PERGAMON Journal of the Franklin Institute 336 (1999) 361–378

Journal
of The
Franklin Institute

The voting as a way to increase the decision reliability

Yu.A. Zuev[a,*], S.K. Ivanov[b]

[a] *Computing Center of Russian Academy of Sciences, Vavilova 40, 117967 Moscow, Russia*
[b] *Main Division of Information Resources for State Organs of Russia, Moscow, Russia*

Abstract

The class of weighted voting decision rules are studied. The estimates for probability of erroneous decision are obtained for a number of cases. The "accelerated perceptron" algorithm for weights correction is proposed and studied in learning and self-learning modes. The computer simulation results are provided. © 1998 The Franklin Institute. Published by Elsevier Science Ltd.

1. Introduction

The majority voting had been used by humanity to make decisions for thousands of years, and its scientific investigation also has a long history. Condorset [1] was the first to study systematically the majority rule from a probabilistic point of view and apply it to the jury practice. Later his affection to apply straightforwardly the abstract probabilistic results to real life of human cases was criticized by Poincaré [2]. But in Condorset's time it was difficult to find the field of application for his investigations apart from humanities.

The situation has changed radically in the 20th century when the question of reliability arose acutely in almost every problem related to the reception, recognition, decision making or other kinds of information processing. A general approach to solve these problems, which involves doubling of the information channels and the use of the majority voting to make decisions, appeared for the first time apparently in the work by Von Neumann [3], who considered the problem of improving the operating reliability of logical automata. Pierce [4] studied the general weighted voting, found the optimal decision rule for the case of independent voters and suggested the self-learning principle for the weighted voting.

* Corresponding author.

PII: S0016-0032(98)00025-8

The approaches described may be used in many similar situations such as: (a) expert decisions or evidence of witnesses analysis, (b) pattern recognition when different recognition algorithms are simultaneously used, (c) logic circuits which use standby units to improve reliability, and (d) multichannel telemetry and radio systems with channel diversity, for example, with respect to carrier frequency, etc.

Here we use the ideas by Von Neumann and Pierce to develop and study the general model which covers all the above mentioned areas. The main goals of the investigations were to obtain quantitative characteristics of reliability of both majority and general weighted voting (in particular, to estimate probability of error for voting rules) and also to find a rather simple and powerful algorithm of weights adaptation to make Pierce's self-learning principle practical. We propose such an algorithm which we call "accelerated perceptron" and demonstrate its good convergence properties both theoretically and by computer simulation.

In Section 2 we introduce the model used and do some preliminary analysis leading us to the majority voting decision rule whose probabilistic properties are investigated in Section 3. Section 4 is dedicated to general weighted voting. In that section we characterize the majority rule as a saddle point of weighted voting error probability P_{err}, find an optimal decision rule, obtain estimates for P_{err} both for optimal and for not necessarily optimal voting, and also study P_{err} as a function of weights permutations. In Section 5, algorithms for adaptive weights tuning are considered. Section 6 presents the results of a computer simulation.

The results presented was obtained by the authors in [5–8].

2. *Multichannel information system*

Not specifying the subject area and not confining ourselves to the majority rule, we will consider an abstract model in the form of a multichannel system, which transmits the same data in each channel. In our analysis, we consider the data as a sequence of binary symbols. The real nature of a channel is not of interest. It may be an individual expert, witness, recognition algorithm, radio channel, etc.

So the symbol $z \in \{-1, 1\}$ is transmitted along each of the n channels, and the noisy symbol $y_i \in \{-1, 1\}$ not generally the same as z, is received at the output of the channel i. The problem is to find the Boolean decision function

$$f(\mathbf{y}) = f(y_1, \ldots, y_n) : \{-1, 1\}^n \Rightarrow \{-1, 1\}$$

which gives the most reliable reconstruction of z.

Having no information about probabilities of errors in channels we can however do some qualitative analysis.

First, if we have no information about channels and there is no reason to prefer one channel to another, we need to choose $f(\mathbf{y})$ which does not change the value under any permutation of its arguments $y_1, \ldots, y_n$. That is $f(\mathbf{y})$ must be *symmetrical*.

Second, if we consider a multichannel information system to be symmetrical relative to $+1$ and -1 we need to choose $f(\mathbf{y})$ satisfying an equation $f(-\mathbf{y}) = -f(\mathbf{y})$. That is $f(\mathbf{y})$ must be *self-dual*.

Third, if our channels are not "liars", we need to choose $f(\mathbf{y})$ which does not decrease when some of y_i changes from -1 to $+1$. That is $f(\mathbf{y})$ must be *monotonic*.

Theorem 1. *For odd n the only Boolean function which is symmetrical, self-dual and monotonic, is the majority function*

$$f(\mathbf{y}) = \mathrm{sign}(y_1 + \cdots + y_n) \tag{1}$$

For even n such a function does not exist.

Proof. For each $\mathbf{y} \in \{-1, 1\}^n$ let $|\mathbf{y}|$ be the number of the components which are equal to 1. Then let $L_i = \{\mathbf{y} : \mathbf{y} \in \{-1, 1\}^n, |\mathbf{y}| = i\}$, $i = 0, \ldots, n$.

Since $f(\mathbf{y})$ is symmetrical, it is constant on each L_i. Furthermore $f(\mathbf{y})$ is monotonic so if $f(\mathbf{y}) = 1$ on L_i for some i, then $f(\mathbf{y}) = 1$ on L_j for any $j > i$. And finally from self-duality of $f(\mathbf{y})$ it follows that $n = 2k + 1$, $f(\mathbf{y}) = -1$ on L_k, and $f(\mathbf{y}) = 1$ on L_{k+1}. □

This result explains to some extent why the majority function is so widely used.

In order to obtain more interesting quantitative results it is necessary to make some additional assumptions about our multichannel system. Let

(1) *a priori* probabilities of both symbols transmitted be identical:

$$\Pr\{z = 1\} = \Pr\{z = -1\} = \tfrac{1}{2}$$

(2) for each channel the probability of correct transmission is not dependent on the input symbol, and is at least $\frac{1}{2}$:

$$\Pr\{y_i = 1 | z = 1\} = \Pr\{y_i = -1 | z = -1\} = p_i \geqslant \tfrac{1}{2}$$

(3) errors in channels are statistically independent.

Though the above constraints are rather severe and may be not satisfied in many practical situations, the classical model is very interesting and useful. It highlights the main features of voting and does not lead to overwhelming technical obstacles. So the meaningful theory can be derived.

3. The probabilistic properties of majority rule

The theorems presented in this section describe some interesting properties of majority voting. Everywhere in this section $n = 2k + 1$.

Theorem 2. *If $p_1 = \cdots = p_n$, then majority decision is an optimal decision, minimizing error probability.*

The result is a direct corollary of more general Theorem 7 which we prove in the next section.

Theorem 3. *The majority decision rule minimizes the error probability* $\bar{P}_{\mathrm{err}}$, *uniformly randomized over all* $n!$ *permutations* σ *of* $\{p_1, \dots, p_n\}$.

