Delineation of geologic facies with statistical learning theory

Daniel M. Tartakovsky and Brendt E. Wohlberg
Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA

Received 28 June 2004; revised 26 July 2004; accepted 10 August 2004; published 25 September 2004.

[1] Insufficient site parameterization remains a major stumbling block for efficient and reliable prediction of flow and transport in a subsurface environment. The lack of sufficient parameter data is usually dealt with by treating relevant parameters as random fields, which enables one to employ various geostatistical and stochastic tools. The major conceptual difficulty with these techniques is that they rely on the ergodicity hypothesis to interchange spatial and ensemble statistics. Instead of treating deterministic material properties as random, we introduce tools from machine learning to deal with the sparsity of data. To demonstrate the relevance and advantages of this approach, we apply one of these tools, the Support Vector Machine, to delineate geologic facies from hydraulic conductivity data. *INDEX TERMS:* 1829 Hydrology: Groundwater hydrology; 1869 Hydrology: Stochastic processes; 3210 Mathematical Geophysics: Modeling. **Citation:** Tartakovsky, D. M., and B. E. Wohlberg (2004), Delineation of geologic facies with statistical learning theory, *Geophys. Res. Lett.*, *31*, L18502, doi:10.1029/2004GL020864.

1. Introduction

[2] Our knowledge of the spatial distribution of the physical properties of geologic formations is often uncertain because of ubiquitous heterogeneity and the sparsity of data. Geostatistics has become an invaluable tool for estimating such properties at points in a computational domain where data are not available, as well as for quantifying the corresponding uncertainty. Geostatistical frameworks treat a formation's properties, such as hydraulic conductivity $K(\mathbf{x})$, as random fields that are characterized by multivariate probability density functions or, equivalently, by their joint ensemble moments. Thus, $K(\mathbf{x})$ is assumed to vary not only across the real space coordinates $\mathbf{x}$, but also in probability space (this variation may be represented by another coordinate ξ, which is usually suppressed to simplify notation). Whereas spatial moments of K are obtained by sampling $K(\mathbf{x})$ in real space (across $\mathbf{x}$), its ensemble moments are defined in terms of samples collected in probability space (across ξ). Since in reality only a single *realization* of a geologic site exists, it is necessary to invoke the ergodicity hypothesis in order to substitute the sample spatial statistics, which can be calculated, for the ensemble statistics, which are actually required. Ergodicity cannot be proved and requires a number of modeling assumptions [*Rubin*, 2003, section 2.7, and references therein].

[3] Machine learning provides an alternative to the geostatistical framework, allowing one to make predictions in the absence of sufficient data parameterization, without treating geologic parameters as random and, hence, without the need for the ergodicity assumptions. Intimately connected to the field of pattern recognition, machine learning refers to a family of computational algorithms for data analysis that are designed to automatically tune themselves in response to data. Neural networks [*Bishop*, 1995] are an example of such a class of algorithms that has found its way into hydrologic modeling. While versatile and efficient for many important applications, such as the delineation of geologic facies [*Moysey et al.*, 2003], the theory of neural networks remains to a large extent empirical in this context.

[4] Here we introduce another subset of the machine learning techniques — the Support Vector Machine (SVM) and its mathematical underpinning, the Statistical Learning Theory (SLT) of *Vapnik* [1998] — which is ideally suited for the problem of facies delineation in geologic formations. While similar to neural networks in its goals, the SVM is firmly grounded in rigorous mathematical analysis, which allows one not only to assess its performance but to bound the corresponding errors as well. Like other machine learning techniques, the SVM and SLT enable one to treat the subsurface environment and its parameters as deterministic. Uncertainty associated with insufficient data parameterization is then represented by treating sampling locations as a random subset of all possible measurement locations. Since such a formulation is ideally suited for hydrologic applications, the use of the SVM in the context of subsurface imaging deserves to be fully explored. This letter is the first step in this direction.

[5] We consider an idealized problem of identifying a boundary between two geologic facies from a sparsely sampled parameter K. We formulate the problem in Section 2, and provide a brief description of a geostatistical approach to its solution in Section 3. Section 4 introduces Support Vector Machines, which we use in Section 5 to estimate the boundary between the two heterogeneous facies in a simulated problem.

