Feature Over-Selection

Sarunas Raudys

Vilnius Gediminas Technical University
Sauletekio 11, Vilnius, LT-10223, Lithuania
raudys@ktl.mii.lt

Abstract. We propose probabilistic framework for analysis of inaccuracies due to feature selection (FS) when flawed estimates of performance of feature subsets are utilized. The approach is based on analysis of random search FS procedure and postulation that joint distribution of true and estimated classification errors is known *a priori*. We derive expected values for the FS bias, a difference between actual classification error after FS and classification error if ideal FS is performed according to exact estimates. The increase in true classification error due to inaccurate FS is comparable or even exceeds a training bias, a difference between generalization and Bayes errors. We have shown that there exists overfitting phenomenon in feature selection, entitled in this paper as feature over-selection. The effects of feature over-selection could be reduced if FS would be performed on basis of positional statistics. Theoretical results are supported by experiments carried out on simulated Gaussian data, as well as on high dimensional microarray gene expression data.

1 Introduction

Well known peaking (over-fitting) phenomenon relates generalization error of pattern recognition algorithm and a number of features in finite learning sample situations: the generalization error decreases at first with an increase in feature dimensionality. Then it saturates and starts increasing afterwards. After discovery [1], this phenomenon was transferred to proper selection of the complexity of a classifier: in small training-set cases, often it is preferable to use simple structured classification rules than the complex ones, and, vice versa, in large training-set cases, complex classifiers can be used more efficiently (the scissors' effect, [2, 3], see also [4], Section 1.5]). In neural network training, this effect is known under a name of overtraining (overfitting) [5]: with an increase in the number of training iterations the generalization error decreases at first, saturates and starts increasing afterwards. Like in the problem with input feature dimensionality, here we face an increase in complexity of the classifier with a progress of learning procedure. If before training the single layer perceptron based classifier, a data mean is shifted to a centre of coordinates, one starts training from initial weight vector with zero components and training sample sizes in two pattern classes $N_2 = N_1 = N/2$, then after the first iteration performed in a batch mode, one obtains simple Euclidean distance classifier. Next, iterative training process gradually moves the perceptron to six more complex classifiers [4] (for an introduction into statistical pattern recognition, see e.g. [6]).

D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109, pp. 622 – 631, 2006.

Peaking phenomenon requires adjusting the dimensionality of input features to training sample size and the complexity of the classification algorithm. To reduce the number of features, FS procedures are utilized usually. There are four examples: a) evaluate the quality of p original features independently and select r best ones, b) forward selection, c) backward selections and d) random search where from p original features one generates *a group of m random feature subsets* composed of r features ($r < p$). Then one evaluates the quality of all m subsets and selects the best.

From point of view of a complexity, the algorithm "a" is the simplest. An answer which algorithm is more complex, "b" or "c", depends on p, r and the data. The complexity of random search feature selection algorithm is determined by number m. In spite of algorithmic simplicity, often random search is comparable in performance with more sophisticated FS algorithms. Therefore, algorithm "d" could be utilized as an undemanding model to study the complexity of FS problem.

If the features are selected incorrectly, generalization error of the classification system increases. Main factors that are affecting FS success in finite sample size situations are: 1) correctness of determination of the number of final features, r, in dependence on complexity of the classifier and training set size, 2) the accuracy of the criterion and validations sample size utilized to evaluate the quality of feature subset and 3) an excellence of the feature selection algorithm.

Determination of optimal dimensionality was considered in [1, 4, 6, 7]. Accuracy of the criteria (a bias, a variance) was considered while comparing methods to estimate the classification error [4, 6]. Comparative complexities of various FS schema have been studied in [8, 9] and references therein. Very often inaccuracies of feature quality determination were ignored. Exceptions are few, papers [10-16].

In order to separate effects of FS from that of training, *we do not study training sample size and complexity relations*. We assume there that a variety of already trained classifiers exist. Each of them is based on individual feature subset of the same dimensionality. On a basis of independent validation set one needs to select the best feature subset (classifier). We investigate both the accuracy of performance estimate (variance) and the complexity of the FS schema. We use probabilistic framework suggested in [10, 11], improve computer simulation tools, derive equations for an increase in expected classification error due inexact FS and show that with an increase in complexity of the feature subset selection schema, classification error rate exhibits peaking behavior. Theoretical and experimental analysis show that while applying random search FS schema, in order to obtain better result, one needs consider smaller amount of feature subsets and do not select apparently the best subset of features. Ng [13] gave reasons for not selecting the hypothesis with the lowest validation error. He demonstrated this by analyzing very artificial schema. Presently, we demonstrate such effect both analytically and experimentally for realistic feature selection tasks.

