A Relevance Vector Machine Prediction Method Based on the Biased Wavelet Kernel Function

Fang LIU[1], Fei ZHAO[1], Zhen-Hao YU[1], Cui ZHANG [1]

[1]*National Engineering Laboratory for Fiber Optic, Sensing Technology, Wuhan University of Technology, Wuhan, China;*
School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China;
Key Laboratory of Fiber Optic Sensing Technology and Information Processing of Ministry of Education, Wuhan University of Technology, Wuhan, China
fangliu@whut.edu.cn, Knight_zf@126.com

Abstract—Relevance Vector Machine (RVM) is an important learning method in the field of machine learning for its sparsity, global optimality and the ability to solve nonlinear problems by using kernel functions. In this paper, a family of biased wavelets was used to construct the kernel functions of RVM. Biased wavelet have adjustable nonzero mean which makes the kernel of RVM more flexible. With the kernel method of the Centered Kernel Target Alignment (CKTA), the biased parameter was selected to improve the prediction performance of RVM model. The algorithm based on the biased wavelet kernel showed an increased prediction accuracy compared to using wavelet kernel and Cauchy kernel. In short, Relevance Vector Machine with the biased wavelet kernel is a flexible prediction algorithm with high prediction accuracy.

1 Introduction

RVM proposed by Tipping [1] is a Bayes probability model, and its kernel does not need to satisfy the Mercer conditions [2]. Based on the favorable features including its sparseness, Bayesian properties, and kernel characteristics [3], RVM is one of the famous sparse Bayesian learning models [4, 5]. Similar to the support vector machine (SVM) model [6], the effect of the RVM depends on the kernel function and kernel parameters [7]. At present, the methods for choosing an effective kernel function and reasonable kernel parameters are still imperfect [8, 9].

The kernel matrix can be learned from data via semi-definite programming (SDP) techniques in [10]. Cristianini, Shawe-Taylor, Elisseeff and Kandola [11] proposed a quantity measure named as 'kernel target alignment' (KTA) to adapt the kernel matrix to sample labels, and a series of algorithms are derived for clustering, transduction, kernel combination [12] and kernel selection [13]. However, KTA is only a sufficient condition to be a good kernel matrix, but not a necessary condition. It is possible for a kernel matrix to have a very good performance even though its KTA is still low [14]. CKTA proposed by Marina [15] is better than KTA with several experiments.

Based on the characteristics of the biased wavelet, it is suitable to be the kernel function of RVM. Wavelet analysis, which can efficiently overcome the shortcomings of Fourier analysis and other analysis tools, is becoming a focus point of many sciences. The zero-mean characteristic of wavelets often drives the phenomenon that a large number of multiresolution levels are needed to reduce the L^2 norm of the approximation error. In order to reduce the redundancy, biased wavelet was proposed by Galvao [16].

In this paper, the biased wavelet was constructed as the kernel function of RVM. CKTA was used to optimize the biased parameters with a fixed scale parameter. We evaluated the prediction accuracy with the biased wavelet kernel, Cauchy kernel and wavelet kernel by using the data from fiber Bragg grating (FBG) temperature sensor system.

With a short introduction of RVM and biased wavelet, the biased parameters were filtrated by CKTA to build the biased wavelet kernel in section 2. Experiments are presented in section 3 and the conclusions in section 4 concludes the paper.

2 Methods

RVM is a fully probabilistic framework and introduce a prior over the model weights governed by a set of hyperparameters, one associated with each weight, whose most probable values are iteratively estimated from the data.

Let x be the input data and y the output data. A point $y(x)$ can be predicted by:

$$y(x)=\sum_{i=1}^{n} w_n \cdot k(x,x_n)+w_0 \tag{1}$$

Where $\{w_n\}$ is the weight vector and $k(x,x_n)$ is a kernel function. n is the length of weight vector and w_0 the measurement noise.

The noise vector w_0 is assumed to be normally distributed with zero mean and a variance of σ^2. Using Bayes' rule, the posterior distribution of w ,α ,σ^2 with y can be computed:

$$P\left(w,\alpha,\sigma^2 \middle| y\right)=\frac{P\left(y \middle| w,\alpha,\sigma^2\right)P\left(w,\alpha,\sigma^2\right)}{P(y)}=\mathrm{N}(\mu,\Sigma) \tag{2}$$

Where

$$\mu=\sigma^{-2}\Sigma\Phi^T y \tag{3}$$

$$\Sigma=\left(A+\sigma^{-2}\Phi^T\Phi\right)^{-1} \ \omega\iota\tau\eta \ A=diag(\alpha) \tag{4}$$

The parameters a corresponds a zero-mean Gaussian distribution over w , which avoids overfitting. The hyperparameters (α and σ^2) is optimized iteratively by maximizing the posterior probability $P\left(w,\alpha,\sigma^2 \middle| y\right)$.

