Urszula LIBAL

Wroclaw University of Technology

Wavelet Decomposition of Signal and Feature Selection by LASSO for Pattern Recognition

Abstract. *There is searched the balance between an increase of pattern recognition risk and a decrease of a model size. The experiments are performed for noisy signals, decomposed in wavelet bases. Wavelet representation of signals, i.e. representation by wavelet coefficients called signal features, constitutes the full model. The presented feature selection method is based on the Lasso algorithm (Least Absolute Shrinkage and Selection Operator). The aim of the experiment is to find an optimal model size and investigate the relations between the risk, the number of signal features and the noise level. A new criterion of feature selection is proposed that minimizes both the risk and the number of signal features. The experimental risk of classification is analysed for all possible reduced by Lasso models and for several values of noise levels.*

Streszczenie. *Poszukiwana jest równowaga pomiędzy wzrostem ryzyka rozpoznawania obrazów oraz zmniejszeniem rozmiaru modelu. Badania przeprowadzono dla zaszumionych sygnałów, zdekomponowanych w bazach falkowych. Falkowa reprezentacja sygnałów, czyli reprezentacja za pomocą współczynników falkowych zwanych cechami sygnału, stanowi pełny model. Przedstawiona metoda selekcji cech jest oparta o algorytm Lasso (Least Absolute Shrinkage and Selection Operator). Celem eksperymentu jest znalezienie optymalnego rozmiaru modelu oraz zbadanie zależności pomiędzy ryzykiem, liczbą cech sygnału oraz poziomem szumu. Zaproponowano nowe kryterium selekcji cech, które minimalizuje ryzyko oraz liczbę cech sygnału. Eksperymentalne ryzyko błędnej klasyfikacji jest badane dla wszystkich możliwych zredukowanych za pomocą Lasso modeli oraz kilku wartości poziomu szumu.* ***(Falkowa dekompozycja sygnału oraz selekcja cech za pomocą LASSO w zadaniu rozpoznawania wzorców)***

Keywords: risk, pattern recognition, feature selection, lasso, wavelets, signal decomposition
Słowa kluczowe: ryzyko, rozpoznawanie wzorców, selekcja cech, lasso, falki, dekompozycja sygnału

Introduction

The main task of classification rules is to assign the examined object to the correct class. The quality of the rule is measured by a risk of classification to a wrong class [3]. The minimization of the risk states our purpose. Theoretically, the misclassification risk can be minimized by basing on a full model. But the large number of features, describing a signal, implies an increase of computation time or worse, increases the misclassification risk (e.g. the empty space phenomenon). This causes that in real-life applications the classification algorithms might be ineffective until the number of features is reduced.

In this article a feature selection method, based on Lars/Lasso [7] (*Least Angle Regression / Least Absolute Shrinkage and Selection Operator*), is proposed. The presented methods are tested on (kNN) k-Nearest Neighbors classifier. The experimental risk for two-class pattern recognition problem is calculated for all sizes of reduced model. The results are shown in the last section. In the section titled 'Criteria of Threshold Choice' is introduced a new intuitive criterion MIN of model selection, that minimizes the risk and the number of features in the model. The risk values for models chosen by Lasso with the criteria: MIN (minimizing the risk), BIC (*Bayesian Information Criterion*) and AIC (*Akaike Information Criterion*) are compared. The pattern recognition is preceded by: 1. *signal pre-processing* - the signal $s(t)$ is approximated by $W(s(t))$ in wavelet bases, 2. *feature selection* - the thresh λ is chosen with the criterion MIN, BIC or AIC, then the Lasso algorithm performs thresholding on the signal features.

