Incorporating Spatial Contiguity into the Design of a Support Vector Machine Classifier

Murat Dundar
Computer Aided Diagnosis and Therapy
Siemens Medical Solutions Inc, USA
Malvern, PA 19355
Email: murat.dundar@siemens.com

James Theiler
Space and Remote Sensing Sciences
Los Alamos National Laboratory
Los Alamos, NM 87544
Email: jt@lanl.gov

Simon Perkins
Space and Remote Sensing Sciences
Los Alamos National Laboratory
Los Alamos, NM 87544
Email: s.perkins@lanl.gov

***Abstract*— We describe a modification of the standard support vector machine (SVM) classifier that exploits the tendency for spatially contiguous pixels to be similarly classified. A quadratic term characterizing the spatial correlations in a multispectral image is added into the standard SVM optimization criterion. The mathematical structure of the SVM programming problem is retained, and the solution can be expressed in terms of the ordinary SVM solution with a modified dot product. The spatial correlations are characterized by a "contiguity matrix" Ψ whose computation does not require labeled data; thus, the method provides a way to use a mix of labeled and unlabeled data. We present numerical comparisons of classification performance for this contiguity-enhanced SVM against a standard SVM for two multispectral data sets.**

I. Introduction

In remote sensing imagery, what is of interest on the ground – especially for broad terrain categories, like lakes, forests, beaches, vegetation canopies, agricultural fields, *etc.* – is often much larger than a single pixel element size. As a consequence, two pixels that are close to each other are more likely to yield similar classifier outputs than are two pixels chosen at random from the image. While there are other spatial cues that can be exploited in classifying multispectral imagery, such as texture and shape, this tendency for nearby pixels to be classified alike is almost universal in images. In this paper, we present a methodology for incorporating this domain knowledge into the design of the classifier.

In particular, we describe an approach for incorporating contiguity into the regularization term of the loss function, and although this can be applied in a number of different classifiers, we will investigate its use in the linear support vector machine. We will show that the contiguity term can be expressed as a conventional support vector machine with a slightly modified kernel.

Given a limited supply of training data (and in real-world settings, this supply is always limited), it is easy to find models that overfit the data – that is, the model fits the training data but does poorly on out-of-sample test data. This problem is typically alleviated by reducing the "capacity" of the set of models that are considered – and this is often implemented in terms of a regularization penalty. Complicated models are more heavily penalized than simple models, so the chosen model is one that optimizes a trade-off between the error on the training data and model complexity. This formalizes the principle of Occam's razor, which seeks the simplest model that explains the data.

A Bayesian interpretation of this regularization is that it incorporates the *prior* information. We are given a training dataset $\{(x_i, y_i)\}_{i=1}^{\ell}$, where $x_i \in \Re^d$ are input variables and $y_i \in \{-1, 1\}$ are class labels. We consider a class of models of the form $f(x) = \alpha^T x + \alpha_0$, with the sign of $f(x)$ predicting the label associated with the point x. The Bayes maximum a posteriori (MAP) estimate for α can be obtained by maximizing the following posterior density.

$$p(\alpha|y) = p(y|\alpha)p(\alpha) \tag{1}$$

Let $\mathcal{L}(z) \equiv -\log p(z)$ be the negative log likelihood associated with the probability density p. Then, the MAP estimate is the estimate that minimizes the loss function $\mathcal{L}(\alpha|y)$ given by

$$\mathcal{L}(\alpha|y) = \mathcal{L}(y|\alpha) + \mathcal{L}(\alpha). \tag{2}$$

The first term on the right-hand side of Eq. (2) is the negative log likelihood of the data given the model, and can be interpreted as a measure of how well the model "fits" the data. The second term is is the penalty for models α that are *a priori* unlikely (small p, or large $\mathcal{L}$); this is the term that can be used to penalize complexity. This Bayesian interpretation suggests that the penalty term is an appropriate place to incorporate domain (or *prior*) knowledge. Different choices for $\mathcal{L}(\alpha)$ correspond to different constraints on α, and lead to different optimization problems.

