A Classification of Remote Sensing Image Based on Improved Compound Kernels of Svm

Jianing Zhao[1], Wanlin Gao[1,4,*], Zili Liu[1], Guifen Mou[2], Lin Lu[3], and Lina Yu[1]

[1] College of Information and Electrical Engineering, China Agricultural University, Beijing,
P.R. China 100083
[2] Yunnan Xin Nan Nan Agricultural Technology Co., Ltd. Yunnan Province,
P.R. China 650214
[3] Kunming Agriculture machinery research institute, Yunnan Province, P. R. China 650034
[4] College of Information and Electrical Engineering, China Agricultural University,
No. 17, Qinghua Dong Lu, Haidian, Beijing 100083, P.R. China,
Tel.: +86-010-62736755; Fax: +86-010-62736755
gaowlin@cau.edu.cn

Abstract. The accuracy of RS classification based on SVM which is developed from statistical learning theory is high under small number of train samples, which results in satisfaction of classification on RS using SVM methods. The traditional RS classification method combines visual interpretation with computer classification. The accuracy of the RS classification, however, is improved a lot based on SVM method, because it saves much labor and time which is used to interpret images and collect training samples. Kernel functions play an important part in the SVM algorithm. It uses improved compound kernel function and therefore has a higher accuracy of classification on RS images. Moreover, compound kernel improves the generalization and learning ability of the kernel.

Keywords: compound kernel, remote sensing, image classification, support vector machine.

1 Introduction

Classification on remote sensing images is very important in the field of remote sensing processing. Among the classification algorithms, the accuracy of SVM (C. J. Burges, 1998) classification is one of the highest methods. SVM is short for Support Vector Machine which is a machine learning method proposed by Vapnik according to statistical leaning theory (Vapnik, 1995; Vapnik, 1998), it integrates many techniques including optimal hyperplane (Cristianini, 2005), mercer kernel, slave variable, convex quadratic programming and so on. Support Vector Machine successfully solves many practical problems such as small number of samples, non linear, multi dimensions, local minimum and so on. In several challenging applications, SVM achieves the best performance so far (ZHANG Xue Gong, 2000). In this paper, a

[*] Corresponding author.

D. Li and C. Zhao (Eds.): CCTA 2009, IFIP AICT 317, pp. 15–20, 2010.

compound kernel is proposed which is better than a single kernel to improve generalization and learning ability.

2 SVM Method

2.1 SVM Basis

Support Vector Machine is a machine learning algorithm based on statistical learning theory, using the principle of structural risk minimization, minimizing the errors of sample while shrinking the upperbound of generalization error of the model, therefore, improving the generalization of the model. Compared with other machine learning algorithms which are based on the principle of empirical risk minimization, statistical learning theory proposes a new strategy: the mechanism of SVM: find a optimal classification haperplane which satisfies the requirement of classification; separate the two classes as much as possible and maximize the margin of both sides of the hyperplane, namely, make the separated data farthest from the hyperplane. A training sample can be separated by different hyperplanes. When the margin of the hyperplane is largest, the hyperplane is the optimal hyperplane of separability (Cristianini et al. 2005).

2.1.1 Linear Separability

A two-class classification problem can be stated in following way: N training samples can be represented as a set of pairs (x_i ,y_i), i=1,2…n with y_i the label of the class which can be set to values of ± 1 and $x \in R^d$ stands for feature vector with d components. The hyperplane is defined as $g(x) = w \cdot x + b = 0$.

Find the optimal hyperplane which leads to maximization of the margin. The optimal question is then translated to seek the minimization of the following function (1) and (2):

$$\Phi(w,\varepsilon) = \frac{1}{2}(w \cdot w) + C(\sum_{i=1}^{n} \varepsilon_i \ ) \tag{1}$$

$$y_i[w \cdot x_i + b] - 1 + \varepsilon_i \geq 0 \ i = 1,\ldots,n \tag{2}$$

Where: w is normal to the hyperplane, C is a regularisation parameter, b is the offset.

The dual problem of the above problem is searching the maximization of the following function:

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{3}$$

which is constrained by

$$\sum_{i=1}^{n} y_i \alpha_i = 0, 0 \leq \alpha_i \leq C, i = 1,2,\cdots,n \tag{4}$$

The classification rule based on optimal hyperplane is to find the optimal classification function:

$$f(x) = \operatorname{sgn}\{(w^* \cdot x) + b^*\} = \operatorname{sgn}\{\sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^*\} \tag{5}$$

through solving the above problems, the coefficient α_i^* of a non support vector is zero.

2.1.2 Non-linear Separability

For a non-linear problem, we use kernel functions which satisfy Mercer's condition to project data onto higher dimensions where the data are considered to be linear separable. With kernel functions introduced, non-linear algorithm can be implemented without increasing the complexity of the algorithm. If we use inner products $K(x, x') =< \varphi(x), \varphi(x') >$ to replace dot products in the optimal hyperplane, which equals to convert the original feature space to a new feature space, therefore, majorized function of function (3) turns to:

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{6}$$

And the corresponding discrimination (5) turns to:

$$f(x) = \operatorname{sgn}\{\sum_{i=1}^{n} \alpha_i^* y_i K(x_i \cdot x) + b^*\} \tag{7}$$

2.2 Kernel Function

Kernel functions have properties as follows:

Property 1: If K_1 , K_2 are two kernels, and a_1 , a_2 are two positive real numbers, then K(u , v) = a_1* K_1 (u , v) + a_2 * K_2 (u, v) is also a kernel which satisfies Mercer's condition.