Two Patterns of Signal

The two-class pattern recognition problem is considered. There is assumed the existance of a generic pattern $f(t)$ for each class. There is introduced the following form of a signal, disturbed by uniform and Gaussian noise:

$$s(t_i) = f(t_i) + cU_i + \epsilon Z_i, \tag{1}$$

where $t_i = \frac{i}{p_0}$, $i = 0, 1, \ldots, p_0 - 1$. The both sets of random variables $\{Z_i\}$ and $\{U_i\}$ are independent and identically distributed, from Gaussian distribution $Z_i \sim \mathcal{N}(0; 1)$ and uniform $U_i \sim \mathcal{U}(-1; 1)$, respectively. The uniform noise expresses a randomness of signals in a given class, and the Gaussian noise constitutes a measuring distortion. Both random variables are centered, i.e. $EU_i = 0$, $EZ_i = 0$, so the expected value is $E\, s(t_i) = f(t_i)$ and the joint variance is $\sigma^2 = Var\, s(t_i) = E(s(t_i) - E\, s(t_i))^2 = c^2 EU_i^2 + \epsilon^2 EZ_i^2 = c^2/3 + \epsilon^2$.

The patterns used in experiments are a sine wave $f_1(t)$ in *class 1* and a triangular wave $f_2(t)$ in *class 2*. They are shown in the first row of Figure 1. The measuring distortion was set on the fixed level with $\epsilon = 0.05$. The noisy versions of signals shown in the next two rows are obtained for $c = 0.7$ (SNR = 7.8 [dB]) and $c = 1.6$ (SNR = 0.7 [dB]).

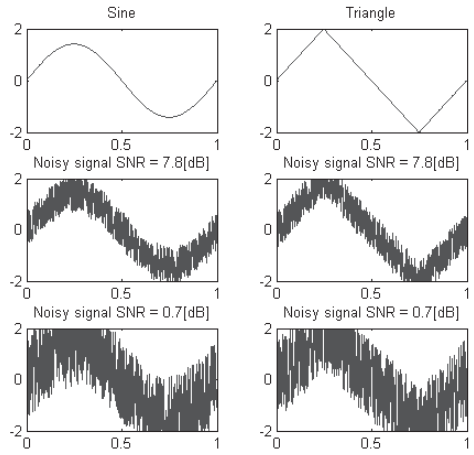

Fig. 1. Sine and triangular signal and their noisy versions with SNR = 7.8 [dB] and 0.7 [dB].

Wavelet Representation of Signal

The transformation of signal from a time domain to a wavelet representation in a time-frequency domain allows to achieve good recognition results [6]. Wavelet decomposition of signals was motivated by this fact. There is assumed that $\phi(t)$ is a *scaling function* and $\psi(t)$ is a proper *mother wavelet*. Let $\phi_{jk}(t) = 2^{j/2}\phi\left(2^j t - k\right)$ be the basic function of approximation space V_j and $\psi_{jk}(t) = 2^{j/2}\psi\left(2^j t - k\right)$ be the basic function of detail space W_j for scale j [2]. The signal approximation for scale j_1 (i.e. for $J = j_1 - j_0$ levels of

decomposition) has the form

$$s(t) \approx \sum_k c_{j_0 k}\phi_{j_0 k}(t) + \sum_{j=j_0}^{j_1-1}\sum_k d_{jk}\psi_{jk}(t) \tag{2}$$

where wavelet coefficients $c_{j_0 k}$ and d_{jk} are given by the formulas $c_{j_0 k} = \int_{\mathbb{R}} s(t)\,\phi_{j_0 k}(t)\,dt$ and $d_{jk} = \int_{\mathbb{R}} s(t)\,\psi_{jk}(t)\,dt$.

Noisy signal (1) transformed by a wavelet filtration to a time-frequency domain is represented by the sequence of wavelet coefficients

$$W(s(t)) = (c_{j_0},\, d_{j_0},\, d_{j_0+1},\, \ldots,\, d_{j_1-1}) = \underline{x}, \tag{3}$$

where $c_{j_0} = (c_{j_0 k})_k$ - a sequence of wavelet approximation coefficients on level J (i.e. for coarse scale j_0), $d_j = (d_{jk})_k$ - a sequence of wavelet detail coefficients on level $j_1 - j$ (i.e. for scale j), for $j = j_0, j_0+1, \ldots, j_1-1$. From now on, it will be denoted with p the total number of wavelet coefficients. The coefficients $\underline{x} = (x_1, x_2, \ldots, x_p)$ will be called *features*.

