fuzzy systems, Bayesian inference,
ECG signal

Alina MOMOT[*], Michał MOMOT[**], Jacek ŁĘSKI[**,***]

THE FUZZY RELEVANCE VECTOR MACHINE AND ITS APPLICATION TO NOISE REDUCTION IN ECG SIGNAL

The paper presents new method called the Fuzzy Relevance Vector Machine (FRVM), a modification of the relevance vector machine, introduced by M. Tipping, applied to learning Takagi-Sugeno-Kang (TSK) fuzzy system. Moreover it describes application of the FRVM to noise reduction in ECG signal. The results of the process are compared to those obtained using both Least Squares method for learning output functions in TSK rules and commonly used method using a low-pass moving average filter.

1. INTRODUCTION

Fuzzy systems play an important role in many disciplines of engineering and science. The major areas of their application are control, system identification, pattern recognition and data mining. Identification of systems from input-output measurements is an important topic of scientific research with a wide range of practical applications. From the input-output view, fuzzy systems are flexible mathematical functions which can approximate other functions with a desired accuracy. This property is called general function approximation. Compared to other well-known approximation techniques such as artificial neural networks, fuzzy systems provide a more transparent representation of the underlying system, which is mainly due to the possible linguistic interpretation in the form of if-then rules.

Inferring a functional mapping based on given set of input-output data is the main goal of supervised learning, which can be formalized as the problem inferring a function $t = f(x)$, based on a training set $T = \{(x_1, t_1), (x_2, t_2), ..., (x_N, t_N)\}$. Usually the inputs are *d*-dimensional real vectors, $x \in \Re^d$ and outputs might be real values, e.g. $t \in \Re$ (in regression) or might be categorical nature, e.g. binary $t \in \{0,1\}$ (in classification). Usually the function *f* is assumed to have a fixed structure and to depend on a set of parameters *w* and the goal becomes to estimate the parameters from the training data. A flexible and popular set of candidates for *f* is that of the form:

$$f(x; w) = \sum_{i=1}^{N} w_i K(x_i, x) + w_0 \qquad (1)$$

[*] Silesian University of Technology, Institute of Computer Science, 16 Akademicka St., 44-101 Gliwice, Poland
[**] Institute of Medical Technology and Equipment, 118 Roosevelt St., 41-800 Zabrze, Poland
[***] Silesian University of Technology, Institute of Electronics, 16 Akademicka St., 44-101 Gliwice, Poland

where $w=(w_0,w_1,...,w_N)^T$ is set of parameters (or weights) and $K(x_i,x)$ is a kernel function, effectively defining one basis function for each example in training set.

It is known that, to achieve good generalization ability, it is necessary to control the complexity of the learning function [14]. A Bayesian approach to complexity control consists in using a prior $p(w|\alpha)$ favouring simplicity or smoothness, in some sense, of the function to be learned (where α is a vector of hyperparameters). The usual choice is a zero-mean Gaussian prior as can be seen in the Relevance Vector Machine [12], where in the set of hyperparameters, one associated with each weight, whose most probable values are iteratively estimated from the training data.

For the last few years, there has been also an increasing interest in fuzzy systems which incorporate tools well-known from the support vector machine methods, being the part of the statistical learning theory [14]. The support vector fuzzy regression machine has been introduced in [4]. The support vector fuzzy clustering method is described in [2]. An ε-insensitive approach to learning of neuro-fuzzy systems has been introduced in [6] and similar approach to learning a classifier, called a fuzzy support vector machine has been independently introduced in [5].

In this paper a new Bayesian learning method, based on the Relevance Vector Machine method, is presented and applied to learning the Takagi-Sugeno-Kang fuzzy system [10], [11]. This approach, called the Fuzzy Relevance Vector Machine method [8], [9], is applied to reduction of artificial Gaussian noise added to electrocardiographic (ECG) signal.

