ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

A Ridgelet Kernel Regression Model using Genetic Algorithm

Shuyuan Yang[1], Min Wang[2], Licheng Jiao[1]*

[1] Institute of Intelligence Information Processing, Department of Electrical Engineering
Xidian University Xi'an, China, 710071
[2] National Lab of Radar Signal Processing, Department of Electrical Engineering
Xidian University Xi'an, China, 710071

Abstract. In this paper, a ridgelet kernel regression model is proposed for approximation of high dimensional functions. It is based on ridgelet theory, kernel and regularization technology from which we can deduce a regularized kernel regression form. Taking the objective function solved by quadratic programming to define the fitness function, we use genetic algorithm to search for the optimal directional vector of ridgelet. The results indicate that this method can effectively deal with high dimensional data, especially those with certain kinds of spatial inhomogeneities. Some illustrative examples are included to demonstrate its superiority.

1 Introduction

In Machine Learning(ML), many problems can be reduced to the tasks of multivariate function approximation (MVFA). MVFA is an active subject that has attracted lots of researching interests in many science and engineering communities[1,2]. Depending on the community involved, it goes by different names including nonlinear regression, function learning, system identification and others. The numeric methods for MVFA have been deeply studied in mathematics and computer science[3,4]. As we all know, approximation of a function from sparse samples is ill-posed, so one often assumes the function to be smooth to obtain a certain solution. However, in practical cases such as industrial control systems, fault detection, system identification and intelligent predicting, most systems are very complex MIMO(multi-input and multi-output) systems. They are equivalent to non-smooth mapping in high dimension, that is, they are MVFAs with spatial inhomogeneities. So the classical mathematical methods cannot estimate or approximate them efficiently by sparse samples.

Reconstructing a function by a superposition of some basis functions is a very inspiring idea on which many regressor are based, such as Fourier Transform (FT)[5], Wavelet Transform(WT)[6], Neural Network(NN)[7] and Projection Pursuit Regression (PPR)[8]. However, FT has a poor performance for singular functions; WT can deal with one-dimension singularity, but it can't extend to higher dimension. NN allows for non-smooth mapping in high dimension, but it inevitably has some disadvantages such as overfitting, slow convergence and too much reliance on experience. PPR is a regression way for smooth functions in a greedy fashion. For PPR converges

* This work was supported by the National Science Foundation (grant no. 60372050).

according to norm, it is of slow convergence. In 1996, *E.J.Candes* developed a new system to represent arbitrary functions by a superposition of specific ridge functions in a more stable way, the ridgelet[9]. Ridgelet proves to be good basis in high dimension, and it has optimal property for functions with spatial inhomogeneities. To adaptively estimate high dimensional functions, ridgelets can be used in PPR or NN to construct a regressor. However, they only minimize the experience risk, which leads to some drawbacks such as overfitting and bad generalization[10].

Recently Kernel Machine(KM) has been a standard tool in ML, which has stricter mathematical foundation than NN and PPR[11]. Based on ridgelet and KM, a ridgelet kernel regression model for MVFA is proposed. The minimum squared error(MSE) based on kernels and regularization technology, or the regularized kernel form of MSE, is adopted[12]. Employment of ridgelet can accomplish a wider range of MVFA, and the regularized items in objective function are used to improve the generalization of solutions. To get the directions of ridgelets, genetic algorithm (GA) is used.

2 Ridgelet Kernel Regression based on Genetic Algorithm

2.1 Ridgelet Regression

Given a pair of sample set **S**={(x, y)}(perhaps polluted by noise) coming from the model *Y=f*(*X*), a nonparametric regression is to estimate the unknown function *f* from the sparse data set **S**. Kernel smoothing, nearest-neighbor and spline smoothing are often used in regression. However, their performances decrease sharply in high dimension due to the 'curse of dimensionality', that is, if samples are not dense enough, one will get a bad mean squared error. Ridgelet is a new harmonic analysis tool, and it proves to be optimal for estimating multivariate regression surfaces especially for those exhibiting specific sorts of spatial inhomogeneities, with a speed rapid than FT and WT. As an extension of wavelet to higher dimension, ridgelet has attracted more and more attention of researches. It is defined as:

