MULTICLASS LEAST SQUARES AUTO-CORRELATION WAVELET SUPPORT VECTOR MACHINES

Yongzhong Xing, Xiaobei Wu and Zhiliang Xu

School of Automation
University of Nanjing University of Science and Technology
Nanjing, 210094, P. R. China
xyz-1971@hotmail.com; wuxb@mail.njust.edu.cn

Abstract. *In this paper, combining the auto-correlation wavelet kernel with multiclass least squares support vector machine (MLS-SVM), a novel notion of multiclass least squares support vector machine with universal auto-correlation wavelet kernels (MLS-AWSVM) is proposed. The translation invariant property of the kernel function enhances the generalization ability of the LS-SVM method and the spiral multiclass classification experimental results show some advantages of MLS-AWSVM over MLS-SVM on the classification and the generalization performance.*
Keywords: Auto-correlation wavelet kernel, LS-SVM, Multiclass classification

1. **Introduction.** Support vector machine (SVM) has been a standard tool in the machine learning community because it nicely deals with high dimensional data, provides good generalization properties and determines the classifier architecture once kernel function and the parameters are chosen by user [1], [2]. SVM uses a kernel function to project the input data onto a high dimension feature space, and then constructs an optimal separating hyper-plane in that space. SVM is a kernel based method, which allows the use of Gaussian, polynomial and wavelet kernels and so on that satisfy Mercer's condition [3]. Least squares support vector machines (LS-SVM) is a SVM version which involves equality instead of inequality constraints and works with a least squares cost function [4]. In the LS-SVM, Mercer's condition is still applicable. A straightforward extension of LS-SVM to multiclass problems (MLS-SVM) has been proposed in [5], where the Gaussian kernel function is used. One issue with the kernel methods is finding an appropriate kernel for the given problem because different kernel function and parameters can have widely varying performance.

The wavelet theory and associated multiresolution techniques [6] has had tremendous impact not only in signal and image processing but also in science and engineering. Presently, there are several contributions to the theoretical development of wavelet kernel reported in the literature such as reproducing wavelet kernel[7],constructing translation invariant wavelet kernel[8]and auto-correlation wavelet kernel[9]. Auto-correlation of a compactly supported wavelet satisfies the translation invariant property. This property is very important in signal processing. The wavelet has a limitation on this. The wavelet transform generates very different wavelet coefficients even if the input signal is shifted a little bit. This limitation can be overcome by taking the auto-correlation on the wavelet function. Based on the property, any compactly supported wavelet function can be chosen to construct auto-correlation wavelet kernel, Daubechies-4 (D4) wavelet has been proven to perform the best for signal regression [9]. In this paper,combining MLS-SVM with the auto-correlation D4 wavelet kernel function, the multiclass least squares auto-correlation wavelet support vector machines (MLS-AWSVM) is proposed. The goal of

the MLS-AWSVM is to find the optimal classification in the space spanned by multivariable wavelet kernel. The spiral multiclass classification experimental results show some advantages of MLS-AWSVM over MLS-SVM on the classification and the generalization performance.

2. **Prepare Knowledge.**

2.1. **SVM for pattern recognition.** Given an independent identically distributed (i.i.d) training data set $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_k, y_k)\}$, where $\mathbf{x} \in R^n, y \in \{-1, 1\}$. For pattern recognition problem, by introducing Lagrange multiplier technique, the resulting decision function of the SVM takes the form

$$f(\mathbf{x}) = sgn[\sum_{i=1}^{k} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + \beta] \tag{1}$$

where α_i, are the Lagrange multipliers and β is the bias, a kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$ is called a support vector (SV) kernel if it satisfies a certain conditions, $\varphi(\cdot) = R^n \rightarrow R^h$, is a nonlinear projection function that dot products between projected vectors is computed by means of a kernel function. According to the different classification problems, the different kernel function can be selected to obtain the optimal classification results.[1]

