Model Selection and Assessment for Classification Using Validation

Wojciech Jaworski

Faculty of Mathematics, Informatics and Mechanics,
Warsaw University, Banacha 2, 02-097 Warsaw, Poland
wjaworski@mimuw.edu.pl

Abstract. We address the problem of determination of the size of the test set which can can guarantee statistically significant results in classifier error estimation and in selection of the best classifier from a given set. We focus on the case of the 0-1 valued loss function and we provide one and two sides optimal bounds for Validation (known also as Hold-Out Estimate and Train-and-Test Method). We also calculate the smallest sample size, necessary for obtaining the bound for given estimation accuracy and reliability of estimation, and we present the results in tables. Finally, we propose strategies for classifier design using the bounds derived.

Keywords: Computational learning theory, Model Selection, Model Assessment, Hold-Out Estimate, Train-and-Test, Validation.

1 Introduction

The ability to act properly in a partially unknown environment is one of the most important properties of an intelligent system. In the case of classification, this 'proper act' is a generalization ability — an ability to classify new samples correctly.

In a classifier design cycle, there are two aspects which concern the classifier behaviour on new samples: Model Selection and Model Assessment. During the Model Selection process, we try to choose the best classifier from a given set. For example, in a rough set theory, this phrase refers to choosing the minimal support for decision rules. During the Model Assessment, we estimate the generalization ability of the classifier.

There are several methods for performing Model Selection and Assessment. However, we restrict ourselves to the analysis of Validation (also known as Train-and-Test Method or Hold-Out Estimate). The reason is that the quality of Validation estimation is independent from the classifying algorithm. Hence, an efficient universal bound can be obtained.

We derive optimal bounds in a probabilistic model of a learning process, based on independence of samples. In this model, we restrict ourselves to the case of the 0-1 valued loss function. Since the 0-1 valued loss function is the one used most often in pattern recognition, this case has multiple applications.

D. Ślęzak et al. (Eds.): RSFDGrC 2005, LNAI 3641, pp. 481–490, 2005.

Using the bounds, the smallest number of samples, needed for performing of the model selection and assessment with statistically significant results, can be determined. The 'optimality' of bound assures that the size of a testing sample, assessed by it, is necessary and it cannot be decreased.

We describe the model and give a formal definition of Validation in Sect. 2. In Sect. 3 we present results concerning the classifier error estimation using Validation. We also provide the tables, where the smallest sample size necessary for obtaining the bound for given estimation accuracy and reliability of estimation is calculated. We discuss the bound in the case of testing many classifiers with the same sample. In Sect. 4, we present the Model Selection and Assessment strategies based on the bounds.

2 The Problem of Learning from the Statistical Point of View

In this section, the fundamental concepts of the learning theory are introduced.

Let X be the **set of examples (attribute value vectors)**, Y be the **set of decisions (labels)**, and ρ be a Borel probability measure on $Z = X \times Y$. ρ plays an important role in sampling as it describes the probability of getting a given sample as well as distribution of decision for any example. Unfortunately, ρ is unknown to us.

We are given a finite sequence $\mathbf{z} = \big((x_1, y_1), \ldots, (x_m, y_m)\big)$, where x_i is an example and y_i – a decision for $i = 1, \ldots, m$. The sequence $\mathbf{z}$ will be called a **sample** of the length m; $\mathbf{z}$ is randomly got by m independent draws according to the probability measure ρ; $\mathbf{z}$ describes all our knowledge about ρ.

An algorithm $A_m : Z^m \to (X \to Y)$ is also such that for each sample $\mathbf{z}$ of the length m, A_m yields a **classifier** (i.e., a function) $f_{\mathbf{z}} : X \to Y$.