If $\psi: R^d \to R$ satisfies the condition：

$$K_\psi = \int \frac{|\hat{\psi}(\xi)|^2}{|\xi|^d} d\xi < \infty \tag{1}$$

Then we call the functions $\psi_\tau(x) = a^{-1/2}\psi(\frac{u \cdot x - b}{a})$ as ridgelets. Parameter $\tau = (a,u,b)$ belongs to a space of neurons $\Gamma = \{\tau = (a,u,b), a,b \in R, a > 0, u \in S^{d-1}, \|u\|^2 = 1\}$. Denote the surface area of unit sphere $S^{d\text{-}1}$ in *d*-dimension as σ_d. Then for any function $f \in L^1 \cap L^2(R^d)$, it can be expanded as a superposition of ridgelet functions:

$$f = c_\psi \int <f, \psi_\tau> \psi_\tau \mu(d\tau) = c_\psi (2\pi)^{-d} K_\psi^{-1} \int <f, \psi_\tau> \psi_\tau \sigma_d da / a^{d+1} du db \tag{2}$$

For function $f(x): R^d \to R^m$, it can be divided into *m* mappings of $R^d \to R$. Select the ridgelet as the basis, then the following approximation equation is obtained:

$$\hat{y}_i = \sum_{j=1}^{N} c_{ij} \psi(\frac{u_j \cdot X - b_j}{a_j}) \quad i = 1, \cdots, m \tag{3}$$

where $\hat{Y} = [y_1, \cdots, y_m]; \|u_j\|^2 = 1, u_j \in R^d; X \in R^d$; c_{ij} is the superposition coefficients of ridgelet. The ridgelet regressor based on PPR[13] and NN with ridgelet neurons are both based on the minimization of experience risk. In 1988, *Vapnik* developed a new learning machine-SVM on the concept of VC dimension, which solves all the theoretic problems of NN[14]. Its strength lies on its minimum of structure risk experience and a consequent protruding optimization. Its good generalization and avoidance of local minimum in learning is unachievable for NN and PPR. SVM helps to build up a family of KM based methods in ML. KM is a powerful technology extending the standard linear methods to nonlinear cases. A foundational idea behind KM is that the kernel function can be interpreted as an inner product in the feature space under certain conditions. This idea, commonly known as the "kernel trick", has been used extensively in generating nonlinear versions of conventional linear supervised and unsupervised learning algorithms, such as SVM, kernel fisher decision (KFD) etc. To get better ridgelet regressor, in the following we discussed a kernel-based regressor with ridgelet being the kernel function.

2.2 Ridgelet kernel regression model

Consider training set **S**={(x_1, y_1),…, (x_P, y_P)} with number P, where x_i is d-dimension and y_i is one-dimension(i=1,..,P). An unknown mapping with noise n can be described by Y=f(X)+n, and our goal is to reconstruct f from **S**. Denoting X=[x_1,., x_P] and Y=[y_1,.,y_P], ridgelet ψ_r is a nonlinear mapping about the input samples. After X going through the directional vector of l ridgelets, we get **R**=[$\boldsymbol{r}_1$,.., $\boldsymbol{r}_P$]$^{\mathrm{T}}$ =[r_{ij}] with $r_{ij} = u_j \cdot x_i$ (i=1,..,P; j=1,..,l). Then the ridgelet regression of $R^d \to R$ becomes an $R^l \to R$ wavelet mapping. Considering the latter mapping, for the linear regressor in feather space, we construct such a linear function with weight w and threshold β: $\hat{f}^{\psi}(r) = w^{\psi} \cdot \psi(r) + \beta$ (4)