The standard penalty for the support vector machine takes a quadratic penalty $\mathcal{L}(\alpha) = \lambda\|\alpha\|^2 = \lambda\sum_{i=1}^{d}\alpha_i^2$. But other penalty functions are possible: for instance, a common approach in feature selection is to choose a penalty function, such as $\mathcal{L}(\alpha) = \lambda\sum_{i=1}^{d}|\alpha_i|^p$, with $p = 0$ or $p = 1$, since this penalizes solutions with many nonzero components, and prefers sparse solutions in which many of the α_i vanish.

Our domain knowledge tells us to prefer solutions for which the discriminant function $f(x)$ exhibits a contiguity property – that is, $f(x)$ tends to be more well conserved over neighboring pairs of pixels than over random pairs of pixels. This corresponds to saying that nearby pixels should align with the hyperplane that separates positive samples from negative ones.

If we construct a quadratic penalty function that penalizes dis-contiguity, then the regularization that penalizes overfitting will at the same time encourage solutions with the contiguity property. This penalty is characterized by a "contiguity matrix" whose computation does not require labeled data; thus the method also provides a way to use a mix of labeled and unlabeled data.

II. Support Vector Machine (SVM)

In a conventional support vector machine [1], the solution is expressed in terms of a linear decision function; the label estimated for a point x is given by the sign of this function evaluated at x. Note that if $y_i f(x_i)$ is positive, then the i^{th} point is correctly classified by the discriminant function $f(x)$. In fact, the larger $y_i f(x_i)$, the greater "margin" by which the point is correctly classified. The SVM cost function promotes large margin classification but does this in a way that bounds the magnitude of the coefficients. The parameters α and α_0 are chosen to optimize the cost function,

$$\begin{array}{cccc} \min\limits_{\alpha,\,\alpha_0,\,\xi_i} & \frac{1}{2}\alpha^T\alpha & + & C\sum_{i=1}^{\ell}\xi_i \\ \text{s.t.} & y_i(\alpha^T x_i+\alpha_0) & \geq & 1-\xi_i \\ & \xi_i \geq 0 & & \end{array} \tag{3}$$

where the constant C expresses the cost of misclassification, relative to the penalty for large coefficients α in the discriminant function and ξ_i, $i = 1, \ldots, \ell$ is the amount by which the prediction is on the wrong side of its margin.

To use the support vector machine to learn nonlinear functions $f(x)$, the "kernel trick" is invoked [2]. Here, a transformation $\phi(x)$ maps x to a (usually) higher dimensional space, and $f(x)$ is a linear combination of of the components of $\phi(x)$. The "trick" is that the optimization of this function depends on $\phi(x)$ *only* through dot products $\phi(x)^T\phi(x')$; so rather than deal with $\phi(x)$ directly, one can define a kernel function

$$K(x, x') = \phi(x)^T\phi(x'). \tag{4}$$

Note that the linear SVM is the special case where $K(x, x') = x^T x'$. The optimization can be performed entirely in terms of this kernel, and the solution can be written

$$f(x) = \sum_{i=1}^{\ell} \theta_i y_i K(x_i, x) + \theta_o. \tag{5}$$

where the θ_i are scalar coefficients that are optimized over. In this sum over the data samples i, the function $f(x)$ depends only on the samples for which $\theta_i \neq 0$; these are the so-called "support vectors."

