

Support vector pattern recognition and AVO classification

Jiakang Li, School of Geology and Geophysics, University of Oklahoma and John Castagna, School of Geology and Geophysics, University of Oklahoma

Summary

The support vector (SV) learning method can be used to classify seismic data patterns for exploration and reservoir characterization applications. The SV method is particularly good at classifying data with nonlinear characteristics. As an example the method is applied to AVO gas sand classification.

Introduction

The purpose of this paper is to present a learning algorithm to classify data with nonlinear characteristics. The support vector (SV) algorithm is a novel type of learning machine based on statistical learning theory (Vapnik, 1998). The support vector (SV) machine implements the following idea: It maps the input vectors 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.

Statistic learning theory and pattern recognition

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. Statistics learning theory 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.

The problem of learning is that of choosing from the 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 $F(x, y) = F(x)F(y|x)$:

$$(x_1, y_1), \dots, (x_\ell, y_\ell) \quad (1)$$

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

$$R(\alpha) = \int L(y, f(x, \alpha)) dF(x, y) \quad (2)$$

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).

In pattern recognition applications we let the supervisor's output y take only two values $y = \{0, 1\}$ and let $f(x, \alpha)$ be a set of indicator functions (functions which take only two values: zero and one). Consider the following loss function:

$$L(y, f(x, \alpha)) = \begin{cases} 0 & \text{if } y = f(x, \alpha) \\ 1 & \text{if } y \neq f(x, \alpha) \end{cases} \quad (3)$$

For this loss function, the functional (2) determines the probability of different answers given by the supervisor and by the indicator function $f(x, \alpha)$. We call the case of different answers a classification error. The problem, therefore, is to find a function that minimizes the probability of classification error when the probability measure $F(x, y)$ is unknown, but the data (1) are given.

In the general learning problem we let the probability measure $F(z)$ be defined on the space Z and consider the set of functions $Q(Z, \alpha)$. The goal is to minimize the risk functional

$$R(\alpha) = \int Q(z, \alpha) dF(z) \quad (4)$$

where the probability measure $F(z)$ is unknown, but an independent and identically distributed sample

$$z_1, \dots, z_\ell \quad (5)$$

is given.

Seismic pattern recognition learning problems are particular cases of the general problem of minimizing the risk functional (4) on the basis of empirical data (5), where z describes a pair (x, y) and $Q(z, \alpha)$ is the specific loss function. For the pattern recognition problem, the functional (2) evaluates the probability of error for any function of the admissible set of functions. The problem is to use the sample to find the function from the set of admissible functions that minimizes the probability of error. This is exactly what we want to obtain.

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

$$R(\alpha) = \int L(y - f(x, \alpha)) dp(x, y) \quad (6)$$

where $L(u)$ is a given loss function if the probability measure, $P(x, y)$ is unknown but the data

$$(x_1, y_1), \dots, (x_\ell, y_\ell)$$

are given.

The empirical risk minimization principle suggests minimizing the functional

$$R_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i - f(x_i, \alpha)) \quad (7)$$

instead of the functional (6).

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

$$S_i \subset \dots \subset S_n \quad (8)$$

and functional (7) is minimized on the approximately chosen element S_k of this structure.

We now consider a new basic function instead of the empirical risk functional (7) and use this functional in the structural risk minimization scheme.

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

$$V(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i - \frac{1}{v_i} \int_{v(x_i)} f(x, \alpha) dx) \quad (9)$$

Minimizing functional (9) instead of functional (6) is called the vicinal risk minimization (VRM) method.

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

$$y = \text{sign}[f(x, \alpha)] \quad (10)$$

where we minimized the empirical functional (7) with the loss function $|y - f(x, \alpha)|$.

The input data may be nonlinear and difficult to separate in input space (Fig.1). Passing transform, we desire that the data to be linear and separable in the feature space (Fig.2). We put input vectors, x , into feature vectors, z , and in the feature space construct a hyperplane

$$(w, z) + b = 0 \quad (11)$$

that separates data

$$(x_1, y_1), \dots, (x_\ell, y_\ell),$$

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

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

$$y_i \int f(x, \alpha) p(x | x_{i,r}) dx \geq 1 - \xi_i \quad (12)$$

(ξ_i are nonnegative slack variables) whose image in the feature space is a linear function

$$l(z) = (w^*, z) + b \quad (13)$$

that minimizes the functional

$$W(w) = (w, w) + C \sum_{i=1}^{\ell} \xi_i \quad (14)$$

(here C is a given upper boundary value) subject to constraint (12).

Computation Approach

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

$$(x_1, y_1), \dots, (x_\ell, y_\ell)$$

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

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

$$\Phi(w) = \frac{1}{2} (w \cdot w) \quad (15)$$

under the constraints of inequality type

$$y_i [(x_i \cdot w) - b] \geq 1, \quad i = 1, \dots, \ell \quad (16)$$

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

$$L(w, b, \alpha) = \frac{1}{2} (w \cdot w) - \sum_{i=1}^{\ell} \alpha_i \{ [(x_i \cdot w) - b] y_i - 1 \} \quad (17)$$

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

The optimal hyperplane (Fig.3) has the following properties:

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

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i = 0, \quad \alpha_i^0 \geq 0, \quad i = 1, \dots, \ell \quad (18)$$

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

$$w_0 = \sum_{i=1}^{\ell} y_i \alpha_i^0 x_i, \quad \alpha_i^0 \geq 0, \quad i = 1, \dots, \ell \quad (19)$$

(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 (16) achieves equality. Therefore, for support vectors (s.v.), we obtain

$$w_0 = \sum_{s.v.} y_i \alpha_i^0 x_i, \quad \alpha_i^0 \geq 0, \quad (20)$$

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

$$\alpha_i^0 \{ [(x_i \cdot w_0) - b_0] y_i - 1 \} = 0 \geq 0, \quad i = 1, \dots, \ell \quad (21)$$

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

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j), \quad (22)$$

under the constraint

$$\alpha_i^0 \geq 0, \quad i = 1, \dots, \ell \quad (23)$$

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i = 0. \quad (24)$$

Thus, to construct the optimal hyperplane we have to solve a quadratic programming problem: Maximize the quadratic form (22) under constraints (23) and (24).