Property 2: If K_1 , K_2 are two kernels, then K(u , v) = K_1 (u , v) * K_2 (u , v) is also a kernel which satisfies Mercer's condition.

Property 3: If K_1 is a kernel, the exponent of K_1 is also a kernel, that is, K(u , v) = exp (K1 (u , v)) (XIA Hongxia, 2009).

There are 4 types of kernels which are often used: linear kernels, polynomial kernels, Gauss RBF kernels, sigmoid kernels.

(1)linear kernels : $K(x_i \cdot x_j) = x_i^T x_j$	
(2)polynomial kernels : $K(x_i \cdot x_j) = (\gamma x_i^T x_j + r)^d$	
(3)RBF kernels $K(x_i \cdot x_j) = \exp(-\gamma \|x_i - x_j\|^2)$	
(4)sigmoid kernel: $K(x_i \cdot x_j) = \tanh(\gamma x_i^T x_j + r)$	

3 Compound Kernels

Currently, there are so many kernels each of which has individual characteristic. But they can be classified into two main types, that is local kernel and global kernels (Smits et al. 2002).

(1)local kernels

Only the data whose values approach each other have an influence on the kernel values. Basically, all kernels based on a distance function are local kernels. Typical local kernels are:

RBF: $K(x_i \cdot x_j) = \exp(-\gamma\left\|x_i - x_j\right\|^2)$	

(2)global kernels

Samples that are far away from each others still have an influence on the kernel value. All kernels based on the dot-product are global:

Linear: $K(x_i \cdot x_j) = x_i^T x_j$	
Polynomial: $K(x_i \cdot x_j) = (\gamma x_i^T x_j + r)^d$	
sigmoid: $K(x_i \cdot x_j) = \tanh(\gamma x_i^T x_j + r)$	

The upperbound of the expected risk of SVM is the proportion of the average number of support vector in the training sample to the total number of training samples: E[P(error)]≤E[number of support vector]/(total number of support vector in train samples-1)

We can get that if the number of support vector is reduced, the ability of generalization of SVM can be improved. Therefore, Gauss kernels can be improved as follows:

$$K(x_i \cdot x_j) = a * \exp(-\gamma\left\|x_i - x_j\right\|^2)\gamma > 0 \tag{8}$$

Through adding a coefficient a which is a real number greater than 1, the absolute value of the coefficient of the quadratic term of quadratic programming function in equation (8) is increased. Hence, the optimal value of α is reduced, and the number of support vector is reduced, the ability of generalization therefore is improved.

If the total number of train samples is fixed, the error rate of classification can be reduced by decreasing the number of support vectors.

If kernel functions satisfy Mercer's condition, the linear combination of them are eligible for kernels. Examples are:

$$K(x_i \cdot x_j) = a * K_1(x_i \cdot x_j) + b * K_2(x_i \cdot x_j) \tag{9}$$

Where both a and b are real numbers greater than 1, K_1 , K_2 can be any kernel. Global kernels have good generalization, while local kernels have good learning ability. Hence, combining the two kernels will make full use of their merits, which achieves good learning ability and generalization.

For a compound kernel, four parameters need to be confirmed. The values of the parameters have great effect of the accuracy of classification. Optimal (C,α,a,b) is needed for the compound kernel.

4 Experimetal Results and Discussion

In this paper, a remote sensing image which resolution is 30 meters of rice paddy in Guangdong province is selected to test the results, the program of classification is written based on libsvm. The results are showed in table 1 as follows:

Table 1. The classification accuracy using different kernels

Number of pixels	kernel	Number of sv	Accuracy
350	RBF C=1 γ=0.1	310	91.2%
350	linear	210	85.3%
350	Compound of Linear and RBF C=2 γ=0.5 a=1 b=3	270	93.1%

Table 1. shows result of the classification of the remote sensing image. The classification method of compound kernel has the highest accuracy, and needs less number of sv.

From the result, we can get that the compound kernel has good ability of generalization and learning. Compared to the RBF kernel, the compound kernel has a higher accuracy of classification but the number of support vector is lower than it. Hence, the compound kernel achieves good generalization ability. In the test, the value of (C,α,a,b) has a great effect on the accuracy of classification. For different remote sensing image and different, different compound kernels and (C,α,a,b) should be selected. The optimal set of compound and value of (C,α,a,b) should be fixed through repeated trails.

5 Conclusion

In conclusion, the compound kernels yield better results than the single kernel. The compound kernels need less number of support vectors, which means that the kernels have good generalization ability and achieve higher classification accuracy than single kernels.

References

Burges, C.J.: A tutorial on support vector machines for pattern recognition. In: Fayyad, U. (ed.) Data mining and knowledge discovery, pp. 1–43. Kluwer Academic, Dordrecht (1998)

Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines and other Kernel-based learning methods: House of Electronics Industry (2005)

Smits, G., Jordaan, E.: Improved SVM regression using mixtures of kernels. In: IJCNN (2002)

Vapnik, V.N.: The nature of statistical learning theory. Springer, New York (1995)

Vapnik, V.N.: Statistical Learning theory. Wiley, New York (1998)

Xia, H., Ding, Z., Li, Z., Guo, C., Song, H.: A Adaptive Compound Kernel Function of Support Vector Machines. Journal of Wuhan University of Technology, 2 (2009)

Zhang, X.G.: Introduction statistical learning theory and support vectormachines. Act Autom atica Sinica 1(26), 32–42 (2000)