Support Vector Machine Training of HMT Models for Multispectral Image Classification

Reda A. El-Khoribi†

Faculty of Computers and Information, Cairo University, Giza, Egypt

Summary

This paper introduces a novel approach to supervised classification of multispectral images. The approach uses a new discriminative training algorithm for discrete hidden Markov tree (HMT) generative models applied to the multi-resolution ranklet transforms. System is implemented and tested on a set of Landsat 7-band images containing eight different land cover classes. Experimental results of the system show significant improvement over the baseline HMT system and give a superior performance in land cover classification.

Key words:

HMT, SVM, land cover classification, discriminative training.

1. Introduction

Land cover is one of the crucial elements for scientific research and real-life earth science applications. Remotely sensed data acquired by the Earth observation satellites provides a number of benefits for studying the Earth's surface, such as continuous acquisition of data, broad regional coverage, cost effective data, map-accurate data, and large archive of historical data. Knowledge of land cover is important in a variety of natural resource applications.

Much research has been done on the subject of automatically determining land use and land cover both in rural and urban areas from images. Several classification algorithms have been developed, and successfully implemented, for land cover classification from multispectral and hyperspectral data ([1]-[3]). Classical statistical methods of classification have been worked on for several decades. Recently there have been many new developments in pattern classification research, and many new applications have been studied.

Remote-sensing classification is a complex process and requires consideration of many factors. The major steps of image classification may include determination of a suitable classification system, selection of training samples, image preprocessing, feature extraction, selection of suitable classification approaches, post-classification processing, and accuracy assessment. The user's need, scale of the study area, economic condition, and analyst's skills are important factors influencing the selection of remotely sensed data, the design of the classification procedure, and the quality of the classification results [4].

HMTs, introduced by (Craouse et al) in [5], have been used in image de-noising, segmentation and classification. As classifiers, although being a good model of wavelet coefficients, they have a major shortcoming. They are trained non-discriminatively using maximum likelihood estimation to model the joint probability of the observation and label trees. On the other hand support vector machines have their powerful discriminative training engine. Several research efforts have been exerted on building hybrid classifiers that benefit from the advantages of generative models and discriminative models ([6]-[9]). This paper introduces a merger between HMTs and SVMs to be applied to the field of land cover classification. The rest of the paper is organized as follows:

Section 2 is about the support vector machines. In section 3, a brief overview of HMTs is introduced. Section 4 describes the hybrid system. Section 5 contains a description of the data set and methodology used. Section 6 contains the results of some experiments to justify the proposed improved training.. The conclusion is then presented in section 7.

2. Support Vector machines

Support vector machine (SVM) ([10], [11]), is a powerful machine learning tool that has been widely used in the field of pattern recognition. The support vector machine optimization problem attempts to obtain a good separating hyper-plane between two classes in the higher dimensional space. The equation of the hyper-plane is:

$$\langle w, x \rangle + b = 0 \qquad (1)$$

where w is a weight vector and b is the bias and $\langle .,$ denotes the inner product. Non-linearity in SVM is

Manuscript received September 5, 2008.
Manuscript revised September 20, 2008.

satisfied by mapping the input features x into higher dimensions using a function $\phi(x): R^d \rightarrow R^p, p >$ and hence the hyperplane equation becomes:

$$\langle w, \phi(x) \rangle + b = 0 \qquad (2)$$

This leads to the following optimization problem:

$$\min_{\xi, w, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \qquad (3)$$

Subject to

$$y_i(\langle w, \phi(x_i) \rangle + b) \geq 1 - \xi_i, \quad i = 1.. ,$$

$$\xi_i \geq 0, \quad i = 1..$$

C is some constant determined by a cross validation process.

The dual formulation of this problem is:

$$\max_{\lambda} \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j K(x_i, x_j) \qquad (4)$$

Subject to:

$$\sum_{i=1}^{N} \lambda_i y_i = 0$$

,

$$0 \leq \lambda_i \leq C, i = 1..N$$

The function $K(x_i, x_j) = \langle \phi(x_i), \phi(x$ is called a kernel function. In SVM literature, there are many forms of the kernel function. If the probability density function of the feature vectors in both classes is known, there is a possibility of defining natural kernels derived from these distributions [9].

