5

Classification of support vector machine and regression algorithm

CAI-XIA DENG[1], LI-XIANG XU[2] and SHUAI LI[1]
[1]Harbin University of Science and Technology, [2]Hefei University
China

1. Introduction

Support vector machine (SVM) originally introduced by Vapnik. V. N. has been successfully applied because of its good generalization. It is a kind of learning mechanism which is based on the statistical learning theory and it is a new technology based on the kernel which is used to solve the problems of learning from the samples. Support vector machine was presented in 1990s, and it has been researched deeply and extensively applied in some practical application since then, for example text cataloguing, handwriting recognition, image classification etc. Support vector machine can provide optimal learning capacity, and has been established as a standard tool in machine learning and data mining. But learning from the samples is an ill–posed problem which can be solved by transforming into a posed problem by regularization. The RK and its corresponding reproducing kernel Hilbert space (RKHS) play the important roles in the theory of function approach and regularization. However, different functions approach problems need different approach functional sets. Different kernel's SVM can solve different actual problems, so it is very significant to construct the RK function which reflects the characteristics of this kind of approach function. In kernel-based method, one map which put the input data into a higher dimensional space. The kernel plays a crucial role during the process of solving the convex optimization problem of SVM. How to choose a kernel function with good reproducing properties is a key issue of data representation, and it is closely related to choose a specific RKHS. It is a valuable issue whether a better performance could be obtained if we adopt the RK theory method. Actually it has caused great interest of our researchers. In order to take the advantage of the RK, we propose a LS-SVM based on RK and develop a framework for regression estimation in this paper. The Simulation results are presented to illustrate the feasibility of the proposed method and this model can give a better experiment results, comparing with Gauss kernel on regression problem.

2. Small Sample Statistical Learning Theory

In order to avoid the assumption that the distribution of sample points and sample purpose of the request created a new principle of statistical inference ---- structured risk minimization principle.

We discussed the two classification problems, that is

$$(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l) \in R^n \cdot Y$$

where $Y = \{-1, +1\}$; $x_i (i = 1, 2, \cdots, l)$ is the independent and identical distribution data based on distribution density function $p(x, y)$.

Suppose f to be classifier, which is defined as the expectations of risk

$$R(f) = \int |f(x) - y| p(x, y) dxdy . \tag{1}$$

Experience of risk is defined as

$$R_{emp}(f) = \frac{1}{l} \sum_{i=1}^{l} |f(x_i) - y_i| . \tag{2}$$

Since the distribution density function $p(x, y)$ is unknown, it is virtually impossible to calculate the risk expectations $R(f)$.

If $l \to \infty$, we have $R(f) \to R_{emp}(f)$. Accordingly, the process from control theory modeling method to the neural network learning algorithm always constructs model with minimum experience risk. It is called as Empirical Risk Minimization principle.

If $R(f)$ and $R_{emp}(f)$ converge to the same limitation $\inf R(f)$ in probability, that is,

$$R(f) \xrightarrow[n \to \infty]{P} \inf R(f) , \qquad R_{emp}(f) \xrightarrow[n \to \infty]{P} \inf R(f) .$$

Then Empirical Risk Minimization principle (method) has the consistency.

Unfortunately, as early as in 1971 Vapnik had proved that the minimum of experience of risk may not converge to the minimum of expectations of risk, that is, the experience of risk minimization principle is not established.

Vapnik and Chervonenkis proposed structural risk minimization principle, laid the foundation for small sample statistical theory. They studied the relationship between experience of risk and expectations of risk in-depth, and obtained the following inequality, that is

$$R(f) \le R_{emp}(f) + \sqrt{\frac{h(\ln \frac{2l}{h} + 1) - \ln \frac{\eta}{4}}{l}} , \tag{3}$$

which l ----- samples number; η ----- Parameters $(0 \le \eta \le 1)$; h ----- Dimension of function f, for short VC-dimension.

The importance of formula (3): the right side of inequality has nothing to do with the specific distribution of the sample, that is, Vapnik's statistical learning theory do not need the assumption about the distribution of samples, it overcomes the problem of the high-dimensional distribution to the demand of samples number as exponential growth with the dimension growth. This is essence distinction with the classic statistical theory and the reasons for we call the Vapnik statistical methods for small samples of statistical theory.

From the formula (3), if l/h is large, the expectations of risk (real risk) is decided mainly by the experienced of risk, and this is the reason of the experience of risk minimization principle can often give good results for large sample set. However, if l/h is small, the small value of the experience of risk $R_{emp}(f)$ has not necessarily a small value of the actual risk. In this case, in order to minimize the actual risk, we must consider two items of the right in formula (3): the experience of risk $R_{emp}(f)$ and confidence range (called the VC

dimension confidence). VC dimension h play an important role, in fact, confidence range is an increasing function about h. When fixed the number l of points in the sample, the more complex the classifier, that is, the greater the VC dimension h, the greater the range of confidence, leading to the difference between the actual risks and experience gets greater. Therefore, in order to ensure the actual risk to be the smallest, to make sure experience risk minimization, but also to make the VC classifier peacekeeping function as small as possible, this is the principle of structural risk minimization.

With Structural risk minimization principle, the design of a classifier has two-step process:

(1) Choice of model classifier to a smaller VC dimension, that is, small confidence range.

(2) Estimate the model parameters to minimize the risk experience

3. Classification of support vector machine based on quadratic program

3.1 Solving quadratic programming with inequality constraints

On the target of finding a classifying space H which can exactly separate the two-class sample, and maximize the spacing of classification. The classifying space is called optimal classifying hyper plane.