III. Contiguity-Enhanced Support Vector Machine (CE-SVM)

For features having broader spatial extent, neighboring pixels have the tendency to have similar classifier outputs. To encourage solutions with this property, we add an extra term to the loss function which penalizes deviations in the classifier output at pixel i from classifier output given by pixels that are neighbors of i. We introduce a contiguity penalty

$$L(\alpha) = \frac{1}{8\ell}\sum_{i}^{\ell}\sum_{j}^{8}\|f(x_i) - f(x_{ij})\|^2 \tag{6}$$

where we write x_{ij} as the jth neighbor of x_i; here we consider the eight pixels surrounding x_i as its neighbors. Since $f(x) = \alpha^T x + \alpha_0$, we can write this as

$$\begin{aligned} L(\alpha) &= \frac{1}{8\ell}\sum_{i}^{\ell}\sum_{j}^{8}\|f(x_i) - f(x_{ij})\|^2 \\ &= \alpha^T \underbrace{\frac{1}{8\ell}\sum_{i}^{\ell}\sum_{j}^{8}(x_i - x_{ij})(x_i - x_{ij})^T}_{\Psi}\alpha. \end{aligned} \tag{7}$$

We call Ψ the contiguity matrix, and note that the computation of this matrix does not require labeled data and can easily be computed for any image. If we add λ times the expression in (7) to the cost function in (3), the above optimization problem can be rewritten as

$$\begin{array}{cccc} \min\limits_{\alpha,\,\alpha_0,\,\xi_i} & \frac{1}{2}\alpha^T(I+\lambda\Psi)\alpha & + & C\sum_{i=1}^{\ell}\xi_i \\ \text{s.t.} & y_i(\alpha^T x_i+\alpha_0) & \geq & 1-\xi_i \\ & \xi_i \geq 0 & & \end{array} \tag{8}$$

Note that λ is the weighting factor determining the trade-off between two possibly conflicting goals when optimizing the hyperplane that separates the two classes: "encouraging neighboring pixels of x_i to align along the hyperplane" and "keeping the coefficients small". The added term nominally alters the mathematical structure of the SVM cost function, so a standard SVM optimizer cannot directly be used.

However, we can use the fact that Ψ is symmetric and positive definite to effect a change of coordinates:

$$z = (I+\lambda\Psi)^{-1/2}x \tag{9}$$

$$\beta = (I+\lambda\Psi)^{1/2}\alpha. \tag{10}$$

The optimization in Eq. (8) then becomes

$$\begin{array}{cccc} \min\limits_{\beta,\,\alpha_0,\,\xi_i} & \frac{1}{2}\beta^T\beta & + & C\sum_{i=1}^{\ell}\xi_i \\ \text{s.t.} & y_i(\beta^T z_i+\alpha_0) & \geq & 1-\xi_i \\ & \xi_i \geq 0 & & \end{array} \tag{11}$$

which is the standard formulation of the SVM, and therefore standard SVM software packages can be used.

Another interpretation of the CE-SVM algorithm is that the contiguity can be incorporated into the definition of the kernel. Instead of $K(x, x') = x^T x'$ for the standard linear SVM, we write

$$K(x, x') = x^T(I+\lambda\Phi)^{-1}x. \tag{12}$$

The use of a different kernel to enhance contiguity-preserving solutions suggest the interpretation, consistent with the philosophy in Chapter 11 of Schölkopf and Smola [2], of engineering kernels to produce invariance-preserving solutions.

In the next section we test our algorithm with two multi-spectral datasets.