Having a classifier, we want to evaluate its quality. The quality of a classifier f is determined by its **generalization error** defined by

$$\mathcal{E}(f) = \int_Z V\big(y, f(x)\big) d\rho(x, y),$$

where $V : Y \times Y \to \mathbb{R}_+$ is called the **loss function**. For example, the loss function can be defined by:

$$V\big(y, f(x)\big) = (y - f(x))^2,$$

$$V\big(y, f(x)\big) = |y - f(x)|,$$

or

$$V\big(y, f(x)\big) = \begin{cases} 0 & \text{if } y = f(x) \\ 1 & \text{if } y \neq f(x). \end{cases}$$

For a finite set of decisions $Y = \{d_1, d_2, \ldots, d_l\}$, the last case may be generalized to

$$V\big(d_i, d_j\big) = a_{i,j}$$

where $a_{i,i} = 0$ and $0 \leq a_{i,j} \leq 1$. Such a loss function allows us to express the fact that we prefer one type of the classifier error to another. In this paper, we concern only with the 0-1 valued loss function, i.e., we assume that

$$V : Y \times Y \to \{0, 1\}.$$

We want to estimate $\mathcal{E}(f_{\mathbf{z}})$, which cannot be calculated directly. To this end, we use the generalization error evaluators such as Validation.

The idea of **Validation** is to divide a given sample $\mathbf{z}$ into two distinct parts $\mathbf{z}_1, \mathbf{z}_2$. The sample $\mathbf{z}_1$ will be used to learn the classifier and the sample $\mathbf{z}_2 = \big((x'_1, y'_1), \ldots, (x'_{m'}, y'_{m'})\big)$ to test it by calculation of

$$\mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1}) = \frac{1}{m'} \sum_{i=1}^{m'} V\big(y'_i, f_{\mathbf{z}_1}(x'_i)\big).$$

$\mathcal{E}_{\mathbf{z}}(f)$ is called the **empirical error** of the function f on the sample $\mathbf{z}$. Having calculated $\mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1})$, we claim that its value is similar to the value of the **generalization error** of $f_{\mathbf{z}_1}$.

$$\mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1}) \sim \mathcal{E}(f_{\mathbf{z}_1})$$

In the next sections, we will try to express this similarity by numeric means.

3 Bounds for Classifier Error Estimation

The simplest way to obtain the quality of estimation is to assess

$$|\mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1}) - \mathcal{E}(f_{\mathbf{z}_1})|$$

or at least

$$\mathcal{E}(f_{\mathbf{z}_1}) - \mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1})$$

if we are interested only in how bad the estimation can be.

According to [8], we may use the following inequalities:

Theorem 1. *Let m denote the size of $\mathbf{z}_2$, and let $\varepsilon > 0$. If $V\big(f_{\mathbf{z}_1}(x), y\big) \in \{0, 1\}$, then the least δ such that*

$$\mathrm{P}\big(\mathcal{E}(f_{\mathbf{z}_1}) - \mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1}) > \varepsilon\big) < \delta \tag{1}$$

has the value

$$\delta = \max_{k<(1-\varepsilon)m} \sum_{i=0}^{k} \binom{m}{i} \left(\varepsilon + \frac{k}{m}\right)^i \left(1 - \varepsilon - \frac{k}{m}\right)^{m-i}.$$

The behaviour of the bound is shown in Table 1.