2. A Problem of Facies Delineation

[6] Consider the problem of reconstructing a boundary between two heterogeneous materials (geologic facies) from parameter data, say conductivity measurements $K_i = K(\mathbf{x}_i)$, collected at selected locations $\mathbf{x}_i = (x_i, y_i)^T$, where $i \in \{1,\ldots, N\}$. The first step to the facies delineation consists of analyzing a data histogram to assign to each data point a value of the indicator function,

$$I(\mathbf{x}_i) = \begin{cases} 1 & \mathbf{x}_i \in M_1 \\ 0 & \mathbf{x}_i \in M_2, \end{cases} \tag{1}$$

where M_1 and M_2 are the two facies.

0094-8276/04/2004GL020864$05.00

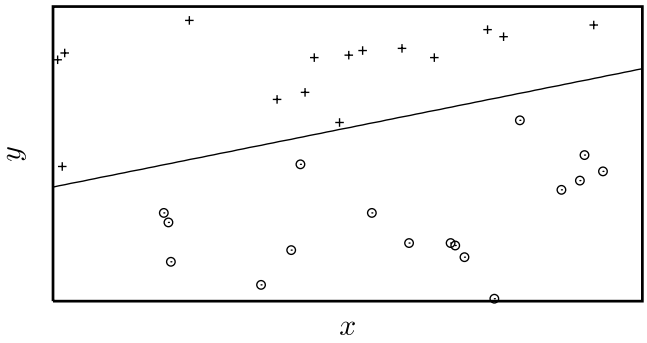

Figure 1. A schematic representation of the boundary between two heterogeneous facies M_1 and M_2 (located above and below the boundary, respectively) in a perfectly stratified geologic formation. The + and ⊙ signs indicate the locations where a parameter K is sampled.

[7] Let $\mathcal{I}(\mathbf{x}, \alpha)$ be an estimate of a "true" indicator field $I(\mathbf{x})$, whose adjustable parameters α are consistent with, and determined from, the available data $\{I(\mathbf{x}_i)\}_{i=1}^N$. One would like to construct an estimate, that is as close to the true field as possible, i.e., to minimize the difference between the two, $\|I - \mathcal{I}\|$. In general, both the indicator field I and the choice of sampling locations $\{\mathbf{x}_i\}_{i=1}^N$ can be modeled as random, and be described by a joint probability distribution $P(I, \mathbf{x})$ or, equivalently, a joint probability density function $p(I, \mathbf{x})$. Then the problem of obtaining the best estimate of the indicator field is equivalent to minimizing the functional

$$R = \int \| I - \mathcal{I} \| \, dP(I, \mathbf{x}) = \int \| I - \mathcal{I} \| \, p(I, \mathbf{x}) dI d\mathbf{x}. \quad (2)$$

Unfortunately, since in reality only a single geologic formation exists, there is no direct way to evaluate $P(I, \mathbf{x})$. Geostatistical and Statistical Learning techniques provide two alternatives for evaluating equation (2).

[8] To compare these alternative approaches, we construct a synthetic perfectly stratified geologic formation consisting of two highly contrasting heterogeneous layers M_1 and M_2, and then reconstruct the linear boundary based on the values of the indicator function I (as inferred from conductivity measurements) at a number of randomly selected sample points. Figure 1 illustrates the form of the boundary and an example set of sample points, designated by the + and ⊙ signs for locations where $I = 1$ and $I = 0$ respectively. Even though this setup is somewhat simplistic, it is ideal for demonstrating the main concepts of the SVM. We will also comment on generalizations that are necessary for applying the SVM to more general problems of facies delineation.

3. Geostatistical Approaches

[9] Geostatistical approaches use the L^2 norm in equation (2), and treat

1. the indicator function I is a random field, and
2. the choice of sampling locations $\{\mathbf{x}_i\}_{i=1}^N$ as deterministic.

