Proc. IGARSS'99, vol. 1, pp. 348-350, 1999.

Support Vector Machines for Land Usage Classification in Landsat TM Imagery

Lothar Hermes ⋆, Dieter Frieauff ‡, Jan Puzicha ⋆, and Joachim M. Buhmann ⋆
⋆ Institut für Informatik III, Rhein. Friedr.-Wilh.-Universität, Römerstr. 164, 53117 Bonn, Germany
Tel: ++49228734102, Fax: ++49228734382, Email: {*hermes, jan, jb*}@cs.uni-bonn.de
‡ Dornier Satellitensysteme GmbH, 88039 Friedrichshafen, Germany
Tel: ++49754582256, Fax: ++49754585650, Email: Dieter.Frieauff@dss.dornier.dasa.de

ABSTRACT – Land usage classification is an essential part of many remote sensing applications for mapping, inventory, and yield estimation. In this contribution, we evaluate the potential of the recently introduced support vector machines for remote sensing applications. Moreover, we expand this discriminative technique by a novel Bayesian approach to estimate the confidence of each classification. These estimates are combined with a priori knowledge about topological relations of class labels using a contextual classification step based on the iterative conditional mode algorithm (ICM). As shown for Landsat TM imagery, this strategy is highly competitive and outperforms several commonly used classification schemes.

INTRODUCTION

It is a key observation that the overall system performance sensitively depends on the choice of a good classifier. The majority of remote sensing approaches are based on classical pattern recognition techniques, mostly maximum likelihood classifiers [4], K-nearest neighbor [6] or combinations of maximum likelihood and clustering [2]. Recently, several neural network based classifiers have been proposed [6, 5]. However, to make automatic classification tools feasible, a further algorithmic improvement in classification accuracy is mandatory in many applications.
Support vector machines (SVM) provide a very promising classification technology that led to remarkable improvements in handwritten digit recognition, face detection in images, object recognition, text categorization and nonlinear time-series prediction. SVM have been developed on a solid base of statistical learning theory [7] and are especially designed to provide high flexibility for approximating class boundaries, yet to avoid over-fitting phenomena. SVM can be understood as novel learning algorithms for neural networks that completely avoid the problem of selecting the number of layers and the number of hidden units. In essence, SVM lead to linearly constrained quadratic programs that are solved by numerical methods.
In this contribution, the potential of SVM for remote sensing applications is examined in a benchmark study. The classification of land usage in Landsat TM imagery is chosen as a representative example. SVM are shown to improve the pixel classification accuracy compared to Gaussian maximum likelihood, 1-nearest neighbor and cluster-based maximum likelihood classifiers. A novel Bayesian approach is proposed which allows us to obtain estimates for the class-likelihoods instead of pure binary classification results. These estimates are combined with contextual a-priori-knowledge about topological relations of class-labels resulting in a further increase in classification accuracy. An efficient implementation of the ICM-algorithm for maximum a posteriori model estimation is discussed. Results are presented for Landsat TM imagery of a rural area in Germany.

SUPPORT VECTOR MACHINES

First, the two class classification problem is considered. Assume that ℓ training examples $\mathbf{x}_i \in \mathbb{R}^d$ labeled by $y_i \in \{+1, -1\}$ are given. Support vector machines [3] separate both classes by so-called *optimal hyperplanes*, where the optimality of a hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ is defined as the maximal distance between the hyperplane and its nearest training example. To enforce a unique parameterization, one usually demands

$$y_i \cdot \left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 \quad \forall i = 1, \ldots, \ell \ , \tag{1}$$

and minimizes $\frac{1}{2}\|\mathbf{w}\|^2$ under these constraints. This concept can be extended to the case when the classes are not linearly separable, i.e. when (1) has no solution. Introducing *slack variables* $\xi_i \geq 0$, $i = 1, \ldots, \ell$, the constraints are weakened to

$$y_i \cdot \left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i \quad \forall i = 1, \ldots, \ell \ , \tag{2}$$

while the objective function is supplemented to keep the constraint violation as small as possible:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \sum_{i=1}^{\ell} \xi_i \stackrel{!}{=} \min \ . \tag{3}$$

This results in a quadratic optimization problem which is solved by a standard numerical optimization package. In

algorithm	scenario (a)	scenario (b)
1-NN classifier	81.9	78.2
ML classifier	74.8	82.8
GMM classifier	77.4	82.2
SVM	85.8	83.7
prob. SVM	87.3	84.1
prob. SVM plus ICM	90.5	89.7

Table 1: κ-coefficient of different classification strategies. Scenario (a): 630 training vectors, 5 classes. Scenario (b): 2230 training vectors, 13 classes. The test set consisted of 13170 data vectors.