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

$$f(x) = \underset{s.v.}{\text{sign}}(\sum y_i \alpha_i^0 (x_i \cdot x) - b), \quad (25)$$

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

AVO classification

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

$$R(\theta) = A + B \sin^2 \theta$$

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.

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.

Table 1

		Inputs		class
		A	B	
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	

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 Fig 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).

Now we use entries 1-4 as an input dataset to train the SV classification machine, and then recognize the pattern of entries 5-8 to classify them. The classified results are shown in table 2 and Fig. 6.

Table 2

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

Conclusions

We present an approach to classify seismic attributes using a Support Vector (SV) machine based on statistical learning theory. The SV 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 SV classification is a useful tool for recognizing non-linear seismic patterns.

REFERENCES

- Castagna, J.P., Swan, H.W., and Foster, D.J., 1998, Framework for AVO gradient and intercept interpretation, *Geophysics*, 63, 948-956.
- Russell B., and Ross C., Lines L., 2002, *Neural Networks and AVO: 72th Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts*.
- Rutherford, S.R. and Williams, R.H., 1989, Amplitude-versus-offset variations in gas sands: *Geophysics*, 54, 680-688.

Authors:

Dr. Jiakang Li,
School of Geology and Geophysics,
University of Oklahoma, 810 Sarkeys Energy Center,
100 East Boyd Street, Norman, Oklahoma 73019-1009.
Email: jkli2000@yahoo.com

Dr. John Castagna,
School of Geology and Geophysics,
University of Oklahoma, 810 Sarkeys Energy Center,
100 East Boyd Street, Norman, Oklahoma 73019-1009.
Email: castagna@ou.edu

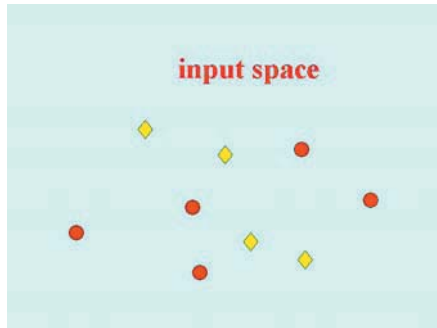


Fig.1 The nonlinear nonseparable data in input data

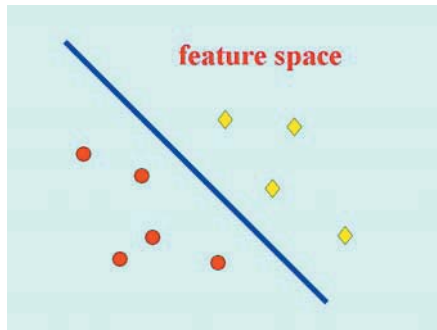


Fig.2 The linear separable data in feature data

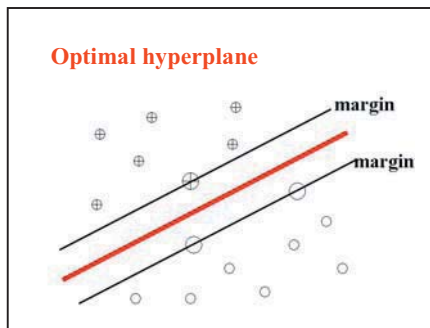


Fig.3 The optimal hyperplane and support vectors. The middle red line is optimal hyperplane, the examples placed on margin are support vectors.

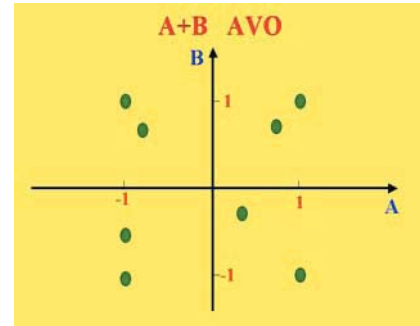


Fig4. All input data.

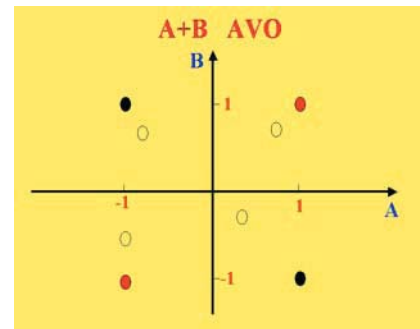


Fig5. The training data their classes are known.

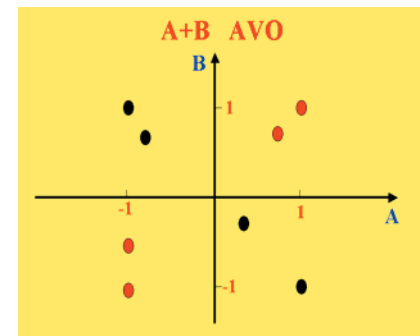


Fig 6. The final classified result.