In mathematics, the equation of classifying space is

$$\langle w, x \rangle + b = 0 ,$$

where $\langle w, x \rangle$ is the inner product of the two vector, w is the weight number, b is a constant.

So we can conclude that the problem which maximizes the spacing of classification between the two-class samples corresponds with an optimization problem as followed:

$$\min \Phi(w) = \min_{w,b} \frac{1}{2} \|w\|^2 = \min_{w,b} \frac{1}{2} \langle w, w \rangle . \tag{4}$$

The constraint condition is

$$y_i \left[\langle w, x_i \rangle + b \right] \geq 1 \qquad i = 1, 2, \cdots, l . \tag{5}$$

The (4) and (5) are current methods which describes the date sample is separated by the rule of the support vector machine. Inherently it's a quadratic program problem solved by inequality constraint.

We adopt the Lagrange optimization method to solve the quadratic optimization problem. Therefore, we have to find the saddle point of a Lagrange function

$$L(w, b, \alpha) = \frac{1}{2} \langle w, w \rangle - \sum_{i=1}^{l} \alpha_i \left[y_i \left(\langle w, x_i \rangle + b \right) - 1 \right] , \tag{6}$$

where $\alpha_i \geq 0$ is the Lagrange multiplier.

By extremal condition, we can obtain

$$Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle . \tag{7}$$

Then we have already changed the symbol from $L(w, b, \alpha)$ to $Q(\alpha)$ for reflecting the final transform.

The expression (7) is called Lagrange dual objective function. Under the constraint condition

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \,, \tag{8}$$

$$\alpha_i \geq 0, \quad i = 1,2,\cdots,l \,, \tag{9}$$

we find that α_i which can maximize the function $Q(\alpha)$. Then, the sample is the support vector when α_i are not zero.

3.2 Kernel method and its algorithm implementation

When the samples are not separated by liner classification, the way of the solution is using a liner transforms $\phi(x)$ to put the samples from input data space to higher dimensional character space, and then we separate the samples by liner classification in higher dimensional character space, and finally we use the $\phi^{-1}(x)$ to put the samples from higher dimensional character space to input data, which is a nonlinear classification in input data. The basic thought of the kernel method is that, for any kernel function $K(x,x_i)$ which satisfies with the condition of Mercer, there is a character space $(\phi_1(x),\phi_2(x),\cdots,\phi_l(x),\cdots)$ and in this space the kernel function implies inner product. So the inner product has been replaced by kernel in input space.

The advantage of the kernel method is that, the kernel function of input space is equivalent to the inner product in character space, so we only choose the kernel function $K(x,x_i)$ without finding out the nonlinear transforms $\phi(x)$.

Considering the Lagrange function

$$L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i\left(y_i\left(\langle w,\phi(x_i)\rangle + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{l}\gamma_i\xi_i \,, \tag{10}$$

$$\alpha_i, \gamma_i \geq 0, \quad i = 1,\cdots,l \,.$$

Similar to the previous section, we can get the dual form of the optimization problem

$$\max \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x,x_i) \,. \tag{11}$$

The constraint condition is

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \,, \tag{12}$$

$$0 \leq \alpha_i \leq C, \quad i = 1,\cdots,l \,. \tag{13}$$

Totally, the solution of the optimization problem is characterized by the majority of α_i being zero, and the support vector is that the samples correspond with the α_i which are not zero.

We can obtain the calculation formula of b from KKT as followed

$$y_i\left(\sum_{j=1}^{l}\alpha_j y_j K\left(x_j,x_i\right)+b\right)-1=0,\quad \alpha_i\in\left(0,C\right). \tag{14}$$

So we can find the value of b from anyone of the support vector. In order to stabilization, we can also find the value of b from all support vectors, and then get the average of the value.

Finally, we obtain the discriminate function as followed

$$f\left(x\right)=\operatorname{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K\left(x_i,x\right)+b\right). \tag{15}$$

3.3 One-class classification problem

Let a sample's set be

$$\left\{x_i,i=1,\cdots,l\right\},\quad x_i\in R^d.$$

We want to find the smallest sphere with a as its center and R as the radius and can contain all samples. If we directly optimize the samples, the optimization area is a hyper sphere. Allowing some data errors existed, we can equip with slack variable ξ_i to control, and find a kernel function $K\left(x,y\right)$ which satisfies that $K\left(x,y\right)=\left\langle\phi\left(x\right),\phi\left(y\right)\right\rangle$, and the optimization problem is

$$\min F\left(R,a,\xi_i\right)=R^2+C\sum_{i=1}^{l}\xi_i. \tag{16}$$

The constraint condition is

$$\left(\phi\left(x_i\right)-a\right)\left(\phi\left(x_i\right)-a\right)^T\le R^2+\xi_i\qquad i=1,\cdots,l, \tag{17}$$

$$\xi_i\ge 0,\quad i=1,\cdots,l. \tag{18}$$

Type (16) will be changed into its dual form

$$\max\sum_{i=1}^{l}\alpha_i K\left(x_i,x_i\right)-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K\left(x_i,x_j\right). \tag{19}$$

The constraint condition is

$$\sum_{i=1}^{l}\alpha_i=1, \tag{20}$$

$$0\le\alpha_i\le C,\quad i=1,\cdots,l. \tag{21}$$

We can get α by solving (19). Usually, the majority of α will be zero, the samples corresponded with $\alpha_i\ne 0$ are still so-called the support vector.