3. Discrete hidden Markov tree generative model

Crouse et al introduced a parametric model for wavelet coefficients called hidden Markov tree [5]. They devised a training algorithm, named upward-downward algorithm, similar to the known Baum Welch algorithm for the hidden Markov models. The overflow problems appearing in the computations lead to introducing another upward downward algorithm based on smoothed probabilities in [12]. In [13], a discrete HMT multi-observation was introduced and used in classifying breast tumors. In this method the ranket transform coefficients (a multi-resolution transform [14] and [15]) were viewed as a vector quad tree in which each node contains a vector of 3-components corresponding to vertical, horizontal and diagonal bands. The k-means algorithm was then applied to the vectors and a codebook was constructed. The codebook was then used to quantize the vectors of the vector quad trees transforming them into scalar quad trees. The discrete HMT was applied to scalar trees.

To describe the discrete HMT, we will use the following notation:

A tree is represented as a set of nodes $\{x_1, x_2, \dots, x$ with a partial order such that if $v >$ then is in the same level or a next level of . The notation $\rho($ means parent of u while $c($ means the set of children of u. the notation means the sub-tree starting at node u so denotes the whole tree.

In the context of discrete HMT models, the inputs are called observation trees $\{x_1, x_2, \dots, x$. The properties of the discrete HMT are as follows:

[1] Each tree node is related to a K-valued discrete random variable representing its hidden state. That's we have an additional tree, namely, the hidden state tree $\{q_1, q_2, \dots, q$, that could be represented in the same way as the observation tree.

[2] The distribution of the state variable of the root node is represented by a vector $\pi = [$ with $\pi_i = P(q_1 =$ and $i \in \{1, 2, \dots,$

[3] The state variable of a child node is statistically independent on each other variable given its parent node. This child parent dependency is represented by a state transition matrix $A = [a_i$ with $a_{ij} = P(q_u = j | q_{\rho(u)} =$ and $i, j \in \{1, 2, \dots, l$ for all nodes u. We assume, for simplicity, that the matrix does not change from level to level, i.e., the coefficients are constants.

[4] The conditional distribution of the feature element of the node u given its state is represented by the emission matrix $B = [$ with $b_{ij} = P(x_u = j | q_u =$ for $i \in \{1, 2, \dots, K\}$ *and* $j \in \{1, 2, \dots,$ and all nodes u where M is the number of observation symbols (codebook size).

The discrete HMT parameters are thus the triplet $\theta = \{\pi, A,$

4. Improved HMT training:

Given a set of classes $\{\omega_i | i = 1, 2, \dots,$ represented by N corresponding baseline HMTs with parameters $\{\theta_i, i = 1, 2, \dots,$. The Bayesian form of the Viterbi algorithm estimates the unknown class $\hat{\omega} = \omega$ using:

$$\hat{m} = \arg\max_m \log P(\omega_m) P(\tilde{x}_1, \tilde{q}_1^{(m)} | \theta_m) = \arg\max_m \left(\log P(\omega_m) + \log P(\tilde{x}_1, \tilde{q}_1^{(m)} | \theta_m)\right) \quad (5)$$

Using the properties of the HMT, $\log P(\tilde{x}_1, \tilde{q}_1|$ can be expressed as:

$$\log P(\tilde{x}_1, \tilde{q}_1|\theta) = \log\left(\pi_{q_1} b_{q_1}(x_1) \prod_{u \text{ non-leaf}} \prod_{v \in c(u)} a_{q_u q_v} b_{q_v}(x_v)\right)$$

$$= \log \pi_{q_1} + \log b_{q_1}(x_1) + \sum_{u \text{ is non-leaf}} \sum_{v \in c(u)} \left(\log a_{q_u q_v} + \log b_{q_v}(x_v)\right) \quad (6)$$