2 Statement of the Problem

In this section we will elucidate the factors influencing FS accuracy by considering as simple pattern recognition problem as possible. Consider two class problem with multivariate p-dimensional Gaussian classes with different means, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and sharing common covariance matrix $\boldsymbol{\Sigma}$. In this demonstrative example, p=150; only several features were "really good": $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ = [1.45 1.15 0.95 0.80 0.70 0.60 0.55 0.50 0.45

0.42 0.40 0.375 0.370 0.3679 0.3657 0.3635 …. 0.0776 0.0755]^T. All variances were equal to 1.0 and correlations between all pairs of features, ρ=0.667. The designer needs to create standard linear Fisher classifier based on a best r-dimensional feature subset (r=8<< p). Note, that in this *example with equal correlations*, a subset of eight individually best features is not the best: this subset results Bayes error P_B=0.1830, while one of randomly formed subset composed of 1, 2, 3, 54, 95, 113, 127, and 113th features gives much better, P_B=0.0983.

In our analysis we assume that there exists a variety of already trained classifiers. The classes are Gaussian. Therefore, the Bayes errors, $\Phi(-½\delta)$ are known to me (δ stands for Mahalanobis distance). The designer does not know performances of feature subsets, however, he/she has independent validation set. On a basis of his information the designer needs to select a best subset (classifier) from $C_p^r \approx 5.26\times10^{12}$ potentially possible ones. We consider in this section that the designer uses the sample based Mahalanobis distance, $\hat{\delta}$, as a measure of feature subset's quality (the classification error, $\hat{P}_{error} = \Phi(-½\,\hat{\delta})$.

In Figure 1a we present a histogram of Bayes errors obtained in M=50,000 random generations of 8-dimensional feature subsets. Such multimodal density is rather typical for many real-world pattern recognition problems where one has a small number of relatively good features. In this example we used a single 150D validation set composed of N = 10+10 vectors. Very small validation set size (N=20 vectors), is specially tailored to to-day's microarray gene expression data experiments to be discussed later. In Figure 1b we present a scatter diagram of m=5,000 2-dimensional (2D) vectors (P_{Bi}, $\hat{P}_{error\,i}$), i = 1, 2, … , m selected uniformly out of M subsets (having M=50,000 subsets, it is possible to do this in $C_{50000}^{5000} \approx 9.68\times10^{32}$ different ways).

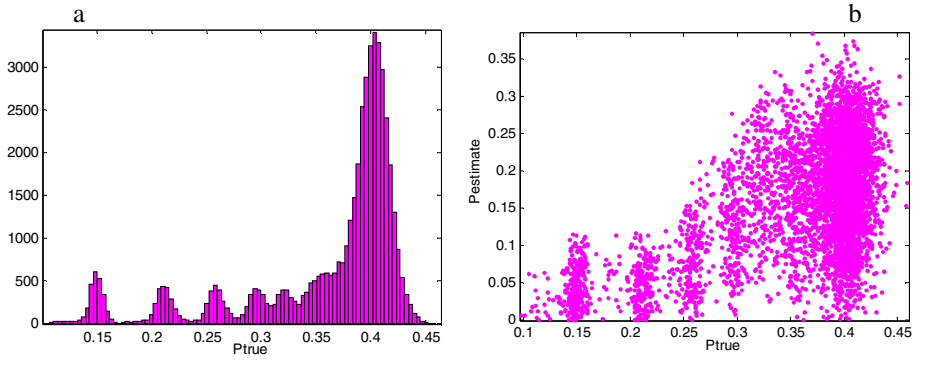

Fig. 1. a – a histogram of 50,000 values of P_B, b – a scatter diagram of vectors (P_B, $\hat{P}_{error}$)