[10] Then the problem of minimizing equation (2) reduces to the minimization of the indicator variance

$$\sigma_I^2 = \int (I - \mathcal{I})^2 dP(I) = \int (I - \mathcal{I})^2 p(I) dI. \quad (3)$$

To approximate $p(I)$, geostatistical approaches assume ergodicity, i.e., that the sample statistics of I, such as mean μ_I, variance σ_I^2, and correlation function ρ_I computed from spatially distributed data $\{I(\mathbf{x}_i)\}_{i=1}^N$ can be substituted for the ensemble statistics. Furthermore, it is necessary to assume that these sampling statistics are representative of the whole field.

4. Support Vector Machines

[11] The statistical learning theory of *Vapnik* [1998] often uses the L^1 norm in equation (2), and treats

1. the indicator function I as deterministic, and
2. the choice of sampling locations $\{\mathbf{x}_i\}_{i=1}^N$ as random.

[12] Then the problem of minimizing equation (2) reduces to the minimization of the *expected risk*

$$R_{\text{exp}} = \frac{1}{2} \int |I - \mathcal{I}| dP(\mathbf{x}) = \frac{1}{2} \int |I - \mathcal{I}| p(\mathbf{x}) d\mathbf{x}. \quad (4)$$

Rather than attempting to estimate probability distributions, such as $p(\mathbf{x})$, from spatially distributed data, statistical learning replaces the expected risk R_{exp} with the *empirical risk*

$$R_{\text{emp}} = \frac{1}{2N} \sum_{i=1}^{N} |I(\mathbf{x}_i) - \mathcal{I}(\mathbf{x}_i)|. \quad (5)$$

These two quantities are related by a probabilistic bound, $R_{\text{exp}} \leq R_{\text{emp}} + C$, where the function C depends on the *Vapnik - Chervonenkis* (*VC*) *dimension* and the number of data points N [*Burges*, 1998; *Cristianini and Shawe-Taylor*, 2000, chap. 4]. The VC dimension represents a measure of the complexity of the family of functions $\mathcal{I}$. Analysis of the tightness of this bound, which while providing a useful theoretical motivation for the SVM described below, is often too loose to be of much practical significance, is an active area of research in the field of statistical learning.

[13] The SVM is a relatively recent technique that has attracted a great deal of interest due to its excellent performance on a wide range of classification problems [*Cristianini and Shawe-Taylor*, 2000; *Burges*, 1998; *Gunn*, 1998]. The theoretical foundation of this technique is grounded in the fact that the maximal margin SVM, which we describe and implement below, provides a bound of the expected risk R_{exp} [*Cristianini and Shawe-Taylor*, 2000, chap. 6, remark 6.7; *Schölkopf and Smola*, 2002, chap. 7].

[14] The problem of locating the boundary between two geologic layers is ideally suited for illustration of basic ideas of the SVM. First, soft data or expert knowledge, e.g., geologic site characterization, is used to provide a rough guess about the shape of the boundary. Then hard data are used to find its optimal location. In this, the SVM is analogous to Bayesian statistical tools.

[15] For stratified geologic media shown in Figure 1, such boundaries can be assumed to be planes, or in two

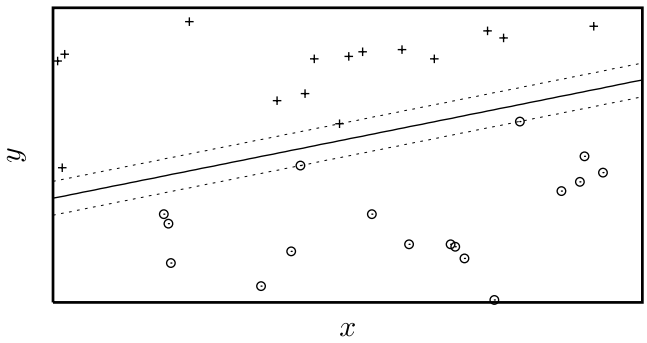

Figure 2. Maximum margin linear classifier for displayed samples. The solid line indicates the decision boundary (the boundary estimate), and the dotted lines denote the margin.

dimensions, straight lines (methods for dealing with more complicated geometries are outlined later in this section). Consider a boundary given by a straight line with equation

$$\mathbf{a} \cdot \mathbf{x} + b = 0. \tag{6}$$

Our goal is to determine the unknown coefficients $\mathbf{a} = (a_1, a_2)^T$ and b from the data set $\{I(\mathbf{x}_i)\}_{i=1}^N$. In machine learning, an algorithm for constructing such a boundary between samples from two classes is known as a linear classifier.

