Semi-Supervised Support Vector Rainfall Estimation Using Satellite Images

Greice Martins de Freitas, Ana Maria Heuminski de Ávila
Meteorological and Climatic Research Center Applied to Agriculture - CEPAGRI/UNICAMP
CEP 13083-886, Campinas - SP, Brazil
{greice,avila}@cpa.unicamp.br

João Paulo Papa
Institute of Computing - UNICAMP
CEP 13084-971, Campinas, SP, Brazil
jpaulo@ic.unicamp.br

Abstract

In this paper we introduce the use of semi-supervised support vector machines for rainfall estimation using images obtained from visible and infrared NOAA satellite channels. Two experiments were performed, one involving traditional SVM and other using semi-supervised SVM (S^{3}VM). The S^{3}VM approach outperforms SVM in our experiments, with can be seen as a good methodology for rainfall satellite estimation, due to the large amount of unlabeled data.

1. Introduction

The precipitation is one of the most important atmospheric phenomena. Some branches of the human activities depends on it, like the management of hydric resources. Its monitoring in agriculture is one of the main determinative elements of the success or the failure of some activities, as the preparation of the ground, plantation of cultures, irrigation and harvest.

Rainfall can be estimated remotely, either from ground-based weather radars or from satellite. Radars are active devices, but its maximum coverage range is limited. Also, radars are prohibitively expensive. Satellite-based measurements offer global coverage and are actually used for rainfall estimation [3, 4]. McCullagh et al. [3] introduced a neural network model for precipitation estimation using both GOES visible and infrared channels. Umehara et al. [4] were one of the first that used support vector machines in rainfall estimation.

However, all the related works cited above used supervised learning machines to perform the rainfall estimation. Despite the large amount of available data provided by satellites, most of them are unlabeled, i.e., we do not know the ground truth to guide a learning process in a supervised classification. Using unsupervised approaches is not well suitable, because them are usually less accurate than supervised methods. A inductive or supervised support vector machine uses labeled data sets to find the optimal hyperplane that best separates the objects in classes. A transductive support vector machine [2] is a semi-supervised approach that uses both labeled and unlabeled data to find this hyperplane. In that way, we propose the use of semi-supervised support vector machines in the rainfall estimation task, using images from visible and infrared NOAA satellite channels. To the best we know, we are the first that used a semi-supervised support vector machine approach in the rainfall estimation. The next sections discusses the semi-supervised support vector machines (S^{3}VM), experimental results and conclusions.

2. Semi-supervised Support Vector Machines

A inductive SVM tries to estimate a classification function $f : \Re^n \mapsto \{\pm 1\}$ using training data from two classes

$$(x_1, y_1), \ldots, (x_l, y_l) \in \Re^n \times \{\pm 1\}. \quad (1)$$

In this work we will limit the discussion to linear classification functions, due to the lack of space. If the points are linearly separable, then exist a vector w and scalar b such that

$$w \cdot x_i - b \geq 1, \; if \; y_i = 1 \; and \quad (2)$$
$$w \cdot x_i - b \leq -1, \; if \; y_i = -1, \; i = 1, \ldots, l \quad (3)$$

or equivalently

$$y_i[w \cdot x_i - b] \geq 1, \; i = 1, \ldots, l \quad (4)$$

The optimal separating hyperplane, $w \cdot x = b$, is the one which is furthest from the closest points in the two classes.

In general the classes will not be separable, so a slack term η_i is added for each point such that if the point is misclassified, $\eta_i \geq 1$. In that way, the SVM algorithm tries to solve

$$\min_{W,b,\eta} C \sum_{i=1}^{l} \eta_i + \frac{1}{2}\|w\|^2 \tag{5}$$

$$s.t.\ \ y_i[w \cdot x + i - b] + \eta_i \geq 1 \tag{6}$$

$$\eta_i \geq 0,\ i = 1, \ldots, l \tag{7}$$

where $C > 0$ is a fixed penalty parameter.