Table 1. Number of samples needed for inequality (1) to hold for given ε and δ

$\varepsilon\backslash\delta$	0.1000	0.0500	0.0200	0.0100	0.0050	0.0020	0.0010	0.0005	0.0002	0.0001
0.005	16624	27255	42379	54319	66549	83038	95695	108475	125522	138510
0.010	4206	6864	10645	13630	16687	20809	23974	27169	31430	34677
0.015	1891	3073	4753	6080	7439	9271	10677	12097	13991	15434
0.020	1076	1741	2686	3432	4197	5227	6018	6817	7882	8694
0.025	697	1122	1727	2205	2694	3353	3860	4371	5053	5572
0.030	489	785	1205	1537	1876	2334	2686	3041	3514	3875
0.035	364	581	889	1133	1383	1719	1977	2238	2586	2851
0.040	281	448	684	871	1062	1319	1517	1717	1983	2186
0.045	225	356	543	690	841	1045	1201	1359	1569	1729
0.050	184	291	442	561	683	848	975	1103	1273	1403
0.055	154	242	367	465	566	703	807	913	1054	1161
0.060	131	205	310	392	477	592	680	768	887	977
0.065	112	175	265	336	408	505	580	656	757	833
0.070	98	152	230	290	353	437	501	566	653	720
0.075	86	134	201	254	308	381	438	494	570	628
0.080	77	118	177	224	272	336	385	435	502	552
0.085	69	105	158	199	241	298	342	386	445	490
0.090	62	95	141	178	216	267	306	345	398	438
0.095	56	86	127	160	194	240	275	310	357	393
0.100	51	78	115	145	176	217	249	280	323	355
0.105	47	71	105	132	160	197	226	255	293	323
0.110	43	65	96	121	146	180	206	233	268	294
0.115	40	60	88	111	134	165	189	213	245	270
0.120	37	55	82	102	123	152	174	196	226	248
0.125	34	51	76	95	114	140	161	181	208	229
0.130	32	48	70	88	106	130	149	168	193	212
0.135	30	45	65	82	98	121	138	156	179	197
0.140	28	42	61	76	92	113	129	145	167	183
0.145	26	39	57	71	86	105	120	135	156	171
0.150	25	37	54	67	80	99	113	127	146	160
0.155	24	35	50	63	75	93	106	119	137	150
0.160	22	33	48	59	71	87	99	112	128	141
0.165	21	31	45	56	67	82	94	105	121	133
0.170	20	29	42	53	63	77	88	99	114	125
0.175	19	28	40	50	60	73	84	94	108	118
0.180	18	27	38	47	57	69	79	89	102	112
0.185	17	25	36	45	54	66	75	84	97	106
0.190	17	24	35	43	51	63	71	80	92	101
0.195	16	23	33	41	49	60	68	76	87	96
0.200	15	22	31	39	46	57	65	72	83	91

Theorem 2. *Let $\varepsilon > 0$ and m be such that $\frac{1}{4\varepsilon^2} + 1 \leq m$. The least δ such that*

$$\mathrm{P}\big(|\mathcal{E}(f_{\mathbf{z}_1}) - \mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1})| > \varepsilon\big) < \delta \tag{2}$$

satisfies

$$\delta = \max_{0 \le k < m(1-\varepsilon)} \sum_{i=0}^{k} \binom{m}{i} \left(\varepsilon + \frac{k}{m}\right)^i \left(1 - \varepsilon - \frac{k}{m}\right)^{m-i} +$$

$$+ \sum_{i=k+\lfloor 2m\varepsilon \rfloor + 1}^{m} \binom{m}{i} \left(\varepsilon + \frac{k}{m}\right)^i \left(1 - \varepsilon - \frac{k}{m}\right)^{m-i}.$$

The behaviour of the bound is shown in Table 2 As we can see, the necessary number of samples for the two side bound is only slightly greater that the number of samples for the one side bound for small δ.

Observe that Theorem 1 provides us with the optimal $\delta_{m,\varepsilon}$ for the following inequality:

$$\mathrm{P}(\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f) > \varepsilon) \le \delta_{m,\varepsilon}.$$

Now, we will look for the optimal bound, considering inequalities of the form

$$\mathrm{P}(\mathcal{E}(f) > g(\mathcal{E}_{\mathbf{z}}(f))) \le \delta,$$

where $g : \{\frac{0}{m}, \frac{1}{m}, \ldots, \frac{m}{m}\} \to [0,1]$ is monotonically increasing and $g(1) = 1$. Let

$$\mathcal{G}_{m,\delta} = \{g : \{\frac{1}{m}, \frac{2}{m}, \ldots, \frac{m}{m}\} \to [0,1] \mid \mathrm{P}(\mathcal{E}(f) > g(\mathcal{E}_{\mathbf{z}}(f))) \le \delta \wedge$$

$$\wedge \bigwedge_{x,y} \; x < y \Rightarrow g(x) \le g(y) \wedge g(1) = 1\}.$$

In order to compare the quality of inequalities, we introduce a partial order on $\mathcal{G}_{m,\delta}$. For any $g_1, g_2 \in \mathcal{G}_{m,\delta}$, let

$$g_1 \preceq g_2 \text{ iff } \bigwedge_x \; g_1(x) \le g_2(x).$$

$g_1 \preceq g_2$ means that the bound estimated using g_1 is better than the one estimated by g_2. The optimal bound is the one corresponding to the $\preceq$-least element.

Definition 1. *Let $k < m$ and $g_{m,\delta}(\frac{k}{m}) = p$ be such that*

$$\sum_{i=0}^{k} \binom{m}{i} p^i (1-p)^{m-i} = \delta$$

and $g_{m,\delta}(1) = 1$.