[16] A maximum margin linear classifier is illustrated in Figure 2 — the boundary estimate is indicated with the solid line, and the dotted lines indicate the extent of the *margin*, i.e., the region within which the boundary could be shifted orthogonally without misclassifying any of the data samples. If d_1 and d_2 designate the perpendicular distances from the estimated boundary (solid line) to the nearest data point(s) in materials M_1 and M_2, respectively, then the size of the margin (dotted lines) is $d = d_1 + d_2$, and the sample points determining the position of the margin are called the *support vectors*. Since the lines bounding the margin are parallel to the boundary equation (6), their normal is also $\mathbf{a}$. The maximal margin SVM determines the coefficients $\mathbf{a}$ and b in equation (6) by maximizing the size of this margin. While any choice of straight line that lies within the margin provides the same empirical risk R_{emp}, the maximum margin straight line is a principled choice for minimizing the expected risk R_{exp}.

[17] The maximal margin SVM is constructed as follows. Since the boundary equation (6) separates the two materials, all data points satisfy either $\mathbf{a} \cdot \mathbf{x}_i + b \geq +1$ or $\mathbf{a} \cdot \mathbf{x}_i + b \leq -1$. Mapping the indicator function $I(\mathbf{x})$ onto an indicator function $I^\star(\mathbf{x})$, so that $I^\star(\mathbf{x}) = -1$ whenever $I(\mathbf{x}) = 0$ and $I^\star(\mathbf{x}) = 1$ whenever $I(\mathbf{x}) = 1$, and denoting $I_i^\star = I^\star(\mathbf{x}_i)$ allows one to combine these to sets of inequalities into one set,

$$(\mathbf{a} \cdot \mathbf{x}_i + b) I_i^\star \geq 1 \qquad \text{for} \qquad i \in \{1, \ldots, N\}. \tag{7}$$

The inequalities (7) become equalities for the $\mathbf{x}_i$ that are support vectors. Since the distance from the coordinate origin to the line $\mathbf{a} \cdot \mathbf{x}_i + b = 1$ is $-(b+1)/\|\mathbf{a}\|$ and the distance to $\mathbf{a} \cdot \mathbf{x}_i + b = -1$ is $-(b-1)/\|\mathbf{a}\|$, the distance between these two lines, i.e., the margin d, is given by $d = 2/\|\mathbf{a}\|$, where $\|\mathbf{a}\| \equiv \sqrt{a_1^2 + a_2^2}$ is the Euclidean length of $\mathbf{a}$. Thus the SVM can be formulated as a problem of maximizing d (or, equivalently, minimizing $\|\mathbf{a}\|$) subject to the linear constraints (7). Introducing Lagrangian multipliers $\lambda_i \geq 0$ ($i \in \{1, \ldots, N\}$) gives the following minimization problem,

$$\min_{\mathbf{a},b,\lambda} \left\{ \frac{1}{2} \| \mathbf{a} \|^2 - \sum_{i=1}^{N} \lambda_i \left[(\mathbf{a} \cdot \mathbf{x} + b) I_i^\star - 1 \right] \right\}. \tag{8}$$

Its solution defines $\mathbf{a}$ and b and thus, in accordance with equation (6), the boundary between the two layers, located at the center of the margin.

[18] While the procedure described above provides the best estimate of the boundary location, it does not quantify the corresponding predictive uncertainty. When it is required, as is the case in random domain decompositions [*Winter and Tartakovsky*, 2000], one can employ probabilistic SVM approaches [*Platt*, 1999].

[19] The SVM approach described above is applicable to the delineation of geologic facies in perfectly stratified layered media. It can be generalized to account for a variety of more realistic settings, some of which are described below.

[20] When the properties of adjacent geologic facies cannot be well differentiated or data are too noisy, it is possible to mislabel the data points in the process of assigning the values of the indicator function. Then even for the case of perfectly stratified aquifer it might not be possible to fit a straight line to such data following the procedure outlined above. This complication is accounted for by replacing the optimization problem (8) with the so-called soft margin optimization [*Cristianini and Shawe-Taylor*, 2000, section 6.1.2]. It introduces slack variables $\xi_i \geq 0$ to replace the constraints (7) with the constraints $(\mathbf{a} \cdot \mathbf{x}_i + b) I_i^\star \geq 1 - \xi_i$ for $i \in \{1, \ldots, N\}$.