If one chooses to start all the state trees at a fixed state, then $\log \pi_{q_1} =$ and hence eq.6 can be written as:

$$\log P(\tilde{x}_1, \tilde{q}_1|\theta) = \sum_{i=1}^{K} \sum_{j=1}^{K} \varphi_{ij}(\tilde{q}_1) \log a_{ij} + \sum_{i=1}^{K} \sum_{j=1}^{M} \psi_{ij}(\tilde{x}_1, \tilde{q}_1) \log b_{ij} \quad (7)$$

where $\varphi_{ij}(\tilde{q}$ is the count of state transitions from state i to state j in the state tree and $\psi_{ij}(\tilde{x}_1, \tilde{q}$ is the number of times the symbol j of the observation tree is emitted by the state i of the state tree .

By constructing the two vectors:

$$T(\tilde{x}_1, \tilde{q}_1) = [\varphi_{ij}, \psi_{jk}; i = 1 \dots K, j = 1 \dots K, k = 1 \dots M] \quad (8)$$

$$W_\theta = [\log a_{ij}, \log b_{jk}; i = 1 \dots K, j = 1 \dots K, k = 1 \dots M] \quad (9)$$

Equation (7) can be written as an inner product:

$$\log P(\tilde{x}_1, \tilde{q}_1|\theta) = \langle T(\tilde{x}_1, \tilde{q}_1), W_\theta \rangle \quad (10)$$

Hence Equation (5) becomes:

$$\hat{m} = \arg\max_m \left(\langle W_{\theta_m}, T_m \rangle + \beta_m\right) \quad (11)$$

where $\beta_m = \log(P(\omega_m$ and the suffix m denotes the class m.

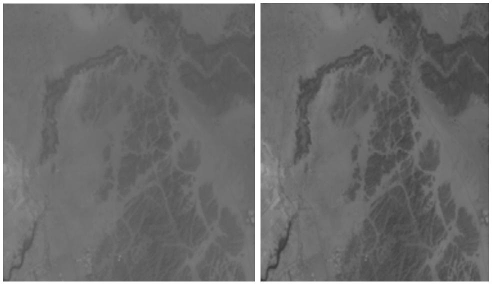
As done in [9] for hidden Markov models (HMMs), and noting that this equation is linear in the T variables with weights W and bias , we propose an improvement of the HMT training by trying to find the weights W and bias of the hyperplane with maximum separation between two classes. This is essentially a support vector machine problem. Hence our proposed system is summarized in the following steps:

1. Train the HMTs using the upward-downward procedure with smoothed probabilities ([12]) to obtain a set of baseline HMTs with parameters: $\{\theta_i, i = 1, 2, \dots,$
2. For each training sample apply the Veterbi decoding algorithm (see [12]) to get the best state tree $\tilde{q}_1^* = \arg\max_{\tilde{q}_1} P(\tilde{q}_1 | \tilde{x}_1, \theta$ where m is the class to which belongs and hence transform the samples into a fixed length vector $T(\tilde{x}_1, \tilde{q}$ using eq. 8.

3. Use the obtained fixed length vectors as inputs to linear discriminant classifiers (e.g. support vector machines or back propagation neural networks) to train the weights W and bias b of the classifier $\langle W, T \rangle +$

5. Data Sets and Methodology

Landsat7 ETM+ was used to acquire the images required for this study from Jeddah, Kingdom of Saudi Arabia. Figure 1 shows the infrared spectral bands 1-7 which were used as raw data to extract discriminate features for each class of eight land cover types, namely, Farm, grass, sea, rock1, rock2, soil1,soil2 and soil3.