its *dual form* the optimization problem is posed in terms of Lagrange multipliers λ_i, $i = 1, \ldots, \ell$. It can be shown that the distance of a vector $\mathbf{x}$ from the optimal separating hyperplane is proportional to

$$g(\mathbf{x}) = \sum_{i=1}^{\ell} \lambda_i y_i \mathbf{x}_i^T \mathbf{x} + b \quad . \tag{4}$$

The training vectors $\mathbf{x}_i$ are solely used in inner products which can be replaced by a kernel function $K(\mathbf{u}, \mathbf{v})$ that obeys Mercer's condition [7]. This replacement can be interpreted as mapping the data vectors into a high-dimensional feature space before using a hyperplane classification there. Depending on the selected kernel function several well-known classification schemes like polynomial classifiers, radial basis function (RBF) classifiers or two-layer neural nets can be emulated this way. In our experiments, we used a RBF-kernel with a standard deviation σ. For a K-class problem, K individual support vector machines are trained to separate one class ω_k from all other classes. A given vector $\mathbf{x}$ is then assigned to the class with maximal distance g_k as given by (4).

PROBABILISTIC SVM

SVM have been shown to produce excellent results in various applications. However, there remain some important shortcomings: (i) SVM crucially depend on the correct choice of their free parameters (σ and C in our case). (ii) SVM do not provide any estimation of their classification confidence. Thus, SVM do not allow us to incorporate any a-priori information. However, in many remote sensing applications a priori information about likely class label configurations is available and it is crucial to integrate this information into the classification process to yield reliable classification results. The second shortcoming can be avoided by the probabilistic SVM variant developed in the sequel. For any vector $\mathbf{x}$, let d_k be defined by $d_k = g_k(\mathbf{x})$. In addition to the described SVM strategy, we now pay attention to the K^2 class-conditional density functions $p(d_k|\omega_j)$, i.e. the densities of d_k, when $\mathbf{x}$ is restricted to samples that belong to class ω_j. These densities are approximated by Gaussian mixture models using a strategy similar to the one described in [2]. In essence, we use the minimum description length (MDL) criterion to control the number of normal distributions within the mixture model, while their individual parameters are adjusted by a maximum likelihood strategy. Additionally, the class probabilities are substituted by the class frequencies in the sample set:

$$p(\omega_i) = \frac{|\{\mathbf{x}_i | \mathbf{x}_i \in \omega_i\}|}{\ell} \quad . \tag{5}$$

To consider the confidence of assigning a data vector $\mathbf{x}$ to class ω_k, we compute

$$h_k(\mathbf{x}) = \frac{P(d_k|\omega_k)\, p(\omega_k)}{P(d_k|\omega_k)\, p(\omega_k) + \sum_{j \neq k} (1 - P(d_k|\omega_j))\, p(\omega_j)} \quad ,$$

where $P(d_k|\omega_j) = \int_{-\infty}^{d_k} p(t|\omega_j)\, dt$. Basically, this formula implements the Bayes rule to estimate the a-posteriori probability $p(\omega_k|d_k)$ if we disregard the difference that *cumulative likelihoods* $P(d_k|\omega_j)$ are used instead of the estimated likelihoods $p(d_k|\omega_j)$. We can thus ensure that $h_k(\mathbf{x})$ increases monotonously in d_k as intuitively expected. As an additional advantage, the confidence estimation becomes more robust. A pixel–wise classification rule is given by $k(\mathbf{x}) = \text{argmax}_k \{h_k(\mathbf{x})\}$.

CONTEXTUAL CLASSIFICATION

Assume that the image contains n pixels $\mathbf{x}_i$. Let their final classification results be represented by a n-dimensional vector $\mathbf{z}$ initialized by the pixel–wise classification results. For the joint contextual classification we propose the cost function

$$E(\mathbf{z}) = E_l(\mathbf{z}) + \rho \cdot E_p(\mathbf{z}) \quad . \tag{6}$$

The likelihood part $E_l(\mathbf{z})$ uses the confidence values provided by the probabilistic SVM:

$$E_l(\mathbf{z}) = -\sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{z_i,\omega_k} \cdot h_k(\mathbf{x}_i) \quad . \tag{7}$$

The prior part $E_p(\mathbf{z})$ is defined by a local *neighborhood* $\mathcal{N}_i$ for each pixel i and class–specific thresholds T_k. Whenever a pixel $\mathbf{x}_i$ is assigned to a class ω_k that less than T_k pixels in its neighborhood are assigned to, we assume that the log–probability of a correct decision is proportional to the difference to T_k. This can be formalized by

$$E_p(\mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{z_i,\omega_k} \left(T_k - \sum_{j \in \mathcal{N}_i} \delta_{z_j,\omega_k} \right)_+ \quad . \tag{8}$$

We use the ICM-algorithm [1] to find a solution of small costs. Although the ICM does not guarantee to find a global cost minimum, it is extremely fast and produces excellent results due to the good initial classification provided by the probabilistic SVM.