After learning the mean and variance, the results of (3) is applied to (1):

$$\hat{y}=\mu^T\Phi \ , \ \hat{\sigma}^2=\sigma_{MP}^2+\Phi^T\Sigma\Phi \tag{5}$$

The predicted variance $\hat{\sigma}^2$ is the sum of the variance caused by the measurement noise σ_{MP}^2 and the uncertainly in the prediction of w.

Kernel functions are important for prediction performance. Methods of selecting the kernel functions, such as experience [17], comparison [18] and multi-kernels [19, 20], have been developed in practical application.

Due to the characteristics of the biased wavelet, it can describe the whole data, as well as the details. Therefore, when the biased wavelet is used as a kernel of RVM, the feature space is able to be close to the target space by dynamically adjusting the biased parameters for different types of data, which can improve the prediction accuracy of the RVM model.

A biased wavelet function u satisfying

i. $u \in L^2(R)$

ii. $u(0) \neq 0$

iii. $u(\tau)$ is rapidly decreasing to zero when $\tau \to \infty$

iv. $\hat{u}(\omega)$ is rapidly decreasing to zero when $\omega \to \infty$

In this paper, we used the third type of the biased wavelet (6) and the Mexican Hat (7) as the mother wavelet.

$$u(x)=\exp\left(-\frac{x^2}{2}\right) \tag{6}$$

$$\psi(x)=\frac{2}{\sqrt{3}}\pi^{-1/4}\left(1-x^2\right)\cdot\exp^{-x^2/2} \tag{7}$$

Then, a set of biased wavelet kernels was defined by

$$k(\tau,\mathrm{b})=|\sigma|^{-1/2}\left[\psi\left(\frac{\tau-\mathrm{b}}{\sigma}\right)+cu\left(\frac{\tau-\mathrm{b}}{\sigma}\right)\right] \tag{8}$$

Where σ is scale parameter, b is translation parameter, c is biased parameter and τ is a continuous real variable.

Since there is no explicit form for the mapping function, the learning algorithm of RVM get the information of the feature space, model, the training data and their relationship from the Gram matrix. KTA based on Gram matrix is considered as an effective way to filter biased parameters.

Let $\mathrm{X}=\{\mathrm{x}_k\}_{k=1}^{N}$ be the input data and the corresponding output data was $Y=\{y_k\}_{k=1}^{N}$. The range of the index was $\mathrm{i,j}\in[1,N]$ and the Gram matrix of kernel was defined by

$$[MK]_{i,j}=k(x_i,x_j) \tag{9}$$

The target matrix was defined by

$$[\mathrm{Y}]_{i,j}=y_i\cdot y_j \tag{10}$$

KTA was defined by

$$A(MK,Y)=\frac{\langle MK,Y\rangle_F}{\sqrt{\langle MK,MK\rangle_F\cdot\langle Y,Y\rangle_F}} \tag{11}$$

Where

$$\left\langle MK, Y \right\rangle_F = \sum_{\mathrm{i,j=1}}^{n} MK\left(x_i, x_j\right) Y\left(x_i, x_j\right) \qquad (12)$$

The purpose of the KTA is to calculate the degree of alignment between the Gram matrix and the target matrix. However, if, in the feature space, the origin is far away from the convex hull of the data, then the elements of MK have about the same value and, as a result, the matrix MK is ill-conditioned. Therefore, CKTA, a better method, was proposed by Marina.

From given kernel k, the centered kernel was defined by

$$\begin{aligned} k_C(x_i, x_j) &= \left\langle \Phi(x_i) - \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i), \Phi(x_j) - \frac{1}{n}\sum_{j=1}^{n}\Phi(x_j) \right\rangle \\ &= k\left(x_i, x_j\right) - \frac{1}{n}\sum_{i=1}^{n} k\left(x_i, x_j\right) - \frac{1}{n}\sum_{j=1}^{n} k\left(x_i, x_j\right) \qquad (13) \\ &+ \frac{1}{n^2}\sum_{i,j=1}^{n} k\left(x_i, x_j\right) \end{aligned}$$

The centered Gram matrix was defined by

$$[MK_C]_{i,j} = k_C(x_i, x_j) \qquad (14)$$

Similar to KTA, CKTA was defined by

$$A(MK_C, Y) = \frac{\langle MK_C, Y\rangle_F}{\sqrt{\langle MK_C, MK_C\rangle_F \cdot \langle Y, Y\rangle_F}} \qquad (15)$$

Compared with KTA, CKTA has the advantage of solving the problem of unbalanced data and the invariance of the linear transform.