Since $\sum_{i=0}^{k} \binom{m}{i} p^i (1-p)^{m-i}$ is strictly monotonically decreasing with growing p, $g_{m,\delta}$ is well-defined.

Theorem 3. *Let $0 < \delta < 1$, $m \in \mathbb{N}$. $g_{m,\delta}$ is the $\preceq$-least element of $\mathcal{G}_{m,\delta}$, i.e.,*

$$\mathrm{P}(\mathcal{E}(f) > g_{m,\delta}(\mathcal{E}_{\mathbf{z}}(f))) < \delta$$

is the optimal bound.

Table 2. Number of samples needed for inequality (2) to hold for given ε and δ

$\varepsilon\backslash\delta$	0.1000	0.0500	0.0200	0.0100	0.0050	0.0020	0.0010	0.0005	0.0002	0.0001
0.005	27100	38500	54200	66400	78800	95500	108300	121200	138400	151400
0.010	6800	9650	13550	16600	19700	23900	27100	30300	34600	37850
0.015	3034	4300	6034	7400	8767	10634	12034	13467	15400	16834
0.020	1700	2425	3400	4150	4925	5975	6775	7575	8650	9475
0.025	1100	1540	2180	2660	3160	3820	4340	4860	5540	6060
0.030	767	1084	1517	1850	2200	2667	3017	3367	3850	4217
0.035	558	786	1115	1358	1615	1958	2215	2486	2829	3100
0.040	425	613	850	1038	1238	1500	1700	1900	2163	2375
0.045	345	478	678	823	978	1189	1345	1500	1712	1878
0.050	280	390	550	670	790	960	1090	1220	1390	1520
0.055	228	328	455	555	655	791	900	1010	1146	1255
0.060	192	275	384	467	550	667	759	842	967	1050
0.065	162	231	324	400	470	570	647	724	824	900
0.070	143	200	279	343	408	493	558	622	708	772
0.075	127	174	247	300	354	427	487	540	620	674
0.080	113	157	213	263	313	375	425	475	544	594
0.085	100	136	189	236	277	336	377	424	483	524
0.090	89	123	173	206	245	300	339	378	428	467
0.095	79	111	153	190	222	269	300	337	385	422
0.100	70	100	140	170	200	240	275	305	345	380
0.105	67	91	124	153	181	220	248	277	315	343
0.110	60	82	114	141	164	200	228	250	287	314
0.115	57	74	105	127	153	183	209	231	261	287
0.120	50	71	96	117	138	167	192	213	242	263
0.125	44	64	88	108	128	156	176	196	224	244
0.130	43	58	81	100	120	143	162	181	204	227
0.135	41	56	78	93	112	134	152	167	189	208
0.140	36	50	72	86	104	125	140	158	179	193
0.145	35	49	66	80	97	114	132	145	166	180
0.150	34	47	64	77	90	107	120	137	154	170
0.155	33	42	59	71	84	100	113	126	146	159
0.160	29	41	57	66	79	94	107	119	135	147
0.165	28	37	52	64	73	88	100	113	128	140
0.170	27	36	50	59	71	86	95	106	121	133
0.175	23	35	46	58	66	80	89	100	115	123
0.180	23	31	45	53	62	75	84	95	109	117
0.185	22	30	41	52	60	71	82	90	103	111
0.190	22	29	40	48	56	69	77	85	98	106
0.195	21	29	36	47	54	65	72	80	93	100
0.200	20	25	35	43	50	60	68	75	88	95

Proof. First, we prove that $g_{m,\delta} \in \mathcal{G}_{m,\delta}$. It is obvious that $g_{m,\delta}(\frac{k}{m}) \leq g_{m,\delta}(\frac{k+1}{m})$. Let $g_{m,\delta}(\frac{-1}{m}) = 0$. We show that the inequality holds.

