Support Vector Machine (SVM) pattern recognition to AVO classification

Jiakang Li
University of Oklahoma, School of Geology and Geophysics, USA

John Castagna
University of Oklahoma, School of Geology and Geophysics, USA

Received 31 July 2003; revised 7 December 2003; accepted 30 December 2003; published 28 January 2004.

[1] The purpose of this paper is to present a learning algorithm to classify data with nonlinear characteristics. The Support Vector Machine (SVM) is a novel type of learning machine based on statistical learning theory [*Vapnik*, 1998]. The support vector machine (SVM) implements the following idea: It maps the input vector X into a high-dimensional feature space Z through some nonlinear mapping, chosen *a priori*. In this space, an optimal separating hyperplane is constructed to separate data groupings. The support vector machine (SVM) learning method can be used to classify seismic data patterns for exploration and reservoir characterization applications. The SVM is particularly good at classifying data with nonlinear characteristics. As an example the SVM method is applied to AVO classification of gas sand and wet sand. *INDEX TERMS:* 0902 Exploration Geophysics: Computational methods, seismic; 3220 Mathematical Geophysics: Nonlinear dynamics; 3299 Mathematical Geophysics: General or miscellaneous. **Citation:** Li, J., and J. Castagna (2004), Support Vector Machine (SVM) pattern recognition to AVO classification, *Geophys. Res. Lett.*, *31*, L02609, doi:10.1029/2003GL018299.

1. Introduction

[2] In geophysical data interpretation the sample population to which training can be applied is often too small for statistically significant prediction. Conventional statistical pattern classification doesn't perform well in this case. The Support Vector Machine (SVM) based on statistical learning theory [*Vapnik*, 1998] deals with the problem of small sample statistics. The theory for controlling the generalization ability of learning machines is devoted to constructing an inductive principle for minimizing the risk functional using a small sample of training instances.

2. Learning Machine Principle

[3] The problem of learning is that of choosing from a given set of functions $f(x, \alpha)$ the one that best approximates the supervisor's response. The selection of the desired function is based on a training set of L independent and identically distributed observations drawn according to $F(x, y) = F(x)F(y|x)$:

$$(x_1, y_1), \cdots, (x_L, y_L) \tag{1}$$

[4] In order to choose the best available approximation to the supervisor's response, one measures the loss, or discrepancy $L(y, f(x, \alpha))$ between the response y of the supervisor to a given input x and the response $f(x,\alpha)$ provided by the learning machine. Consider the expected value of the loss, given by risk functional

$$R(\alpha) = \int L(y, f(x, \alpha))dF(x, y) \tag{2}$$

[5] The goal is to find the function $f(x,\alpha_0)$ that minimizes the risk functional $R(\alpha)$ (over the class of functions $f(x,\alpha)$) in the situation where the joint probability distribution function $F(x,y)$ is unknown and the only available information is contained in the training set (1).

[6] For a set of functions $f(x, \alpha)$, statistical approaches minimize the functional

$$R(\alpha) = \int L(y - f(x, \alpha))dp(x, y) \tag{3}$$

Where $L(u)$ is a given loss function if the probability measure $P(x, y)$ is unknown.

[7] In the case of a small population, the empirical risk minimization principle suggests minimizing the functional

$$R_{emp}(\alpha) = \frac{1}{L}\sum_{i=1}^{L} L(y_i - f(x_i, \alpha)) \tag{4}$$

instead of the functional (3).The structural risk minimization method defines where a structure on a set of functions $f(x, \alpha)$ has been defined as

$$S_1 \subset \cdots \subset S_n \tag{5}$$

and functional (4) is minimized on the approximately chosen element S_k of this structure.

[8] We now consider a new basic function instead of the empirical risk functional (4) and use this functional in the structural risk minimization scheme.

[9] We construct (using data) vicinity functions $v(x_i)$ of the vectors x_i for all training vectors and then using these vicinity functions we construct the vicinal risk functional

$$V(\alpha) = \frac{1}{L}\sum_{i=1}^{L} L\left(y_i - \frac{1}{v_i}\int_{v(x_i)} f(x, \alpha)dx\right) \tag{6}$$

Minimizing functional (6) instead of functional (3) is called the vicinal risk minimization (VRM) method.