According to the KKT condition, the samples corresponded with $0<\alpha_i<C$ are satisfied

$$R^2-\left(K\left(x_i,x_i\right)-2\sum_{j=1}^{l}\alpha_j K\left(x_j,x_i\right)+a^2\right)=0, \tag{22}$$

$a = \sum_{i=1}^{l} \alpha_i \phi(x_i)$. Thus, according to the (22), we can find the value of R by any support vector. For a new sample z, let

$$f(z) = (\phi(z) - a)(\phi(z) - a)^T = K(z,z) - 2\sum_{i=1}^{l} \alpha_i K(z, x_i) + \sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j K(x_i, x_j) .$$

If $f(z) \le R^2$, z is a normal point; otherwise, z is an abnormal point.

3.4 Multi-class support vector machine

1. One-to-many method

The idea is to take samples from a certain class as one class and consider the remaining samples as another class, and then there is a two-class classification. Afterward we repeat the above step in the remaining samples. The disadvantage of this method is that the number of training sample is large and the training is difficult.

2. One-to-one method

In multi-class classification, we only consider two-class samples every time, that is, we design a model of SVM for every two-class samples. Therefore, we need to design $\frac{k(k-1)}{2}$ models of SVM. The calculation is very complicated.

3. SVM decision tree method

It usually combines with the binary decision tree to constitute multi-class recognizer, whose disadvantage is that if the classification is wrong at a certain node, the mistake will keep down, and the classification makes nonsense at the node after that one.

4. Determine the multi-class objective function method

Since the number of the variables is very large, the method is only used in small problem.

5. DAGSVM

John C.Platt brings forward this method, combining DAG with SVM to realize the multi-class classification.

6. ECC-SVM methods

Multi-class classification problem can be changed into many two-class classification problems by binary encoding for classification. This method has certain correction capability.

7. The multi-class classification algorithm based on the one-class classification

The method is that we first find a center of hyper sphere in every class sample in higher dimensional character space, and then calculate the distance between every center and test the samples, finally, judge the class based on the minimum distance a point on it.

4. Classification of support vector machine based on linear programming

4.1 Mathematical background

Considering two hyper plane of equal rank on R^d, $H_1 : \langle \omega, x \rangle + b_1 = 0$ and $H_2 : \langle \omega, x \rangle + b_2 = 0$. Based on L_p two the hyper plane distance of norm is:

$$d_p(H_1,H_2) := \min_{\substack{x\in H_1\\ y\in H_2}} \|x-y\|_p , \tag{23}$$

and

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{\frac{1}{p}} . \tag{24}$$

Choose a $y \in H_2$ arbitrarily, then two hyper plane's can be write be

$$d_p(H_1,H_2) = \min_{x\in H_1} \|x-y\|_p . \tag{25}$$

Moves two parallel hyper plane to enable H_2 to adopt the zero point, can be obtain the same distance hyper plane:

$$H_1 : \langle \omega, x\rangle + (b_1,b_2) = 0 , \quad H_2 : \langle \omega, x\rangle = 0 .$$

If chooses y spot is the zero point, then the distance between two hyper plane is

$$d_p(H_1,H_2) = \min_{x\in H_1} \|x\|_p . \tag{26}$$

If L_p is the L_q conjugate norm, that is p and q satisfy the equality

$$\frac{1}{p} + \frac{1}{q} = 1 . \tag{27}$$

By the Holder inequality may result in

$$\|x\|_p \|\omega\|_q \geq |\langle x, \omega\rangle| . \tag{28}$$

Regarding $x \in H_1$, we have $\langle \omega, x\rangle = b_1 - b_2$. Therefore

$$\min_{x\in H_1} \|x\|_p \|\omega\|_q = b_1 - b_2 . \tag{29}$$

So, the distance between two hyper plane is

$$d_p(H_1,H_2) = \min_{x\in H_1} \|x\|_p = \frac{|b_1 - b_2|}{\|\omega\|_q} , \tag{30}$$

4.2 Classification algorithm of linear programming

1 norm formula of L_1

The two hyper-plane $H_1 : \langle \omega, x\rangle + b_1 = 0$ and $H_2 : \langle \omega, x\rangle + b_2 = 0$, through the definition of the norm of the distance between them

$$d_1(H_1,H_2) = \frac{|b_1 - b_2|}{\|\omega\|_\infty} , \tag{31}$$

Where, $\|\omega\|_\infty$ expressed as a norm of L_∞, it is the dual norm of L_1, defined as

$$L_\infty = \max_j |\omega_j| . \tag{32}$$

Supposes $H^+ : \langle \omega, x\rangle + b = 1$, $H^- : \langle \omega, x\rangle + b = -1$, established through the two types of support Vector distance between the hyper-plane as follow

$$d_1(H^+,H^-) = \frac{|(b+1)-(b-1)|}{\|\omega\|_\infty} = \frac{2}{\max_j |\omega_j|} . \tag{33}$$

Therefore the optimized question's equation is

$$\min_{\omega,b} \max_j \left|\omega_j\right|. \tag{34}$$

The restraint is

$$y_i\left(\left\langle \omega, x_i \right\rangle + b\right) \geq 1, i = 1, \cdots, l. \tag{35}$$

Therefore obtains the following linear programming

$$\min a. \tag{36}$$

The restraint is

$$y_i\left(\left\langle \omega, x_i \right\rangle + b\right) \geq 1, i = 1, \cdots, l, \tag{37}$$

$$a \geq \omega_j, j = 1, \cdots, d, \tag{38}$$

$$a \geq -\omega_j, j = 1, \cdots, d, \tag{39}$$

$$a, b \in R, \omega \in R^d. \tag{40}$$

This is a linear optimization question, must be much simpler than the quadratic optimization.