2.2. **Wavelet theory and wavelet kernel.** The idea of wavelet analysis is to approach a signal or function using a family of functions which are produced by dilation and translation of the mother wavelet $\psi(x)$

$$\psi_{a,b}(x) = |a|^{-1/2} \psi(\frac{x-b}{a}) \tag{2}$$

where $x, a, b \in R, a \neq 0$ is dilation factor, and b is a translation factor. The wavelet translation of any function $f(x)$ can be expressed as

$$W_{a,b}(f) = \langle f(x), \psi_{a,b}(x) \rangle, f(x) \in L_2(R) \tag{3}$$

where the notation $\langle \cdot, \cdot \rangle$ refers the inner product in $L_2(R)$. Equation (3) means that any function $f(x)$ can be decomposed on wavelet basis $\psi_{a,b}(x)$ if it satisfies the condition [8][11].

$$C_f = \int_{0}^{+\infty} \frac{|H(\omega)|^2}{|\omega|} d\omega < \infty \tag{4}$$

where $H(\omega)$ is Fourier transform of $\psi_{a,b}(x)$. The function $f(x)$ can be reconstructed as follows

$$f(x) = (C_f)^{-1} \int_{-\infty}^{+\infty} \int_{0}^{+\infty} W_{a,b}(f) \psi_{a,b}(x) \frac{da}{a^2} db \tag{5}$$

To approximate Equation (5) [6], then the finite can be written as

$$\hat{f}(x) = \sum_{i=1}^{k} W_i \psi_{a_i, b_i}(x). \tag{6}$$

Here, $f(x)$ is approximated by $\hat{f}(x)$.

For a common multidimensional wavelet function, the mother wavelet can be given as the product of one-dimensional (1-D) wavelet functions [11]:

$$\psi_n(\mathbf{x}) = \prod_{i}^{n} \psi(x_i). \tag{7}$$

where $\mathbf{x} = (x_1, \cdots, x_n) \in R_n$. So, every 1-D mother wavelet $\psi(x)$ must satisfy (4). The wavelet kernel is defined as

$$K(\mathbf{x}, \mathbf{x}^{\cdot}) = \prod_{i}^{n} \psi(\frac{x_i - x_i^{\cdot}}{a}). \tag{8}$$

2.3. **Auto-correlation wavelet kernel.** Wavelets satisfy a multiresolution analysis and they also obey the following relations:

$$\varphi(x) = 2^{1/2} \sum_{m=0}^{N-1} h_m \varphi(2x - m), \tag{9}$$

and

$$\psi(x) = 2^{1/2} \sum_{m=0}^{N-1} g_m \varphi(2x - m). \tag{10}$$

where $g_m = (-1)^m h_{N-m-1}, m = 0, \cdots, N-1$. The auto-correlation is defined by

$$\Phi(x) = \int_{-\infty}^{+\infty} \varphi(t)\varphi(t - x)dt, \tag{11}$$

$$\Psi(x) = \int_{-\infty}^{+\infty} \psi(t)\psi(t - x)dt. \tag{12}$$

It can be derived that

$$\Phi(x) = \Phi(2x) + \frac{1}{2} \sum_{n=1}^{N/2} a_{2n-1}(\Phi(2x - 2n + 1) + \Phi(2x + 2n - 1)), \tag{13}$$

$$\Psi(x) = \Phi(2x) - \frac{1}{2} \sum_{n=1}^{N/2} a_{2n-1}(\Phi(2x - 2n + 1) + \Phi(2x + 2n - 1)). \tag{14}$$

where $\{a_m\}$ are the auto-correlation coefficients of the filter $\{h_1, \cdots, h_{N-1}\}$,
$a_m = 2 \sum_{n=0}^{N-m-1} h_n h_{n+m}$, for $m = 1, \cdots, N-1$ and $a_{2m} = 0$, for $m = 1, \cdots, N/2 - 1$.
It is not difficult to find that both Φ and Ψ have support of $[-N + 1, N - 1]$.