0094-8276/04/2003GL018299

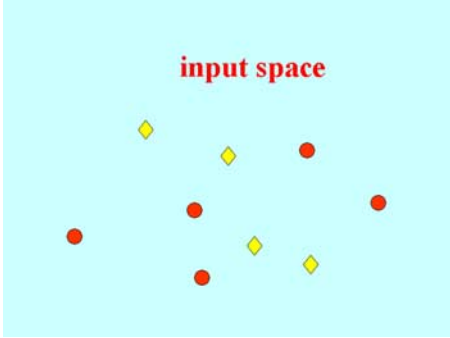

Figure 1. The nonlinear nonseparable data in input space.

[10] We apply the VRM method to the two class $\{-1, 1\}$ pattern recognition problem. Consider the set of indicator functions

$$y = sign[f(x, \alpha)] \tag{7}$$

where we minimized the empirical functional (4) with the loss function $|y - f(x, \alpha)|$.

[11] The input data may be nonlinear and difficult to separate in input space (Figure 1). After transformation, we desire the data to be linear and separable in the feature space (Figure 2). We put input vectors x into feature vectors z and in the feature space construct a hyperplane

$$(w, z) + b = 0 \tag{8}$$

the separates data

$$(x_1, y_1), \cdots, (x_L, y_L),$$

which are image in the feature space of the training dataset (1).

[12] Our goal is to find the function $f(x, \alpha)$ satisfying the constraints

$$y_i \int f(x, \alpha) p(x|x_i, r_i) dx \geq 1 - \xi_i \tag{9}$$

(here $p(x|x_i, r_i)$ are distribution functions where they define the parameters of position and width, ξ_i are nonnegative slack variables) whose image in the feature space is a linear function

$$l(z) = (w^*, z) + b \tag{10}$$

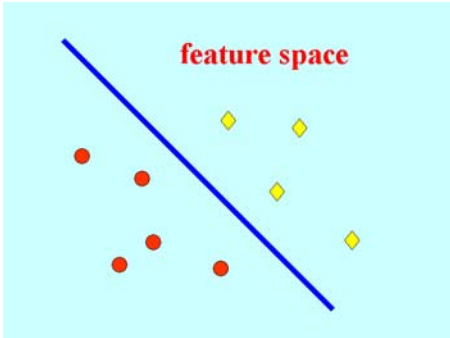

Figure 2. The linear separable data in feature space.

that minimizes the functional

$$W(w) = (w, w) + C \sum_{i=1}^{L} \xi_i \tag{11}$$

(here C is a given upper boundary value) subject to constraint (9).

3. Computation Approach

[13] To construct the optimal hyperplane one has to separate the vector x_i of the training set

$$(x_1, y_1), \cdots, (x_L, y_L)$$

belonging to two different classes $y = \{-1, 1\}$ using the hyperplane with the smallest norm of the coefficients.

[14] To find this hyperplane we have to solve the following quadratic programming problem: minimize the functional

$$\Phi(w) = \frac{1}{2}(w \cdot w) \tag{12}$$

under the constraints of inequality type

$$y_i[(x_i \cdot w) - b] \geq 1, \quad i = 1, \cdots, L \tag{13}$$

[15] The solution to this optimization problem is given by the saddle point of the Lagrange functional (Lagrangian):

$$L(w, b, \alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{L} \alpha_i \{[(x_i \cdot w) - b]\, y_i - 1\} \tag{14}$$

where the α_i are Lagrange multipliers. The Lagrangian has to be minimized with respect to w and b and maximized with respect to $\alpha_i > 0$.