2. THE FUZZY RELEVANCE VECTOR MACHINE

Given a data set of input-output pairs $\{x_i,t_i\}_{i=1}^N$, considering scalar-valued target functions only it is assumed that the targets are samples from the model with additive noise:

$$t_i = f(x_i;w)+\varepsilon_i \qquad \forall i\in\{1,2,...,N\} \tag{2}$$

where ε_i are independent samples from some noise process which is further assumed to be zero-mean Gaussian with variance σ_i^2 for each i respectively and f is prediction function in the form (1). Thus, the likelihood of the complete data set can be written as 1

$$p(t\mid w,\beta)=(2\pi)^{-\frac{N}{2}}\mid B\mid^{\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(t-\Phi w)^T B(t-\Phi w)\right\}, \tag{3}$$

where $t=(t_1,t_2,...,t_N)^T$, $w=(w_0,w_1,...,w_N)^T$, $\beta=(\sigma_1^{-2},\sigma_2^{-2},...,\sigma_N^{-2})^T$, $B=\mathrm{diag}(\sigma_1^{-2},\sigma_2^{-2},...,\sigma_N^{-2})$ and Φ is the $N\times(N+1)$ matrix with elements $\Phi_{nm}=K(x_n,x_{m-1})$ or $\Phi_{n1}=1$ respectively.

Using a Bayesian approach for controlling the complexity of the prediction function the prior $p(w|\alpha)$ is taken in form:

$$p(w\mid\alpha)=(2\pi)^{-\frac{N+1}{2}}\mid A\mid^{\frac{1}{2}}\exp\left\{-\tfrac{1}{2}w^TAw\right\}, \tag{4}$$

1 For simplicity, it will be omitted to notate the implicit conditioning upon the set of input vectors $\{x_i\}$ in (3) and subsequent expressions.

where $\alpha = \left(\alpha_0^2, \alpha_1^2, ..., \alpha_N^2\right)^T$ and $A = \mathrm{diag}\left(\alpha_0^2, \alpha_1^2, ..., \alpha_N^2\right)$. Importantly, like in the Relevance Vector Machine method, there is an individual hyperparameter associated independently with every weight, moderating the strength of the prior.

Using Bayes' theorem it can be computed the posterior distribution over the weights $p(w|t,\alpha,\beta)$. It is Gaussian distribution with mean m and covariance C, given by [8]

$$m = C\Phi^T Bt, \quad C = \left(\Phi^T B\Phi + A\right)^{-1}. \tag{5}$$

The distribution $p(t|\alpha,\beta)$ is also computable and given by Gaussian distribution with zero mean and covariance equal $\left(B^{-1} + \Phi A^{-1}\Phi^T\right)$.

Values of α and β which maximize $p(t|\alpha,\beta)$ cannot be obtained in closed form, hence below there are presented formulas for their iterative re-estimation [8]:

$$\alpha_i^{new} = \left(C_{ii} + m_i^2\right)^{-1}, \qquad \beta_i^{new} = \left(\phi(x_i)C\phi(x_i)^T + \left(t_i - \phi(x_i)m\right)^2\right)^{-1}, \tag{6}$$

where $\phi(x_i) = \left(1, K(x_i, x_1), ..., K(x_i, x_N)\right)$ and m as well as C are given by (5).

Given a new test point x_* and σ_*, prediction is made for the corresponding target t_*, in terms of the predictive distribution:

$$p(t_* \mid t, \alpha_{MP}, \beta_{MP}) = \int p(t_* \mid w, \beta_{MP})\, p(w \mid t, \alpha_{MP}, \beta_{MP}) dw, \tag{7}$$

where α_{MP} and β_{MP} are the most-probable values of term $p(t|\alpha,\beta)$. Since both terms in the integrand are Gaussian, this is readily computed, giving Gaussian distribution with mean given by (1) for weight vector w equal vector m (5) and variance equal $\sigma_*^2 + \phi(x_*)C\phi(x_*)^T$. Thus the vector of mean of the posterior distribution over the weights $p(w|t,\alpha,\beta)$ can be taken as the wanted weight vector w.

It is also possible to take another strategy for computing components of vector β_{MP}. It can be assumed that there are given initial values $\hat{\beta}$ of vector β and iteration procedure changes only their scale s ($\beta = s\hat{\beta}$):

$$s^{new} = \frac{N}{\mathrm{Tr}\left(\hat{B}\Phi C\Phi^T\right) + \left(t - \Phi m\right)^T \hat{B}\left(t - \Phi m\right)}, \tag{8}$$

where $\hat{B} = \mathrm{diag}(\hat{\beta})$. It is worth mentioning that in case when all components of $\hat{\beta}$ are equal one, this strategy leads to original algorithm for learning the Relevance Vector Machine proposed in [13].