A translation invariant kernel $K(\mathbf{x}, \mathbf{x}^{\cdot}) = K(\mathbf{x} - \mathbf{x}^{\cdot})$ is an admissible SV kernel if and only if its Fourier transformation is non-negative [8]. This can be satisfied by defining the following auto-correlation wavelet kernel [9]:

$$K(\mathbf{x}, \mathbf{x}^{\cdot}) = \prod_{i=1}^{l} (\Psi(\frac{x_i - x_i^{\cdot}}{a})). \tag{15}$$

where l is the dimension of the input feature vector and a is the scale factor. It should be mentioned that we can choose any compactly supported wavelet function to construct auto-correlation wavelet kernel $K(\mathbf{x}, \mathbf{x}^{\cdot})$.

The wavelet function used here does not have an explicit form. In order to generate it, we need to set one wavelet coefficient to 1 and all the rest coefficients to 0. An inverse wavelet transform generates the desired wavelet function depending on the selected input wavelet filter. Since the wavelet function has an implicit form, we save it in memory as one dimensional array with a relatively large number of sample points. This array needs to be generated only once and then saved for later use. It can be easily proved that this kernel

$K(\mathbf{x}, \mathbf{x}^{\prime})$ is an admissible SV kernel. Fig.1 shows the D4 wavelet and its auto-correlation kernel.
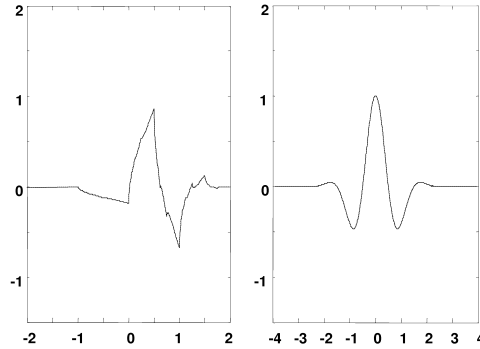

FIGURE 1. Daubechies-4 wavelet and its auto-correlation kernel

3. **Multiclass Least Squares Auto-correlation Wavelet Support Vector Machines.** Let $\{\mathbf{x}_i, y_i^{(k)}\}_{i=1, ;k=1}^{i=N, k=m}$ be training data set, m be number of classification, N be number of training data set, $\mathbf{x}_i = \{x_{i1}, \cdots, x_{in}\}$ input index, $y_i^{(k)}$ output index, and $y_i^{(k)} = 1$ means the ith input vector belongs to the kth class, $y_i^{(k)} = -1$ not. The derivation of the multi-class wavelet LS-SVM is based upon the formulation [5]

$$\min_{W_k, \beta_k, \xi_{ik}} J^{(m)}(W_k, \beta_k, \xi_{ik}) = \frac{1}{2}(\sum_{k=1}^{m} W_k^T W_k + C \sum_{k=1}^{m} w_k \sum_{i=1}^{N} \xi_{ik}^2) \qquad (16)$$

with the equality constraints

$$\begin{cases} y_i^{(1)}[W_1^T \varphi_1(x_i) + \beta_1] = 1 - \xi_{i1} \\ y_i^{(2)}[W_2^T \varphi_2(x_i) + \beta_2] = 1 - \xi_{21} \\ \cdots, \cdots \\ y_i^{(m)}[W_m^T \varphi_m(x_i) + \beta_m] = 1 - \xi_{im} \end{cases} \qquad (17)$$

where $i = 1, \cdots, N$, $W \in R^h$ is the weight vector, $C > 0$ is the regularization factor [10], $W_k \in R$ is the weight of kth classification error. $\xi_{ik} \in R$ is the classification eror and $\beta_k \in R$ is the bias. The corresponding Lagrange equation is

$$L^{(m)}(W_k, \beta_k, \xi_{ik}, \alpha_{ik}) = J^{(m)} - \sum_{i,k} \alpha_{ik}\{y_i^k[W_k^T \psi_{(k)}(x_i) + \beta_k] - 1 + \xi_{ik}\} \qquad (18)$$