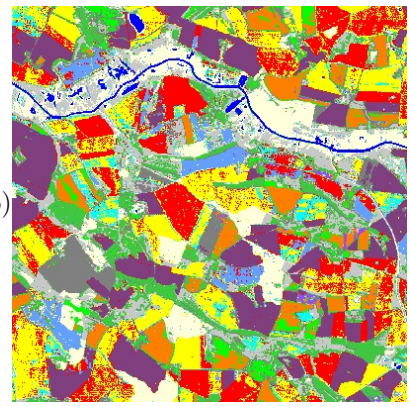
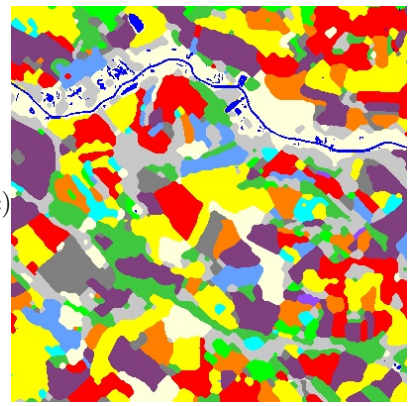
(a)  (b)  (c) 

Figure 1: Classification of a Landsat TM image using a probabilistic SVM: (a) original image, (b) pixel–wise classification and (c) contextual classification by ICM.

RESULTS

The described methods have been tested for a Landsat TM image which is depicted in Fig. 1. A training set of 2230 pixels and a test set of 13170 pixels has been defined. All pixels have been manually classified into 5 joined classes (forests, water, villages, pastures or grasslands, farming) to obtain ground truth information. Two different experiments have been carried out. In the first experiment the training set has been reduced to little more than a quarter to investigate the influence of small training set sizes. Especially the class "farming" exhibits a very heterogeneous spectral characteristic. Therefore, in the second experiment the 5 classes have been further split into its 13 subclasses and all 2230 training samples have been used for training.

Several classification schemes have empirically been compared. The results are summarized in Tab. 1. As expected, the maximum likelihood classifier (ML classifier) performed poorly in the first experiment, because the highly heterogenous class "farming" could not be modeled by a single Gaussian. The performance of the Gaussian mixture model (GMM classifier) turned out to be only slightly better. As the main reason, too few training samples have been available to estimate the joint density in the 6-dimensional feature space. The SVM approaches showed significantly superior performances. Obviously the 1-dimensional density estimations performed by the probabilistic SVM does not require large amounts of data and thus results in high classification accuracies.

In the second experiment, the performance gap between SVM techniques and the approaches based on density estimation narrows down. This trend can be explained by the fact that the large amount of training data and the more homogeneous classes facilitate the density estimation problem. Still, the SVM approaches exhibit slightly superior results to the classical techniques. The probabilistic SVM has the additional advantage of being less dependent on the optimal choice of its parameters. Furthermore, the contextual classification by ICM provides a remarkable increase in classification accuracy. As shown in Fig. 1, the classification result appears less noisy after ICM has been applied.

SUMMARY AND CONCLUSION

In this paper we have developed a strategy to classify Landsat TM imagery on the basis of support vector machines. In a benchmark study, this strategy has been demonstrated to outperform several other widely used techniques and has been shown to produce classification results of high accuracy. Support vector machines appear to be especially advantageous if dealing with heterogenous classes for which only a small number of training samples are available. Their classification performance can be considerably increased by using a joint contextual classification step on the basis of ICM. The presented algorithms are part of the FACTS image processing system developed by Dornier Satellitensysteme GmbH.

REFERENCES

[1] J. Besag. On the statistical analysis of dirty pictures. *J. Royal Stat. Soc.*, 48(3):259–302, 1986.

[2] C. Bouman and M. Shapiro. A multiscale model for Bayesian image segmentation. *IEEE Transactions on Image Processing*, 3(2):162–177, 1994.

[3] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:1–25, 1995.

[4] G. M. Foody and R. A. Hill. Classification of tropical forest classes form Landsat TM data. *Int. J. Remote Sensing*, 17(2):2352–2367, 1996.

[5] D. Miller, E. Kaminsky, and S. Rana. Neural network classification of remote-sensing data. *Computers and Geosciences*, 21(3), 1995.

[6] S. Serpico, L. Bruzzone, and F. Roli. An experimental comparison of neural and statistical non-parametric algorithms for supervised classification of remote-sensing images. *Patt. Rec. Letters*, 17, 1996.

[7] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.