Feature Selection - LARS/LASSO Technique

Feature selection is an important initial step of signal analysis. Well performed feature selection can increase the efficiency of further recognition [7, 8, 9].

The LASSO Technique (*Least Absolute Shrinkage and Selection Operator*) was proposed by Tibshirani [11], who noticed its similarity to soft thresholding technique, e.g. used for wavelet-based signal de-noising [4]. Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_p)$ be the LASSO estimate of

$$\hat{\beta} = \arg\min \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 \tag{4}$$

$$\text{subject to} \sum_j |\beta_j| \le \lambda. \tag{5}$$

The soft thresholding method has equivalent effect as LASSO regularization for the case of an orthonormal design $X^T X = I$, i.e.

$$\hat{\beta}_j = \text{sign}(\beta_j)(|\beta_j| - \lambda)^+ \tag{6}$$

for a threshold λ determined by the condition $\sum_j |\hat{\beta}_j| = \lambda$.

The j^{th} wavelet coefficient of i^{th} signal is denoted by x_{ij}. Let the matrix X of size $n \times p$ be composed of sequences of coefficients $\underline{x_i}$ in every row, i.e.

$X = \begin{bmatrix} \underline{x_1} \\ \vdots \\ \underline{x_n} \end{bmatrix}$, and $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ be a vector of centered and normalized class labels, $y_i \in \{-1, 1\}$.

Efron, Hastie, Johnstone and Tibshirani [5] presented a modification of LARS (*Least Angle Regression*) algorithm giving all possible LASSO estimates in an efficient, algorithmic way. The plots of all LASSO estimate coordinates β_j versus the sum of absolute value of the estimate coordinates $\sum |\beta_j|$ are shown in the Figure 2.

The LASSO modification of the LARS algorithm [5] has more than p steps, in contrast to LARS algorithm with exactly p steps. In the first step, there is chosen a coordinate β_i being the most correlated with the output y, i.e. $c_i = max(\widehat{c})$, where $\widehat{c}$ is the correlation vector $\widehat{c} = X^T y$, and the index i is put into the empty active set. The calculated value of shift, in the current direction, updates the predictor $\widehat{y}$. In each step, there is chosen a next coordinate β_j, the most correlated with the difference between actual output and the predictor, i.e. $c_j = max(\widehat{c})$, where the correlation vector $\widehat{c} = X^T(y - \widehat{y})$, and the next index j goes to the active set. However, the modification of LARS / LASSO can remove from the active set an index of the coordinate that changes the sign of the step in the calculated direction.

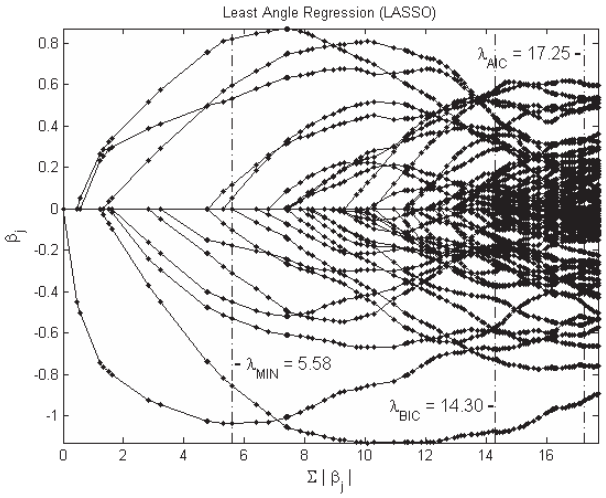

Fig. 2. LASSO estimates β_i vs. $\sum |\beta_i|$. SNR = 7.8 [dB].