Based on CKTA, the filtering strategy of biased parameters was shown in Fig. 1. The target biased parameter of the maximum CKTA was used to construct the final selected kernel. Our experiments showed that the relationship between biased parameters and values of CKTA was not monotonic, which meant that the target biased parameter could be found within a certain range.

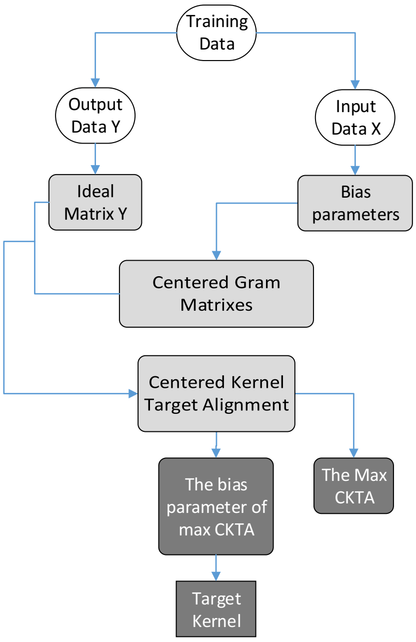

Figure 1. Flowchart representation of the selection of the bias parameter

3 Experiments

3.1 Data Set and Error Measures

The dataset contains 1440 instances from 120 hours of responses from FBG temperature sensor system for the Second Yangtze River Bridge in Wuhan.

The prediction algorithms were evaluated with respect to the mean relative error (MRE), the mean absolute error (MAE) and the root-mean-square error (RMSE).

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - \hat{y}_i\right|}{\left|y_i\right|} \qquad (16)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \qquad (17)$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}\frac{\left(y_i - \hat{y}_i\right)^2}{n}} \qquad (18)$$

where y_i is the true value, $\hat{y}_i$ is the predictive value and n is the length of test samples involved.

DOI: 10.1051/itmconf/20171205002

3.2Selection Method of Biased Wavelet Kernel

The wavelength responses from sensors were selected as the output data, and time series were the input data.

Fig. 2 showed the relationship between the CKTA value and biased parameters c, when 48 hours of data were used as the training data and the next 50 samples were test data. It could be seen from the figure that the CKTA reached its maximum when the Biased Parameter was -1.3. Because of the monotonic of the connection, the best biased parameter was generally found in the range $[-10,10]$. Therefore, the target biased wavelet kernel was filtered out.

Compared with the biased wavelet kernel, Cauchy kernel and the Wavelet kernel by using the Mexican Hat (7) as the mother wavelet, TABLE I. listed the prediction results of the MAE, MRE and RMSE with different length of training set from 96 hours of data. The next 50 samples after every length of training set were taken as test set. Test responses indicated that performance enhancement could be obtained by using the biased wavelet kernel.

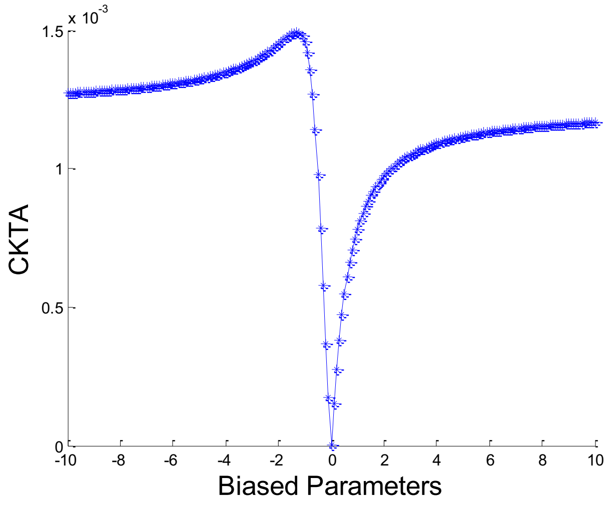

Figure 2. The connection between CKTA and values of biased parameters.

TABLE I. Wavelength variation estimation performances of rvm models

Hours		24h	48h	72h	96h
CKTA	Wavelet	6.842e-06	3.305e-06	2.384e-06	1.847e-06
	Cauchy	2.664e-03	1.111e-03	6.460e-04	4.809e-04
	Biased Wavelet	2.820e-03	1.495e-03	6.836e-04	5.094e-04
MAE (pm)	Wavelet	35.699	9.077	4.186	4.534
	Cauchy	13.016	7.541	4.148	2.587
	Biased Wavelet	10.407	6.861	3.738	2.555
MRE (pm)	Wavelet	0.143	0.036	0.019	0.012
	Cauchy	0.051	0.030	0.017	0.011
	Biased Wavelet	0.041	0.027	0.015	0.010
RMSE (pm)	Wavelet	36.820	9.305	5.359	3.475
	Cauchy	14.755	7.776	5.009	2.958
	Biased Wavelet	12.229	7.042	4.389	2.948

Conclusions

In this paper, the kernel function of RVM was constructed by biased wavelet and the optimization of the kernel parameters was investigated. For the adjustable nonzero mean of biased wavelet, the biased wavelet kernel is a flexible function. The CKTA method was used to optimize the parameters of RVM kernel. The biased wavelet kernel function can be adjusted by changing the parameters to maximum the CKTA. Experimental results showed a higher prediction accuracy by the biased wavelet kernel function.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61402345). This work was supported by the Fundamental Research Funds for the Central Universities (WUT:2017IVB072).