$$\mathrm{P}(\mathcal{E}(f) > g(\mathcal{E}_{\mathbf{z}}(f))) = \sum_{i=0}^{m} \mathrm{P}(\mathcal{E}_{\mathbf{z}}(f) = \frac{i}{m})\mathrm{P}(\mathcal{E}(f) > g(\mathcal{E}_{\mathbf{z}}(f))|\mathcal{E}_{\mathbf{z}}(f) = \frac{i}{m}) =$$

$$= \sum_{i=0}^{m} \binom{m}{i} \mathcal{E}(f)^i (1-\mathcal{E}(f))^{m-i} \mathrm{P}(\mathcal{E}(f) > g_{m,\delta}(\frac{i}{m})) \leq$$

$$\leq \max_{k\in\{-1,0,\ldots,m-1\}} \sup_{p\in(g_{m,\delta}(\frac{k}{m}),g_{m,\delta}(\frac{k+1}{m})]} \sum_{i=0}^{m} \binom{m}{i} p^i (1-p)^{m-i} \mathrm{P}(p > g_{m,\delta}(\frac{i}{m})) =$$

$$= \max_{k\in\{-1,0,\ldots,m-1\}} \sup_{p\in(g_{m,\delta}(\frac{k}{m}),g_{m,\delta}(\frac{k+1}{m})]} \sum_{i=0}^{k} \binom{m}{i} p^i (1-p)^{m-i} < \delta$$

Now, we show that $g_{m,\delta}$ is the smallest in $\mathcal{G}_{m,\delta}$. Let $g \in \mathcal{G}_{m,\delta}$. Assume that $\mathcal{E}(f) = g(\frac{k}{m}) + \varepsilon$. Then,

$$\lim_{\varepsilon\to 0^+} \mathrm{P}(\mathcal{E}(f) > g(\mathcal{E}_{\mathbf{z}}(f))) = \lim_{\varepsilon\to 0^+} \sum_{i=0}^{k} \binom{m}{i} \mathcal{E}(f)^i (1-\mathcal{E}(f))^{m-i} =$$

$$= \sum_{i=0}^{k} \binom{m}{i} g(\frac{k}{m})^i (1-g(\frac{k}{m}))^{m-i}.$$

Since $g \in \mathcal{G}_{m,\delta}$,

$$\sum_{i=0}^{k} \binom{m}{i} g(\frac{k}{m})^i (1-g(\frac{k}{m}))^{m-i} \leq \delta.$$

Thus, from monotonicity of $\sum_{i=0}^{k} \binom{m}{i} p^i (1-p)^{m-i}$,

$$g(\frac{k}{m}) \geq g_{m,\delta}(\frac{k}{m}).$$

Using Theorem 3, we derive an efficient algorithm for that approximation of $g_{m,\delta}$ for given m and δ. Let k_p be the largest k such that

$$\sum_{i=0}^{k} \binom{m}{i} p^i (1-p)^{m-i} \leq \delta$$

and $n \in \mathbb{N}$. We calculate $k_{\frac{j}{n}}$ for $j \in \{0, 1, \ldots, n\}$. Values $g_{m,\delta}$ satisfy following inequality:

$$\min\{\frac{j}{n} : k_{\frac{j}{n}} \geq k\} - \frac{1}{n} < g_{m,\delta}\big(\frac{k}{m}\big) \leq \min\{\frac{j}{n} : k_{\frac{j}{n}} \geq k\}.$$

Function $g(\frac{k}{m}) = \min\{\frac{j}{n} : k_{\frac{j}{n}} \geq k\}$ generates a bound that is worse than the best one less than $\frac{1}{n}$.

Table 3 illustrates the behaviour of the bound.

Remark 1. If we consider function $g(\frac{k}{m}) = g_{m,\delta}(\frac{k-1}{m})$, where $p_{-1} = -1$, than we will obtain the inequality

$$\mathrm{P}(\mathcal{E}(f) > g(\mathcal{E}_{\mathbf{z}}(f))) \geq \delta.$$

As $\sum_{i=0}^{m} p_k - p_{k-1} = 1$, the average distance between lower and upper bounds is $\frac{1}{m}$.