The solution concludes in a constrained optimization with the equality conditions, with the expunction of W_k and ξ_{ik}, one can get the linear system:

$$\begin{bmatrix} 0 & Y_M^T \\ Y_M & \Omega_M \end{bmatrix} \begin{bmatrix} \beta_M \\ \alpha_M \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{1} \end{bmatrix} \qquad (19)$$

with given matrices
$\bar{1} = [1, \cdots\cdots, N]; \beta_M = [\beta_1, \cdots\cdots, \beta_m]; \alpha_M = [\alpha_{11}, \cdots\cdots, \alpha_{N1}, \cdots\cdots, \alpha_{1m}, \cdots\cdots \alpha_{Nm}];$
$\Omega_M = blockdiag\{\Omega^{(1)}, \cdots\cdots, \Omega^{(m)}\}; \Omega_{il}^{(k)} = y_i^{(k)} y_l^{(k)} \varphi_k^T(x_i)\varphi_k(x_l) + C^{-1}I;$

$$Y_M = blockdiag\{ \begin{bmatrix} y_1^{(1)} \\ \vdots \\ y_N^{(1)} \end{bmatrix}, \cdots\cdots, \begin{bmatrix} y_1^{(m)} \\ \vdots \\ y_N^{(m)} \end{bmatrix} \}.$$

Let $K_k(x_i, x_l) = \varphi_k^T(x_i)\varphi_k(x_l)$ be Daubechies-4 auto-correlation wavelet kernel function, then

$$K_k(x_i, x_l) = \prod_{t=1}^{n}(\Psi(\frac{x_{it} - x_{lt}}{a})), \; k, i, l = 1, \cdots\cdots, N \tag{20}$$

Decision function of multiclass auto-correlation wavelet kernel support vector machines (MLS-AWSVM) is

$$f(x) = sgn[\sum_{i=1}^{N} \alpha_{ik}y_i^k \prod_{t=1}^{n}(\Psi(\frac{x_{it} - x_{lt}}{a_i^t})) + \beta_k], \; k = 1, \cdots\cdots, m \tag{21}$$

Here, MLS-SVM can adopt the Daubechies-4 auto-correlation wavelet kernel as its kernel function. It is difficult to determine $N \times n$ parameters , for the sake of simplicity, let $a_i^t = a_k$, the parameter a_k of the auto-correlation kernel can be obtained by means of cross-validation.

4. **Numerical Example.** Now, we validate the performance of MLS-AWSVM by a small illustrative example on a simple spiral problem for which four classes have been defined. Figure.2 shows $N = 60$ training data points with 4 classes that each contain 15 data points (outputs equal to $[+1; +1], [+1; -1], [-1; +1], [-1; -1]$). The four classes have been encoded by taking $m = 2$. Note that the range of display is limited to $[-0.7, 0.7] \times [-0.7, 0.7]$. We let $w_k = 1(k = 1, \cdots, m)$ and $C = 1$ for all the following examples.

For comparison, we show the results obtained by the auto-correlation wavelet kernel and Gaussian kernel, respectively. The Gaussian kernel's expression is $K_k(x_i, x_l) = \exp\{-\|x_i - x_l\|_2^2/2\sigma_k^2\}$, where σ_k is kernel's parameter chosen by user. Figure. 2 show the estimated separating hyperplanes based on the Gaussian kernel (MLS-GSVM) using parameters $\sigma_1^2 = \sigma_2^2 = 0.5$. Figure. 3 show the estimated separating hyperplanes based on auto-correlation wavelet kernel (MLS-AWSVM) using parameters $a_1 = a_2 = 2$. Table 1 shows the misclassification ratio in leave-one-out cross-validation of each condition.

