Data-oriented composite kernel-based support vector machine for image classification

Jiakui Tang*[a], Xianfeng Zhang[b], Xiuwan Chen[b], Jie Zhang[c],
Xiaohu Wen[a], Zhidong Zhang [a], De Wang [a]

[a]Yantai Institute of Coastal Zone Research for Sustainable Development
Chinese Academy of Sciences, Yantai, 264003, China
[b]Institute of Remote Sensing & GIS, Peking University, Beijing, 100871, China
[c]The First Institute of Oceanography, State Oceanic Administration, Qingdao, 266061, China
E-mail: jktang@yic.ac.cn

ABSTRACT

One novel composite kernel based support vector machine (SVM), which is called DOCKSVM (Data Oriented Composite Kernel based Support Vector Machine) is proposed in the paper. SVM have been proved good potential in various studies, and tried to application for pattern classification problems such as text categorization, image classification, objects detection etc. Recently, more and more researches show that SVM is promising in remote sensing image classification. Unlike traditional SVM method, DOCKSVM could integrate the bio-geophysical character into final classification through the composite kernels, which lead to the accuracy improvement of classification results. Firstly method of DOCKSVM is described in detail, then the novel method according to information entropy of training data to evaluate the weighted value of kernels is proposed, finally, preliminary results of application to remote sensing image classification is given which show that it's good potential tool for remote sensing image classification.

Keywords: Support Vector Machine, Composite Kernel, mage Classification, Remote Sensing

1. INTRODUCTION

The Support Vector Machine (SVM) is a relatively recent approach introduced by Boser [1] and Vapnik [2] for solving supervised classification and regression problems. SVM provides a new way to design classification algorithms which learn from examples (supervised learning) and generalize when applied to new data. SVM approach does not suffer from the Hughes phenomenon, which is for a limited number of training samples data, the classification accuracy rate decreases as the dimension of samples data increases. For the linearly separable training data, SVM is to find the optimal separation surface between classes thanks to the identification of the most representative training samples of the side of the class, for the no-linearly separable, kernel method is used to project the data into high dimension space, where the classes are linearly separable.

Composite kernels based SVM have demonstrated excellent performance for classification problems in the terms of accuracy and robustness. Joachims et al. (2001)[3] create two kernels based on text and hyperlink features respectively, and linearly combine them by assigning the same weight into a composite kernel for web page classification, the final composite kernel has good performance than the single kernel. Sun et al. (2004) [4] proposed an way to optimize the linear combination of kernels, experiments results of the method for classifying Web pages show that composite kernel has a better generalization performance. Jang et.al.,2007[5]proposed a composite kernel method using polynomial kernel and RBF kernel, the results of application to time serials prediction problems show that good accuracy the single kernel.

Recently, composite kernel method is more and more tried in remote sensing image classification. Mercer and Lennon (2003) [6] propose mixture of kernel method, by combining multiple kernels linearly, spectral and spatial information can be well integrated into classification, the results of application to CASI image classification shows better accuracy. Camps-vall et al. (2006) [7]present a framework of composite kernels method for hyperspectral images classification, which construct a family of composite kernels that easily combine spatial and spectral information, the results experiments for application to AVIRIS data show good performance while the optimal weight between kernels is

Second International Conference on Earth Observation for Global Changes,
edited by Xianfeng Zhang, Jonathan Li, Guoxiang Liu, Xiaojun Yang,
Proc. of SPIE Vol. 7471, 74711X · © 2009 SPIE · CCC code: 0277-786X/09/$18

selected. Gu et al. (2007) [8] demonstrated composite kernels could integrate spatial and spectral information for hyperspectral image classification, experiments results show that better results compared with Maximum Likelihood Classifier (MLC) and the introduction of spectral-spatial kernels can greatly improve classification accuracies.

Composite kernels method can well integrate different features information of the samples data, which results into better classification accuracy then single kernel. The important factors to affect the performance of Composite kernels method not only include the kernels combination structure but also kernel's function structure, which should be concerned when using Composite kernels method. More research carried have shown that the kernel function structure selection have an influence on SVM [9], modification of kernel function structure to take into bio-physical meaning of training data account have been demonstrated to improve SVM performance[6][10]. For kernels combination structure, the weight value of kernel act as efficient tradeoff parameter to account for the contribution of different features information to the classification result, which mean the weight value selected is one of the important parameters other than kernel function structure, however, until now how to decide the weight value of kernel remains a problem, step search method is commonly used while time-consuming, furthermore, the same weight value of kernels for all classifiers, which means that all the SVM classifiers between classes are assigned the same weight value of kernels.