[21] In most practical problems, boundaries between geologic facies are significantly more complex than a straight line or a plane. The kernel technique [*Cristianini and Shawe-Taylor*, 2000, chap. 3; *Schölkopf and Smola*, 2002] allows the use of non-linear decision functions, based on Mercer kernels, while retaining the quadratic optimization of the linear SVM.

5. Simulation Results

[22] We construct a simple simulation of the facies identification problem as follows. First, we generate two uncorrelated Gaussian random fields on a 128×256 grid, one with mean of 1 and a standard deviation of 0.1, and the other with a mean of 3 and a standard deviation of 0.2. Correlated log-normal random fields are then derived from these fields by convolving with a lowpass filter and then exponentiating the result. The final random field is constructed by combining these two fields at a fixed linear boundary (as indicated in Figure 1), with the first component above the boundary, and the second below it. The marginal histogram of the resulting field is displayed in Figure 3.

[23] We compare the accuracy of the boundary reconstruction by means of the SVM with that obtained by a geostatistical (GS) approach due to *Ritzi et al.* [1994]. This approach consists of the following steps: First, we use

Kriging [*Deutsch and Journel*, 1992] to construct a map of the ensemble average of the indicator function $\langle I(\mathbf{x})\rangle$ from the data $\{I(\mathbf{x}_i)\}_{i=1}^{N}$. The ensemble mean $I(\mathbf{x})$ is the probability that a point $\mathbf{x}$ lies in Material 1, $\langle I(\mathbf{x})\rangle = P[\mathbf{x} \in M_1]$. Then we define a boundary between the two materials as an isoline $P[\mathbf{x} \in M_1] = c$, where c is a number of data points in Material 1 (or 2) relative to the total number of data points, after accounting for data clustering. This geostatistical approach to facies delineation assumes that the relative volumes occupied by the two materials obtained from a sample are representative of the whole field.

[24] The comparison of the performance of the GS and SVM approaches consisted of 20 trials for each of 6 sampling densities. For each trial, a fixed number of sampling points was selected according to a uniform distribution (an example is illustrated in Figure 1). These sampling points and the value of the random field at these points were then used within the GS approach, and as a training set for a SVM (A. Schwaighofer, unpublished data, 2002) to estimate the boundary between the two random fields. The results of this comparison are displayed in Figure 4 — the sampling density indicates the number of sample points as a fraction of the total grid points, and the error is the percentage of the grid misclassified according to the estimated boundaries. The SVM method outperforms the GS approach by a very significant margin. This is because the knowledge (often derived from soft data or expert opinion) that the geologic formation in Figure 1 is perfectly stratified—i.e., that the two heterogeneous layers are separated by a plane, or a straight line in two dimensions — has been explicitly incorporated into our SVM procedure.

[25] The relative performance of the two approaches under more general conditions, as well as of other geostatistical approaches, remains to be investigated. Nevertheless, we argue that the SVM method has great potential, especially when some information (e.g., geologic and other soft data) regarding the shape of the boundary is available.

6. Discussion and Conclusions

[26] The main goal of this letter is to introduce a general framework of the statistical learning theory to the fields of hydrology and subsurface imaging. One of the ubiquitous features of these applications is the sparsity of data, which results in parameter uncertainty, leading, in turn, to predictive uncertainty in subsurface modeling. Prevailing approaches to quantifying these uncertainties rely on geostatistical and stochastic methods, which treat the subsurface and its parameters as random. An attractive feature of the statistical learning theory is that it does not require ergodicity and other assumptions associated with such approaches.

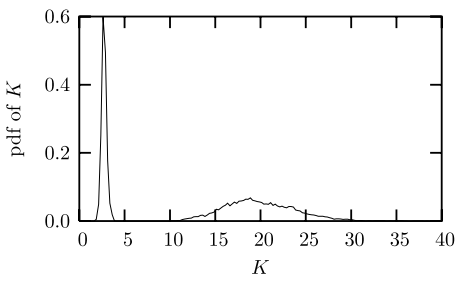

Figure 3. Histogram of the parameter K sampled at the locations denoted by the + and ⊙ signs in Figure 1.