2 norm formula of L_∞

If defines L_∞ between two hyper-planes the distances, then we may obtain other one form linear optimization equation. This time, between two hyper-planes distances is

$$d_\infty\left(H_1, H_2\right) = \frac{\left|b_1 - b_2\right|}{\left\|\omega\right\|_1}. \tag{41}$$

Regarding the linear separable situation, two support between two hyper-planes the distances is

$$d_\infty\left(H^+, H^-\right) = \frac{\left|\left(1-b\right) - \left(-1-b\right)\right|}{\left\|\omega\right\|_1} = \frac{2}{\sum_j \left|\omega_j\right|}. \tag{42}$$

Maximized type (42) is equivalent to

$$\min_{\omega,b} \sum_j \left|\omega_j\right|. \tag{43}$$

The restraint is

$$y_i\left(\left\langle \omega, x_i \right\rangle + b\right) \geq 1, i = 1, \cdots, l. \tag{44}$$

Therefore the optimized question is

$$\min \sum_{j=1}^{d} a_j. \tag{45}$$

Bound for

$$y_i\left(\left\langle \omega, x_i \right\rangle + b\right) \geq 1, i = 1, \cdots, l, \tag{46}$$

$$a_j \geq \omega_j, j = 1, \cdots, d, \tag{47}$$

$$a_j \geq -\omega_j, j = 1, \cdots, d. \tag{48}$$

4.3 One-class classification algorithm in the case of linear programming

The optimized question is

$$\min \frac{1}{2}\|\omega\|_2^2 - \rho + C\sum_{i=1}^{l}\xi_i \,, \tag{49}$$

Restrain for

$$\left\langle \omega, \phi\left(x_i\right)\right\rangle \geq \rho - \xi_i \,, \quad \xi_i \geq 0 \,, \quad i = 1, \cdots, l \,. \tag{50}$$

Introduces Lagrange the function

$$L = \frac{1}{2}\|\omega\|^2 - \rho + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i\left(\left\langle \omega, \phi\left(x_i\right)\right\rangle - \rho + \xi_i\right) - \sum_{i=1}^{l}\beta_i\xi_i \,, \tag{51}$$

in the formula $\alpha_i \geq 0, \beta_i \geq 0, i = 1, \ldots, \tilde{l}$.

The function L's extreme value should satisfy the condition

$$\frac{\partial}{\partial\omega}L = 0, \frac{\partial}{\partial\rho}L = 0, \frac{\partial}{\partial\xi_i}L = 0 \,. \tag{52}$$

Thus

$$\omega = \sum_{i=1}^{l}\alpha_i\phi\left(x_i\right) \,, \tag{53}$$

$$\sum_{i=1}^{l}\alpha_i = 1 \,, \tag{54}$$

$$C - \alpha_i - \beta_i = 0, i = 1, \cdots, l \,. \tag{55}$$

With (53)~(55) replace in Lagrange function (51). And using kernel function to replace inner product arithmetic in higher dimensional space, finally we may result in the optimized question the dual form is

$$\min \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j k\left(x_i, x_j\right) \,. \tag{56}$$

Restrain for

$$0 \leq \alpha_i \leq C \,, \quad i = 1, \cdots, l \,, \tag{57}$$

$$\sum_{i=1}^{l}\alpha_i = 1 \,. \tag{58}$$

After solving the value of α we may get the decision function

$$f(x) = \sum_{i=1}^{l}\alpha_i k\left(x_i, x\right) \,. \tag{59}$$

While taking the Gauss kernel function , we may discover that the optimized equation (56) and a classification class method's of the other form ----- type (19) is equal.

We may obtain its equal linear optimization question by the reference

$$\min\left(-\rho + C\sum_{i=1}^{l}\xi_i\right) \,. \tag{60}$$

Restrain for

$$\left\langle \omega, \phi(x_i)\right\rangle \geq \rho - \xi_i \,, \quad \xi_i \geq 0 \,, \quad i = 1, \cdots, l \,, \tag{61}$$

$$\|\omega\|_1 = 1 \,. \tag{62}$$

Using kernel expansion $\sum_{j=1}^{l}\alpha_j k\left(x_j, x_i\right)$ to replace the optimized question type (60) the

inequality constraint item $\langle \omega, \phi(x_i) \rangle$, so we can obtain the following linear programming form:

$$\min\left(-\rho + C\sum_{i=1}^{l}\xi_i\right). \tag{63}$$

Restrain for

$$\sum_{i=1}^{l}\alpha_i k(x_i, x) \geq \rho - \xi_i, i = 1, \cdots, l, \tag{64}$$

$$\sum_{i=1}^{l}\alpha_i = 1, \tag{65}$$

$$\alpha_i, \xi_i \geq 0, i = 1, \cdots, l. \tag{66}$$

Solving this linear programming may obtain the of value α and ρ, therefore we can obtain a decision function:

$$f(x) = \sum_{i=1}^{l}\alpha_i k(x_i, x). \tag{67}$$

According to the significance of optimization problems, regarding the majority of training samples will meet $f(x) \geq \rho$, the significance of parameter C satisfies the condition $f(x) < \rho$ to control the sample quantity , the larger parameter C will cause all samples to satisfy the condition ,and the geometry significance of parameter C will give in the 5th chapter. Hyper plane of the decision-making to be as follows:

$$\sum_{i=1}^{l}\alpha_i k(x_i, x) = \rho. \tag{68}$$

After Hyper plane of the decision-making reflected back to the original space, the training samples will be contained in the regional compact. Regarding arbitrary sample x in the region, satisfies $f(x) \geq \rho$, and for region outside arbitrary sample y to satisfy $f(x) < \rho$ In practical application, the value of parameter σ^2 in kernel function is smaller , which obtains the region to be tighter in the original space to contain the training sample , this explained that the parameter σ^2 will decide classified precisely .