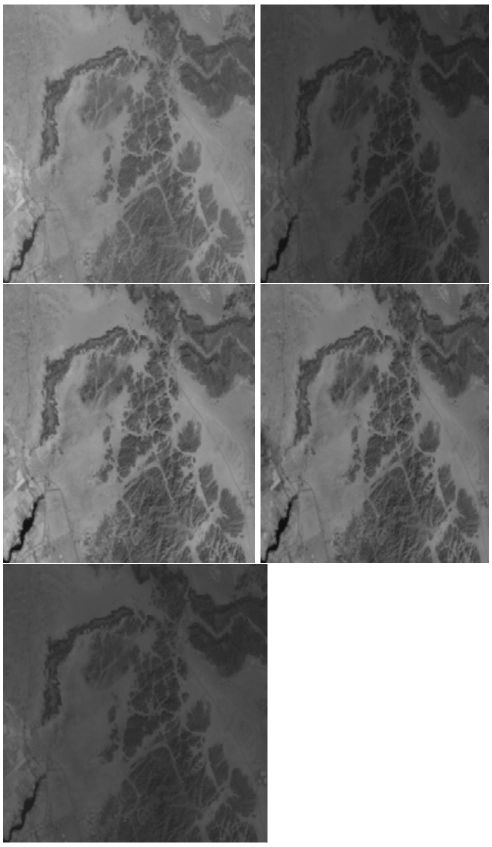


Figure 1: The 7 band images of the LANDSAT ETM+

These images were used for subsequent analysis and classification problem involved in identification of eight land cover types (i.e. farm, grass, rock1, rock2, rock3, sea, soil1, soil2 and soil3). A total of 5800 pixels were selected for all eight classes using stratified random sampling. The pixels collected were divided into two subsets, one of which was used for training and the second for testing the classifiers, so as to remove any bias resulting from the use of the same set of pixels for both training and testing. Also, because the same test and training data sets are used for each classifier, any difference resulting from sampling variations was avoided. A total of 2700 training and 3100 test pixels were used.

To test the effectiveness of the new proposed system, eight baseline HMT models were trained using the training samples from each land cover type. Each sample contains 3 layers of images corresponding to spectral bands 1, 4 and 7. Ranklet transform was applied to each image to construct a quad tree of 3-component vectors corresponding to the horizontal, vertical and diagonal components. The vector trees of the spectral bands are combined in one vector tree of 9 components. Vector quantization is then used to transform the vector tree into a scalar tree by quantizing the 9-component vectors. A codebook of size 512 was used in the vector quantization module. The scalar trees were then used with the upward downward algorithm with smoothed probabilities to train the eight baseline-discrete-HMTs. The number of states used in the HMTs was 5. The sample scalar trees of different classes were then converted into fixed length vectors as described above. The fixed length vectors were used in training multiclass SVM which uses 1 against all strategy. The simulations for SVM are carried out using compiled C-coded SVM packages: LIBSVM4 [16].

6. Results

The purpose of the present study is to evaluate the performance of the new training method and comparing its performance with the baseline HMT classifiers. Table 1 shows the confusion matrix resulting from applying the baseline HMTs while table 2 shows the confusion matrix of the proposed system with improved training. These Results justify the fact that the discriminative training of HMT generative models improves the classification accuracy by taking into account all the class samples at once.

Table 1 : Results of the baseline system

	Farm	Grass	Sea	Rock1	Rock2	Soil1	Soil2	Soil3
Farm	84	6	0	5	3	2	0	0
Grass	0	97	0	0	0	0	3	0
Sea	0	0	98	0	0	0	0	2
Rock1	7	1	0	80	1	4	0	7
Rock2	0	0	0	0	93	0	0	7
Soil1	0	0	0	3	0	88	5	4
Soil2	0	5	0	0	0	4	90	1
Soil3	8	5	0	0	0	2	14	71

Table 2: Result of the proposed system

	Farm	Grass	Sea	Rock1	Rock2	Soil1	Soil2	Soil3
Farm	97	3	0	0	0	0	0	0
Grass	1	97	0	0	0	0	2	0
Sea	0	0	100	0	0	0	0	0
Rock1	4	0	0	94	0	2	0	0
Rock2	0	0	0	1	99	0	0	0
Soil1	0	0	0	5	0	95	0	0
Soil2	0	1	0	0	0	5	93	1
Soil3	0	3	0	0	0	0	7	90

7. Conclusion

In this paper we introduced a new training algorithm for the discrete HMT. The algorithm uses the sufficient statistics of the HMT generative model to form a fixed length training vector to be used in linear discriminant classifiers (like SVM). The algorithm proves considerable amount of improvement over the baseline HMT when applied to land cover images.