[16] The optimal hyperplane (Figure 3) has the following properties:

[17] (1) The coefficients α_i^0 for the optimal hyperplane should satisfy the constructs

$$\sum_{i=1}^{L} \alpha_i^0 y_i = 0, \quad \alpha_i^0 \geq 0, \quad i = 1, \cdots, L \tag{15}$$

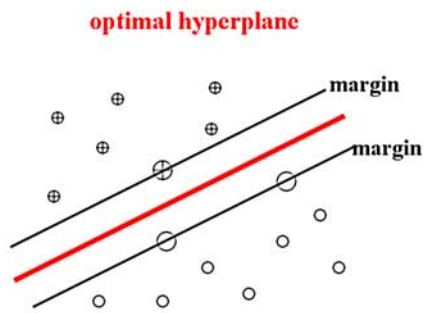

Figure 3. The optimal hyperplane and support vectors. The middle red line is the optimal hyperplane. The samples occur on margins are support vectors.

Table 1.

		Inputs		
		A	B	class
1	Top Gas	−1	−1	+1
2	Base Wet	−1	+1	−1
3	Top Wet	+1	−1	−1
4	Base Gas	+1	+1	+1
5		−1	−0.5	
6		−0.8	0.8	
7		0.75	0.75	
8		0.25	−0.25	

[18] (2) The optimal hyperplane (vector w_0) is a linear combination of the vectors of the training set

$$w_0 = \sum_{i=1}^{L} y_i \alpha_i^0 x_i, \quad \alpha_i^0 \geq 0, \quad i = 1, \cdots, L \tag{16}$$

[19] (3) Moreover, only the so-called support vectors can have nonzero coefficient α_i^0 in the expansion of w_0. The support vectors are the vectors making (13) achieve equality. Therefore, for support vectors (*s.v.*), we obtain

$$w_0 = \sum_{s.v.} y_i \alpha_i^0 x_i, \quad \alpha_i^0 \geq 0, \tag{17}$$

The necessary and sufficient conditions of the optimal hyperplane are that the separating hyperplane satisfy the conditions

$$\alpha_i^0 \{[(x_i \cdot w_0) - b_0]\, y_i - 1\} = 0, \quad i = 1, \cdots, L \tag{18}$$

[20] Putting the expression for w_0 into the Lagrangian and taking into account the Kuhn-Tucker conditions, one obtains the functional

$$W(\alpha) = \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i,j}^{L} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j), \tag{19}$$

under the constraint

$$\alpha_i^0 \geq 0, \quad i = 1, \cdots, L \tag{20}$$

$$\sum_{i=1}^{L} \alpha_i^0 y_i = 0. \tag{21}$$

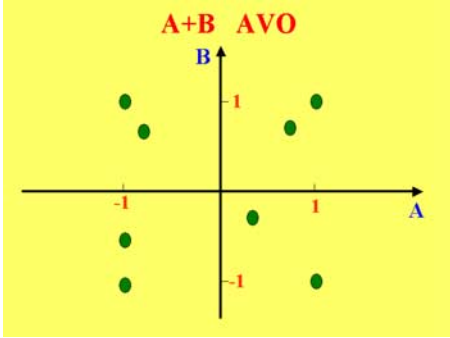

Figure 4. All input data.

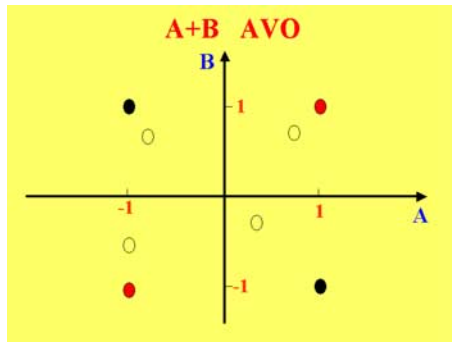

Figure 5. The training data their classes are known.

Thus, to construct the optimal hyperplane we have to solve a quadratic programming problem: Maximize the quadratic form (19) under constraints (20) and (21).

[21] The separating rule, based on the optimal hyperplane, is the following indicator function

$$f(x) = sign\left(\sum_{s.v.} y_i \alpha_i^0 (x_i \cdot x) - b\right), \tag{22}$$

where x_i are the support vectors, α_i^0 are the corresponding Lagrange coefficients, and b is the constraint (threshold).