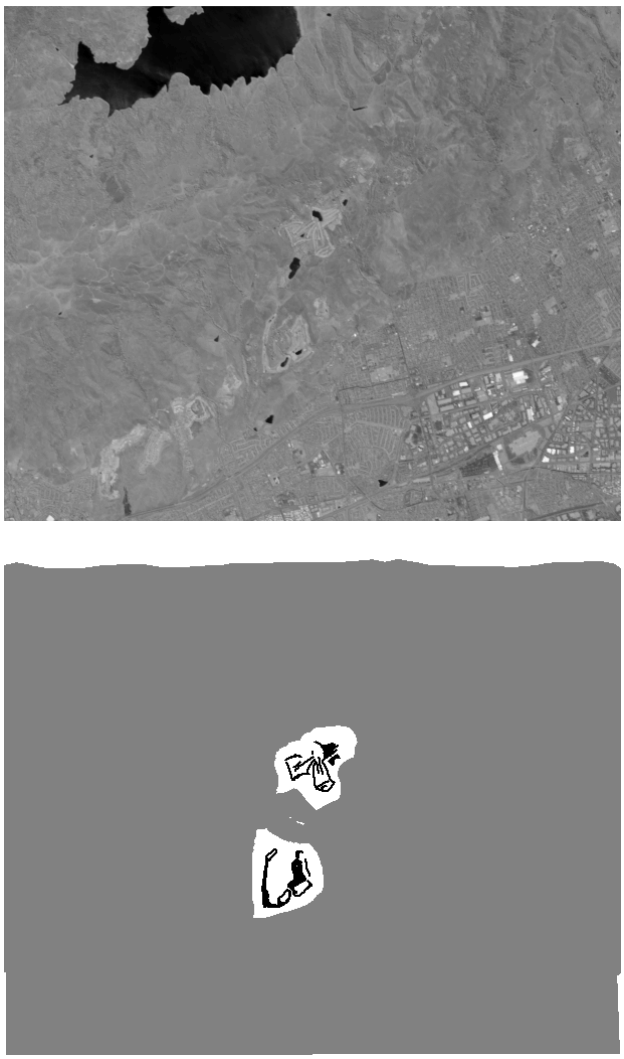

Fig. 1. **(a)** Broadband image of Moffet field. This is a sum of the ten channels used in the multispectral dataset that was derived from AVIRIS data. **(b)** Markup for golf courses on the Moffet field image; here black indicates the pixels where the golf courses are, gray indicates where the golf course are not, and white is not marked up.

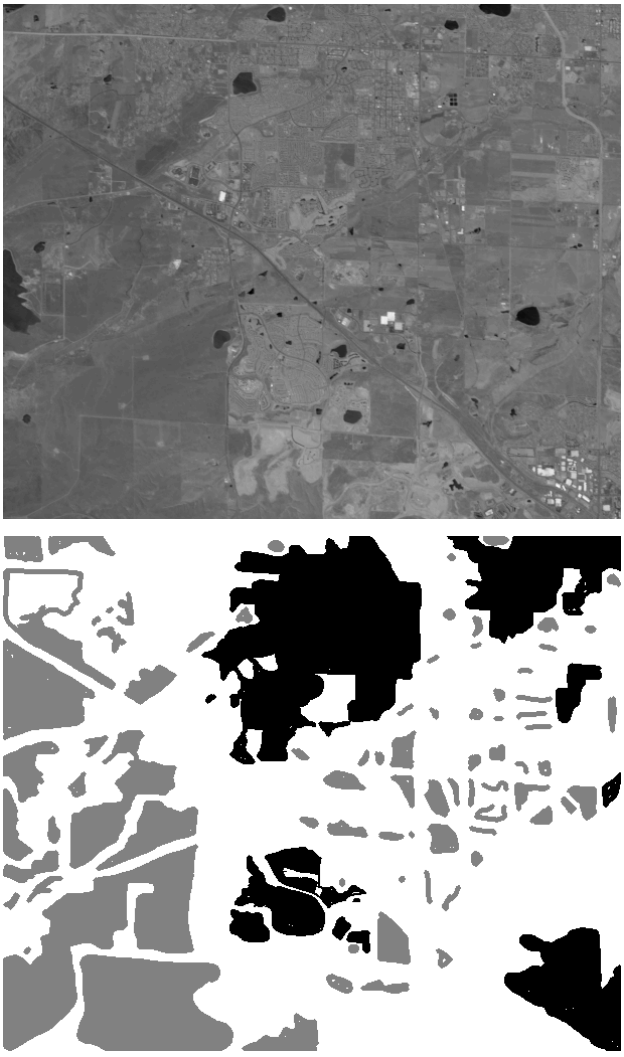

Fig. 2. **(a)** Broadband image of the Denver area. **(b)** Markup for urban areas; black indicates the pixels where the urban areas are, gray indicates where the urban areas are not, and white is not marked up.

