is acceptable when arbitrarily long learning samples are available and estimator $\theta^{est}: Z \to \Theta$ is consistent. If the learning sample is of limited size then the approach gives no guarantee for subsequent recognition. Indeed, the approach is not deduced from any risk-oriented requirement. Reasonable requirements to the quality of post-learning recognition imply the decision of the form

$$y^* = \operatorname*{argmin}_{y' \in Y} \sum_{\theta \in \Theta} \tau(\theta) p_Z(z; \theta) \sum_{y \in Y} p_{XY}(x, y; \theta) w(y, y') \tag{15}$$

that differs from (14). Moreover, any decision that differs from (15) can be replaced with a decision of the form (15) with better recognition quality.

There is nothing in decision (15) that could be treated as selecting some best model from the model set and so no question stands on what estimator $\theta^{est}: Z \to \Theta$ has to be used. No model has to be selected. On the contrary, all models should take part with their weights in making the decision. It is essential that the weights do not depend on learning data, they are determined by the requirement for the desired strategy in a particular applied situation. The paper shows a way for computing these weights for minimax deviation strategy that is appropriate for learning samples of any length and in such a way fills the gap between maximum likelihood and minimax strategies.

Minimax deviation strategy is not at all a single strategy that is reasonable in such or other application. Many other strategies are appropriate too, for example,

strategies of the form

$$\operatorname*{argmin}_{q\in Q}\max_{\theta\in\Theta}\frac{R(q,\theta)-\alpha(\theta)}{\beta(\theta)} \tag{16}$$

with predefined numbers $\alpha(\theta)$ and $\beta(\theta)>0$. The minimax strategy is a special case of (16) when $\alpha(\theta)=0$, $\beta(\theta)=1$, and the minimax deviation strategy is a case when $\alpha(\theta)=\min\limits_{q\in Q} R(q,\theta)$, $\beta(\theta)=1$. A reasonable modification of the minimax deviation strategy is a case when $\alpha(\theta)=0$, $\beta(\theta)=\min\limits_{q\in Q} R(q,\theta)$. The numbers $\alpha(\theta)$ may be risks of some other previously developed strategy and this is a case when the developer wants to check whether another better strategy is possible. At last, numbers $\alpha(\theta)$ may simply be desired values of risk in a particular applied situation.

Requirements of the form (16) together with various loss functions determine various applied situations. The obtained results show the way to cope with all of them. It has become quite clear now that each strategy of the form (16) may be represented in the form (15) because, obviously, none of them are improper. Obtained results imply an unexpected conclusion that learning data take part in the decision (15) in a unified form that depends neither on the applied situation nor on the object under ecognition, no question stands anymore on how the learning data have to influence the decision about the current state when the current signal is observed. Learning data influence the decision via and only via probabilities $p_Z(z;\theta)$, not via a choice of some best model from the model set.

REFERENCES

1. Robbins H. Asymptotically subminimax solutions of compound statistical decision problems. *Proc. the Second Berkeley Symposium on Mathematical Statistics and Probability* (31 July – 12 August 1950, Berkeley, USA). Berkeley, 1950. P. 131–148.
2. Duda R.O., Hart P.E., Stork D.G. Pattern classification. New York: Wiley, 2000. 688 p.
3. Webb A.R. Statistical pattern recognition. Wiley, 2002. 514 p.
4. Borwein J.M., Lewis A.S. Convex analysis and nonlinear optimization. New York: Springer Verlag, 2000. XII, 310 p.
5. Boyd S., Vandenberghe L. Convex optimization. New York: Cambrige University Press, 2004. 730 p.
6. Hiriart-Urruty J.-B., Lemaréchal C. Fundamentals of convex analysis. Berlin; Heidelberg: Springer Verlag, 2002. 269 p.
7. Shor N.Z. Nondifferentiable optimization and polynomial problems. Springer, 1998. 413 p.

М.І. Шлезінгер, Є.В. Водолазський

МІНІМАКСНІ СТРАТЕГІЇ МАШИННОГО НАВЧАННЯ І РОЗПІЗНАВАННЯ ОБРАЗІВ НА ОСНОВІ КОРОТКИХ НАВЧАЛЬНИХ ВИБІРОК

Анотація. Виконано аналіз задач розпізнавання образів і машинного навчання у випадку, коли якість стратегій для їхнього розв'язання визначається ризиком під час їхнього використання. Спираючись на поняття багатокритерійної оптимізації, визначено клас стратегій, непридатних для розв'язання задач, і виведено загальний вигляд усіх інших стратегій. Показано, що застосування окремих широковживаних підходів призводить до непридатних у визначеному сенсі стратегій. Зокрема, це стратегії, що ґрунтуються на найвірогіднішому оцінюванні, особливо у разі використання навчальних вибірок фіксованого і малого обсягу. Сформульовано задачі розпізнавання і навчання в єдиній уніфікованій формі, яка охоплює увесь спектр обсягів навчальних вибірок, що включає вибірки нульового обсягу. Доведено, що розв'язання задач у наведеному формулюванні виключає отримання непридатної стратегії. Сформульовано поняття стратегій розпізнавання і навчання, що мінімізують максимальне відхилення досягнутої якості від бажаної, яка, можливо, є недосяжною. Наведено приклади побудови таких стратегій та їхнього порівняння з широковживаними методами, що ґрунтуються на найвірогіднішому оцінюванні.

Ключові слова: розпізнавання образів, машиннне навчання, короткі навчальні вибірки.

Надійшла до редакції 27.07.2022