4.4 Multi-class Classification algorithm in the case of linear programming

The following linear programming will be under the classification of a class which is extended to many types of classification. Using the methods implement a classification class operation to each kind of samples, then obtains a decision function to each kind. Then input the wait for testing samples in each decision function, according to the decision function to determine the maximum-point belongs to the category. The concrete algorithm is as follows stated.

Supposes the training sample is:

$$\{(x_1, y_1), \ldots, (x_l, y_l)\} \subset R^n \times Y, Y = \{1, 2, \cdots, M\},$$

where, n is the dimension of input samples ; M is a category number . Sample is divided into M -type, and various types of classifications are written separately:

$$\left\{\left(x_1^{(s)}, y_1^{(s)}\right), \ldots, \left(x_{l_s}^{(s)}, y_{l_s}^{(s)}\right), s = 1, \cdots, M\right\}$$

where $\left\{\left(x_i^{(s)}, y_i^{(s)}\right), s=1,2,\cdots,l_s\right\}$ represents the s -th type of training samples $l_1+l_2+\cdots+l_M=l$. A kind of classification thought according to 2.3 section, made the following linear programming:

$$\min\left(-\rho+C\sum_{s=1}^{M}\sum_{i=1}^{l_s}\xi_{si}\right). \tag{69}$$

Restrain for

$$\sum_{j=1}^{l_s}\alpha_j^{(s)}k\left(x_j^{(s)}, x_i^{(s)}\right)\geq\rho-\xi_{si}, s=1,\cdots,M, i=1,\cdots,l_s\ , \tag{70}$$

$$\sum_{j=1}^{l_s}\alpha_j^{(s)}=1, s=1,\cdots,M\ , \tag{71}$$

$$\alpha_i^{(s)}, \xi_{si}\geq 0, s=1,\cdots,l_s\ . \tag{72}$$

Solving this linear programming, may obtain M decision functions

$$f_s(x)=\sum_{j=1}^{l_s}a_j^{(s)}k\left(x_j^{(s)}, x\right), s=1,\cdots,M\ . \tag{73}$$

Assigns treats recognition sample z, calculated $\gamma_i=f_s(z), i=1,\cdots,M$. Compared with the size, find the largest γ_{k}, then z belongs to the k -th type. At the same time, the definition of the classification results can trust is as follows:

$$\mathrm{B_k}=\begin{cases}1 & \gamma_k\geq\rho\\ \dfrac{\gamma_k}{\rho} & otherwise\end{cases}. \tag{74}$$

When the difference of the number among samples of various types is large, we can introduce the different ρ value in optimized type (69). And using quadratic programming the similar processing methods, here no longer relates in details.

Another alternative way is to directly compare the new sample size in all decision function, and then the basis maximum value to determine where a new category of sample was taken. As a result of the decomposition algorithm to the optimization process is an independent, it can also be carried out in parallel computing.

5. The beat-wave signal regression model based on least squares reproducing kernel support vector machine

5.1 Support Vector Machine

For the given sample's set

$$\{(x_1, y_1), \ldots, (x_l, y_l)\}$$

$x_i\in R^d$, $y_i\in R$, l is the samples number, d is the number of input dimension. In order to precisely approach the function $f(x)$ which is about this data set, For regression analysis, SVM use the regression function as following

$$f(x)=\sum_{i=1}^{l}w_i k(x_i, x)+q\ , \tag{75}$$

w_i is the weight vector, and q is the threshold, $k(x_i, x)$ is the kernel function.
Training a SVM can be regarded as minimizing the value of $J(w, q)$

$$\min J(w, q) = \frac{1}{2}\| w \|^2 + \gamma \sum_{k=1}^{l} \left(y_k - \sum_{i=1}^{l} w_i k(x_i, x_k) - q \right)^2 . \quad (76)$$

The kernel function $k(x_i, x)$ must satisfy with the condition of Mercer. When we define the kernel function $k(x_i, x)$, we also define the mapping which is from input datas to character space. The general used kernel function of SVM is Gauss function, defined by

$$k(x, x') = \exp(-\|x - x'\|^2 / 2\sigma^2) . \quad (77)$$

For this equation, σ is a parameter which can be adjusted by users.

5.2 Support Vector's Kernel Function

1. The Conditions of Support Vector's Kernel Function
In fact, if a function satisfies the condition of Mercer, it is the allowable support vector's kernel function.
Lemma 2.1 The symmetry function $k(x, x')$ is the kernel function of SVM if and only if: for all function $g \neq 0$ which satisfies the condition of $\int_{R^d} g^2(\xi) d\xi < \infty$, we need to satisfy the condition as following

$$\iint_{R^d \otimes R^d} k(x, x') g(x) g(x') dx dx' \geq 0 . \quad (78)$$

This Lemma proposes a simple method to build the kernel function.
For the horizontal floating function, we can give the condition of horizontal floating kernel function.
Lemma 2.2 The horizontal floating function $k(x, x') = k(x - x')$ is a allowable support vector's kernel function if and only if the Fourier transform of $k(x)$ need to satisfy the condition as following

$$\hat{k}(\omega) = (2\pi)^{-\frac{d}{2}} \int_{R^d} \exp(-jwx) k(x) dx \geq 0 . \quad (79)$$