Proof. We have

$$\bar{P}_{\mathrm{err}} = \frac{1}{n!}\sum_{\sigma} P_{\mathrm{err}}(\sigma) = \frac{1}{n!}\sum_{\sigma}\sum_{\mathbf{y}\in\{-1,1\}^n} \Pr(z = -f(\mathbf{y}); \sigma)$$

$$= \sum_{\mathbf{y}\in\{-1,1\}^n} \frac{1}{n!}\sum_{\sigma} \Pr(z = -f(\mathbf{y}); \sigma)$$

$$= \sum_{\mathbf{y}\in\{-1,1\}^n} \Pr_{\mathrm{rand}}(z = -f(\mathbf{y}))$$

where $\Pr_{\mathrm{rand}}$ is a randomized distribution, obtained by averaging over all $n!$ permutations of $\{p_1, \dots, p_n\}$. For any $\mathbf{y}$ with $|\mathbf{y}| = i$, where $|\mathbf{y}|$ is the number of positive coordinates of $\mathbf{y}$,

$$\Pr_{\mathrm{rand}}\{\mathbf{y} \mid z = 1\} = \frac{1}{n!}\sum_{\sigma} p_{\sigma(1)} \cdots p_{\sigma(i)}(1 - p_{\sigma(i+1)}) \cdots (1 - p_{\sigma(n)})$$

$$\Pr_{\mathrm{rand}}\{\mathbf{y} \mid z = -1\} = \frac{1}{n!}\sum_{\sigma} p_{\sigma(1)} \cdots p_{\sigma(n-1)}(1 - p_{\sigma(n-i+1)}) \cdots (1 - p_{\sigma(n)})$$

$p_i \geqslant \frac{1}{2}$ for $i = 1, 2, \dots, n$, so $\Pr_{\mathrm{rand}}\{\mathbf{y} \mid z = 1\} \geqslant \Pr_{\mathrm{rand}}\{\mathbf{y} \mid z = -1\}$ for $i > n - i$ and $\Pr_{\mathrm{rand}}\{\mathbf{y} \mid z = 1\} \leqslant \Pr_{\mathrm{rand}}\{\mathbf{y} \mid z = -1\}$ for $i < n - i$, and assertion of the theorem follows from $\Pr_{\mathrm{rand}}\{z = 1 \mid \mathbf{y}\}/\Pr_{\mathrm{rand}}\{z = -1 \mid \mathbf{y}\} = \Pr_{\mathrm{rand}}\{\mathbf{y} \mid z = 1\}/\Pr_{\mathrm{rand}}\{\mathbf{y} \mid z = -1\}$. □

Theorem 4. *If average probability* $p = (\sum_{i=1}^{n} p_i)/n$ *is constant and* $p \geqslant (k + 1)/(2k + 1)$ *then the maximum of the majority rule error probability (considered as a function of* $p_1, \dots, p_n$*) is reached when* $p_1 = \cdots = p_n = p$.

The theorem directly follows from Hoeffding's results [9]. See also [10, 11].

Theorem 5. *For* $p_1 = \cdots = p_n = p > \frac{1}{2}$, *the error probability of majority rule is asymptotically equal to*

$$P_{\mathrm{err}}^{\mathrm{maj}} \sim \frac{(4p(1 - p))^{k+1}}{2(2p - 1)\sqrt{\pi k}}, \quad k \to \infty \tag{2}$$

Proof. $P_{\mathrm{err}}^{\mathrm{maj}} = \sum_{i=0}^{k} b_i$, where $b_i = \binom{n}{i} p'(1 - p)^{n-1}$. As follows from Stirling's estimate for factorial,

$$b_k = \binom{2k + 1}{k} p^k (1 - p)^{k+1} \sim \frac{1}{\sqrt{\pi k}} 2^{2k+1} p^k (1 - p)^{k+1}, \quad k \to \infty$$

It is easy to see that

$$\frac{b_{i+1}}{b_i}=\frac{1}{n-i+1}\frac{1-p}{p}\sim\frac{1-p}{p},\quad \text{for } k-\sqrt{k}\leqslant i\leqslant k,\ k\to\infty$$

So we have

$$p_{\mathrm{err}}^{\mathrm{maj}}=\sum_{i=0}^{k}b_i=\sum_{i=0}^{k-\sqrt{k}}b_i+\sum_{i=k-\sqrt{k}+1}^{k}b_i\sim\sum_{i=k-\sqrt{k}+1}^{k}b_i\sim b_k\sum_{i=k-\sqrt{k}+1}^{k}\left(\frac{1-p}{p}\right)^{k-i}$$

$$\sim b_k\sum_{i=0}^{\infty}\left(\frac{1-p}{p}\right)^{i}=b_k\frac{p}{2p-1}\sim\frac{1}{\sqrt{\pi k}}2^{2k+1}p^k(1-p)^{k+1}\frac{p}{2p-1}$$

$$=\frac{(4p(1-p))^{k+1}}{2(2p-1)\sqrt{\pi k}}\qquad\square$$

The quantity of interest is the error probability in the worst case when channels are not statistically independent. The answer is given by the following theorem.

Theorem 6. *In general case when statistical independence may not be satisfied the error probability rule is not greater than the average channel error probability multiplied by* $(2k+1)/(k+1)$.

$$p_{\mathrm{err}}^{\mathrm{maj}}\leqslant\frac{2k+1}{k+1}\frac{1}{n}\sum_{i=1}^{n}(1-p_i)\tag{3}$$

and the equality can be satisfied.

Proof. Let x_i be equal to 1 if there is an error in the i-th channel, else, x_i be equal to 0. Then

$$p_{\mathrm{err}}^{\mathrm{maj}}=\Pr\left\{\sum_{i=1}^{n}x_i>k+1\right\}\leqslant\frac{1}{k+1}E\left(\sum_{i=1}^{n}x_i\right)=\frac{2k+1}{k+1}\frac{1}{n}\sum_{i=1}^{n}(1-p_i)$$

In the following example the equality takes place. Consider a finite probabilistic space, whose elements are vectors $\mathbf{x}^j=(x_1^j,\dots,x_n^j), j=1,\dots,(n+m)$. Let $(k+1)$ components of any $\mathbf{x}^j, j=1,\dots,n$, are equal to 1 and other k components are equal to 0; $\mathbf{x}^j=0$, $j=n+1,\dots,n+m$; and $\Pr\{\mathbf{x}^j\}=1/(n+m)$ for any j. As above $(x_i^j=1)\Leftrightarrow$(there is an error in the ith channel). It is easy to see that the average channel error probability is equal to $(k+1)/(n+m)$ and the majority rule error probability is equal to $n/(n+m)$. □

4. General weighed voting

We consider the generalization of the majority rule known as the weighed voting rule. In that case some positive number a_i called the weight is prescribed to each channel, and the decision with maximal summary weight is chosen. Formally, it may be expressed as

$$f(\mathbf{y}) = \operatorname{sign}(a_1 y_1 + \cdots + a_n y_n) = \operatorname{sign}(\mathbf{a}, \mathbf{y}) \tag{4}$$

Such a Boolean function is called the threshold function (see [12, 13] for details). The weighed voting error probability is a function of both weights and channels probabilities of correct transmission: $P_{\text{err}}(a_1, \ldots, a_n, p_1, \ldots, p_n)$. Theorems 2 and 4 show that when the average probability $p = (\sum_{i=1}^{n} p_i)/n$ is constant and $p \geqslant (k+1)/(2k+1)$; the error probability $P_{\text{err}}(a_1, \ldots, a_n, p_1, \ldots, p_n)$ has a saddle point when $a_1 = \cdots = a_n$ (majority decision rule is used) and $p_1 = \cdots = p_n = p$ (case of equal channel probabilities). In other words, when the decision rule is changed, the error probability increases; and when the probabilities are changed, the error probability decreases.