Scatter diagram 1b shows that a great number of subsets with practically zero estimate $\hat{P}_{error}$ of classification error exist. True classification error for these subsets varies between 0.12 and 0.40. In mimicking random search FS strategy performed by

classifier designer, we formed C_{50000}^{m} virtual groups composed of m feature subsets (m = 1, 3, 6, 10, …, 10000). In each group we found a subset (say s-th subset) with smallest estimate $\hat{P}_{\text{error}\,s}$ and this subset's true error, $P_{\text{B}}s$. An average of C_{50000}^{m} values of $P_{\text{B}}s$ we call "*a mean of the true classification error after feature selection*". The average was calculated by specially combinatory algorithm developed by Pikelis (see Appendix A.4 in [4]). In Figure 2a we have a graph, True1, of dependence of the mean of the true error in feature selection on m, the group size, the number of feature subsets in the group. At the same time we calculated average of C_{50000}^{m} values of $\hat{P}_{\text{error}\,s}$ which is called "*a mean of an apparent classification error after feature selection*", graph Apparent in Figure 2a. In similar way, we can find average of "*ideal classification error after feature selection*" where we seek for subset with smallest $P_{\text{B}}g$ in the group (graph Ideal in Figure 2a). We see, a bias between True1 and Ideal is rather wide. The bias increases with an increase in m, the size of the group.

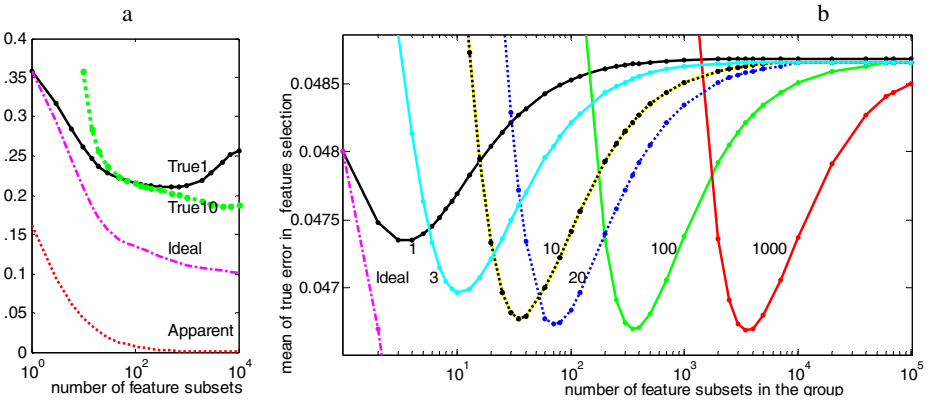

Fig. 2. Dynamics of true, ideal and apparent errors in feature selection: a - experiments with artificial Gaussian data; b – theoretical graphs

Graph True1 in Fig. 2a demonstrates peaking behavior: with an increase in the group size, true classification error after FS decreases at first, saturate and then starts increasing - we fit to validation set too much. We name this effect *feature over-selection*. *In order to obtain better result*, one need to consider smaller amount of feature subsets and/or do not select apparently the best subset of features. While inspecting 2D vectors in each group, we select the j-th positional statistics, i.e. the j-th feature subset according to estimates $\hat{P}_{\text{error}\,s}$. In order to find averages from virtually formed $C_{50000-j+1}^{m-j+1}$ groups of feature subsets, a new combinatory algorithm was developed. The FS performed according to positional statistics helped in many pattern recognition tasks. Graph True10 in Figure 2a shows a mean value of true error after

FS if the j-th positional statistics (j=10) was used to select best feature subset. We see "*not choosing the best*" strategy allowed to reduce classification error substantially. This selection strategy firstly was analyzed by Ng in [13], for rather simplified artificial model of the hypotheses selection. Now we demonstrate its usefulness for standard pattern recognition task with dependent variables. In next section, we will suggest probabilistic framework to analyze feature over-selection theoretically.