Table 3. Values of $g_{m,\delta}(\frac{k}{m})$ for a chosen values of k and $m = 1000$ (In sup row are the maximum values)

$\frac{k}{m}\backslash\delta$	0.1000	0.0500	0.0200	0.0100	0.0050	0.0020	0.0010	0.0005	0.0002	0.0001
0.01	0.0054	0.0070	0.0088	0.0101	0.0113	0.0129	0.0140	0.0151	0.0165	0.0176
0.02	0.0070	0.0090	0.0113	0.0129	0.0145	0.0164	0.0177	0.0191	0.0208	0.0220
0.05	0.0101	0.0129	0.0162	0.0185	0.0206	0.0231	0.0250	0.0268	0.0290	0.0307
0.10	0.0133	0.0170	0.0213	0.0242	0.0269	0.0302	0.0326	0.0348	0.0376	0.0397
0.15	0.0155	0.0199	0.0249	0.0282	0.0313	0.0351	0.0378	0.0403	0.0435	0.0458
0.20	0.0172	0.0220	0.0275	0.0311	0.0345	0.0387	0.0416	0.0444	0.0479	0.0504
sup	0.0208	0.0266	0.0330	0.0373	0.0413	0.0460	0.0494	0.0526	0.0565	0.0593

We construct a two sides bound, combining the one side ones:

Theorem 4. *Let* $0 < \delta < 1$ *and* $m \in \mathbb{N}$.

$$\mathrm{P}(\mathcal{E}(f) > g_{m,\delta}(\mathcal{E}_{\mathbf{z}}(f)) \cup \mathcal{E}(f) < 1 - g_{m,\delta}(1 - \mathcal{E}_{\mathbf{z}}(f))) < 2\delta.$$

As Remark 1 is valid for the two sides inequality, we see that it is quite strict.

Now, we deal with another important question: What does it happen, when one uses the same test sample for testing many classifiers?

Assume that we have k classifiers $f_1, \ldots, f_k$ and we want to estimate probability that $\mathcal{E}(f_i) \in G(\mathcal{E}_{\mathbf{z}}(f_i))$ for each of them, i.e.,

$$\mathrm{P}(\mathcal{E}(f_1) \in G(\mathcal{E}_{\mathbf{z}}(f_1)) \wedge \cdots \wedge \mathcal{E}(f_k) \in G(\mathcal{E}_{\mathbf{z}}(f_k))).$$

The trivial bound uses the fact that $\mathrm{P}(A \vee B) \leq \mathrm{P}(A) + \mathrm{P}(B)$ for any random events A and B:

$$\mathrm{P}(\mathcal{E}(f_1) \in G(\mathcal{E}_{\mathbf{z}}(f_1)) \wedge \cdots \wedge \mathcal{E}(f_k) \in G(\mathcal{E}_{\mathbf{z}}(f_k))) \geq$$

$$1 - k + \sum_{i=0}^{k} \mathrm{P}(\mathcal{E}(f_i) \in G(\mathcal{E}_{\mathbf{z}}(f_i))) \tag{3}$$

On the other hand, if $\mathcal{E}_{\mathbf{z}}(f_1), \ldots, \mathcal{E}_{\mathbf{z}}(f_k)$ are independent, then

$$\mathrm{P}(\mathcal{E}(f_1) \in G(\mathcal{E}_{\mathbf{z}}(f_1)) \wedge \cdots \wedge \mathcal{E}(f_k) \in G(\mathcal{E}_{\mathbf{z}}(f_k))) = \prod_{i=0}^{k} \mathrm{P}(\mathcal{E}(f_i) \in G(\mathcal{E}_{\mathbf{z}}(f_i))).$$

Note that unseemingly, the independence of $\mathcal{E}_{\mathbf{z}}(f_1), \ldots, \mathcal{E}_{\mathbf{z}}(f_k)$ is possible when $\mathcal{E}(f_i)$ is small, whereas the classifiers $f_1, \ldots, f_k$ are similar.

If we assume $\mathrm{P}(\mathcal{E}(f_i) \in G(\mathcal{E}_{\mathbf{z}}(f_i))) = 1 - \delta$, we can easily calculate the difference between the trivial bound and the case of independence.

$$\prod_{i=0}^{k} \mathrm{P}(\mathcal{E}(f_i) \in G(\mathcal{E}_{\mathbf{z}}(f_i))) - 1 + k - \sum_{i=0}^{k} \mathrm{P}(\mathcal{E}(f_i) \in G(\mathcal{E}_{\mathbf{z}}(f_i))) =$$

$$= (1 - \delta)^k - 1 + k\delta \leq \frac{1}{2}(k\delta)^2$$

As we can see, there is no big difference between the both cases, so the trivial bound is near to the optimal one in the interesting cases.