2. Reproducing Kernel Support Vector Machine on the Sobolev Hilbert space $H^1(R : a, b)$
Let $F(E)$ be the linear space comprising all complex-valued functions on an abstract set E. Let H be a Hilbert (possibly finite-dimensional) space equipped with inner product $(\cdot, \cdot)_H$. Let $h : E \to H$ be a Hilbert space H-function on E. Then, we shall consider the linear mapping L from H into $F(E)$ defined by

$$f(q) = (Lg)(p) = (g, h(p))_H . \quad (80)$$

The fundamental problems in the linear mapping (80) will be firstly the characterization of the images $f(p)$ and secondly the relationship between g and $f(p)$.
The key which solves these fundamental problems is to form the function $K(p, q)$ on $E \times E$ defined by

$$K(p, q) = (g(q), g(p))_H . \quad (81)$$

We let $R(L)$ denote the range of L for H and we introduce the inner product in $R(L)$ induced from the norm

$$\|f\|_{R(L)} = \inf\{\| g \|_H ; \quad f = Lg\} . \tag{82}$$

Then, we obtain

Lemma 2.3 For the function $K(p,q)$ defined by (81), the space $\left[R(L),(\cdot,\cdot)_{R(L)}\right]$ is a Hilbert (possibly finite dimensional) space satisfying the properties that

(i) for any fixed $q \in E$, $K(p,q)$ belongs to $R(L)$ as a function in p;

(ii) for any $f \in R(L)$ and for any $q \in E$,

$$f(q) = (f(\cdot) , K(\cdot,q))_{R(L)} .$$

Further, the function $K(p,q)$ satisfying (i) and (ii) is uniquely determined by $R(L)$.

Furthermore, the mapping L is an isometry from H onto $R(L)$ if and only if $\{h(p); p \in E\}$ is complete in H.

On the Sobolev Hilbert space $H^1(R:a,b)$ on R comprising all complex valued and absolutely continuous functions $f(x)$ with finite norms

$$\left\{\int_{-\infty}^{\infty}(a^2\left|f'(x)\right|^2 + b^2\left|f(x)\right|^2)dx\right\}^{\frac{1}{2}} < \infty , \tag{83}$$

where a, $b > 0$.

The function

$$G_{a,b}(x,y) = \frac{1}{2ab}e^{-\frac{b}{a}|x-y|} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{i\omega(x-y)}}{a^2\omega^2 + b^2}d\omega . \tag{84}$$

is the RK of $H^1(R:a,b)$.

On the Hilbert space, we construct this horizontal floating kernel function:

$$k(x,x') = k(x-x') = \prod_{i=1}^{d} G_{a,b}(x_i - x_i') . \tag{85}$$

Theorem 2.1 The horizontal floating function of Sobolev Hilbert space $H^1(R:a,b)$ is defined as

$$G_{a,b}(x,x') = \frac{1}{2ab}e^{-\frac{b}{a}|x-x'|} , \tag{86}$$

and the Fourier transform of this function is positive.

Proof. By (86), we have

$$\hat{\mathrm{G}}_{a,b}(\omega) = \int_R \exp(-j\omega x)\cdot G_{a,b}(x)dx = \int_R \exp(-j\omega x)\cdot\frac{1}{2ab}e^{-\frac{b}{a}|x|}dx$$

$$= \frac{1}{2ab}\int_R e^{-\frac{b}{a}|x|-j\omega x}dx = \frac{1}{b^2 + a^2\omega^2} \geq 0$$

Theorem 2.2 The function

$$\hat{k}(\omega) = (2\pi)^{-\frac{d}{2}}\int_{R^d}\exp(-j\omega x)k(x)dx = (2\pi)^{-\frac{d}{2}}\cdot\prod_{i=1}^{d}\left(\frac{1}{2ab}\int_{-\infty}^{+\infty}e^{-\frac{b}{a}|x_i|-j\omega_i x_i}dx_i\right) , \tag{87}$$

is a allowable support vector kernel function.

Proof. By the Lemma 2.2, we only need to prove

$$\hat{k}(\omega) = (2\pi)^{-\frac{d}{2}}\prod_{i=1}^{d}\frac{1}{b^2 + a^2\omega_i^{\,2}} = (2\pi)^{-\frac{d}{2}}\prod_{i=1}^{d}\hat{G}_{a,b}(\omega_i) \geq 0 .$$

That is

$$\hat{k}(\omega) = (2\pi)^{-\frac{d}{2}} \int_{R^d} \exp(-j\omega x) k(x) dx$$
$$= (2\pi)^{-\frac{d}{2}} \cdot \prod_{i=1}^{d} \left(\frac{1}{2ab} \int_{-\infty}^{+\infty} e^{-\frac{b}{a}|x_i| - j\omega_i x_i} dx_i \right)$$

By Theorem 2.1, we have

$$\hat{k}(\omega) = (2\pi)^{-\frac{d}{2}} \prod_{i=1}^{d} \frac{1}{b^2 + a^2 \omega_i^2} = (2\pi)^{-\frac{d}{2}} \prod_{i=1}^{d} \hat{G}_{a,b}(\omega_i) \geq 0$$

For regression analysis, the output function is defined as

$$f(x) = \sum_{i=1}^{l} w_i \prod_{j=1}^{d} \frac{1}{2ab} e^{-\frac{b}{a}|x_j - x_j^i|} + q \,, \tag{88}$$

x_j^i is the value of the i-th training sample's j-th attribute.

5.3 Least Squares RK Support Vector Machine

Least squares support vector machine is a new kind of SVM. It derives from transforming the condition of inequation into the condition of equation. Firstly, we give the linear regression algorithm as follows.