The special role of threshold functions in decision making is explained by the next theorem, well-known from the beginning of the 1960s. [4, 14–16].

Theorem 7. *The optimal decision rule, minimizing probability of error at the output of a system, is given by the function of type* (4), *where*

$$a_i = \log \frac{p_i}{1 - p_i}, \quad i = 1, \ldots, n \tag{5}$$

Proof. The probabilities of the two values of z can be expressed as

$$\Pr\{z = 1 | \mathbf{y}\} = C \prod_{i=1}^{n} \Pr\{y_i | z = 1\}, \quad \Pr\{z = -1 | \mathbf{y}\} = C \prod_{i=1}^{n} \Pr\{y_i | z = -1\}$$

where $C = (2 \Pr\{\mathbf{y}\})^{-1}$.

Following the maximum likelihood principle we need to choose the value $z = 1$ if $\Pr\{z = 1 | \mathbf{y}\} / \Pr\{z = -1 | \mathbf{y}\} \geqslant 1$ and $z = -1$ in the contrary case.

But

$$\Pr\{z = 1 | \mathbf{y}\} / \Pr\{z = -1 | \mathbf{y}\} \geqslant 1 \iff \sum_{i=1}^{n} \log(\Pr\{y_i | z = 1\} / \Pr\{y_i | z = -1\}) \geqslant 0$$

and

$$\log(\Pr\{y_i | z = 1\} / \Pr\{y_i | z = -1\}) = y_i \log(p_i / (1 - p_i)) \qquad \square$$

The next theorem gives a two-sided estimate for the probability of error at the output of a system when the optimal decision rule is used.

Theorem 8. *Let P_{err} be the probability of error at the output of a system with optimal channel weights. If $\frac{1}{2} < m \leqslant p_i \leqslant M < 1$, $i = 1, \ldots, n$, then*

$$\frac{1 - M}{M} \binom{n}{[n/2]} \prod_{i=1}^{n} \sqrt{p_i(1 - p_i)} \leqslant P_{\text{err}} \leqslant \frac{m}{2m - 1} \binom{n}{[n/2]} \prod_{i=1}^{n} \sqrt{p_i(1 - p_i)} \tag{6}$$

We need some preliminary results to the proof of the theorem. Let us now consider $\mathbf{y} \in \{-1, 1\}^n$ as the vertices of n-cube. Vertex $\mathbf{y}$ is called positive if $f(\mathbf{y}) = 1$. The pair of vertices $[\mathbf{y}, \mathbf{x}]$ is an edge of n-cube if and only if there is unique j such that $y_i = x_i$ for $i \neq j$, and $y_j \neq x_j$. In that case, the jth coordinate is called the leading coordinate of the edge. The edge $[\mathbf{y}, \mathbf{x}]$ is considered as positive if $y_j = -1$, otherwise it is negative.

Let π be a partial order on the set of the vertices: $\mathbf{y}_1 \, \pi \, \mathbf{y}_2 \Leftrightarrow$ (the sequence of positive edges $[\mathbf{y}_1, \mathbf{x}_1], [\mathbf{x}_1, \mathbf{x}_2], \ldots, [\mathbf{x}_{m-1}, \mathbf{x}_m], [\mathbf{x}_m, \mathbf{y}]$ exists). Subset A of pairwise incomparable (in the sense of above order) vertices is called Sperner's set or anti-chain. According to the well-known Sperner's theorem [17] the cardinal number of any anti-chain A satisfies the inequality $\mathrm{card}(A) \leqslant \binom{n}{[n/2]}$. If $n = 2k$, there is only one anti-chain for which

$$\mathrm{card}(A) = \binom{n}{[n/2]}$$

It consists of all the vertices with k positive coordinates. If $n = 2k + 1$, there is two such anti-chains: the subset of all the vertices with k positive coordinates and subset of all the vertices with $k + 1$ positive coordinates.

We also need some new definitions. Let border set $\Gamma(f)$ of monotonic threshold function f be defined as a set of its positive vertices $\mathbf{y}$ for which exists the negative edge $[\mathbf{y}, \mathbf{x}]$ and $f(\mathbf{x}) = -1$; first adjoining layer $\Gamma_1(f)$ of monotonic threshold function f be defined as a set of its positive vertices $\mathbf{y}$ such that $f(\mathbf{x}) = -1$ for all negative edges $[\mathbf{y}, \mathbf{x}]$; and ith adjoining layer $\Gamma_i(f)$ of monotonic threshold function f be defined as a set of its positive vertices $\mathbf{y}$ such that for all negative edges $[\mathbf{y}, \mathbf{x}]$ either $f(\mathbf{x}) = -1$ or there exists $j \in [1, \ldots, i-1]$ such that $\mathbf{x} \in \Gamma_j(f)$. It is obvious that every positive vertex belongs to only one adjoining layer, and $\Gamma_1(f) \subseteq \Gamma(f)$. If f is a majority function then $\Gamma_1(f) = \Gamma(f)$.

Lemma 1. *If f is a monotonic self-dual threshold function then for any i,*

$$\mathrm{card}(\Gamma_i(f)) \leqslant \binom{n}{[n/2]} \leqslant \mathrm{card}(\Gamma(f)) \tag{7}$$

Equality signs in Eq. (7) are valid only for the majority function and first adjoining layer.