References

[1] Kamata, S., Eason, R., Perez, A., and Kawaguchi, E. (1992) A Neural Network Classifier for LANDSAT Image Data. IAPR. The Hague, The Netherlands: IEEE, 573-576.

[2]Haung, C., Davis, L. S., and Townshend, J. R. G. (2002) An Assessment of Support Vector Machines for Land Cover Classification. International Journal of Remote Sensing, 23, 725–749.

[3]Bazi, Y., and Melgani, F. (2006) Toward an Optimal SVM Classification System for Hyperspectral Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 44, 3374-3384.

[4]P. Watanachaturaporn, P. K.Varshney, and M. K. Arora(2003) Evaluation of Factors Affecting Support Vector Machines for Hyperspectral Classification. NASA, (NAG5-11227).

[5] M.S. Crouse, R.D. Nowak, and R.G. Baraniuk. Wavelet-Based Statistical Signal Processing Using Hidden Markov Models. IEEE Transactions on Signal Processing, 46(4):886-902, 1998.

[6] Y. Altun, I. Tsochantaridis, and T. Hofmann, “Hidden Markov support vector machines,” presented at the 20th Int. Conf. Mach. Learn.(ICML), Washington, DC, Aug. 2003.

[7] L. Xu, D. Wilkinson, F. Southey, and D. Schuurmans, “Discriminative unsupervised learning of structured predictors,” inProc. 23rd Int. Conf. Mach. Learn., Pittsburgh, PA, Jun. 2006, pp. 1057–1064.

[8] W. Xu, J. Wu, and Z. Huang, “A maximum margin discriminative learning algorithm for temporal signals,” in Proc. 18th Int. Conf. Pattern Recogn. (ICPR’06), Hong Kong, Aug. 2006, vol. 2, pp. 460–463.

[9] A. Sloin and D. Burshtein, Support Vector Machine Training for Improved Hidden Markov Modeling IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 56, NO. 1, JANUARY 2008

[10] V. N. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.

[11] C. Burges, “A tutorial on support vector machines for pattern recognition,” Data Mining Knowl. Discov., vol. 2, no. 2, pp. 121–167, 1998.

[12] J. B. Durand, P. Goncalvès, Y. Guédon. Computational methods for hidden Markov tree models - An application to wavelet trees. IEEE Transactions on Signal Processing, Volume 52(9):2551-2560, 2004.

[13] A. S. Mohamed, R. A. El-Khoribi, L. Fekry, Discrete Hidden Markov Tree Modelling of Ranklet Transform for Mass Classification in Mammograms, ICGST/GVIP, special issue on mammograms, vol. 7, pp61-68.

[14] F. Smeraldi. Ranklets: Orientation selective nonparametric features applied to face detection. In Proceedings of the 16th International Conference on Pattern Recognition, Quebec, QC, 2002.

[15] Masotti, M. (2005) Exploring ranklets performances in mammographic mass classification using recursive feature elimination, Research Repor 930, Department of Physics, University of Bologna, Italy

[16] C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines 2001 [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

Reda A. El-Khoribi, Associate Prof., Faculty of Computers and Information, Cairo University, Egypt. PhD in Electronics and Communications Engineering, Cairo University 1998. MSC in Electronics and Communication Engineering, Faculty of Engineering, Cairo University, 1993. BSC in Electronics and Communication Engineering, Faculty of Engineering, Cairo University, 1988. He is currently in the Information Technology department of the faculty of Computers and Information, Cairo University. His research interests include: Image processing and Computer vision, pattern classification, signal processing, computer graphics and artificial intelligence.