Data Oriented Composite Kernel Based Support Vector Machine (DOCKSVM) is proposed in this paper, which develop the method of construct composite kernels for SVM according to the classified data with different weight value of kernels for SVM classifiers, and the method of kernels weight value evaluation based entropy of training data is firstly proposed.

The paper is outlined as follows. Section 2 briefly reviews the theory basis of SVM classifiers. Section 3 described the method of DOCKSVM in detail. Section 4 presents the preliminary experiments of application to remote sensing classification. In Section 5, the conclusions and final remarks are given.

2. THE SVM CLASSIFIERS

2.1 Theoretical background of SVM

A comprehensive introduction to SVM is given in [11]. The theoretical background of SVM is just reviewed briefly as follows. SVM is originally formulated to construct binary classifiers from a set of training examples, given a labeled training samples dataset consisting of pairs of class labels and n -dimensions feature vectors as (y_i, x_i) , i=1,…, l, $y_i \in \{1,-1\}$, $x_i \in R^n$, the SVM approach is to find the separating hyperplane $<(w, x>+b) = 0$, $x \in R^n$, $b \in R^n$, so that:

1) Samples with labels $y_i = \pm 1$ are located on each side of the hyperplane;

2) The margin defined as the distance of the closest vectors to the hyperplane in each side to the hyperplane is maximized.

The classifiers are allowed to be defined as: $f(x) = Sgn(< w, x > +b)$, the support vectors lie on two optimal hyperplanes of the equations: $< w, x > +b = \pm 1$.

The maximization of the margin leads to the following constrained optimization problem:

$$\text{Min } \{\frac{1}{2}\|w\|^2\} \text{ with } y_i(<(w, x_i > +b) \geq 1, \ i=1,\ldots, l \tag{1}$$

where w and b define a linear classifier in the feature space, the maximal margin classifier can be generalized to nonlinearly separated data via two approaches. One is introduce a soft margin parameter C to relax the constraint that all the training vectors of a certain class lie on the same side of the optimal hyperplane, which is effective to noisy data. The other is to transform input vectors into a higher dimension feature space by a map function ϕ that is performed in accordance with Cover's theorem [12], which guarantees more likely to be linearly separable in the resulting space, then get the form as follow:

$$\text{Min}\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i\} \tag{2}$$

Constrained to:

$$y_i(<\phi(x_i, w> +b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0,\ i=1,\ldots,l$$

C is the regularization parameter that control the capabilities of the classifier and it must be selected by the user, and ξ_i are positive slack variable enabling to deal permitted errors.

By introducing the Lagrange multiples $\alpha_i \geq 0$, the optimization problem can be translated into the following form:

$$\text{Max } w(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j < \phi(x_i), \phi(x_j) > \tag{3}$$

$$\text{Subject to: } \sum_{i=1}^{l}\alpha_i y_i = 0 \text{ and } C \geq \alpha_i \geq 0,\ i=1,\ldots,l$$

Only a small number of multiples α_i of nonzero values which are associated with the so-called support vectors form the boundaries of the classes. Kernel function can be introduced to reduce the expensive computation of inner product of mapping function ϕ:

$$k(x_i, x_j) = < \phi(x_i) \cdot \phi(x_j) > \tag{4}$$

Then a nonlinear SVM can be constructed using only the kernel function without having to considering the mapping function ϕ explicitly. After solving the dual problem by introducing Equation (4) into (3), the decision function for any test vector x can be expressed in terms of kernel expansion as:

$$f(x) = \text{sgn}(\sum_{i=1}^{l} y_i \alpha_i k(x_i, x) + b) \tag{5}$$

Where b can be easily computed from α_i as explained in [13]

2.2 Multiple class separation strategy for SVM

SVM is designed to solve two-class classification problems, two popular approaches can be used for M-class problems:

1) One against one (OAO): each pair classes have to be computed for a SVM classifier, so $\frac{M(M-1)}{2}$ classifiers should be acquired.
2) One against all (OAA): only M classifiers should be obtained by iteratively computation each class against all the others.

The OAO approach is used in this paper, although more SVM classifiers needed to be applied, the computing time can be improved because the complexity of the algorithm depends on strongly the number of training samples.