Proof. Since every adjoining layer is anti-chain, the left inequality follows from Sperner's theorem [17]. The proof of the right inequality is more tedious. It may be found in [8]. □

Proof of Theorem 8. Let $f(\mathbf{y})$ be optimal. Then for $\mathbf{y}$ such that $f(\mathbf{y}) = 1$,

$$\Pr\{\mathbf{y} \mid z = -1\} \leqslant \sqrt{\Pr\{\mathbf{y} \mid z = -1\} \Pr\{\mathbf{y} \mid z = 1\}} = \prod_{i=1}^{n} \sqrt{p_i(1 - p_i)}$$

$$P_{\mathrm{err}} = \sum_{\mathbf{y} \in \{-1,1\}^n} \Pr\{\mathbf{y} \mid z = -1\} = \sum_{l} \sum_{\mathbf{y} \in \Gamma_i(f)} \Pr\{\mathbf{y} \mid z = -1\}$$

For every $l > 1$ and $\mathbf{y} \in \Gamma_l(f)$ we can find $\mathbf{x} \in \Gamma_{l-1}(f)$ such that there exists positive $[\mathbf{x}, \mathbf{y}]$. Let the jth coordinate be leading. Then

$$\Pr\{\mathbf{y} | z = -1\} \leqslant \frac{1-p_j}{p_j} \Pr\{\mathbf{x} | z = -1\} \leqslant \frac{1-m}{m} \Pr\{\mathbf{x} | z = -1\}$$

and

$$P_{\text{err}} \leqslant \sum_l \left(\frac{1-m}{m}\right)^{l-1} \operatorname{card}(\Gamma_l(f)) \prod_{i=1}^{n} \sqrt{p_i(1-p_i)}$$

$$\leqslant \binom{n}{[n/2]} \prod_{i=1}^{n} \sqrt{p_i(1-p_i)} \sum_{l=1}^{\infty} \left(\frac{1-m}{m}\right)^{l-1} \leqslant \frac{m}{2m-1} \binom{n}{[n/2]} \prod_{i=1}^{n} \sqrt{p_i(1-p_i)}$$

Now let $\mathbf{y} \in \Gamma(f)$. Then there exists positive $[\mathbf{x}, \mathbf{y}]$ with $f(\mathbf{x}) = -1$. Let the jth coordinate be leading. Then

$$\Pr\{\mathbf{y} | z = -1\} \geqslant \frac{1-p_j}{p_j} \Pr\{\mathbf{x} | z = -1\} \geqslant \frac{1-M}{M} \Pr\{\mathbf{x} | z = -1\}$$

$$\geqslant \frac{1-M}{M} \prod_{i=1}^{n} \sqrt{p_i(1-p_i)}$$

$$P_{\text{err}} = \frac{1-M}{M} \operatorname{card}(\Gamma(f)) \prod_{i=1}^{n} \sqrt{p_i(-p_i)}$$

$$\geqslant \frac{1-M}{M} \binom{n}{[n/2]} \prod_{i=1}^{n} \sqrt{p_i(1-p_i)} \qquad \square$$

In general case of not necessary optimal weights the next more rough estimates for error probability may be obtained.

Theorem 9. *For every threshold function* (2) *the next estimates takes the place:*

1. *if* $(\mathbf{a}_\beta, \mathbf{a}) \geqslant 0$, *then* $P_{\text{err}} \leqslant \exp\{-(\mathbf{a}_\beta, \mathbf{n})^2/2\}$,
2. *if* $(\mathbf{a}_\pi, \mathbf{a}) \geqslant 0$, *then*

$$P_{\text{err}} \leqslant \prod_{i=1}^{n} 2\sqrt{p_i(1-p_i)} \exp\{\mathbf{a}_\pi^2/2\} \exp\{-(\mathbf{a}_\pi, \mathbf{n})^2/2\} \tag{8}$$

where

$\mathbf{a}_\beta = (2p_1 - 1, \ldots, 2p_n - 1)$ *baricentric weight vector*
$\mathbf{a}_\pi = (\frac{1}{2}\log(p_1/1 - p_1), \ldots, \frac{1}{2}\log(p_n/1 - p_n)$ *optimal weight vector*
$\mathbf{n} = \mathbf{a}/|\mathbf{a}|$ *normalized weight vector.*

Proof. We use the exponential version of Techebyshev's inequality, also known as Bernstein's inequality. Let $z = 1$. For $h > 0$ we have

$$P_{\text{err}} = \Pr\left\{\sum_{i=1}^{n} a_i y_i < 0\right\} = \Pr\left\{\exp\left(-h\sum_{i=1}^{n} a_i y_i\right) > 1\right\}$$

$$\leqslant E\left\{\exp\left(-h\sum_{i=1}^{n} a_i y_i\right)\right\} = \prod_{i=1}^{n} E\{\exp(-ha_i y_i)\}$$

$$E\{\exp(-ha_i y_i)\} = 2\sqrt{p_i(1-p_i)}\,\mathrm{ch}(ha_i - \tfrac{1}{2}\ln(p_i/(1-p_i)))$$

1. *First estimate.* Let $L_i(h) = \ln(2\sqrt{p_i(1-p_i)}\,\mathrm{ch}(ha_i - \frac{1}{2}\ln(p_i(1-p_i))))$. Then $L_i(h) = L_i(0) + L_i(0)h + \frac{1}{2}L_i''(h_l)h^2$, $0 < h_l < h$, $L_i(0) = 0$, $L_h'(h) = \mathrm{th}(ha_i - \frac{1}{2}\ln(p_i/(1-p_i)))$, $L_i(0) = -(2p_i - 1)a_i$. $L_i''(h) = a_i^2/\mathrm{ch}^2(ha_i - \frac{1}{2}\ln(p_i/(1-p_i))) \leqslant a_i^2$.

It gives the estimates

$$L_i(h) \leqslant -h(2p_i - 1)a_i + \tfrac{1}{2}h^2 a_i^2$$

and

$$P_{\text{err}} \leqslant \exp\left(-h\sum_{i=1}^{n}(2p_i - 1)a_i + \tfrac{1}{2}h^2\sum_{i=1}^{n} a_i^2\right) = \exp(R(h))$$

$$\min\{R(h)\} = R(h_0) = -\left(\sum_{i=1}^{n}(2p_i - 1)a_i\right)^2 \Big/ \sum_{i=1}^{n} a_i^2$$

where

$$h_0 = \sum_{i=1}^{n}(2p_i - 1)a_i \Big/ \sum_{i=1}^{n} a_i^2 > 0$$

from which the first estimate follows.

2. *Second estimate*: Using the inequality $\ln(\mathrm{ch}(x)) \leqslant \frac{1}{2}x^2$ we have:

$$L_i(h) = \ln(2\sqrt{p_i(1-p_i)}) + \ln(\mathrm{ch}(ha_i - \tfrac{1}{2}\ln(p_i/(1-p_i))))$$

$$\leqslant \ln(2\sqrt{p_i(1-p_i)}) + \tfrac{1}{2}(ha_i - \tfrac{1}{2}\ln(p_i(1-p_i)))^2$$

$$= \ln(2\sqrt{p_i(1-p_i)}) + \tfrac{1}{2}(h\mathbf{a} - \mathbf{a}_\pi)^2$$

The minimum of the expression on the right-hand side of the above inequality is achieved for $h = (\mathbf{a}_\pi, \mathbf{a})/\mathbf{a}^2$ and is equal to $\ln(2\sqrt{p_i(1-p_i)}) + \frac{1}{2}(\mathbf{a}_\pi^2 - (\mathbf{a}_\pi, \mathbf{n})^2)$. That gives the second estimate. □

Note that the first estimate also follows from the more general Hoeffding's result [18].