The algorithm for finding prediction function $t = f(x)$ in form of (1), described above, based on a training set $T = \left\{(x_1, t_1), (x_2, t_2), ..., (x_N, t_N)\right\}$ can be, taking inner product as kernel function, applied to computing output of individual rule function in Takagi-Sugeno-Kang fuzzy models [10], [11]. The identification of system from input-output measurements could be based on fuzzy c-mean clustering of the input space preserving similarity of input data, assuming that each cluster corresponds to a fuzzy if-then rule in the Takagi-Sugeno-Kang form:

$$R^{(i)}: \text{IF } x \text{ is } A^{(i)}, \text{THEN } y = \left(w^{(i)}\right)^T x', \quad i = 1,2..., c \tag{9}$$

where $x \in \Re^d$ is the input variable, $y \in \Re$ is the output variable, $x' = (1, x^T)^T$ is the augmented input vector and $w^{(i)} = \left(w_0^{(i)}, w_1^{(i)}, ..., w_d^{(i)}\right)^T$ is the vector of consequent parameters of the *i*th rule. Presented approach assumes that the antecedent fuzzy set of *i*th rule $A^{(i)}$ has a Gaussian membership function. For the input *x* the overall output of the fuzzy model is completed by a weighted averaging aggregation of outputs of individual rules (9) as

$$y = f(x, w^g) = \sum_{i=1}^{c} \overline{A^{(i)}(x)} \left(w^{(i)}\right)^T x' \qquad (10)$$

where $\overline{A^{(i)}(x)}$ is the normalized firing strength of the *i*th rule for the input *x* and $w^g = \left(\left(w^{(1)}\right)^T, \left(w^{(2)}\right)^T, \left(w^{(c)}\right)^T\right)^T$ denotes consequent parameters vector.

Consequent parameters $\{w^{(i)}\}_{i=1}^{c}$ can be computed in two different approaches: to solve one global problem, for all if-then rules, called global learning, or to solve *c* independent weighted problems, one for each if-then rule, called local learning. In the case of local learning, the variances of output values of the training set, required by the learning algorithm, can be established using following formula:

$$\sigma_n^2 = \left(A^{(i)}(x_n)\right)^{-p} \qquad \forall n \in \{1,2,...,N\} \;\; \forall i \in \{1,2,...,c\} \qquad (11)$$

where parameter *p* needs to be estimated during learning phase, using cross-validation procedure for instance.

3. CHARACTERIZATION OF ECG SIGNAL

The modern cardiology possesses wide knowledge about heart diseases and methods of its successful curing and preventing. The non-invasive electrocardiography became one of the basic methods of investigating electrical heart activity and analyzing the processes of treatment cardiovascular diseases. The electrocardiogram (ECG signal) is the mathematical representation of the voltage generated by heart muscle on the surface of body as a function of time (Figure 1).

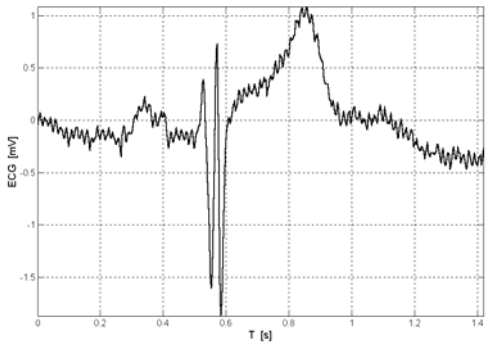

Fig. 1 The example of ECG physiological signals.

Most of the medical electrocardiographic systems involve the automatic processing of ECG signal delivered from the surface of chest. The quality of the signal often suffers from various

noises such as: the baseline wander (caused by varying electrode-skin impedance, patient's breath or movements), 50 or 60 Hz sinusoidal noise (caused by powerline interference) or muscle noise (caused by electrical activity of non-heart muscles). Thus using methods of noise reduction are very important and they have decisive influence on performance of all electrocardiographic (ECG) signal processing systems [1], [3], [7].

4. NUMERICAL EXPERIMENTS

The performances of the presented algorithm were investigated using the ECG test signal named ANE20000 from CTS-ECG database [15], which was proposed by the international electrotechnical commission (IEC) within the European project "common standards for quantitative electrocardiography" in order to test an accuracy of signal processing methods. The signal contains single heart beat sampled at 1000Hz with approximate duration 0.6 second.

To the original signal (Figure 2, on the left), denoted by t_n, there was added Gaussian noise s_n with standard deviation equal 0.2*std* (Figure 2, in the middle) and 1.0*std* (Figure 2, on the right), where *std* is the standard deviation of original signal.