3 Probabilistic Framework to Analyze Feature Selection Bias

A main declaration utilized in our analysis, is consideration of random search feature selection procedure. Then we may assume that m 2D vectors ($P_{\mathrm{B}}{}^{i}$, $\hat{P}_{\mathrm{error}}{}^{i}$) utilized to select certain subset of features are random vectors extracted from 2D population. Following a standard probability theory, probability density function of random vector (P_{B} , $\hat{P}_{\mathrm{error}}$) may be expressed as a product of conditional and unconditional densities

$$f_1(P_{\mathrm{B}}, \hat{P}_{\mathrm{error}}) = f_2(\hat{P}_{\mathrm{error}} \mid P_{\mathrm{B}})\, f_3(P_{\mathrm{B}}) = f_4(P_{\mathrm{B}} \mid \hat{P}_{\mathrm{error}})\, f_5(\hat{P}_{\mathrm{error}}). \tag{1}$$

In derivation of expected value of true classification error after FS we postulate that conditional density, $f_2(\hat{P}_{\mathrm{error}} \mid P_{\mathrm{B}})$, *and unconditional one*, $f_3(P_{\mathrm{B}})$, *are known a priori*. As a result, final conclusions are conditioned by $f_1(P_{\mathrm{B}}, \hat{P}_{\mathrm{error}})$. Standard theory gives

$$f_5(\hat{P}_{\mathrm{error}}) = \int f_2(\hat{P}_{\mathrm{error}} | P_{\mathrm{B}})\, f_3(P_{\mathrm{B}})\, dP_{\mathrm{B}} \text{ and} \tag{2}$$

$$f_4(P_{\mathrm{B}} \mid \hat{P}_{\mathrm{error}}) = f_2(\hat{P}_{\mathrm{error}} \mid P_{\mathrm{B}})\, f_3(P_{\mathrm{B}}) / f_5(\hat{P}_{\mathrm{error}}), \tag{3}$$

where integration is performed over all interval of varying P_{B}. If one defines the distribution of P_{B} over a set of discrete values, integration is substituted by summation.

Equation (3) could be used in order to evaluate a mean of true classification error if certain estimate, $\hat{P}_{\mathrm{error}\,k}$, is already picked out of a pool of values, $\hat{P}_{\mathrm{error}\,1}$, $\hat{P}_{\mathrm{error}\,2}$, … , $\hat{P}_{\mathrm{error}\,m}$. In original paper [11], a situation with extreme (minimal) value of the error estimate $\hat{P}_{\mathrm{error}}$ was considered. Inspired by the author's multiple experimental observations that utilization of positional statistics sometimes outperforms usage of minimal values (one of them is presented in previous section) and Ng [13] considerations, in this paper we will move from analysis of minimal value to the k-th positional statistics, $\hat{P}_{\mathrm{error}\,k}$, the k-th value in a ranged sequence $\hat{P}_{\mathrm{error}\,k_1} \le \hat{P}_{\mathrm{error}\,k_2} \le \ldots$, $\hat{P}_{\mathrm{error}\,k_1 m}$. Statistical theory of extreme value distributions gives

$$f_6(\hat{P}_{\mathrm{error}\,k}) = \frac{\Gamma(m+1)}{\Gamma(k)\Gamma(m-k+1)} [F_5(\hat{P}_{\mathrm{error}})]^{k-1} [1 - F_5(\hat{P}_{\mathrm{error}})]^{m-k} f_5(\hat{P}_{\mathrm{error}}), \tag{4}$$

where $F_5(\hat{P}_{\mathrm{error}})$ is cumulative distribution of random variable $\hat{P}_{\mathrm{error}}$.

Use of distribution density of the k-th positional statistics (4) results expected value of true classification after feature selection and that of average of apparent error

$$EP_{\text{true}k} = \int\int f_4(P_{\text{B}}|\hat{P}_{\text{error}^k}) f_6(\hat{P}_{\text{error}^k}) d\hat{P}_{\text{error}^k} dP_{\text{B}} \text{ and} \tag{5}$$

$$E\,\hat{P}_{\text{apparent}k} = \int \hat{P}_{\text{error}^k} f_6(\hat{P}_{\text{error}^k}) d\hat{P}_{\text{error}^k} . \tag{6}$$

The integrations (or summations) in Eq. (5) are performed along intervals of variations of error estimate, $\hat{P}_{\text{error}}$, and true error, P_{B}. Both expected values are conditioned by serial number of positional statistics. If k=1, we deal with extreme (minimal) value. Hypothetical characteristics, an expectation of ideal classification error, we have in case where from m randomly formed subsets of features we select the best one according to true classification error values, $P_{\text{B}1}$, $P_{\text{B}1}$, … , $P_{\text{B}m}$. Probability density function of extreme value is given by equation

$$F_7(P_{\text{B}}) = m\,[1 - F_3(P_{\text{B}})]^{m-1} f_3(P_{\text{B}}), \tag{7}$$

where $F_3(P_{\text{B}})$ is cumulative distribution function of random variable P_{B}.

