Support Vector Machines for shade identification in urban areas

PAJARES G., CRUZ, J.M. and BELMONTE, M.
Arquitectura de Computadores y Automática
Universidad Complutense.- Facultad Informática
Juan del Rosal 8, 28040 Madrid
SPAIN

Abstract: - The aim of this paper is the identification of shades in urban areas for remote sensing images. The final goal consists in the risk minimization for image change detection algorithms. Correct shade identification help us to discard urban shades as urban changes. The main contribution of the paper is to focus the problem as a classification problem using the well founded *Support Vector Machines* theory. A comparative analysis is carried out against other classical existing classification methods where the performance of the proposed approach is verified.

Key-Words: - Support vector machines, classification, density estimation

1 Introduction

The high resolution provided by the remote sensing sensors has opened a new field in remote sensing applications: the urban dynamic analysis. A final goal of any urban dynamic analysis is to detect urban changes. Our shade identification approach is focused under such goal. Indeed, when we try to detect the changes in urban areas, we find that the shades are labelled as urban changes. This is because the images are captured under different illumination conditions (different days and hours) and different view points. Once the shades are identified, they should be discarded as urban changes and the risk for any erroneous decision is minimised.

With such purpose we use the well known Support Vector Machines (SVM) [1,2] theory for shade identification. This implies that the shade identification becomes a classification problem. This is the main contribution of the paper under the performance of the SVM framework. SVMs are one type of large margin classifiers which have proved highly successful in a number of classification studies.

We compare the SVM performance against two well-founded statistical strategies: (1) the Bayesian Statistical Decision (BSD) theory [2,3]; (2) the Parzen´s window decision (PWD) theory [3]

This paper is organised as follows: in section 2 we formulate the classification problem under the SVM theory. In section 3 a comparative analysis between SVM, BSD and PWD is carried out. Finally in section 4 the conclusions are presented.

2 The classification problem

. The classification problem can be viewed as a learning machine problem where the role and the problem of the learning machine is to select a function that best approximates the system's response. The learning machine is limited to observing a finite number (n) example patterns in order to make this selection.

Our goal is concerned with the shade distinction in urban areas. Hence, the unique class of interest in the input images is that produced by the building shades. This implies that the classification problem is a two-classification (c_1, c_2) approach, where the shades belong to the unique class of interest (c_1) and the remainder areas belong to the other class (c_2). The output of the system takes on only two symbolic values $y = \{+1,-1\}$ corresponding to the two mentioned classes respectively. Therefore, we have a trainable pattern classifier that learns to differentiate between patterns from the two classes.

The remote sensing images are panchromatic images acquired by the IKONOS sensor from Madrid with spatial and radiometric resolutions of 1 meter and 11 bits respectively [4]. We use as pattern samples the pixel radiometric information. Hence, each pattern sample i is a 1-dimensional vector $\mathbf{x}_i \equiv \{x_i\}$, where its component is the radiometric pixel value. The finite set of n training data under the SVM formulation is

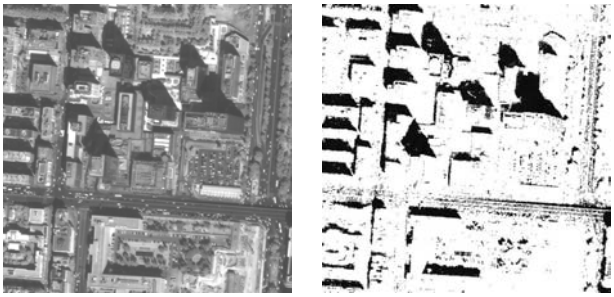
$$(\mathbf{x}_i, y_i), \quad i = 1,...,n \qquad (1)$$

where each $\mathbf{x}_i$ vector denotes a training data, i.e. $\mathbf{x}_i \in \Re$ and $y_i = \{+1,-1\}$.

Hereinafter, the samples shall be denoted as x_i, avoiding the vector notation.

Given the set of training samples defined in (1), the goal is to find a decision function (D) into the classes c_1 and c_2 as follows,

$$D(x) = \sum_{i=1}^{n} \alpha_i y_i H(x_i, x) \qquad (2)$$



(*a*) (*b*)

Fig. 1. (*a*) urban shading area; (*b*) training pattern samples

The equation (2) establishes a representation of the decision function *D* as a linear combination of kernels centred in each data point. Using different kernels $H(x,z)$ [2] we get different functions. We have used Gaussian Radial Basis functions of the form $H(x,z) = \exp\left\{ x - z \right|^2 / \sigma^2 \}$ where σ defines the width of the kernel, set to 2.5 after different experiments.

The parameters α_i, $i = 1,...,n$, in Eq. (2) are the solution for the following quadratic optimisation problem: Maximise the functional

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

subject to constraints

$$\sum_{i=1}^{n} y_i \alpha_i = 0, \quad 0 \le \alpha_i \le \frac{C}{n}, \; i = 1,...,n \qquad (4)$$

given the training data (x_i, y_i), $i = 1,...,n$, the inner product kernel *H*, and the regularization parameter *C*. As stated in [2], at present, there is no well-developed theory on how best to select *C*, in several applications it is set to a large fixed constant value, set to 2000 in our approach.