4. AVO Classification

[22] The AVO reflection coefficient variation with angle of incidence, $R(\theta)$ can be written in Shuey's form:

$$R(\theta) = A + B \sin^2 \theta \tag{23}$$

where A is the AVO intercept, and B is the AVO gradient. Crossplotting AVO intercept (A) and gradient (B) can sometimes reveal anomalous AVO behavior caused by hydrocarbons. Hydrocarbon bearing sands may be classified according to their location in the A-B plane, [*Castagna et al.*, 1998]. In this paper, however, we will attempt to differentiate only two situations - gas sands or wet sands.

[23] Theoretically, gas sands may occur in any quadrant of the A-B plane. We now consider some known gas sand and wet sand normalized pairs of intercepts and gradients (entries 1–4 in Table 1) and some pairs from unknown reflections (entries 5–8). We classify the reflections such that +1 represents a gas sand and −1 represents a wet sand.

[24] The 8 entries are shown in Figure 4. Suppose the classes of examples 1–4 are known *a priori* and their A-B distribution is as displayed in Figure 5. This is a typical class 3 AVO anomaly [*Rutherford and Williams*, 1989]. Obviously, the classes are not linearly separable in input space [*Russell et al.*, 2002].

Table 2.

	Inputs		
	A	B	class
5	−1	−0.5	+1
6	−0.8	0.8	−1
7	0.75	0.75	+1
8	0.25	−0.25	−1

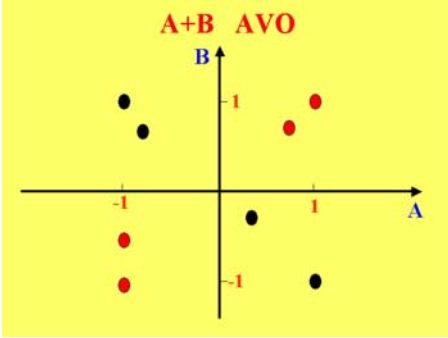

Figure 6. The final classified result.

[25] The classification of entries 1–4 are known, entry 1 is top gas sand which class is +1, entry 2 is wet sand which class is −1, entry 3 is wet sand which class is −1, entry 4 is gas sand which class is +1. From Table 1, we can see that gas sands occur in quadrants I and III, wet sands are in quadrant II and IV.

[26] Now we use entries 1–4 as an input dataset to train the SVM (solve a Quadratic Programming problem, equation (19)–(21)) to classify this prior dataset into gas sand (+1) and wet sand (−1) for finding Lagrange multipliers α_i^0 and support vectors x_i. While a correct classification is made to prior dataset entries 1–4, we obtain a group of Lagrange multipliers α_i^0 and support vectors x_i as well as their classification indicators y_i, (i = 1,. . ., ℓ). Then unknown dataset entries 5–8 are classified by equation (22) to recognize their pattern. The classified results are shown in Table 2. Entry 5 occurs in quadrant III and entry 7 is in quadrant I and are classified as gas sands; entry 6 is in quadrant II and entry 8 is in quadrant IV and are classified as wet sands. Here, for the two class problem in the intercept-gradient plane of AVO analysis, unknown dataset entries 5–8 are correctly classified according to a learned classification from prior data entries 1–4. The final classified result is shown in Figure 6.

5. Conclusions

[27] We present an approach to classify seismic attributes using the Support Vector Machine (SVM) based on statistical learning theory. The SVM algorithm maps nonlinear nonseparable data in input space into a multi-dimensional feature space in which a hyperplane separates the mapped data. To construct the optimal hyperplane, a quadratic programming problem is solved to find support vectors. A simple intercept and gradient AVO classification problem illustrates this approach. The result shows that SVM classification is a useful tool for recognizing non-linear seismic patterns.

References

Castagna, J. P., H. W. Swan, and D. J. Foster (1998), Framework for AVO gradient and intercept interpretation, *Geophysics*, *63*, 948–956.

Russell, B., C. Ross, and L. Lines (2002), Neural Networks and AVO: *72th Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts*.

Rutherford, S. R., and R. H. Williams (1989), Amplitude-versus-offset variations in gas sands, *Geophysics*, *54*, 680–688.

Vapnik, V. (1998), Statistical Learning Theory, John Wiley & Sons Inc.

J. Li and J. Castagna, School of Geology and Geophysics, University of Oklahoma, USA. (jiakang_li@yahoo.com; castagna@ou.edu)