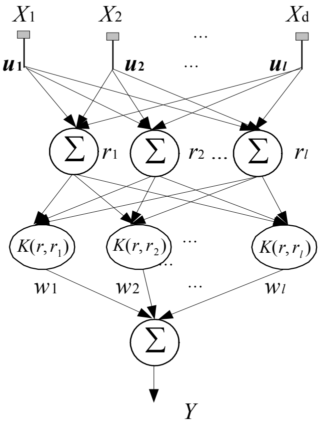

Fig.1 Ridgelet kernel regression

According to the reproducing kernel Hilbert space, the solution to this optimization problem must be in the space formed by the samples, i.e.,w is a linear superposition of all the samples: $w^{\psi} = \sum_{i=1}^{l} \alpha_i \psi(r_i) (\alpha_i \in R)$ (5)

Denote the kernel function K($\boldsymbol{r}_i$,$\boldsymbol{r}_j$)= Ψ($\boldsymbol{r}_i$) Ψ($\boldsymbol{r}_j$), and it should satisfy strict allowance condition. Then the estimate function in the feather space is obtained:

$$\hat{f}(r) = \sum_{i=1}^{l} a_i \psi(r_i) \psi(r) + \beta = \sum_{i=1}^{l} a_i K(r_i, r) + \beta \qquad (6)$$

The ridgelet kernel regression model is shown in Fig.1. $\Psi(r)$=cos(1.75r)exp(-r^2/2) proves to be served as a good kernel function [15]. Then we get:

$$K(r_i, r) = \cos(1.75 \times (r_i - r)/a) \exp(-\| r_i - r \|^2 / 2a^2) \qquad (7)$$

2.3 MSE based regularized kernel form

Vapnik showed that the key to get an effective solution is to control the complexity of the solution. In the context of statistical learning this leads to new techniques known as regularization networks or regularized kernel methods. We consider the regularized kernel methods, and define a generalized estimator for the approximation:

$$\min \quad E=\sum_{i=1}^{P} V(y,\hat{f}(r))/P+\lambda\|f\|_H^2 \quad (\lambda>0) \tag{8}$$

where V is the loss function; H is the Hilbert space of the hypotheses; λ is the regularization parameter. Various kinds of penalty terms can act as loss function, and squared loss function known as the rule of minimum square error (MSE) is most often used. The second term in (8) is a regularized term employed to get a specified solution and better generalization, and they are different in various function spaces. Then the solution is found by minimizing function consisting of the loss function and regularized term in KM. We adopt such form: min $E=\frac{1}{2}\|y-\hat{f}(r)\|^2+\frac{1}{2}\lambda(w.w)$ (9)

Let $e_k=y_k-w\cdot\psi(r_k)-\beta$ and convert the inequality restriction in the standard regression to the equality restriction; take the quadratic programming to represent the MSE based regularized kernel form:

$$\begin{cases} \min E(w,\lambda,\beta)=\sum_{k=1}^{l}\left\|e_k^2\right\|/2+\lambda(w.w)/2 \\ \text{s.t.} \quad y_k=w\cdot\psi(r_k)+\beta+e_k,\ k=1,..,l \end{cases} \tag{10}$$

The corresponding *Lagrange* function is $L=E(w,\lambda,\beta)-\sum_{k=1}^{l}\mu_k[w\cdot\psi_\lambda(r_k)+\beta-y_k+e_k]$ (11)

where μ is Lagrange factor. The optimal solution of this problem is:

$$\begin{aligned} &\partial L/\partial w=0 \Rightarrow w=\sum_{k=1}^{l}\mu_k\psi(r_k), \quad &&\partial L/\partial\beta=0 \Rightarrow \sum_{k=1}^{l}\mu_k=0, \\ &\partial L/\partial e_k=0 \Rightarrow \mu_k=\lambda e_k, &&\partial L/\partial\mu_k=0 \Rightarrow w\cdot\psi(r_k)+\beta-y_k+e_k=0 \end{aligned} \tag{12}$$