IV. Experimental Results and Discussion

A. Data Used in the Experiments

The two images used in this study are derived from the Airborne Visible and Infrared Imaging Spectrometer (AVIRIS) [3], [4]. The AVIRIS sensor collects data in 224 contiguous, relatively narrow (10 nm), uniformly-spaced spectral channels. AVIRIS is an airborne sensor and spatial resolution can vary from a few meters to 20 meters. The studies reported in this paper use a reduced number of relatively wide spectral bands. This reduction in the number of spectral bands was performed to match the bands of a new remote sensing satellite called the Multispectral Thermal imager (MTI) [5]. The MTI satellite was launched in March 2000 and collects data in 15 spectral bands. Ten of these bands sample wavelengths between 0.4 and 2.4 microns, a region covered by the AVIRIS instrument. AVIRIS data were convolved with the MTI spectral filter functions to produce simulated MTI data. This 10-band simulated data was also used in Ref. [6]. In this data, the features of interest are Golf Courses and Urban Areas. These features are chosen because of their particular attributes in multi-spectral data. Both urban areas and golf courses generally encompass quite broad area land-cover distinction. The scenes together with their corresponding test field maps are shown in Figs. 1 and 2.

B. Classification Tasks

We conduct experiments to compare the performance of contiguity enhanced SVM (CE-SVM) with that of conventional SVM. To estimate the parameters C and λ we first coarsely tune these parameters independently and determine a range of values for each parameter. Then for each parameter we consider a discrete set of parameters. For CE-SVM we select the pair of C and λ values that yield the best accuracy on the validation set. For conventional SVM the performance

TABLE I

AVERAGE CLASSIFICATION ERROR ON GOLF COURSES IN FIG. 1 FOR VARYING TRAINING SAMPLE SIZE PER CLASS. VALUES ARE PERCENTAGE ERROR, WITH STANDARD DEVIATIONS IN IN PARENTHESES.

	Training Sample Size			
Classifier	10	20	50	100
SVM	8.7 (1.0)	7.6 (1.0)	6.0 (1.0)	5.1 (0.5)
SVM-CE	6.1 (1.0)	5.7 (1.0)	5.0 (0.6)	4.6 (0.4)

on the validation set is optimized with respect to C only. Note that C values estimated this way may not necessarily be the same. We observe the performance of both classifiers for varying sizes of training set, randomly sampled from the pool of labeled dataset. Each experiment is repeated 30 times and results are shown in Tables I and II.

C. Results and Analysis

As the above experimental results suggest when the training size is small contiguity enhanced Support Vector Machine performs better than its naive version. This is quite intuitive as the added contiguity term acts as a regularizer over the classifier and yields a classifier with better generalizability. The impact of the regularization is more significant when the training data is limited because the classifier is more prone to overfit the training data in this case.

V. REMARKS ON OTHER CLASSIFIERS

We have shown how contiguity can be incorporated into the linear support vector machine, and we remark that the formalism can be extended to nonlinear (kernelized) SVM as well. The derivation in that case, however, is a fair bit more complicated, and since our numerical results are only for the linear SVM, we will not show that derivation here.