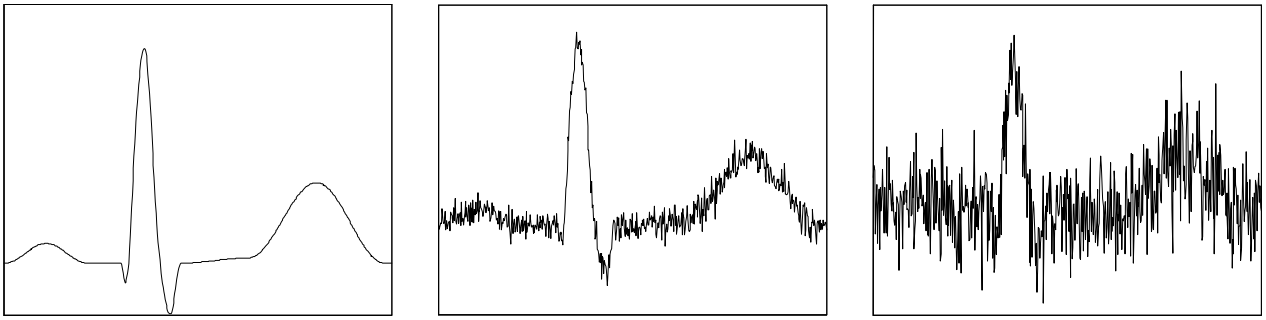

Fig. 2 The example of ECG signals with various level of noise.

For the training stage of the learning algorithm the set of input vectors $\{x_n\}$ was prepared in form: $x_n = (x_{n-M}, x_{n-M+1}, ..., x_n, ..., x_{n+M-1}, x_{n+M})$, where $x_i = t_i + s_i$ and parameter M is controlling width of moving window. Output values were taken as $y_n = t_n$ respectively for each n. The construction of the test set was similar with the same size and newly generated noise (with the same amplitude). In the experiment the parameter M took values from set {5, 10, 15, 20, 25}.The number of if-then rules c took values from set {2, 3, 4, 5, 10}. The experiments were performed using both local and global learning method. In case of local learning the parameter p took values from set {0.001, 0.005, 0.01, 0.015}.

All the experiments were run in the MATLAB environment. The generalization ability of the proposed Fuzzy Relevance Vector Machine algorithms, FRVM.0 (where all components of vector β are iteratively change) and FRVM.1 (where iteration procedure changes only parameter of scale s for components of vector β), were determined as a root mean squared error (RMSE) on the test set. The best results, with the values of parameters for which the lowest RMSE was obtained, for the amplitude of noise equal 0.2*std* are presented in Table 1 and for the amplitude of noise equal 1.0*std* are presented in Table 2. The tables also contain the best results obtained using Least Squares (LS) method for estimation consequent parameters vector $w^g = \left(\left(w^{(1)}\right)^T, \left(w^{(2)}\right)^T, \left(w^{(c)}\right)^T\right)^T$ both in local and global learning.

Table 1. The best results obtained for the amplitude of noise equal 0.2*std*.

Global learning		Local learning	
Learning method	RMSE	Learning method	RMSE
FRVM.0 (M = 5, c = 2)	13.1572	FRVM.0 (M = 5, c = 10, p = 0.001)	12.0076
FRVM.1 (M = 15, c = 2)	13.270e+03	FRVM.1 (M = 15, c = 2, p = 0.01)	11.8199
LS (M = 20, c = 5)	11.4139	LS (M = 20, c = 5)	12.0190

Table 2. The best results obtained for the amplitude of noise equal 0.2*std*.

Global learning		Local learning	
Learning method	RMSE	Learning method	RMSE
FRVM.0 (M = 10, c = 5)	40.6484	FRVM.0 (M = 15, c = 3, p = 0.001)	36.2079
FRVM.1 (M = 25, c = 2)	7.2998e+03	FRVM.1 (M = 20, c = 5, p = 0.015)	34.5200
LS (M = 10, c = 2)	39.8202	LS (M = 10, c = 10)	37.9267

In the case of FRVM.1 for global learning and all considered parameters the matrices, subject to inverse in the learning algorithm, were ill-conditioned (close to singular or badly scaled) which led to unacceptable levels of error. As it can be seen the all errors, both for the amplitude of noise equal 0.2*std* and the amplitude of noise equal 1.0*std*, were at least order of magnitude 10^3.