References

[1] Tipping M E. Sparse bayesian learning and the relevance vector machine[J]. Journal of Machine Learning Research, 2001, 1(3):211-244.

[2] Vapnik, V.N. The Nature of Statistical Learning Theory,2nd ed. ; Springer; New York, NY, USA, 2000.

[3] Son Y, Lee J. Active Learning Using Transductive Sparse Bayesian Regression[J]. Information Sciences, 2016, 374:240-254.

[4] Wu Y, Breaz E, Gao F, et al. Nonlinear Performance Degradation Prediction of Proton Exchange Membrane Fuel Cells Using Relevance Vector Machine[J]. IEEE Transactions on Energy Conversion, 2016:1-1.

[5] Lin Y, Xia K, Jiang X, et al. Landslide Susceptibility Mapping Based on Particle Swarm Optimization of Multiple Kernel Relevance Vector Machines: Case of a Low Hill Area in Sichuan Province, China[J]. 2016, 5(10):191.

[6] Wang X Y, Liang L L, Li W Y, et al. A new SVM-based relevance feedback image retrieval using probabilistic feature and weighted kernel function ☆[J]. Journal of Visual Communication & Image Representation, 2016, 38:256-275.

[7] Fei S W. Kurtosis prediction of bearing vibration signal based on wavelet packet transform and Cauchy kernel relevance vector regression algorithm[J]. Advances in Mechanical Engineering, 2016, 8(9).

[8] Close R, Wilson J, Gader P. A Bayesian approach to localized multi-kernel learning using the relevance vector machine[C]// Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International. IEEE, 2011:1103-1106.

[9] Gönen M, Alpaydın E. Localized algorithms for multiple kernel learning[J]. Pattern Recognition, 2013, 46(3):795-807.

[10] Lanckriet G R G, Cristianini N, Bartlett P, et al. Learning the Kernel Matrix with Semi-Definite Programming[J]. Journal of Machine Learning Research, 2002, 5(1):323-330.

[11] Cristianini N, Shawe-Taylor J, Elisseeff A, et al. On kernel-target alignment[C]// International Conference on Neural Information Processing Systems: Natural and Synthetic. MIT Press, 2001:367--373.

[12] Lei Y. A Relevance Vector Machine Prediction Method Based on Adaptive Multi-kernel Combination and Its Application to Remaining Useful Life Prediction of Machinery[J]. Journal of Mechanical Engineering, 2016, 52(1):87.

[13] Trafalis T B, Malyscheff A M. Optimal selection of the regression kernel matrix with semidefinite programming[M]// Frontiers in Global Optimization. Springer US, 2004:575-584.

[14] Nguyen C H, Ho T B. An efficient kernel matrix evaluation measure[J]. Pattern Recognition, 2008, 41(11):3366-3372.

[15] Marina M A. Data Centering in Feature Space[J]. Ninth International Workshop on Artificial Intelligence & Statistics, 2002.

[16] Galvão R K H, Yoneyama T, Rabello T N. Signal Representation by Adaptive Biased Wavelet Expansions[J]. Digital Signal Processing, 1999, 9(4):225-240.

[17] Fei S W, He Y. Wind speed prediction using the hybrid model of wavelet decomposition and artificial bee colony algorithm-based relevance vector machine[J]. International Journal of Electrical Power & Energy Systems, 2015, 73:625-631.

[18] Zhao C H, Zhang Y, Wang Y L. Relevant Vector Machine Classification of Hyperspectral Image Based on Wavelet Kernel Principal Component Analysis[J]. Dianzi Yu Xinxi Xuebao/journal of Electronics & Information Technology, 2012, 34(8):1905-1910.

[19] Nen M, Alpayd&#, Ethem N. Multiple Kernel Learning Algorithms[J]. Journal of Machine Learning Research, 2011, 12:2211-2268.

[20] Li D, Wang J, Zhao X, et al. Multiple kernel-based multi-instance learning algorithm for image classification[J]. Journal of Visual Communication & Image Representation, 2014, 25(5):1112-1117.