But it is relatively straightforward to produce contiguity-enhanced algorithms both for other pattern recognition tasks, from unsupervised clustering [7] to supervised classification. In particular, we will illustrate how the contiguity matrix can be incorporated into the Fisher discriminant [8]. The standard derivation is based on the within-class covariance

$$S_w = \sum_i (x_i - \mu_{y_i})(x_i - \mu_{y_i})^T, \tag{13}$$

and the between-class covariance

$$S_b = \sum_i (\mu_{y_i} - \mu)\,(\mu_{y_i} - \mu)^T, \tag{14}$$

where μ_y is the mean value of the data samples x_i for which $y_i = y$, and μ is the mean over all the data samples. The decision function is given by $f(x) = \alpha^T x + \alpha_o$, and minimizing the within-class variance ($\alpha^T S_w \alpha$) while maximizing the between-class variance ($\alpha^T S_b \alpha$) leads to the best discriminant(s), which are given by the eigenvectors of the matrix product $S_b S_w^{-1}$. For the Fisher discriminant, this leads to $\alpha = S_w^{-1}(\mu_{+1} - \mu_{-1})$.

TABLE II

SAME AS TABLE I, BUT FOR URBAN AREAS IN FIG. 2.

	Training Sample Size			
Classifier	10	20	50	100
SVM	13.1 (2.5)	10.6 (1.7)	7.6 (0.6)	6.0 (0.5)
SVM-CE	10.7 (1.7)	8.8 (1.7)	7.0 (1.0)	6.0 (0.6)

In the contiguity-enhanced Fisher discriminant, we want both the contiguity penalty $\alpha^T \Psi \alpha$ and within-class variance $\alpha^T S_w \alpha$ to be small. One way to combine these is to replace S_w with $S_w^* = (I + \lambda\Psi)^{1/2} S_w (I + \lambda\Psi)^{1/2}$. Then, the Fisher discriminant is given by $\alpha = {S_w^*}^{-1}(\mu_{+1} - \mu_{-1})$. An advantage of this formulation is that it can be expressed as the standard Fisher discriminant with data scaled according to Eq. (9).

VI. CONCLUSION

Many features of interest in real-world images have a spatial extent; as a consequence, neighboring pixels tend to be similarly classified. We have described an efficient way to exploit this tendency in order to improve the performance of linear support vector machines. Here, the spatial information is directly incorporated into the design of the classifier, using a contiguity matrix that can be computed for any image, without regard to the model that is being fit. Finally, we remark that the CE-SVM requires a single convex optimization, and avoids the expensive iterative steps that are necessary for contiguity-enhancing algorithms based on Markov Random Fields (*e.g.,* see Refs. [9], [10]).

REFERENCES

[1] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[2] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, MA: MIT Press, 2002.

[3] G. Vane, R. O. Green, T. G. Chrien, H. T. Enmark, E. G. Hansen, and W. M. Porter, "The Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)," *Remote Sensing of the Environment*, vol. 44, pp. 127–143, 1993.

[4] Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), Jet Propulsion Laboratory (JPL), National Aeronautics and Space Administration (NASA) `http://aviris.jpl.nasa.gov/`.

[5] P. G. Weber, B. Brock, A. J. Garrett, B. W. Smith, C. C. Borel, W. B. Clodius, S. C. Bender, R. R. Kay, and M. L. Decker, "Multispectral Thermal Imager mission overview," *Proc. SPIE*, vol. 3750, pp. 340–346, 1999.

[6] N. R. Harvey, J. Theiler, S. P. Brumby, S. Perkins, J. J. Szymanski, J. J. Bloch, R. B. Porter, M. Galassi, and A. C. Young, "Comparison of GENIE and conventional supervised classifiers for multispectral image feature extraction," *IEEE Trans. Geosci. and Remote Sens.*, vol. 40, pp. 393–404, 2002.

[7] J. Theiler and G. Gisler, "A contiguity-enhanced k-means clustering algorithm for unsupervised multispectral image segmentation," *Proc. SPIE*, vol. 3159, pp. 108–118, 1997.

[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. San Diego: Academic Press, 1990.

[9] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence PAMI*, vol. 6, pp. 721–741.

[10] J. Besag, "On the statistical analysis of dirty pictures," *J. R. Statis. Soc. B*, vol. 48, pp. 259–302, 1986.