Simulation results show that, compared with MLS-GSVM, the recognition precision and the generalization ability is improved by our MLS-AWSVM.
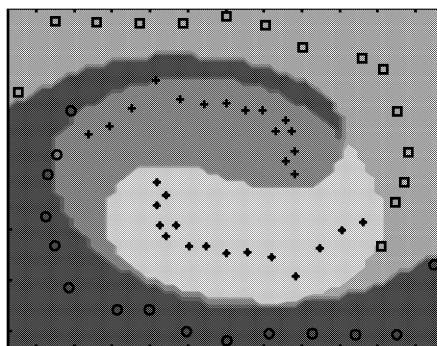

FIGURE 2. The classification result based on the MLS-GSVM for spiral four class problem with $\sigma_1^2 = \sigma_2^2 = 0.5$
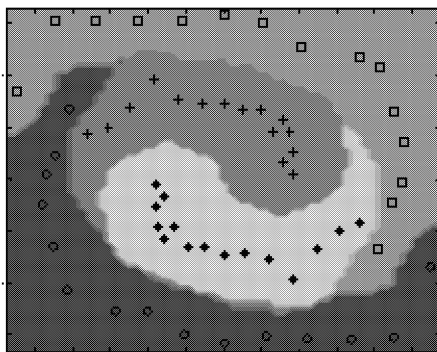
FIGURE 3. The classification result based on the MLS-AWSVM for spiral four class problem with $a_1 = a_2 = 2$

TABLE 1. Misclassification ratio in leave-one-out cross-validation

Condition	Misclassification ratio
MLS-GSVM with $\sigma_1^2 = \sigma_2^2 = 0.5$	3.3
MLS-AWSVM with $a_1 = a_2 = 2$	0.0

5. **Conclusions.** In this paper, we discussed a practical way to construct auto-correlation wavelet kernel using a compactly supported wavelet function. The wavelet kernel is a kind of multidimensional function that can approximate arbitrary functions. A new MLS-SVM version is presented based on MLS-SVM and the auto-correlation wavelet kernel, namely MLS-AWSVM. The MLS-AWSVM is applied to spiral classification problem. Simulation shows that the wavelet kernel has better classification than the Gaussian kernel. Notice that the wavelet kernel is orthonormal approximately, whereas the Gaussian kernel is not. The Gaussian kernel is correlative or even redundancy. Thus, for many real life applications MLS-AWSVM can offer a faster method for obtaining classifier with better generalization performance than MLS-GSVM.

REFERENCES

[1] V. Vapnik , *The Nature of Statistical Learning Theory*, New York: Spring - Verlag, 1995.
[2] N. Y. Deng and Y. J Tian, *The Method of Data Mining-support Vector Machine*, Science Press, Beijing, 2004.
[3] A. Smola, B. Schölkopf and K.-R. Müller, The connection between regularization operators and support vector kernels, *Neural Network*, vol.11, pp.637-649, 1998d.
[4] J. A. K. Suykens and J. Vandewalle, Least squares support vector machines classifiers,*Neural Processing Letters*, vol.9, no.3, pp.293-300, 1999.
[5] J. A. K. Suykens and J. Vandewalle, Multiclass least squares support vector machines, *Proc. of the International Joint Conference on Neural Networks*, Washington DC, USA, pp.900-903, 1999.
[6] G. Z. Liu and Y. Tian, *Wavelet Analysis and Application*, Xi'an, China: Xidian Univ Press, 1992.
[7] A. Rakotomamonjy and S. Canu, Frame, reproducing kernel,regularization, *Journal of Machine Learning Research*, vol.6, pp.1485-1515, 2005.
[8] L. Zhang, W. Zhou and L. Jiao, Wavelet support vector machine, *IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics*, vol.34. no.1, pp.34-39, 2004.
[9] G. Y. Chen and G. Dudek, Auto-correlation wavelet support vector machine and its application to regress, *Proc. of the Second Conference on Computer and Robot Vision*, British, Columbia, 2005.
[10] J. H. Xu, X. G. Zhang and Y. D. Li, Regularized kernel forms of minimum square error methods, *Acta Automat. Sin.*, vol.30, no.1, pp.27-36, 2004.
[11] Q. H. Zhang and A. Benveniste, Wavelet networks, *IEEE Transaction on Neural Networks*, vol.3, pp.889-898, 1992.