Criteria of Threshold Choice

The threshold λ is chosen by minimization of the selected criterion. The tests were performed for the three following criteria AIC, BIC and a new criterion MIN:

1) **Akaike information criterion** [1]

$$AIC(\lambda) = \frac{RSS}{p\widehat{\sigma^2}} + \frac{2}{p}df(\lambda),$$

2) **Bayesian information criterion** [10]

$$BIC(\lambda) = \frac{RSS}{p\widehat{\sigma^2}} + \frac{ln(n)}{p}df(\lambda),$$

3) **Risk minimazing criterion** [8]

$$MIN(\lambda) = R + df(\lambda),$$

where:

$RSS = ||y - X\hat{\beta}||^2 = ||y - \hat{y}||^2$ (prediction error),
$R = R(\hat{\beta})$ (experimental risk for model $\hat{\beta}$),
p - number of wavelet coefficients (features),
n - number of signal samples,
$df(\lambda)$ - size of model for fixed λ (number of coefficient in active set, number of non-zero coordinates of $\hat{\beta}$).

The first two criteria of choosing threshold λ are commonly used in regression estimation problems, also for LASSO [12]. The criteria try to negotiate between the value of RSS and the size of the model df. The λ is chosen when the value of AIC or BIC reaches the minimum. While the model $\widehat{y}$ for λ_{AIC} has optimal RSS value, the regression estimate $\hat{\beta}$ chosen by BIC is consistent with the full size model, as $n \to \infty$. The exemplary values of the λs are marked with vertical line in Figures 2 and 3.

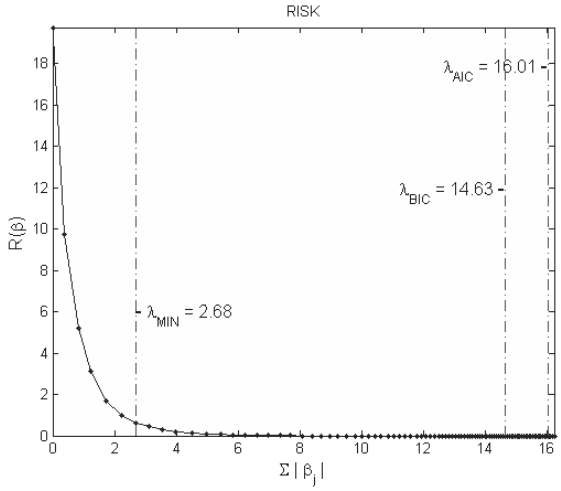

Fig. 3. Experimental risk R [%] for models selected by LASSO estimate β shrinked by $\lambda = \sum |\beta_i|$. SNR = 7.8 [dB].

Experiments

The pattern recognition is preceded by:

1) **signal pre-proccessing** - the signal $s(t)$ is approximated by $W(s(t))$ in wavelet bases,
2) **feature selection** - the thresh λ is chosen with the criterion MIN, BIC or AIC, then the Lasso algorithm performs thresholding on the signal features.

For four noise levels with signal-to-noise ratio $SNR = 7.8,\ 4.7,\ 2.5$ and 0.7 [dB] ($\epsilon = 0.05$, $c = 0.7,\ 1.0,\ 1.3$ and 1.6, respectively in the formula (1)), there were generated $n = 100$ noisy signal samples (50 from *class 1* and *class 2*) in learning set and the same number of test samples. Each signal has $2^{10} = 1024$ samples on the time interval $[0, 1]$.

For recognition there was chosen a pair of generic patterns that seem to be quite similar and hard to classify for the noised versions. The tests were performed for sine (*class 1*) and triangular signal (*class* 2), see Figure 1. The classification was made by 5-Nearest Neighbors classifier. The resultes are in Table 2.

To average the recognition results, all the experiments were performed 10 times. The wavelet decomposition of signals were executed for a *Haar* wavelets. Because of the number of signal samples, the possible decomposition levels were $J = 1, 2, ..., 10$. It means that every signal (2) was approximated in wavelet bases and represented by vectors of coefficients (3) for $J = 10$ levels, i.e. by $(c_{j_0}, d_{j_0}, d_{j_0+1}, \ldots, d_{j_0+9})$.