To compare results presented in Table 1 and Table 2 there were made noise reduction experiments using commonly used a low-pass moving average filter, having the following form:

$$y_n = \frac{1}{2M+1}\sum_{i=-M}^{M} x_{n+i} \tag{12}$$

where M was changed from 1 to 20, to the test dataset. The lowest RMSE for this method 12.121 was obtained for M = 5 while the amplitude of noise was equal 0.2*std* and 38.78 for M = 12 while the amplitude of noise was equal 1.0*std.*

5. CONCLUSION

In this work the new approach to learning Takagi-Sugeno-Kang fuzzy system was presented along with the application to noise reduction in ECG signal. Presented method uses the results of statistical learning theory and Bayesian methodology in the area of training fuzzy systems. The method of designing output functions in TSK system leads to sparsity in individual rules which results in improved generalization ability comparing with alternative methods.

It is also worth mentioning that the result of learning process is not only the obtained vector of parameters of prediction function, but also the predictive distribution $p(t_*|t,\alpha_{MP},\beta_{MP})$ for each new test point x_* and its σ_*. However the primary disadvantage of the sparse Bayesian method is the computational complexity of the learning algorithm. Although the presented update rules are relatively simple in form, they require memory and computation scale respectively with the square and cube of the number of elements in the training set. This implies that the algorithm becomes less practical to apply in case where the training examples number several thousands or more.

The results of numerical experiments show usefulness of the presented methods in the noise reduction in ECG signal in case of 50 or 60 Hz sinusoidal noise caused by powerline interference and muscle noise caused by electrical activity of non-heart muscles due to their similarity to Gaussian noise. Although in the case of the baseline wander the method, described above, can hardly be used because manifestation of this kind of noise usually appears in long period of time, over a few second, and the number of training example might be too large for the algorithm in presented form.

Nevertheless in the case of high level noise the results obtained by using the FRVM.1 for local learning appear to be much better than those obtained by using both the Least Squares method and commonly used low-pass moving average filter. Perhaps there is possibility to improve the results of noise reduction (achieve even lower error) using another set of learning parameters. Thus it seems to be important to research methods for automatic determination of the learning parameters such as number of rules c and in the case of local learning, exponent parameter p, as well as in this application, parameter M controlling width of time window.

BIBLIOGRAPHY

[1] ALSTE J.A. van, ECK W. van, HERRMANN O.E., ECG baseline wander reduction using linear phase filters, Comput. Biomed. Res. 19, pp. 417–427, 1986.
[2] CHIANG J.-H., HAO P.-Y., A New Kernel-Based Fuzzy Clustering Approach: Support Vector Clustering With Cell Growing, IEEE Transactions on Fuzzy Systems 11(4), pp. 518-527, 2003.
[3] FRANKIEWICZ Z., Methods for ECG signal analysis in the presence of noise, Ph.D. Thesis, Silesian Technical University, Gliwice, 1987.
[4] HONG D.H., HWANG C., Support Vector Fuzzy Regression Machines, Fuzzy Sets and Systems, 138(2), pp. 271-281, 2003.
[5] LIN C.-F., WANG S.-D. Fuzzy Support Vector Machine, IEEE Transaction on Neural Networks 13(2), pp. 464-471, 2002.
[6] ŁĘSKI J., Neuro-fuzzy system with learning tolerant to imprecision, Fuzzy Sets and Systems 138(2), pp. 427--439, 2003.
[7] ŁĘSKI J.M., HENZEL N., ECG baseline wander and powerline interference reduction using nonlinear filter bank, Signal Processing 85, pp. 781–793, 2005.
[8] MOMOT A., Uczenie bayesowskie w modelowaniu rozmytym, Ph.D. Thesis, Silesian Technical University, Gliwice, 2004.
[9] MOMOT A., Uczenie systemu rozmytego TSK z wykorzystaniem wnioskowania bayesowskiego, In Bazy Danych Modele, Technologie, Narzędzia: Analiza danych i wybrane zastosowania, pp.127-133, WKŁ, Warszawa, 2005.
[10] SUGENO M., KANG G.T., Structure identification of fuzzy model, Fuzzy Sets and Systems 28, pp. 15-33, 1988.
[11] TAKAGI T., SUGENO M., Fuzzy identification of systems and its application to modeling and control, IEEE Trans. on System, Man and Cybernetics 15(1), pp. 116-132, 1985.
[12] TIPPING M., The Relevance Vector Machine. In Advances in Neural Information Processing Systems 12, pp. 652 - 658, MIT Press, Cambridge, 2000.
[13] TIPPING M., Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research, 1(2), pp. 211 - 244, 2001.
[14] VAPNIK V.N., The nature of statistical learning theory. Springer, New York, 1995.
[15] International Electrotechnical Commission Standard 60601-3-2, 15 December 1999.