2.4 Optimization of directional vector based on GA

From above we see that as long as the directional vector is determined, the above method can be used to get a certain solution. To obtain the direction of ridgelet, here genetic algorithm (GA) is employed. GA is a good optimization tool which has the global searching capacity, as well as inner parallelism and self-learning[16]. Firstly define the reciprocal of regularized object function as the fitness function. Since the directional vector of ridgelet $\boldsymbol{u}=[u_1,..,u_d]$ and $\|\boldsymbol{u}\|^2=1$, the directional vector in $\boldsymbol{u}$ can be described using (d-1) angles $\theta_1,..,\theta_{d-1}: u_1=\cos\theta_1, u_2=\sin\theta_1\cos\theta_2,..,\ u_d=\sin\theta_1\sin\theta_2 ...\sin\theta_{d-1}$ .Code the chromosome using $l(d$-1) angles. The algorithm is described as:

Step 1: Init the iteration times t=0 and a population P with M chromosomes: Generate a set of angles randomly to form P={θ^1,... θ^M} where θ^i=[θ_1^i,...., $\theta_{l(d-1)}^i$] (i=1,..,M);
Step 2: Derive the directional vectors **U**=[$\boldsymbol{u}_1$,...$\boldsymbol{u}_M$] ($\boldsymbol{u}_i \in R^d$) from the population P;
Step 3: The directional vectors in **U** are corresponding to M ridgelet functions. Using them and the quadratic programming algorithm depicted in 2.3 to approximate the input samples. Compute the fitness, that is, the reciprocal of objective function 1/E;
Step 4: Roulette selection is performed to form new population;
Step 5: Judge the stop condition. When t is bigger than the given maximum iteration times T or the error is small than ε , stop, else go on;

Step 6: Repeat such operation M times on the population: Select two individuals randomly and perform crossover with probability p_c;
Step 7: Perform the mutation on each chromosome with probability p_m: $\theta_j^i(t+1)=\theta_j^i(t)+\eta\times rand$, where i=1,., M, j=1,…,$l(d$-1) and *rand* means a random number in [0,1]; η is a constant in [0,180] which determine the mutation degree;
Step 8: Get the new population P(t+1), $t=t+1$, go to Step 2.

3 Simulation Experiments

Experiment 1: Radar Target Recognition
Radar target recognition is a challenging subject in radar signal processing. We applied our method to radar target recognition of three-class planes(B52,J-6,J-7) using one-dimensional image(or radar range profile). Our data are obtained in a microwave darkroom with imaging angle from 0 to 179 degree and we get totally 1084 images of 64 dimension. In this experiment three models are considered-Gaussian kernel LS-SVM(GSVM),Wavelet kernel LS-SVM(WSVM) and our method. Kernel function in (7) is used in the latter two models. The models are under the same condition and in our method M=5, $\eta=10$, T=100, l=6, λ=0.5,a=1, $\varepsilon=10^{-8}$,p_c=0.7, p_m=0.1. We get the recognition results in table 1, from which we can see that our method has both high recognition rates and least training time than GSVM and WSVM.

PLANE	samples		GSVM(%)		WSVM(%)		our method(%)	
	train	test	train	test	train	test	train	test
B-52	60	322	100	98	100	98	100	97
J-6	60	311	100	88	100	90	100	92
J-7	90	451	100	80	100	83	100	91
Time(s)			4.5	2.3	2.8	2.4	1.1	2.8

Table 1: Recognition results of three planes

Experiment 2: Function approximation

RMSE	error	GSVM	WSVM	our method	expression
f^1	train	0.00097011	2.2764e-010	4.247e-019	$f^1(x_1,x_2)=\begin{cases}4-x_1^2-x_2^2 & x_1+4x_2<1.2\\ 0 & otherwise\end{cases}$
	test	1.5928	1.4641	1.1583	
f^2	train	1.5520e-005	6.9517e-012	6.5869e-020	$f^2(x_1,x_2)=\begin{cases}[1-4sinc(4x_1)]\times[1-3sinc(3x_2)] & x_1-x_2>\\ 0 & otherwise\end{cases}$
	test	1.9338	1.9223	1.9040	
f^3	train	1.6862e-010	7.879e-015	2.7152e-020	$f^3(x_1,x_2)=\begin{cases}sinc(x_1)-sinc(x_2) & 12x_1+x_2>0\\ 0 & otherwise\end{cases}$
	test	0.2154	0.1954	0.1459	
f^4	train	4.9328e-005	1.0865e-011	5.5366e-020	$f^4(x_1,x_2)=\begin{cases}\sqrt{x_1^2+(x_2^2+0.5)^2} & 3x_1+x_2>1\\ 1-0.2x_1^2-x_2^2 & x_1+2x_2<0.5\\ 0 & otherwise\end{cases}$
	test	0.4074	0.3759	0.3271	
f^5	train	1.552e-005	6.9517e-012	6.7489e-020	$f^5(x_1,x_2)=\begin{cases}e^{-(x_1^2+x_2^2)} & x_2\geq x_1^2\\ 0 & otherwise\end{cases}$
	test	1.9338	1.9223	1.9046	