Then the expectation of ideal classification error could be found as

$$EP_{\text{ideal}} = \int P_{\text{B}} f_7(P_{\text{B}}) dP_{\text{B}} . \tag{8}$$

Above equations allow to investigate behavior of expected values of true, apparent and ideal classification errors after feature selection, We remind that the conditional density, $f_2(\hat{P}_{\text{error}} | P_{\text{B}})$, and unconditional density, $f_3(P_{\text{B}})$, should be known *a priori*. Due to complexity of the problem with extreme values and positional statistics we do not have explicit expressions. So, further analysis should be performed by numerical integration. In Fig. 2b we depict dependence of expected values of true classification errors when feature selections were performed according to extreme value (graph 1) and the 3$^{\text{th}}$, 10$^{\text{th}}$, 20$^{\text{th}}$, 100$^{\text{th}}$ and 2000$^{\text{th}}$ positional statistics. Moreover, it was postulated that distribution density of 2D vectors (P_{B}, $\hat{P}_{\text{error}}$) was a mixture of two Gaussian densities

$$f_3(P_{\text{B}}, \hat{P}_{\text{error}}) = q_1 \times f_{\text{N}}((P_{\text{B}}, \hat{P}_{\text{error}}), \mathbf{M}_1, \mathbf{S}_1) + (1 - q_2) \times f_{\text{N}}((P_{\text{B}}, \hat{P}_{\text{error}}), \mathbf{M}_2, \mathbf{S}_2),$$

where $f_{\text{N}}((x_1, x_1), \mathbf{M}, \mathbf{S})$ denotes Gaussian probability density function of 2D vector, (x_1, x_1), having mean $\mathbf{M}$ and covariance matrix $\mathbf{S}$. Note that density $f_3(P_{\text{B}}, \hat{P}_{\text{error}})$ depends on random validation set critically. After analysis of two dozens of very left parts of distributions similar to that depicted in Fig. 1a,b, in a variety of situations with single validation sets of size (N=20), we selected for this illustration: q_1=0.1,

$$\mathbf{M}_1 = \begin{bmatrix} 0.03 \\ 0.01 \end{bmatrix}, \mathbf{S}_1 = \begin{bmatrix} 0.0002^2 & 0.0 \\ 0.0 & 0.005^2 \end{bmatrix}, \mathbf{M}_2 = \begin{bmatrix} 0.05 \\ 0.013 \end{bmatrix}, \mathbf{S}_2 = \begin{bmatrix} 0.00025^2 & 0.0 \\ 0.0 & 0.009^2 \end{bmatrix}.$$

We pay readers attention to a fact that for single validation set, the variances of hold out error counting error estimates (right bottom elements of matrices $\mathbf{S}_1$, $\mathbf{S}_2$) are much

smaller as variance, s^2, predicted by theory for diverse independent validation sets, $s^2 = P_B(1- P_B)/N$. In Figure 2b we also depict the very left part of graph "Ideal", the ideal error after feature selection, which decreased rapidly until 0.03 (for $m \approx 30$) and saturated at this level. Apparent classification error after feature selection started to decrease from 0.013 level, and for $m \approx 200$ it became practically zero. Experimentation with artificially generated feature selection problems revealed that a character of distribution $f_1(P_B, \hat{P}_{error})$ greatly depends on a way how dependencies between original variables of the data are generated, individual qualities of the features and, most important, on particular randomly chosen p-dimensional validation set.

4 Analysis of Over-Selection Phenomenon in Real World Task

We performed experiments with 7129-dimensional two class leukaemia data set [16, 17]. In order to select $r = 8$ features we employed standard linear Fisher classifier. From 72 examples we used 35 samples for training. Remaining 37 vectors constituted the validation set. Such large dimensionality/sample size ratio is frequent in many biomedical investigations, especially in to-day's analysis of microarray gene expression data. *Three millions* random feature subsets were generated. We examined the accuracy problem in a situation where training set (re-substitution) error estimates *with certain correction of "training bias"* were used as evaluations of "true" error. The validation set estimates were used to pick up "the best" feature subsets.