3. METHOD OF DOCKSVM

SVM is actually one kernel based method, kernel function is a very important factor for the performance of SVM, any kernel function K which fulfils MERCER'S condition [6] can be used in SVM, however, different kernel function lead into different performance of SVM. In mathematics, dot product for two vectors is meaningfully an assessment of the distance or similarity between them, so that kernel maybe taken as an assessment of the distance or similarity two training samples in feature space, each feature of samples represents deferent bio-physical properties so that each feature had better use specially designated distance or similarity measure function[14][15], the idea of DOCKSVM is to modify the kernel function for each feature by using corresponding distance or similarity measure function according to bio-physical properties of sample data, the composite kernels finally obtained are expected to well integrate different features information of the samples data, and to get better classification accuracy then single kernel. DOCKSVM, one composite kernel based SVM approach is presented in detail in the following:

Given a kernel function K, for x , $z \in X$, there will be:

$$k(x,z) = <\phi(x) \cdot \phi(z)> \tag{6}$$

Where ϕ is the mapping function for X to feature space.

Suppose k_1 and k_2 are kernel function that fulfils Mercer's condition, composite kernels as follows should be kernel function and still fulfils Mercer's condition [11].

$$k(x,z) = k_1(x,z) + k_2(x,z) \tag{7}$$

$$k(x,z) = ak_1(x,z) \text{ , } a \in R \tag{8}$$

According to Equations (7) and (8), Joachims [3] 2001 deduced the valid kernel function in the following and applied to hypertext categorization:

$$k(x,z) = \lambda\, k_1(x,z) + (1-\lambda)k_2(x,z) \tag{9}$$

Four approaches of composite kernels are proposed by Camps-valls in 2006 [7], in which, weighted summation kernel approach is referred, furthermore, according to Equations (7), (8) and (9), weighted summation kernel approach can be generalized in the following:

Given a training samples dataset X with feature space of s dimension, the number of s kernel function $k_p(x_i^p, x_j^p)$, $p = 1$, 2 , …… s, is defined, which are applied on each feature domain, then the composite kernels based feature can be in the following form:

$$\begin{aligned} k(x_i, x_j) &= \sum_{p=1}^{s} \lambda_p k_p(x_i^p, x_j^p) \\ &= \lambda_1 k_1(x_i^1, x_j^1) + \lambda_2 k_2(x_i^2, x_j^2) +,\ldots\ldots,+ \lambda_s k_s(x_s^s, x_j^s) \end{aligned} \tag{10}$$

With $\lambda_1 + \lambda_2 +,\ldots..,+ \lambda_s = 1$

where λ_p is the weight value of the kernel function $k_p(x_i^p, x_j^p)$.

Suppose that the number of training samples is m, kernel matrix $K = (k(x_i, x_j))_{i,j=1}^{m}$ of $m \times m$ dimension should be obtained in the training procedure of SVM, the element of K stand for the value of distance or similarity for one pair of samples. The weight value λ_p stands for the contribution of the corresponding feature information to the classification. Obviously, Single kernel based SVM is the special case in Equation (10)

For composite kernel based SVM, is expected to improve the classified accuracy, however, uncertainty increases when the parameter of weight value λ is introduced. How to evaluate λ is the key for the performance of composite kernels. Step search in [0, 1] is proposed by Camps-valls in 2006 [7], which is time-consuming when a less step is used for good accuracy, so that the tradeoff between accuracy and training speed has to be done, furthermore, the same weight value of kernels for all classifiers, which means that all the SVM classifiers between classes are assigned the same weight value of kernels. Actually, for multiple classes separation problems, one group of SVM classifiers should be applied, different feature maybe have the different contribution on each classifier, so rational λ is expected for each corresponding classifier.

OAO strategy is only taken in DOCKSVM. Given the number of training samples X is N, and the dimension of feature space is s ,then there are N（N-1）/2 classifiers needed to be applied, the form of the classifier q（q=1,2,……,N（N-1）/2）is as follow:：

$$k_q(x_i,x_j)=\sum_{p=1}^{s}\lambda_{qp}k_p(x_i^p,x_j^p)$$
$$=\lambda_{q1}k_1(x_i^1,x_j^1)+\lambda_{q2}k_2(x_i^2,x_j^2)+,\ldots\ldots,+\lambda_{qs}k_s(x_s^s,x_j^s) \tag{11}$$

where， $\lambda_{q1}+\lambda_{q2}+,\ldots..,+\lambda_{qs}=1$ ， q =1,2,……N（N-1）/2

Since the parameter λ is the key factor for DOCKSVM performance, it is worth nothing that to focus on how to obtain the parameter λ , the method of information entropy based composite kernels weights determination is proposed in the following.