Table 2: Approximation result of three methods

Just as described above, our method can deal with not only high dimensional data, but also data with spatial inhomogeneities. Consider approximations of some singular functions f^1- f^5. In the experiment, the numbers of training and testing samples are 36 and 121. To estimate the approximation, the root mean squared error (RMSE) is employed. The approximation results are shown in table 2.

4 Conclusion

Ridgelet is a new geometrical multi-scale analysis tool developed recently, which provides good basis for high dimensional space. Starting from the problem of MVFA, we proposed a ridgelet kernel regression model which can represent a wider range of high dimensional functions more efficiently in a stable way. The regularized items are employed in the object function to improve the generalization of our method. The objective function solved by quadratic programming is used to define the fitness function, and GA is taken for optimizing the directions of ridgelets. Theoretical analysis proves its accurate regression for high dimensional data, especially those with spatial inhomogeneities singularities. Experiment results on pattern recognition and function approximation also prove its efficiency.

Reference

[1] Breiman, L. The II method for estimating multivariate functions from noisy data (with discussion), Technometrics 33: 125-160, 1991.

[2] G. G. Lorentz, Constructive Approximation, Advanced Problems. *New York: Springer-Verlag*,1996.

[3] Wang ReHong, Approximation of multivariate functions.*Beijing publishing company of China*,1988.

[4] Grimm, L.G., Yarnold, P.R.. Reading and understanding more multivariate statistics. *Washington, DC: American Psychological Association*. 2000.

[5] Arfken, G. "Fourier Transforms--Inversion Theorem,"§15.3 in *Mathematical Methods for Physicists, 3rd ed. Orlando, FL: Academic Press*, pp.794-810, 1985.

[6] HÄardle, ect. Wavelets, approximation,and statistical applications.*New York: Springer-Verlag*,1998

[7] Cybenko, G.Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems*, 2, 303- 314, 1989.

[8] Friedman, J. H.,Stuetzle, W. Projection pursuit regression. *J. Amer. Statist. Assoc*.,76, 817-823,1981.

[9] Candes, E. J. Ridgelets: theory and applications. dissertation, Stanford University,1998.

[10] Vu, V. H. On the infeasibility of training neural networks with small mean-squared error. *IEEE Transactions on Information Theory*, vol 44, pp 2892-2900,1998.

[11] Gasser, T. and Muller, H. G. Kernel Estimation of Regression Functions, *New York: Springer-Verlag*, Vol. 757, pp. 23-68, 1979.

[12] Xu JianHu, ect. Regularized Kernel forms of Minimum square methods. *ACTA AUTOMATIC SINICA*, vol 30, No 1: 27 -36, 2004.

[13] A. Rakotomamonjy, Ridgelet Pursuit : Application to regression estimation, Technical Report, Perception Systèmes Information, ICANN 2001.

[14] V. Vapnik. Statistical Learning Theory. Wiley, New York, 1988.

[15] L. Zhang, W. D. Zhou, L. C. Jiao. “Wavelet Support Vector Machine” . *IEEE Trans. On Systems, Man, and Cybernetics. Part B: Cybernetics*. Vol. 34, no. 1, February 2004.

[16] Holland J.H. Genetic algorithms and classifier systems: foundations and their applications. *Proceedings of the Second International Conference on Genetic Algorithms,* pp 82-89, 1987.