A remarkable property of SVMs is that the data points x_i associated with the nonzero α_i are called *support vectors*. If all data points which are not support vectors were to be discarded for the training set the same solution would be found, an interesting perspective on SVMs is to consider its information compression and storage properties. The support vector represent the most informative data points and compress the information contained in the training set. This implies that only the support vectors need to be stored. Once the support vectors have been determined, the SVC decision function has the form

$$f(x) = \sum_{\text{support vectors}} \alpha_i y_i H(x_i, x) \qquad (5)$$

The SVC generates a scalar output $f(x)$ whose polarity, sign of $f(x)$, determines the class membership. The magnitude can usually be interpreted as a measure of belief or certainty in the decision made. As BSD and PWD use posterior probabilities, we use a warping function that maps $f(\mathbf{x})$ to a posterior probability. This is carried out assuming that posterior probabilities take the form of a sigmoid and directly estimating the sigmoid [5],

$$p(x) = \frac{1}{1 + \exp\{-(af(x) + v)\}} \qquad (6)$$

In order to avoid severe bias in the distances for the training data, the parameters *a* and *v* are estimated experimentally and set to 0.2 and 0 in our experiments.

3 Comparative Analysis and performance evaluation

We verify the performance of our SVM approach comparing the results against two classical existing classification approaches: BSD and PWD.

BSD is a parametric estimation method assuming that the data follow a known probability density Gaussian function with two parameters to be estimated: the mean μ and the variance σ^2.

The estimation process is carried out via likelihood minimization and the resulting function is

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma} \right. \right. \qquad (7)$$

In order to avoid severe bias in the distances for the training data, the parameters *a* and *v* are estimated experimentally and set to 0.2 and 0 in our experiments.

PWD is a non-parametric density estimation based on the Parzen's window,

$$p(x) = \sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2h^2}\left(\frac{x - x_j}{\sigma} \right)^2 \right] \qquad (8)$$

h is the smoothing parameter, often expressed as a function of the number of patterns $h = h_0 n^{-k}$ with $0 \le k \le 0.5$. The choice of the band-width h_0 is very

critical in Parzen´s density estimation [6]. An overlay small h gives a spiky or noise estimate of $p(x)$, we have carried out several experiments according to the guidelines in [3] and finally h_0 is set to 4 with $k = 0.3$.

a) *Training phase*

We have used a set of 5 remote sensing images, captured under different illumination conditions (different days and hours) and different view points. Figure 1(*a*) shows an example. We have selected the training samples for the classes c_1 and c_2 according to the following process: *a*) select interactively several samples from different shading areas; *b*) with these samples compute the mean value *m*; *c*) extract the remainder training sample patterns x if $|x - m| < T$; where T is a threshold value set to 40 in our experiments. With this criterion we get $n \approx 25 * 10^4$ training samples for class c_1. We select a similar number of training samples for class c_2 (from the remainder samples), Figure 1(*b*) shows in black and white the sets of samples for classes, c_1 and c_2 respectively. With both sets of training samples we obtain 2868 *support vectors* ($\alpha_i \neq 0$), achieving a considerable reduction with respect the number of initial training pattern samples. With the above n shading pattern samples, we estimate the parameters for the BSD approach, as required by the equation (7), achieving $\mu = 147.8$ and $\sigma = 17.6$. This set of n shading pattern samples is also used in equation (8) for the PWD. We have proved two type of kernels in the equation (3): polynomials of degree 2 and Radial Basis with $\sigma = 2$. The best performance is achieved with the last. The α_i parameters range from −18.3 to +25.8.

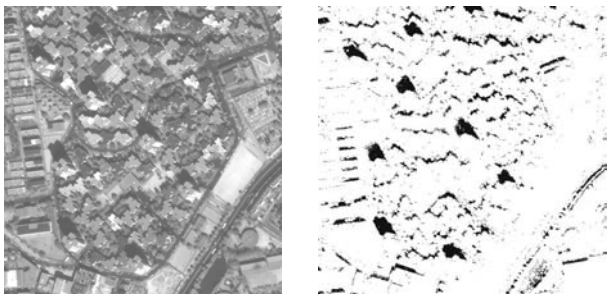
b) *Decision phase*

We have now available the functions given by equations (6), (7) and (8). So, for each new x pattern, we compute the corresponding probability according to such equations and classify x as belonging to class c_1 if $p(x) > 0.5$, i.e. x it belongs to a shading area.

We have used 6 remote sensing images for classification purposes, figure 2(*a*) shows an example of a new remote sensing image and figure 2(*b*) the samples classified as shading areas by the our SVM approach.

Table 1 shows, on average for the 6 remote sensing images, the performance for SVM, BSD and PWD. This is verified under the expert human criterion, which selects the samples interactively, following the criterion explained in the above section.