If the weights of the threshold function are constant, but may be permutated, the reliability of the decision rule may be treated as a function of weight permutations. Let

$a_1 \geqslant a_2 \geqslant \cdots \geqslant a_n$, $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_n$ and G_n be a symmetrical group of permutations on the set $\{1, \ldots, n\}$ of indices of weights. As usual, define the set of disorders $X(\sigma)$ for permutation $\sigma \in G_n$: $\{\sigma(i_1), \sigma(i_2)\} \in X(\sigma)$ if $i_1 < i_2$, $\sigma(i_1) > \sigma(i_2)$.
Let π be partial order on G_n: $\sigma_1 \, \pi \, \sigma_2 \Leftrightarrow X(\sigma_1) \subseteq X(\sigma_2)$

Then the minimal element of G_n is the identical permutation

$$\sigma_I = \begin{pmatrix} 1 & 2 & \ldots & n \\ 1 & 2 & \ldots & n \end{pmatrix}$$

the maximal element is the reverse permutation

$$\sigma_R = \begin{pmatrix} 1 & 2 & \ldots & n \\ n & n-1 & \ldots & 1 \end{pmatrix}$$

Theorem 10. *The error probability P is a monotonic function relative to this order, i.e.*

$$\sigma_1 \, \pi \, \sigma_2 \Rightarrow P(\sigma_1) \leqslant P(\sigma_2)$$

Proof. As follows from [19], if $\sigma_1, \sigma_2 \in G_n$, $\sigma_1 \, \pi \, \sigma_2$, $\sigma_1 \neq \sigma_2$ then there exists the set of transpositions $\tau_i \in G_n$, such that $\sigma_1 \, \pi \, \sigma_1 \, \tau_1 \, \pi \, \sigma_1 \, \tau_1 \tau_2 \pi \ldots \pi \sigma_1 \tau_1, \tau_2, \ldots, \tau_m = \sigma_2$ for some m. So we only need to prove the theorem for a transposition. Suppose τ transposes indices $u = i_k$ and $v = i_l$ which reside at positions k and l, $k < l$ (i.e. $p_k \geqslant p_l$). We have $\sigma_1 \, \pi \, \sigma_1 \tau \Rightarrow u < v$ (i.e. $a_u > a_v$). Then

$$\begin{aligned} P(\sigma_1) = {} & p_k p_l \Pr\left\{\sum a_i y_i < -a_u - a_v\right\} + p_k(1-p_i) \Pr\left\{\sum a_i y_i < a_v - a_u\right\} \\ & + (1-p_k)p_i \Pr\left\{\sum a_i y_i < a_u - a_v\right\} \\ & + (1-p_k)(1-p_l) \Pr\left\{\sum a_i y_i < a_u + a_v\right\} \\ P(\sigma_1\tau) = {} & p_k p_l \Pr\left\{\sum a_i y_i < -a_u - a_v\right\} + p_k(1-p_i) \Pr\left\{\sum a_i y_i < a_u - a_v\right\} \\ & + (1-p_k)p_i \Pr\left\{\sum a_i y_i < a_v - a_u\right\} \\ & + (1-p_k)(1-p_l) \Pr\left\{\sum a_i y_i < a_u + a_v\right\} \end{aligned}$$

where $\sum$ acts on $\{i : i \neq u, \ \neq v\}$,

$$a_u > a_v \Rightarrow \Pr\left\{\sum a_i y_i < a_u - a_v\right\} \geqslant \Pr\left\{\sum a_i y_i < a_v - a_u\right\}$$

and

$$p_k \geqslant p_l \Rightarrow p_k(1-p_l) \geqslant p_l(1-p_k)$$

so $P(\sigma_1) \leqslant P(\sigma_l\tau)$. □

If the values of channels probabilities of correct transmission are known but we don't know which channel each probability corresponds to, the problem of finding the weights, optimal in minimax sense, arises. We consider the choice of weights as a two move game with nature in which the man does the first move, choosing the weights a_i to minimize error probability P, and then the nature does the second and last move, permutating probabilities p_i to maximize P.

Theorem 11. *The optimal policy of the man is to choose equal weights, i.e. use the majority rule. The optimal policy of the nature is to use reverse permutation of probabilities.*

Proof. The first assertion follows from the fact that any function f, minimizing P for worst permutation of probabilities, is not worse than the majority function for any permutation, because the error probability of the majority rule does not depend on permutations at all. So $\bar{P}_{\mathrm{err}}$ (defined in Theorem 3) for f is not greater than that for the majority function. But as shown in Theorem 3, just majority function minimizes $\bar{P}_{\mathrm{err}}$, so f is identical to it. The second assertion follows from Theorem 10, since the error probability P is always maximal on the maximal element of G_n, which is reverse permutation. □

5. Learning and self-learning of weighed voting procedures

If the probabilities p_i are known and channels are independent, the optimal decision rule given by Theorem 7 may be used. Problems arise when the probabilities are unknown or independence condition is not satisfied. In such cases the majority rule may be used. But if among the channels there is a large group of channels with error probabilities close to $\frac{1}{2}$ or a large group of strongly correlated channels, the majority rule will be far from being optimal.

In many cases a general method of choosing weights can be used which imposes no *a priori* restrictions on the statistical properties of the multichannel system (and in particular does not require the channels to be independent). The method involves sending a test sequence of symbols $z_1, z_2, \ldots$ through the multichannel system and choosing a threshold function which separates the n-dimensional patterns of symbols $\mathbf{y}_1, \mathbf{y}_2 \ldots$ on the set of hypercube vertices $\mathbf{y} \in \{-1, 1\}^n$ with no errors, or with a small number of errors. This procedure is referred to as learning with a teacher or, in short, the learning case.

The self-learning principle formulated by Pierce [3] implies using the system output symbol as a reference instead of the true transmitted symbol z which remains unknown. The need for self-learning arises either when it is impossible to transmit a test sequence, or when the transmission environment in the channels varies rapidly. If the characteristics of the channels are changed substantially during test transmission, learning with a teacher has no sense at all. The idea of self-learning is very fruitful in such situations.

In the learning case, the classical approach is the perceptron algorithm [15, 16, 20], which corrects the weights using the rule

$$\mathbf{a}_{k+1} = \begin{cases} \mathbf{a}_k & \text{if } \operatorname{sign}(\mathbf{a}_k, \mathbf{y}_k) = z_k \\ \mathbf{a}_k + z_k y_k & \text{if } \operatorname{sign}(\mathbf{a}_k, \mathbf{y}_k) \neq z_k \end{cases} \tag{9}$$

A remarkable property of the classical perceptron is that after a finite number of weights corrections it correctly finds the hyperplane separated the test sequence, if such a hyperplane exists. In real life problems, however, nothing at all is usually known about separability. At the same time, the classical perceptron has a number of drawbacks. These include:

(1) slow learning, since the weight vector is only corrected when errors are made;
(2) random final position of the separating hyperplane, while common sense and experience suggest that out of the learning context, good separation requires the patterns of test symbols to be as far away from it as possible;
(3) in the case of linear inseparability of the test sequence, the modulus of the weight vector remains small during learning and there is a large change in the threshold function with each correction, leading to strong fluctuations in the quality of the decision rule;
(4) self-learning is impossible.