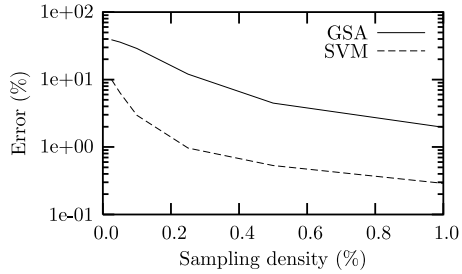

Figure 4. Comparison of boundary estimation errors using the geostatistical approach (GSA) of [*Ritzi et al.*, 1994] and SVM.

[27] It is important to note a key difference between geostatistics and statistical learning. While geostatistics provides a set of *interpolation* tools, SVMs are essentially *regression* techniques. Specifically, in the absence of a nugget effect and measurement errors, Kriging and other geostatistical estimation techniques produce ensemble estimates of a random field that match the data exactly. In contrast, SVMs provide estimates that minimize overall errors without matching the data exactly.

[28] Other useful features of statistical learning theory in general, and support vector machines in particular, include

[29] • The ease with which prior information, e.g., geologic site characterization, can be incorporated. This is accomplished by selecting the shapes of boundaries that are consistent with such prior knowledge.

[30] • The ability to estimate boundaries between geologic facies from poorly differentiable data. This is accomplished by assigning reliability weights to the indicator function, which account for the relative values of conductivity, and may be included in the SVM optimization.

[31] We will explore these and other issues related to the performance of statistical learning techniques for subsurface imaging in future studies.

[32] **Acknowledgments.** This research was performed under the auspices of the U.S. Department of Energy, under contract W-7405-ENG-36. This work was supported in part by the U.S. Department of Energy under the DOE/BES Program in the Applied Mathematical Sciences, Contract KC-07-01-01, and in part by the LDRD Program at Los Alamos National Laboratory. This work made use of shared facilities supported by SAHRA (Sustainability of semi-Arid Hydrology and Riparian Areas) under the STC Program of the National Science Foundation under agreement EAR-9876800.

References

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford Univ. Press, New York.

Burges, C. J. C. (1998), A tutorial on support vector machines for pattern recognition, *Data Min. Knowledge Discovery*, *2*, 121–167.

Cristianini, N., and J. Shawe-Taylor (2000), *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge Univ. Press, New York.

Deutsch, C. V., and A. G. Journel (1992), *Geostatistical Software Library and User's Guide*, Oxford Univ. Press, New York.

Gunn, S. R. (1998), Support vector machines for classification and regression, technical report, Sch. of Electr. and Comput. Sci. Univ. of Southampton, U. K.

Moysey, S., J. Caers, R. Knight, and R. M. Allen-King (2003), Stochastic estimation of facies using ground penetrating radar data, *Stoch. Environ. Res. Risk Assess.*, *17*, 306–318.

Platt, J. C. (1999), Probabilities for SV Machines, in *Advances in Large-Margin Classifiers*, edited by A. Smola et al., chap. 5, pp. 61–74, MIT Press, Cambridge, Mass.

Ritzi, R. W., D. F. Jayne, A. J. Zahradnik Jr., et al. (1994), Geostatistical modeling of heterogeneity in glaciofluvial, buried-valley aquifer, *Groundwater*, *32*, 666–674.

Rubin, Y. (2003), *Applied Stochastic Hydrogeology*, Oxford Univ. Press, New York.

Schölkopf, B., and A. J. Smola (2002), *Learning With Kernels*, MIT Press, Cambridge, Mass.

Vapnik, V. N. (1998), *Statistical Learning Theory*, John Wiley, Hoboken, N. J.

Winter, C. L., and D. M. Tartakovsky (2000), Mean flow in composite porous media, *Geophys. Res. Lett.*, *27*, 1759–1762.

D. M. Tartakovsky and B. E. Wohlberg, Theoretical Division, Group T-7, MS B284, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. (dmt@lanl.gov; brendt@t7.lanl.gov)