Entropy is a measure of the uncertainty associated with a random variable, which come from information theory [16]. Given a discrete random variable X with possible values $\{x_1,\ldots,x_n\}$, If $p=(p_1,p_2,\ldots\ldots p_n)$ denotes the occurrence probability vector of the element of X , the entropy H is:

$$H(p)=H(p_1,p\ \ ,\ldots\ldots,p_n)=-\sum_{i=1}^{n}p_i\times\ln p_i \tag{12}$$

With $\sum_{i=1}^{n}p_i=1$, $0\le p\ \ \le 1$

Information entropy can be used to feature extraction for machine leaning classification problems [17][18], the more less of Information entropy of the feature, the larger the certain information, the more important the feature affect the classification results, i.e. the weight of the feature, and vice versa.

For one classifier of formula (3.5), it is only one classifier for two classes training samples, so two classes (m =1,2) separation problem is considered here. Suppose that there are two classes training samples with s feature Domains, and the i th feature Domains has n_i feature parameters, the sum of feature parameters of all s feature Domains is $n=\sum_{i}^{s}n_i$, for the feature $F_j(j=1,2,\ldots\ldots n)$, the distribution range of the feature F_j can be equally divided into M sub-range, $r_k(j)$, $k=1,2,\ldots\ldots M$,then the probability of samples meet the condition of $F_j\in r_k(j)$ belongs to m th class is:

$$p_{km}(j)=\frac{M_{km}(j)}{M_k(j)} \tag{13}$$

where, $M_k(j)$ is the sum of samples meet the condition of $F_j\in r_k(j)$, $M_{km}(j)$ is the sum of samples $M_k(j)$ belongs to m th class.

$$M_k(j) = \sum_{m=1}^{2} M_{km}(j) \tag{14}$$

Suppose that $p_k(j)$ is the probability of one sample meet the condition of $F_j \in r_k(j)$,then :

$$p_k(j) = \frac{M_k(j)}{M_0} \tag{15}$$

M_0 is the sum of total samples , which is in the form :

$$M_0 = \sum_{k=1}^{M} M_k(j) \tag{16}$$

The entropy of the j th feature F_j can be defined as:

$$H(F_j) = -\sum_{k=1}^{M} p_k(j) \sum_{m}^{2} p_{km}(j) \log_2 p_{km}(j) \tag{17}$$

The less $H(F_j)$ is, the larger the contribution of F_j to classification, i.e., the larger the corresponding kernel weight.

Suppose that: $H^{'}(F_j) = \dfrac{1}{H(F_j)}$, $H_0 = \sum_{j=1}^{n} H^{'}(F_j)$ (18)

The normalization value of feature F_j contributing to kernel weight can be defined as :

$$\omega_j = \frac{H^{'}(F_j)}{H_0} \text{ , j=1,2,……,} n \tag{19}$$

In DOCKSVM, the kernels are applied on feature domain, then the kernel weight λ_i corresponding to the i th feature Domains can be defined as:

$$\lambda_i = \sum_{j=1}^{n_i} \omega_{ij} \text{ , where , } i = 1,2,\ldots\ldots s \tag{20}$$

There are N(N-1)/2 classifiers when DOCKSVM is used to N class separation problem, the kernels weight λ of all N(N-1)/2 classifiers can be obtained after the abovementioned method of information entropy iterately applied to each pair classes training samples.

4. EXPERIMENTS AND RESULTS

DOCKSVM is an approach of composite kernels bases SVM for classification problems, to test its performance, the experiments of application on ship classification from remote sensing images is carried out. Data used in the experiments are IKONOS panchromatic images with 1 meter spatial resolution taken over sea area, in which three kinds of ships training samples and test samples are obtained by manual interpretation. The features used for ship classification include spectral, texture and geometrical shape.

For using the spatial geometrical shape, the segmentation of the image is important for the classification, the experiments are aimed to test the performance of DOCKSVM for classification, and segmentation by hand of the image is used to guarantee the segmentation accuracy.