Re-substitution error estimates are optimistically biased. For Fisher classifier, asymptotically as training sample size, N, and dimensionality, r, are large, expected value of classification error can be found from simple, however, exact asymptotic formula $EP_N \approx \Phi(-\frac{1}{2}\delta / T_{bias})$, where $T_{bias} = \sqrt{(N-p)/N/(1+4p/N/\delta^2)}$ [4, 7].

We are also interested in the bias of re-substitution error, which can be expressed as $E\hat{P}_R \approx \Phi(-\frac{1}{2}\delta \times T_{bias})$. Double use of one-dimensional interpolation for $\hat{P}_R$ value allows to obtain, approximate (restored) value of generalization error, $\hat{P}_g$.

The histograms of "*restored*" generalization, $\hat{P}_g$, and validation, $\hat{P}_V$, error estimates are presented in Figure 3ab. We see that in spite of simplicity of asymptotic formulae, training bias elimination was performed quite correctly: left and middle parts of both histograms are almost identical. Fig. 3c shows a scatter diagram of distribution of 2000 vectors ($\hat{P}_R$, $\hat{P}_V$). Due to very small sample size, many subsets have the same $\hat{P}_R$ and $\hat{P}_V$ values. For better data visualization, a small uniform noise was added to each component. Two dimensional distribution of 10,474 vectors with smallest re-substitution and validation errors is detailed by 2D scatter diagram in Fig. 3d. We see that FS strategies based on the validation set and the modified training set estimates pick different feature subsets. This conclusion follows also from conditional mean of $\hat{P}_V$ presented in Fig. 3e: the smallest generalization error values could be obtained if subsets with *two* validation errors

would be selected! We plot mean values of *restored generalization error* rates conditioned on 17 distinct validation error values. In Figure 3f we present dynamics of true error after feature selection performed according to the smallest validation error (True1) and that performed according to 30th and 300th positional statistics. We also present "Ideal" and "Apparent" classification errors in the same way as it was done in the experiments with artificial Gaussian data (Section 2, Figure 2a).

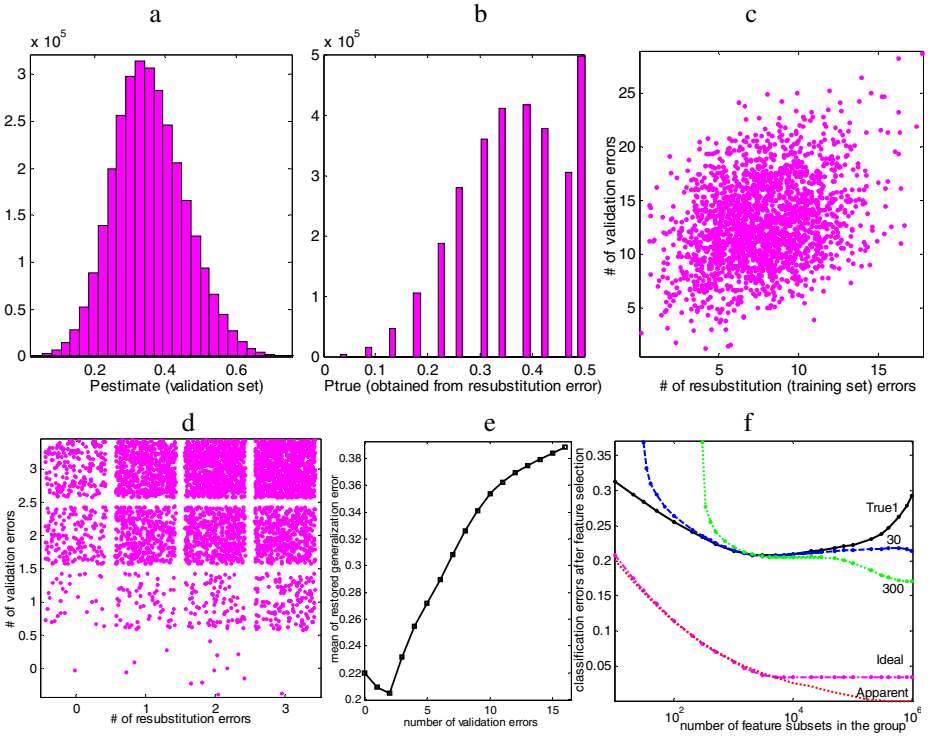

Fig. 3. The histograms of 3,0000,000 values of validation error estimates (a) and restored generalization error (b); c, d - scatter diagrams of distribution of the number of misclassification errors in training and validation sets; e – average "restored" generalization error as a function of estimated error, $\hat{P}_{\text{error}}$, f – dynamics of true, ideal and apparent errors