For the given samples set

$$\{(x_1, y_1), \cdots, (x_l, y_l)\}$$

$x_i \in R^d$, $y_i \in R$, l is the sample's number, d is the number of input dimension. The linear regression function is defined as

$$f(x) = w^T x + q \,. \tag{89}$$

Importing the structure risk function, we can transform regression problem into protruding quadratic programming

$$\min \left\{ \frac{1}{2} \|w\|^2 + \gamma \frac{1}{2} \sum_{i=1}^{l} \xi_i^2 \right\}. \tag{90}$$

The limited condition is

$$y_i = w^T x_i + q + \xi_i \,. \tag{91}$$

We define the Lagrange function as

$$L = \frac{1}{2} \|w\|^2 + \gamma \frac{1}{2} \sum_{i=1}^{l} \xi_i^2 - \sum_{i=1}^{l} \alpha_i (w^T x_i + q + \xi_i - y_i) \,, \tag{92}$$

and we obtain

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{l} \alpha_i x_i \\ \dfrac{\partial L}{\partial q} = 0 \rightarrow \sum_{i=1}^{l} \alpha_i = 0 \\ \dfrac{\partial L}{\partial \xi_i} = 0 \rightarrow \alpha_i = \gamma \xi_i \,; \quad i = 1, \cdots, l \\ \dfrac{\partial L}{\partial \alpha_i} = 0 \rightarrow w^T x_i + q + \xi_i - y_i = 0 \,; \quad i = 1, \cdots, l \end{cases} . \tag{93}$$

From equations (93) , we can get the following linear equation

$$\begin{bmatrix} I & 0 & 0 & -x \\ 0 & 0 & 0 & \Phi^T \\ 0 & 0 & \gamma I & -I \\ x^T & \Phi & I & 0 \end{bmatrix} \begin{bmatrix} w \\ q \\ \xi \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ y \end{bmatrix}, \tag{94}$$

where $x=[x_1,\cdots,x_l]$, $y=[y_1,\cdots,y_l]$, $\Phi=[1,\cdots,1]$, $\xi=[\xi_1,\cdots,\xi_l]$, $\alpha=[\alpha_1,\cdots,\alpha_l]$.
The equation result is

$$\begin{bmatrix} 0 & \Phi^T \\ \Phi & x^T x+\gamma^{-1} I \end{bmatrix} \begin{bmatrix} q \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{95}$$

where $w=\sum_{i=1}^{l}\alpha_i x_i$, $\xi_i=\alpha_i/\gamma$.

For non-linear problem, the non-linear regression function is defined as

$$f(x)=\sum_{i=1}^{l}\alpha_i k(x_i,x)+q\,. \tag{96}$$

The above equation result can be changed into

$$\begin{bmatrix} 0 & \Phi^T \\ \Phi & K+\gamma^{-1} I \end{bmatrix} \begin{bmatrix} q \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{97}$$

$K=\left\{k_{i,j}=k(x_i\ ,\ x_j)\right\}_{i\ ,\ j=1}^{l}$, the function $k(\cdot\ ,\ \cdot)$ is given by (87). Based on the RK kernel function, we get a new learning method which is called least squares RK support vector machine (LS-RKSVM). Since using least squares method, the computation speed of this algorithm is more rapid than the other SVM.

5.4 Simulation results and analysis

We use LS-RKSVM to regress the Beat-wave signal

$$a(t)=A\sin(\omega_1 t)\cdot\sin(\omega_2 t)$$

where $a(t)$ is the Beat-wave signal, A is the Signal amplitude, ω_1 is the higher frequency of Beat-wave frequencies, that is the resonant frequency of resonant Beat-wave, ω_2 is the frequency of Beat-wave, the relationship between ω_1 and ω_2 is that $\omega_2=\omega_1/2n$, where n is the cycle number of sine wave which is included in a beat-wave with a ω_1 frequency.

We assume $A=1.1,\ \omega_2=0.5,\ n=5,\ t=2s$, and 150 sampling points. We can get the result of this experiment which can be described as figure 1, figure 2 and table 1. Figure 1 is the regression result of LS-SVM which uses the Gauss kernel function. Figure 2 is the regression result of LS-RKSVM which uses the RK kernel function.

For regression experiments, we use the approaching error as following

$$E_{ms}=\left(\frac{1}{N}\sum_{t=1}^{N}(y(t)-y^{l}(t))^2\right)^{\frac{1}{2}}. \tag{98}$$

The Simulation results shows that the regression ability of RK kernel function is much better than Gauss kernel function. This reveals RK kernel function has rather strong regression

ability and it can be used for pattern recognition. We can find that the LS-SVM is a very promising method based on RK kernel. The model has strong regression ability.

kernel function	kernel parameter	error
RBF kernel:γ=150	σ =0.01	0.0164
RK kernel: γ=150	a=0.1, b=1	0.0110

Table 1. The regression result for Beat-wave signal

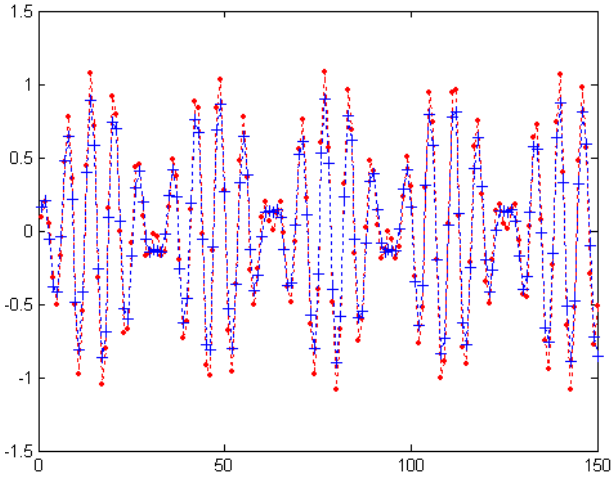


Fig. 1. The regression curve based on Gauss kernel (“.” is true value, “+” is predictive value)