To clarify the behaviour and performance of the SVM approach, we have designed the following test strategy. We have arranged the support vectors so that their absolute values are in increasing order, i.e. from less to greater relevance. Figures 3(*a*) and (*b*) show the performance of the SVM approach (circles), against BSD (triangle down) and PWD (diamond) according to the percentage of successes. In Figure 3(*a*) we have used the number of support vectors in the x-axis starting from the minimum value in the arranged set. In Figure 3(*b*), we have used the indicated number of support vectors but in reverse order, so that now the most relevant support vectors are firstly used.



(*a*) (*b*)

Fig. 2. (*a*) new urban shading area; (*b*) classification results from SVM

Table 1. Percentage of successes for SVM, BSD and PWD for the 6 remote sensing images

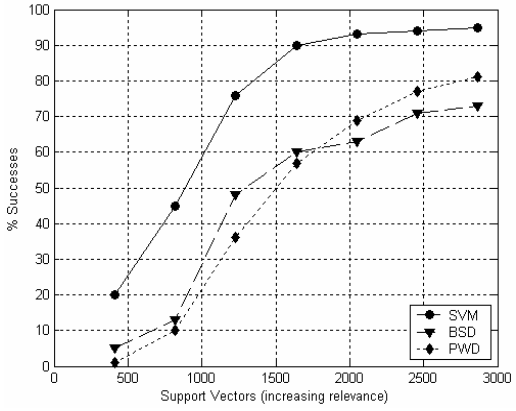
	SVM	**BSD**	**PWD**
% successes	0.95	0.89	0.92

Taking as pattern samples the number of support vectors given in the x-axis, we use them for estimating the functions given in equations (5), (7) and (8), i.e. they are now the training samples for BSD and PWD.

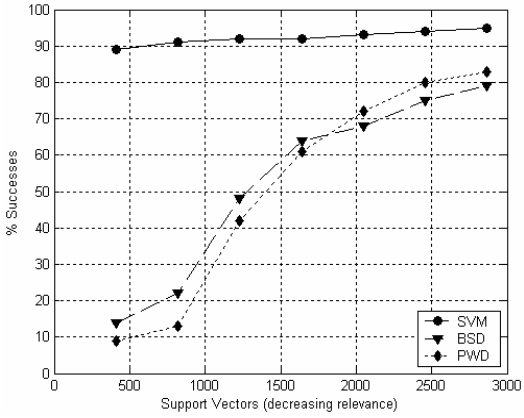
From the results in table 1 and figure 3, the following conclusions can be inferred:

1. The best performance is achieved with SVM. This is obvious in table 1 and also in figure 3, where the percentage of successes for SVM overpasses always the percentage of BSD and PWD.
2. The slope for SVM is greater than the slopes for BSD and WD. This means that SVM achieves quickly a better performance with a reduced number of support vectors.
3. PWD achieves better results than BSD, i.e. the estimation of a density function without the assumption of a know distribution is well suited.

4. The most relevant support vectors are the last 1500 according to the arrangement. This can be derived from figure 3(*a*) where only SVM reaches a high performance once this number is over passed and from figure 3(*b*) where SVM achieves a high performance with the first set of support vectors, which are the most relevant. Then a slight improvement is achieved with the remainder support vectors.
5. The improvement in BSD and PWD is only achieved as the number of training patterns increases, without the influence of relevant support vectors.



(*a*)



(*b*)

Fig. 3. Percentage of successes for SVM, BSD and PWD obtained with the support vectors arranged: (*a*) in increasing order and (*b*) in decreasing order.

4 Conclusion

We have shown the ability of SVM for shade classification in urban areas for remote sensing images, as compared with other classical existing approaches, with encouraging results. We have verified that the relevant support vectors are decisive for this achievement and that using a reduced number of support vectors this performance is still reachable. This implies that SVM only requires a reduced number of sample patterns as compared with the number of pattern samples required by BSD and PWD.

In summary, we can conclude that the SVM method is suitable for shade identification in the panchromatic remote sensing images. So, this is a useful tool that helps to minimize the error in image change detection applications.

We have carried out some previous experiments with a BSD approach [7], and in this paper we have proven the better performance of SVM with respect to the previous BSD results. The performance is also verified against the PWD classical strategy.

References:

[1] V.N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998

[2] V. Cherkassky and F. Mulier, F., *Learning from Data: Concepts, Theory and Methods*, Wiley, New York, 1998.

[3] R.O. Duda, P.E. Hart and D.G. Stork, *Pattern Classification*, Wiley, New York, 2000.

[4] R. Li, Potencial of High-Resolution Satellite Imagery for National Mapping Products, *Photogrammetric Engineering Remote Sensing*, Vol. 64, No 12, pp. 1165-1169

[5] J. Platt, Probabilistic Outputs for Support Vector Machines and comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 2000.

[6] R.P.W. Duin, On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions, *IEEE Trans. Computers*, Vol. 25, 1976, pp. 1175-1179.

[7] G. Pajares, C. Alonso, J.M. Cruz and V. Moreno, Shade identification in urban areas through the bayesian classifier. *Proc. European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems* (eunite 2002), Proceedings, 2002, pp. 476-481, Albufeira, Algarve, Portugal.