4 Model Selection and Assessment

In order to assess the model, we simply need to estimate its generalization error, using one of the bounds presented above. The procedure is the following:

- Divide the data given into the training sample and the test sample. Choose the size of the test sample, m, according to Table 1 or 2, and the total number of samples.
- Generate the classifier f using training sample.
- Test f using the test sample and the bound from Theorem 3 or 4.

To assure the bound to hold true, it has to be chosen before the testing process starts. The test may be performed only once. Any repetition, especially the one performed in order to choose the best bound, causes a rapid decrease in reliability.

While selecting a classifier from a given set, we are interested in its behaviour in comparison to the other ones. We select the classifier which has the smallest empirical error. The question is: How many samples do we need to know that the classifier which has the smallest empirical error is the one that has the smallest generalization error?

When classifiers have very similar generalization errors, they are almost indistinguishable. Fortunately, in this case, it is not really important which one we choose. It is enough to consider the differences bigger than ε.

The most straightforward way is to use Theorem 4 for every classifier from the set. Testing multiple classifiers on the same data will decrease the reliability, as shown in (3). So we will obtain the bound

$$\mathrm{P}(\mathcal{E}(f_1) \in G(\mathcal{E}_{\mathbf{z}}(f_1)) \wedge \cdots \wedge \mathcal{E}(f_k) \in G(\mathcal{E}_{\mathbf{z}}(f_k))) \geq 1 - k\delta, \tag{4}$$

where

$$G(\mathcal{E}_{\mathbf{z}}(f_i)) = [1 - g_{m,\delta}(1 - \mathcal{E}_{\mathbf{z}}(f_i)), g_{m,\delta}(\mathcal{E}_{\mathbf{z}}(f_i))].$$

If $G(\mathcal{E}_{\mathbf{z}}(f_i)) \cap G(\mathcal{E}_{\mathbf{z}}(f_j)) = \emptyset$, then we can decide which one is better with probability $\geq 1 - k\delta$. The procedure is the following:

- Divide the data given into the training sample and the validation sample. Choose the validation sample size, m, according to Table 2, the number of classifiers to be constructed and total number of samples.
- Generate classifiers $f_1, \ldots, f_k$ using the training sample.
- Select the best classifier that has the smallest empirical error on the validation sample. The relation between the generalization errors of classifiers is described by the inequality (4).

As we can see in Table 2, in order to estimate the error of 100 classifiers with the reliability 95%, one needs to have aproximately 4 times the number of samples that is needed to estimate the error of one classifier. The advantage is that all classifiers are already assessed after the selection process. We may combine the model selection and the model assessment and we may use the same sample for both of them. As a consequence, the sample is bigger and the bound is tighter.

Acknowledgment. The research has been supported by the grant 3 T11C 002 26 from Ministry of Scientific Research and Information Technology of the Republic of Poland.

References

1. F. Cucker and S. Smale, *On the mathematical foundations of learning*, Bulletin of AMS, 39:1-49, 2001.
2. R. Duda, P. Hart, D. Stock, *Pattern Classification*, John Wiley & Sons, Inc. 2001
3. J. H. Friedman, T. Hastie, R. Tibshirani, *Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, Heidelberg, 2001
4. K. Fukunaga, R. R. Hayes, *Effects of Sample Size in Classifier Design* IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(8):873-885, 1989
5. K. Fukunaga, R. R. Hayes, *Estimation of Classifier Performance*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(10):1087-1101, 1989
6. I. Guyon, J. Makhoul, R. Schwartz, V. Vapnik, *What size test set gives good error rate estimates?*, IEEE Pattern Analysis and Machine Intelligence, 20:52-64, 1998
7. W. Hoeffding, *Probability Inequalities for Sums of Bounded Random Variables*, JASA 58, 13-30
8. W. Jaworski, *Bounds for Validation*, Fundamenta Informaticae (to appear)
9. D. Michie, D. Spiegelhalter, C. Taylor (Eds.), *Machine Learning, Neural and Statistical Classification*, John Wiley & Sons, Inc. 2001
10. V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.