We proposed a new algorithm, called an "accelerated perceptron", that modifies the weights vector at each step. The accelerated perceptron does not suffer from the drawbacks of the classical perceptron, although it does not guarantee error-free separation in the separate case. One of its properties is that, using Pierce's principle, it can be employed immediately without any modification in the self-learning mode.

The weight correction rule for an accelerated perceptron in learning with a teacher mode is

$$\mathbf{a}_{k+1} = \mathbf{a}_k + z_k\mathbf{y}_k \tag{10}$$

The weight correction rule for an accelerated perceptron in the self-learning mode is

$$\mathbf{a}_{k+1} = \mathbf{a}_k + \operatorname{sign}(\mathbf{a}_k, \mathbf{y}_k)\mathbf{y}_k \tag{11}$$

Theorem 12. *Let $\mathbf{a}_\beta$ be a baricentric vector $\mathbf{a}_\beta = (2p_i - 1, \dots, 2p_n - 1)$ and $(\mathbf{a}_\beta, \mathbf{y}) \neq 0$ for every $\mathbf{y} \in \{-1, 1\}^n$. In the case of learning, the sequence of threshold functions generated by accelerated perceptron stabilizes at baricentric threshold function $f_\beta(\mathbf{y}) = \operatorname{sign}(\mathbf{a}_\beta, \mathbf{y})$ with probability 1.*

Proof. $E(\mathbf{a}_{k+1} - \mathbf{a}_k) = \mathbf{a}_\beta$ so the convergence of the sequence $\mathbf{a}_1/|\mathbf{a}_1|, \mathbf{a}_2/|\mathbf{a}_2|, \dots$ to $\mathbf{a}_\beta/|\mathbf{a}_\beta|$ follows from the strong law of large numbers. □

In order to get an analogous result for the case of self-learning, we need to define two new concepts.

Definition 1. The vector $\mu = \mu(f) = \sum_{\mathbf{y} \in \{-1,1\}^n} \Pr\{\mathbf{y}\} f(\mathbf{y})\mathbf{y}$, is called the moment of $f(\mathbf{y})$, where $\Pr\{\mathbf{y}\} = \frac{1}{2}(\Pr\{\mathbf{y}|z = 1\} + \Pr\{\mathbf{y}|z = -1\})$ is the probability of $\mathbf{y}$.

Definition 2. $f(\mathbf{y})$ is called stable, if $f(\mathbf{y}) = \operatorname{sign}(\boldsymbol{\mu}(f), \mathbf{y})$.

Theorem 13. *If $p_i > 1/2$, $i = 1, \dots, n$, then the set of stable functions is not empty.*

Proof. Let $\mathbf{a}_1 \neq 0$, and all its components be nonnegative. Let $\boldsymbol{\mu}_i$ be the moment of $f(\mathbf{y}; \mathbf{a}_i)$, where $\mathbf{a}_{i+1} = \mathbf{a}_i + \boldsymbol{\mu}_i$, $i = 1, 2, \ldots$ If $f(\mathbf{y}; \mathbf{a}_{i+1}) = f(\mathbf{y}; \mathbf{a}_i)$ then $\boldsymbol{\mu}_{i+1} = \boldsymbol{\mu}_i$. Else there exists $\mathbf{y}_i$, such that $(\mathbf{a}_i, \mathbf{y}_i) < 0$ and $(\mathbf{a}_i + \boldsymbol{\mu}_i, \mathbf{y}_i) > 0$, hence $\boldsymbol{\mu}_{i+1} = \boldsymbol{\mu}_i + 4\Pr\{\mathbf{y}_i\}\mathbf{y}_i$, $(\boldsymbol{\mu}_i, \mathbf{y}_i) > 0$ and $|\boldsymbol{\mu}_{i+1}| > |\boldsymbol{\mu}_i|$. There are only finite number of threshold functions. So there exists j such that $|\boldsymbol{\mu}_i| = |\boldsymbol{\mu}_j|$ for every $i > j$. The function $f(\mathbf{y}; \boldsymbol{\mu}_i)$ is stable. □

Theorem 14. *In the case of self-learning, the sequence of threshold functions* $f(\mathbf{y}; \mathbf{a}_k)$, *generated by the accelerated perceptron, stabilizes at one of the stable functions with probability* 1.

Proof. Let $n_k = \mathbf{a}_k/|\mathbf{a}_k|$. Consider $B_k = (\mathbf{n}_k, \boldsymbol{\mu}_k)$. It is obvious that $0 \leqslant B_k \leqslant \max_{\mathrm{f}} \{|\boldsymbol{\mu}(f)|\}$ and hence $E(B_k)$ is also bounded. Let $\boldsymbol{\mu}_k = E(\mathbf{a}_{k+1} - \mathbf{a}_k) = E(f(\mathbf{y}_k; \mathbf{a}_k)\mathbf{y}_k)$ is the moment of $f(\mathbf{y}_k; \mathbf{a}_k)$. First of all we note that $|\mathbf{a}_k| = O(k)$ with probability 1 as $k \to \infty$. Consider the increment $\Delta B_k = B_{k+1} - B_k$. We have

$$\Delta B_k = (\mathbf{n}_k + \Delta\mathbf{n}_k, \boldsymbol{\mu}_k + \Delta\boldsymbol{\mu}_k) - (\mathbf{n}_k, \boldsymbol{\mu}_k) = (\Delta\mathbf{n}_k, \boldsymbol{\mu}_k) + (\mathbf{n}_k + \Delta\mathbf{n}_k, \Delta\boldsymbol{\mu}_k)$$

where $\Delta\boldsymbol{\mu}_k$ is the possible change of the moment due to the changing of f. Clearly

$$(\mathbf{n}_k + \Delta\mathbf{n}, \Delta\boldsymbol{\mu}_k) \geqslant 0, \text{ so } E(\Delta B_k) \geqslant E(\Delta\mathbf{n}_k, \boldsymbol{\mu}_k)$$

We expand $\mathbf{y}_k$ and $\boldsymbol{\mu}_k$ into tangential and normal components with respect to $\mathbf{a}_k$, and use the upper index n to denote the normal component. Then we have

$$E(\Delta\mathbf{n}_k) \sim E(f(\mathbf{y}_k; \mathbf{a}_k)y_k)/|\mathbf{a}_k| = \boldsymbol{\mu}_k^n/|\mathbf{a}_k|$$

Then $E(\Delta B_k) \sim (\boldsymbol{\mu}_k^n, \boldsymbol{\mu}_k^n)/|\mathbf{a}_k| = |\boldsymbol{\mu}_k|^2 \sin^2 \alpha_k/|\mathbf{a}_k|$, where α_k is the angle between $\boldsymbol{\mu}_k$ and $\mathbf{a}_k$. The series $\sum_k 1/|\mathbf{a}_k|$diverges as $\sum_k 1/k$. So the condition ($\alpha_k \to 0$ with probability 1) is necessary for $E(B_k) = \sum_k E(\Delta B_k)$ to be bounded. But for sufficiently small α_k the vector $\boldsymbol{\mu}_k$ will lie inside the cone of weights of f_k, which is therefore stable. □

Note that the statistical independence of channels was not used in the proof. So the assertion of Theorem 9 is valid regardless dependence or independence.