According to the features include spectrum, texture and geometrical shapes, the composite kernels can be defined as follow:

$$k(x_i,x_j)=\lambda_1 k_{spec}(x_i^{spec},x_j^{spec})+\lambda_2 k_{text}(x_i^{text},x_j^{text})+\lambda_3 k_{shap}(x_i^{shap},x_j^{shap}) \tag{21}$$

With $\lambda_1+\lambda_2+\lambda_3=1$, $k_{spec}, k_{text}, k_{shap}$ stand for the kernels to spectrum, texture and geometrical shapes feature domain respectively,

Modified Radial Basis Function (RBF) kernel is used by introducing different distance or similarity measure function of each feature as follow:

$$\text{RBF kernel：}\quad k(x_i,x_j)=\mathrm{e}^{-\gamma\|x_i-x_j\|^2} \tag{22}$$

$$d(x_i,x)=\|x_i-x_j\|^2$$

For spectral kernel k_{spec}, spectral angle measure function is taken：

$$d(x_i,x_j)=\arccos(\frac{x_i\cdot x_j}{\|x_i\|\|x_j\|})) \tag{23}$$

For texture kernel k_{text}, Canberra distance measure function is taken

$$d(x_i,x_j)=\frac{\|x_i-x_j\|}{\|x_i\|+\|x_j\|} \tag{24}$$

For geometrical shape kernel k_{shap}, keep the Euclidian distance measure function of RBF.

The spectral feature domain includes mean and standard deviation, the texture feature domain includes derived from Gray Level Co-occurrence Matrix, the geometrical shape feature domain includes derived from Hu's moment invariants.

Table 1. The feature data of 3 selected training samples

samples	spectrum		texture			Hu's moment invariants Of geometrical shapes						
	Mean	Standard deviation	Entropy	Second-order moment	Contrast	M1	M2	M3	M4	M5	M6	M7
Class-1	77.5	1106.1	2.1	0.12	2832.8	0.46	1.02	3.41	3.61	7.09	4.13	8.13
Class-2	147.5	1953.2	2.2	0.11	4842.2	0.60	1.45	3.16	3.64	6.66	4.37	7.64
Class-3	79.6	605.1	2.2	0.11	1483.2	0.40	0.89	3.95	4.19	8.21	4.69	8.90

The kernel weight value λ for each DOCKSVM classifier obtained by using the method of information entropy is listed in Table 2.

Table 2. The kernel weight for all the classifiers

DOCKSVM classifier	λ_1	λ_2	λ_3
Class-1 to class-2	0.54	0.18	0.28
Class-1 to class-3	0.21	0.46	0.33
Class-2 to class-3	0.39	0.16	0.45

The penalization parameter C and the kernel width parameter σ are derived by using the method of K-fold Cross Validation which is proposed in [19]. C =128.0 and σ =0.125 is used for the trained DOCKSVM.

Eight test samples is used to test the performance of the trained DOCKSVM, for the sparseness of experiment samples, accuracy of classification is defined as:

$$ASC = \frac{n_{right}}{n_{all}} \times 100\% \text{ , and } OA = \frac{N_{right}}{N_{all}} \times 100\% \qquad (25)$$

where ASC is the accuracy for single class, n_{right} and n_{all} are the sum of correct classified samples and the sum of all samples of single class, respectively. OA is the overall accuracy, N_{right} and N_{all} are the sum of all correct classified samples and the sum of all samples of all classes, respectively.

The data of classification accuracy can be seen in Table 3, not only accuracy for single class but overall accuracy are 100%, which show that the promising performance of DOCKSVM.

Table 3. The accuracy of ship classification using DOCKSVM

Ship classes	correct classified samples	all samples of single class	OSC (%)	OA (%)
Class-1	1	1	100	
Class-2	3	3	100	
Class-3	4	4	100	
Total samples	8	8		100

5. CONCLUSION

DOCKSVM is proposed in this paper, which is aimed to get composite kernels according to bio-physical properties of the feature parameters of training data, and to take advantage of the prior entropy information of training samples, all the SVM classifiers can obtain different kernel weight when OAO separation strategy is taken. Compared with traditional composite kernel based SVM, the novel DOCKSVM generalize the weighted summation kernel approach by exploiting the properties of MERCER'S kernels, which using different modified kernel function according to distance or similarity measure function of each feature domain, so that the feature information can be well combined into the classifiers. Moreover, DOCKSVM realize that each classifier has special kernel weights which accounts for different feature maybe have no the same contribution to the classification results, the proposed method of information entropy to evaluate the kernel weights can be expected to get the rational kernel weights instead of step search method.