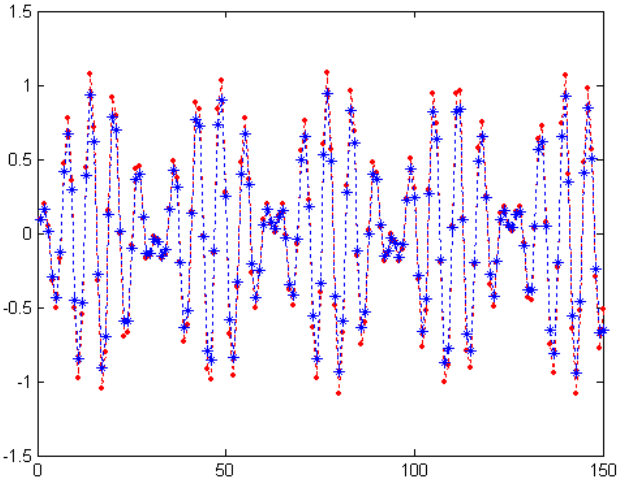


Fig. 2. The regression curve based on RK kernel (“.” is true value, “*” is predictive value)

The SVM is a new machine study method which is proposed by Vapnik based on statistical learning theory. The SVM focus on studying statistical learning rules under small sample.

Through structural risk minimization principle to enhance extensive ability, the SVM preferably solves many practical problems, such as small sample, non-linear, high dimension number and local minimum points. The LS-SVM is an improved algorithm which base on SVM. This paper proposes a new kernel function of SVM which is the RK kernel function. We can use this kind of kernel function to map the low dimension input space to the high dimension space. The RK kernel function enhances the generalization ability of the SVM. At the same time, adopting LS-SVM, we get a new regression analysis method which is called least squares RK support vector machine. Experiment shows that the RK kernel function is better than Gauss kernel function in regression analysis. The RK and LS-SVM are combined effectively. Thereby we can find that the result of regression is more precisely.

6. Prospect

Further study should be started in the following areas:
1. The kernel method provides an effective method which can change the nonlinear problem into a linear problem, that is, the kernel function plays an important role in the support vector machine. Therefore, for practical problems, rational choice of the kernel function and the parameter in it is a problem which should be research.
2. For the massive data of practical problems, a serious problem need to be solved is to propose an efficient algorithm.
3. It is a valuable research direction that fusion of the Boosting and the Ensemble methods are proposed to be a better algorithm of support vector machine.
4. It is significant to put the support vector machine, planning network, Gauss process and neural network into same frame.
5. It is a significant research subject that combines the idea of support vector machine with the Bayes Decision and consummates the maximum margin algorithm.
6. The research on support vector machine still needs to be done extensively.

7. References

Bernhard S. & Sung K.K. (1997). Comparing support vecter machines with Gaussian kernels to radical basis fuction classifiers. IEEE transaction on signal processing

Fatiha M. & Tahar M. (2007). Prediction of continuous time autoregressive processes via the reproducing kernel spaces. Computational mechanics, Vol 41, No. 1, dec

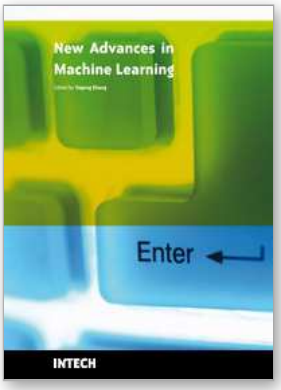
Karen A. Ames; Rhonda J. & Hughes. (2005). Structural stability for ill-posed problems in Banach space, Semigroup forum, Vol 70, No. 1, Jan

Mercer J. (1909). Function of positive and negative type and their connection with the theory of integral equations. Philosophical transactions of the royal society of London, Vol 209, pp. 415~446

O.L.Mangasarian. (1999). Arbitrary-norm Separating Plane. Operation Research Letters, 1(24): 15~23

Saitoh S. (1993). Inequalities in the most simple Sobolev space and convolutions of L_2 functions with weights. Proc. Amer. Math. Soc. Vol 118, pp. 515~520

Smola A. J.; Scholkopf B. & Muller K. R. (1998). The connection between regularization operators and support vector kernels. Neural networks, Vol 11, No. 4, pp. 637~649
Suykens J. & A. K. (2002). Least squares support vector machines, World scientific, Singapore
Vapnik. V. N. (1995). The nature of statistical learning theory. Springer-verlag, New York
Vapnik V N. (1998). Statistical Learning Theory. Wiley, New York
Yu Golubev. (2004). The principle of penalized empirical risk in severely ill-posed problems. Probability theory and related fields, Vol 130, No. 1, sep



New Advances in Machine Learning

Edited by Yagang Zhang

ISBN 978-953-307-034-6
Hard cover, 366 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

The purpose of this book is to provide an up-to-date and systematical introduction to the principles and algorithms of machine learning. The definition of learning is broad enough to include most tasks that we commonly call “learning” tasks, as we use the word in daily life. It is also broad enough to encompass computers that improve from experience in quite straightforward ways. The book will be of interest to industrial engineers and scientists as well as academics who wish to pursue machine learning. The book is intended for both graduate and postgraduate students in fields such as computer science, cybernetics, system sciences, engineering, statistics, and social sciences, and as a reference for software professionals and practitioners. The wide scope of the book provides a good introduction to many approaches of machine learning, and it is also the source of useful bibliographical information.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821