Theorem 15. *If* $p_i \geqslant 1/2 + c$, $c > 0$, $i = 1, \ldots, n$, *then moments of all stable functions in a positive orthant are of the form* $(2P - 1)\mathbf{a}_\beta + \mathbf{e}$, *where* $|\mathbf{e}| < 4\sqrt{n}\exp(-2nc^4)$ *and P is the probability of correct decision.*

Proof. By simple direct calculations [6] we can easily obtain the expression

$$\boldsymbol{\mu} = (2P - 1)\mathbf{a}_\beta + \mathbf{e} \text{ with}$$

$$e_i = 4p_i\left(\Pr\left\{\sum_{1 \leqslant j \leqslant n, j \neq i} a_j y_j > -a_i | z = 1\right\} - \Pr\left\{\sum_{1 \leqslant j \leqslant n} a_j y_j > 0 | z = 1\right\}\right)$$

$0 \leqslant e_i \leqslant 4p_i(1-P) \leqslant 4(1-P)$, so $|\mathbf{e}| \leqslant 4\sqrt{n}(1-P)$ and as follows from Theorem 9:

$$|\mathbf{e}| \leqslant 4\sqrt{n}\exp\{-(\mathbf{a}_\beta, \mathbf{a})^2/2(\mathbf{a}, \mathbf{a})\}. \tag{12}$$

Let $\mathbf{a}^* = t_1\mathbf{a}$, $\boldsymbol{\mu}^* = t_1\boldsymbol{\mu}$ such that $\sum_{i=1}^n a_i^* = 1$, $\sum_{i=1}^n \mu_i^* = 1$. In [6] it was shown by the authors that $P \geqslant \min\{p_1, p_2, \dots, p_n\}$. Using that inequality we can obtain the estimate $4c^2 \leqslant \mu_i \leqslant 1$ and hence $\mu_i^* \leqslant 1/4c^2$.

Let $\max\{a_i^*\} > 1/nc^2$. Then for $\mathbf{d}$ with $d_i = \operatorname{sgn}(a_i^* - \boldsymbol{\mu}_i^*)$ we can obtain the inequality $(\mathbf{a}^*, \mathbf{d}) - (\boldsymbol{\mu}^*, \mathbf{d}) \geqslant 3/4nc^2$. Then sequentially inverting the coordinates of $\mathbf{d}$ (except for the maximal one) we can find $\mathbf{d}_0$ for which $(\mathbf{a}^*, \mathbf{d}_0)$ and $(\boldsymbol{\mu}^*, \mathbf{d}_0)$ have different signs and hence $f(\mathbf{y}; \mathbf{a})$ should not be stable.

Thus $\max_i\{a_i^*\} \leqslant 1/nc^2$, and $\max\{(\mathbf{a}^*, \mathbf{a}^*)\} = 1/nc^2$ is reached when $[nc^2]$ coordinates of $\mathbf{a}^*$ are equal to $1/nc^2$. Then we have $(\mathbf{a}_\beta, \mathbf{a}^*) \geqslant 2c\sum_{i=1}^n a_i^* = 2c$. Using these estimates in Eq. (12) we get the necessary inequality. □

This result may be interpreted as an asymptotic collinearity of weight vectors of stable functions to that of a baricentric function.

Theorem 16. *If $p_i > 1/2$, $i = 1, \dots, n$, for every $\varepsilon > 0$ there is such a constant $C(\varepsilon) > 0$, that for every initial weight vector $\mathbf{a}$ belonging to the positive orthant and such that $|\mathbf{a}| > C(\varepsilon)$, all the weight vectors, generated by the accelerated perceptron in case of self-learning will belong to the positive orthant with probability $(1-\varepsilon)$.*

Proof. The theorem follows directly from a strong law of large numbers. □

On the qualitative level the two last theorems indicate that for a large number of rather good channels the self-learning or the accelerated perceptron should lead to the improvement of reliability.

6. Computer simulation of accelerated perceptron

The following numerical model of a multichannel system discussed above was programmed for a PC and used for numeric experiments.

A random symbol z is generated. An n-dimensional random noise vector $\boldsymbol{\xi}$ is then generated and the vector $\boldsymbol{\zeta}$ of the noisy symbols in channels is determined: $\zeta_i = z + \xi_i$. The components of the vector $\mathbf{y}$ of the output symbols of the channels are obtained by threshold detection of the components of the vector $\boldsymbol{\zeta}$: $y_i = \operatorname{sign}(\zeta_i)$, $i = 1, 2, \dots, n$.

The noise $\boldsymbol{\xi}$ is an n-dimensional Markov Gaussian stochastic process, modeled as in [21]. The recommendations made by Knuth [22] were followed when generating the random numbers.

The model enables to change the amplitude of the noise in the channels and the parameters of the correction matrix function of the process. So we can control probabilities of errors in the channels, make channels independent or correlated, and change time correlation of errors.

The output of the model is the estimate for system error probability, defined by the formula

$$P_{\text{err}} = \Pr\{z = 1\} \sum_{\mathbf{y} \in \{-1,1\}^n, (\mathbf{a}, \mathbf{y}) < 0} \Pr\{\mathbf{y} | z = 1\} + \Pr\{z = -1\}$$

$$\sum_{\mathbf{y} \in \{-1,1\}^n, (\mathbf{a}, \mathbf{y}) > 0} \Pr\{\mathbf{y} | z = -1\} \qquad (13)$$

where **a** is the weight vector of the function. The conditional probabilities $\Pr\{\mathbf{y}|z = 1)$ and $\Pr\{\mathbf{y}|z = -1)$ were determined by the Monte Carlo method before the model experiments.

Two blocks of results obtained for the case of independent channels with time uncorrelated errors are given in Figs 1 and 2. Each figure shows three graphs: for the classical perceptron (a), the accelerated perceptron (b) and the self-learning accelerated perceptron (c). The count of symbols generated is plotted along the abscissa axis, and the current error probability plotted along the ordinate axis. In all cases, modeling starts with equal weights in all channels, that is from a majority function.

Figure 1 shows the results for a 5-channel system. The probabilities of no error in the channel are $p_1 = 0.924$, $p_2 = p_3 - 0.818$, $p_4 = p_5 = 0.622$, the Bayes error level is 0.054. The optimum Bayes decision is $f(3y_1 + 2y_2 + 2y_3 + y_4 + y_5)$, which, in this case is identical to the baricentric function.

Figure 2 gives the results for a 9-channel system. The probabilities of no error in the channel are $p_1 = 0.924$, $p_2 = p_3 = 0.818$, $p_4 = \cdots = p_9 = 0.55$. The optimum Bayes error level is 0.0735.