To introduce the S^{3}VM [1], we will start with the formulations given above. Two constraints, ξ and z, will be added for each point in the working set. One constraint calculates the misclassification error as if the point were in class 1 and other constraint calculates the misclassification error as if the point in class -1. The final class of the points corresponds to the one that results in the smallest error. In that way, we will define the semi-supervised support vector machine problem as

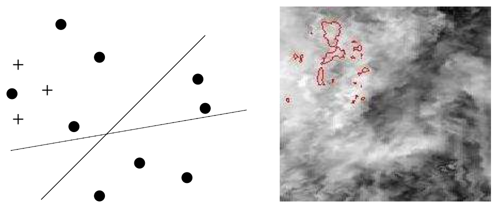
$$\min_{W,b,\eta,\xi,z} C \left[\sum_{i=1}^{l} \eta_i + \sum_{j=l+1}^{l+k} \min(\xi_j, z_j)\right] + \|w\| \tag{8}$$

$$s.t.\ \ y_i[w \cdot x + i - b] + \eta_i \geq 1 \quad \eta_i \geq 0,\ i = 1, \ldots, l \tag{9}$$

$$w \cdot x_j - b + \xi_j \geq 1 \ \ \xi_j \geq 0 \ \ j = l+1, \ldots, l+k \tag{10}$$

$$-(w \cdot x_j - b) + z_j \geq 1 \quad z_j \geq 0 \tag{11}$$

where $l+k$ is the number of unlabeled samples. In that way, the formulation uses both labeled and unlabeled data to estimate the optimal hyperplane. Figure 1a illustrates the problem.



(a) (b)

Figure 1. (a)The maximum margin hyperplanes. Positive/negative examples are marked as +/-, test examples are dots. The dashed line is the solution of the inductive SVM. The solid line shows the S^{3}VM classification. (b) Image obtained from visible channel of the NOAA satellite, covering the area of Bauru, SP-Brazil. The rainfall regions are represented by the bounded locations.

3. Experimental results

We used images obtained from visible and infrared NOAA satellite channels (Figure 1b). Each pixel p_i is represented by its feature vector $\vec{v_i} = (x_i^1, x_i^2)$, where x_i^1 and x_i^2 are, respectively, the brightness value of p_i in the visible and infrared channels. Each image is 199×184 8 bits/pixel. So, we perform two experiments using the SVM^{light} [2] software: one involving a SVM and other using S^{3}VM, both with gaussian kernels. For the SVM experiment, we use one image to train and other to test. Both images were pixel by pixel labeled by a CEPAGRI/UNICAMP meteorologist as being rainfall or not. Note that each image has $199 \times 184 = 36616$ pixels, i.e., the size of training set is 36616. For the S^{3}VM we use the same training and test images, but the training set was increased by 54 unlabeled images, i.e., we have more 54x(199*184) = 1.977.264 samples to be used in the training set. We conduct some experiments using different number of unlabeled images in S3VM training step to evaluate the accuracy of the method by increasing the number of unknown information. We noted that the performance considerably increases until 45 images. After this point, the computational cost is very high and the classifier accuracy appears to stabilize. In that way, the final classification rates obtained for SVM and S^{3}VM are, respectively, 60.51% and 71% in the test set.

4. Conclusion

We propose the use of semi-supervised support vector machines for the rainfall estimation using satellite images. The S^{3}VM methodology outperformed the SVM approach, due to the extra information provided by unlabeled images. We are currently working for better characteristics from other NOAA channels.

References

[1] K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. In *Proc. of the 1998 Conf. on Advances in Neural Information Processing Systems II*, pages 368–374, Cambridge, MA, USA, 1999. MIT Press.

[2] T. Joachims. Making large-scale support vector machine learning practical. In *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge, MA, 1998.

[3] J. McCullagh, K. Bluff, and E. Ebert. A neural network model for rainfall estimation. *ANNES '95: Proc. of the 2nd New Zealand Two-Stream Int. Conf. on Artificial Neural Networks and Expert Systems*, 00:389, 1995.

[4] S. Umehara, T. Yamazaki, and Y. Sugai. A precipitation estimation system based on support vector machine and neural network. *Electronics and Communications in Japan*, 89(3):38–47, 2006.