5 Concluding Remarks

While designing the classifiers from training set we obtain training bias, a difference between generalization and Bayes errors, EP_N - P_B, In present paper we consider feature selection bias (the difference between True1 and Ideal)) which arise when size of validation set is finite. We present further development of probabilistic framework started in [10, 11]. This approach is based on analysis of random search FS procedure and postulation that joint distribution of true and estimated classification errors is known *a priori*. Theoretical and experimental results advocate that feature selection

bias can be very large if the size of feature subset group is very large. Calculation of expected generalization error of Fisher classifier according to $EP_N \approx \Phi(-\frac{1}{2}\delta / T_{\text{bias}})$ for N=20 and p=8 gives that for P_B=0.2 (for m≈5, inspect Fig. 2a), EP_N≈0.3 and for P_B=0.1 (m≈10,000), EP_N ≈0.19. Similar estimates we obtained for gene expression data. It means that FS bias is comparable with training bias provided the training set size is equal to that of validation set and linear Fisher discriminant is utilized as the classifier.

We also showed that there exists overfitting phenomenon in feature selection, named in this paper as feature over-selection. This effect is validation set dependent. It was observed when validation set size was very small. The feature over-selection phenomenon could be diminished if FS would be performed on basis of positional statistics. Development of practical recommendations is a problem of future research.

Acknowledgement. A part of the experiments with gene expression data were performed when the author visited the Institute for Biodiagnostics, NRCC, Winnipeg, Canada (NATO Expert Visit Grant SST.EAP.EV 980950). The author is thankful to Richard Baumgartner and Ray Somorjai for useful and challenging discussions.

References

1. G. F. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory* IT-14:55–63, 1965.
2. S. Raudys. On the problems of sample size in pattern recognition. In V.S. Pugatchiov (editor) *Detection, Pattern Recognition and Experiment Design*, 2:64–76. Proceedings of the 2nd All-Union Conference Statistical Methods in Control Theory. Nauka, Moscow, 1970 (in Russian).
3. L. Kanal, B. Chandrasekaran. On dimensionality and sample size in statistical pattern classification. *Pattern Recognition* 3:238–55, 1971.
4. S. Raudys. *Statistical and Neural Classifiers - An integrated approach to design*. Springer-Verlag London, 2001.
5. S. Haykin. *Neural Networks: A comprehensive foundation*. 2nd edition. Prentice-Hall, Englewood Cliffs, NJ, 1999.
6. K. Fukunaga. *Introduction to Statistical Pattern Recognition*. 2nd Ed. Acad. Press, 1990.
7. S. J. Raudys, A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendation for practitioners. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13 (3) 242-254, 1991.
8. P. Pudil, J. Novovicova and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119-1125, 1994.
9. I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *J. of Machine Learning Research,* 3:1157-1182, 2003.
10. S. Raudys. Classification errors when features are selected. In S. Raudys (editor), *Statistical Problems of Control*, 38: 9-26, 1979. Institute of Mathematics and Informatics, Vilnius, 1979 (in Russian).
11. S. Raudys. Influence of sample size on the accuracy of model selection in pattern recognition. In *Statistical Problems of Control* (S. Raudys, ed., Institute of Mathematics and Informatics, Vilnius) 50 9-30, 1981 (in Russian).
12. G. D. Murray. A cautionary note on selection of variables in discriminant analysis. *Appl. Statist*. 26 (3) 246-250, 1997.

13. A. Ng. Preventing "overfitting" of cross-validation data, *Proc. of the Fourteenth International Conference on Machine Learning*, Morgan Kaufman, 245-253, 1997.
14. J. Ye. On measuring and correcting the effects of data mining and model selection. *J. of American Statistical Association*, 93 (441) 120-131, 1998.
15. P. Domingos. Process-oriented estimation of generalization error. In Proceedings of the Sixteenth International, *Joint Conf on Art. Intell*., Morgan Kaufman, 714-722, 1999.
16. C. Ambroise and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA* 99 (10) 6562-6566, 2002.
17. T.R. Golub et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286 531-537, 1999.