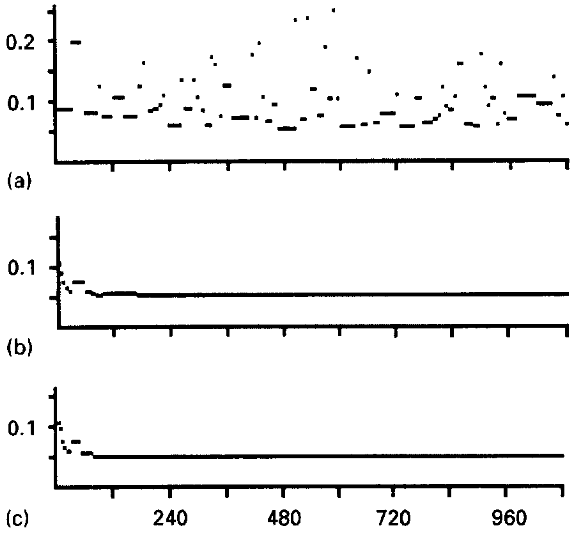

Fig. 1.

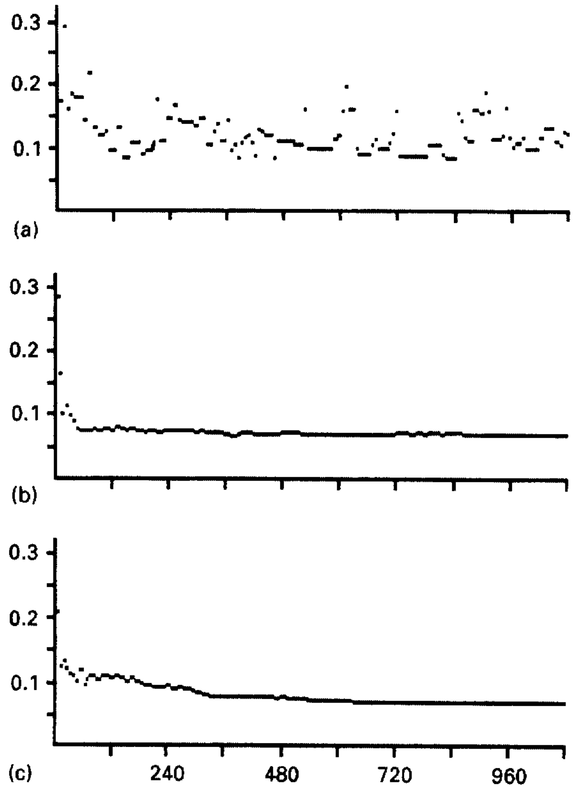

Fig. 2.

In both cases, the quality of the decision made by the classical perceptron was very variable, the system error probability ranges from the Bayes level to three or four times of that.

The accelerated perceptron both in learning and self-learning modes converges, in the first case, to the optimal decision rule and, in second case, to a decision rule of quality close to that of the optimal decision. Note, while self-learning perceptron operates in unsteady conditions, with parameters of the channels varying over the time, the weight vector must be periodically normalized to track the variations.

7. Conclusions

Choosing the voting as a decision rule we stay on a way which inevitably leads us to the great questions: What kind of voting to use? What is the risk to use it? What gain of decision reliably can be obtained? To some extent our results can help to clarify the situation. Both axiomatic characterization of the majority rule (Theorem 1) and characterizations of it as a saddle point of weighted voting (Theorems 2 and 4) and as man's optimal policy in the antagonistic game with nature (Theorem 11) make us recommend its use when there is a lack of sure information about the voters. This is the case for most of humanities. In the fine but rare case of surely known estimates for reliabilities of voters we should use optimal weighted voting. If estimates for reliabilities are unknown but the hypothesis of statistical independence of voters can be accepted as adequate, weighted voting can be successfully used along with the accelerated perceptron to adapt the weights. As shown in Section 5, the process of adaptation converges to some stable threshold function, asymptotically close to the suboptimal baricentric function. Computer simulation also demonstrates a good convergence of the accelerated perceptron. Our estimates for error probability, obtained for both optimal and not necessarily optimal voting, cover all the above cases.

Acknowledgements

Financial support for this research was provided by RFFR Grant 97-01-00024 and INTAS Grant 96-106.

References

[1] De Condorcet, Essai sur l'Application de l'Analyse à la Probabilité des Désisions Rendues à la Pluralité des Vox, Paris, 1785.

[2] H. Poincaré, Science et méthode, Paris, Ernest Flammarion, 1908.

[3] J. Von Neumann, Probabilistic logic and the synthesis of reliable organisms from unreliable components, in: Automata Studies, Princeton University Press, Princeton, NJ, 1956.

[4] W.J. Pierce, Failure-Tolerant Computer Design, Academic Press, New York, 1965.

[5] Yu.A. Zuev, A probabilistic model of a committee of classifiers, Zh. Vychisl. Mat. Mat. Fiz., 26 (1986) 276–292.

[6] Yu.A. Zuev, S.K. Ivanov, Learning and self-learning in weighted voting procedures, Zh. Vychisl. Mat. Mat. Fiz. 35(1) (1995) 104–121.

[7] Yu.A. Zuev, S.K. Ivanov, Weighted voting in multichannel systems of digital signals transmission, Problemy peredachi informacii 31(4) (1995) 22–36.

[8] Yu.A. Zuev, On voting procedure efficiency estimation, Theory Probab. Appl. 42 (1) (1997).

[9] W. Hoeffding, On the distribution of the number of successes in independent trials, Ann. Math. Statist. 27(3) (1956) 713–721.

[10] L.J. Gleser, On the distribution of the number of successes in independent trials. Ann. Probab. 3 (1) (1975) 182–188.

[11] A.W. Marshall, I. Olkin, Inequalities: Theory of Majorization and its Application, Academic Press, New York, 1979.

[12] S. Muruga, Threshold logic and its Applications, Wiley, New York, 1971.

[13] Yu.A. Zuev, Threshold functions and threshold representations of Boolean functions, in: Mathematical Problems of Cybernetics, No. 5, Nauka, Moscow, 1994, pp. 5–61.

[14] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[15] M. Minsky, S. Papert, Perceptrons, MIT Press, Cambridge, MA, 1969.

[16] N.J. Nilsson, Learning Machines, McGraw-Hill, New York, 1965.

[17] E. Sperner, Ein Satz über Untermengen einer endlichen Menge, Math. Z. 27 (1928) 544–548.

[18] W. Hoeffding, Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. 58(301) (1963) 13–30.

[19] T. Yanagimoto, M. Okamoto, Partial Orderings of permutations and monotonicity of a rank correlation statistic, Ann. Inst. Statist. Math. 21 (3) (1969) 489–506.

[20] F. Rosenblatt, Principles of Neurodynamics: Perceptron and the Theory of Brain Mechanisms, Spartan Books, Washington, 1962.

[21] S.M. Yermakov, G.A. Mikhailov, Statistical Modeling, Nauka, Moscow, 1982.

[22] D.E. Knuth, The Art of Computer Programming, vol. 2, Seminumerical Algorithms, Addison-Wesley, Reading, MA, 1969.