Experiments of testing the performance of DOCKSVM is also carried out, results of which show that DOCKSVM have a good classification accuracy when used to object classification from remote sensing image, DOCKSVM is not only designed for remote sensing Image Classification application, however, it maybe Potential Tool for remote sensing Image Classification in the future.

The sparseness of samples restricted the further test of DOCKSVM. More experiments to test the performance of DOCKSVM need to be conducted and other methods of kernel weights evaluation should be studied in the future.

ACKNOWLEDGEMENTS

This study is supported by the projects “Aerosol Retrieval Using Remote Sensing Data Over Costal Zone Based Grid Technology” supported by NSFC, China (40801124) , “Research on Novel Data Oriented SVM for Remote Sensing Information Extraction” supported by CAS, China and "China coastal comprehensive oceanic resource survey and evaluation of 908 program" funded by SOA, China. The authors would like to thank Prof. Peijun Li at Peking University for his valuable advices and discussions.

REFERENCES

[1] Boser, B. E., Guyon, I. M. and Vapnik, V. N., A Training Algorithm for Optimal Margin Classiers, Fifth Annual Workshop on Computational Learning Theory. ACM, 144-152, (1992)

[2] Vapnik, V. N., [The Nature of Statistical Learning Theory], Springer-Verlag, New York, (1995).

[3] Joachims, T, Cristianini, N, Shawe-Taylor, J. Composite kernels for hypertext categorisation, Proc. of the International Conference on Machine Learning, 250-257 (2001)

[4] Sun, J.T., Zhang, B. Y., and Chen, Z., GE-CKO: A method to optimize composite kernels for Web page classification, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04) (2004).

[5] Jiang, T.J., Wang, S. Z., and Wei, R.X. "Support vector machine with composite kernels for time series prediction", LNCS 4493, 350–356 (2007).

[6] Mercier, G., Lennon, M., Support vector machines for hyperspectral image classification with spectral-based kernels, Geoscience and Remote Sensing Symposium, vol.1, p288- 290, (2003)

[7] Camps-Valls, G., Chova, L. G., et al, Composite kernels for hyperspectral image classification, IEEE Trans. Geosci. Remote Sensing, 3(1), 93–97 (2006)

[8] Gu, Y., Liu, Y., and Zhang, Y., A soft classification algorithm based on spectral-spatial kernels in hyperspectral images, 2nd International Conference on Innovative Computing, Information and Control, 5-7 Sept. , 548 – 548(2007)

[9] Roli, F, Fumera G, Support vector machines for remote-sensing image classification, Image and Signal Processing for Remote Sensing, 4170, 160-166 (2001)

[10] Amari, S., Wu, S., Improving support vector machine classifier by modifying kernel function, Neural Networks, 12: 783-789 (1999)

[11]Cristianini, N., and Shawe-Taylor, J., An Introduction to Support Vector Machines. Cambridge, U.K.: Cambridge Univ. Press (2000)

[12]Cover, T. M., Geometrical and statistical properties of systems of linear inequalities with application in pattern recognition, IEEE Trans. Electron. Comput., vol. EC-14, pp. 326–334 (1965)

[13] Scholkopf, B., and Smola, A., Learning With Kernels - Support Vector Machines, Regularization, Optimization and Beyond. Cambridge, MA: MIT Press (2002).

[14] Ma, Y., Lao, S., Takikawa, E., Kawade, M., Discriminant analysis in correlation similarity measure space, Proceedings of the 24th International Conference on Machine learning, Vol. 227, 577 -584 (2007)

[15]Weisberg, A., Najarian, M., Borowski, B., Spectral Angle Automatic CLuster Routine (SAALT): An Unsupervised Multispectral Clustering Algorithm, http ://citeseer. nj. nec.com, (2001)

[16]Thomas, M., Cover, J., Thomas, A., Elements of information theory New York: Wiley, ISBN 0471062596, (1991)

[17]Pang, X., Feng, Y., Jiang, W., an improved document classification approach with maximum entropy and entropy feature selection, International Conference on Machine Learning and Cybernetics, Hong Kong (2007)

[18]Shifei D., Zhongzhi S., Supervised feature extraction algorithm based on improved polynomial entropy, Journal of Information Science, 32:309 (2006)

[19]Fan, R.E., Chen, P.H., and Lin, C.J. Working set selection using the second order information for training SVM. Journal of Machine Learning Research 6, 1889-1918 (2005)