Results and Conclusions

The experiments show that the threshold λ for the new criterion MIN achives the lowest values amoung all investigated criteria (see Table 1 and Figure 3). It results in the strongest reduction of the model and the smallest number of selected features. Generally, the criterion MIN chose from the initial number of 1024 features about 10-30 coefficients. While BIC criterion chose about 60-110, and the AIC about 90-130 features.

Taking into account the model reduction, **the lowest number of features was left by MIN criterion**. The risk values for high noise level (e.g. for $c = 1.6$) for MIN and AIC criteria are comparable (see Table 2). The performance of all criteria was good, i.e. the risk was lower than 1% for the noise levels $c = 0.7,\ 1.0$ and 1.3. The results show that searching for the optimal threshold dedicated to the feature selection for the classification is an open problem. And it should be investigated if other criteria for feature selection can be established.

Table 1. Average values of λ.

SNR [dB]	**7.8**	**4.7**	**2.5**	**0.7**
c	**0.7**	**1.0**	**1.3**	**1.6**
λ_{MIN}	2.7	6.9	13.2	20.1
λ_{BIC}	14.6	19.2	24.2	28.5
λ_{AIC}	16.0	21.2	26.0	30.0

Table 2. Experimental risk [%] for averaged values of λ.

SNR [dB]	**7.8**	**4.7**	**2.5**	**0.7**
c	**0.7**	**1.0**	**1.3**	**1.6**
R_{MIN}	0.5	0.6	0.7	3.6
R_{BIC}	0.0	0.0	0.2	2.9
R_{AIC}	0.0	0.0	0.2	3.8

Acknowledgments

This work is co-financed by the European Union as part of the European Social Fund.

REFERENCES

[1] Akaike, H.: Information theory and an extension of maximum likelihood principle, Proc. 2nd International Symposium on Information Theory, Eds. B.N. Petrov and F. Csaki, Budapest, pp. 267–281, 1973.
[2] Daubechies, I.: Ten Lectures on Wavelets, CBMS-NSF Lecture Notes nr. 61, SIAM, 1992.
[3] Devroye, L., Gyorfi, L., and Lugosi, G.: A probabilistic theory of pattern recognition, Springer-Verlag, New York, 1996.
[4] Donoho, D. L.: De-noising by soft-thresholding, IEEE Transactions on Information Theory, 41(3), pp. 613–627, 1995.
[5] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R.: Least Angle Regression, Annals of Statistics, 32(2), pp. 407–499, 2004.
[6] Kowalski, C.: Zastosowanie analizy falkowej w diagnostyce silników indukcyjnych, Przegląd Elektrotechniczny, 1, pp. 21–26, 2006.
[7] Libal, U.: Feature Selection for Pattern Recognition by LASSO and Thresholding Methods – a Comparison, Proc. 16th IEEE International Conference on Methods and Models in Automation and Robotics - MMAR 2011, Międzyzdroje, 22-25 August 2011, pp. 168–173.
[8] Libal, U.: Kryteria selekcji modelu w eksperymentalnym rozpoznawaniu sygnałów zdekomponowanych w bazach falkowych, Interdyscyplinarność badań naukowych 2012: praca zbiorowa / pod red. Jarosława Szreka, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2012, pp. 297–302.
[9] Osowski, S. and Kurek, J.: Selekcja cech diagnostycznych w zastosowaniu do rozpoznania różnych uszkodzeń prętów maszyny indukcyjnej, Przegląd Elektrotechniczny, 1, pp. 121–123, 2010.
[10] Schwarz, G.: Estimating the dimension of a model, Annals of Statistics, 6(2), pp. 461–464, 1978.
[11] Tibshirani, R.: Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B, 58(1), pp. 267–288, 1996.
[12] Zou, H., Hastie, T., and Tibshirani, R.: On the "degrees of freedom" of the lasso, Annals of Statistics, 35(5), pp. 2173–2192, 2007.

***Author**: Urszula Libal, Institute of Computer Engineering, Control and Robotics, Faculty of Electronics, Wroclaw University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland, email: urszula.libal@pwr.wroc.pl*