ModelCARE 2005

Fifth International Conference on

Calibration and Reliability in Groundwater Modelling

From Uncertainty to Decision Making

TNO

ICGW

IAHS AISH

Delft Cluster

UCG

UNESCO

iah aih

igwmc

NWO

Pre-published Proceedings

The Hague (Scheveningen)
The Netherlands
6 - 9 June 2005

Reconstruction of geologic facies with statistical learning theory

D. M. TARTAKOVSKY

Theoretical Division, Los Alamos National Laboratory, e-mail: dmt@lanl.gov; *and Department of Mechanical & Aerospace Engineering, University of California, San Diego, La Jolla, CA 92093, USA,* e-mail: dmt@ucsd.edu

B. E. WOHLBERG

Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

A. GUADAGNINI

Politecnico di Milano, D.I.I.A.R., Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Abstract A typical subsurface environment is heterogeneous, consists of multiple materials (geologic facies), and is often insufficiently characterised by data. The ability to delineate geologic facies and to estimate their properties from sparse data is essential for modeling physical and biochemical processes occurring in the subsurface. We study the problem of facies delineation in geologic formations by means of a subset of the machine learning techniques - the Support Vector Machine (SVM) and its mathematical underpinning, the Statistical Learning Theory. To demonstrate the potential of the SVM, we randomly generate a two-dimensional porous medium composed of two heterogeneous materials, and then reconstruct boundaries between these materials from a few data points. We analyse the accuracy of the SVM facies delineation, and compare the SVM performance with that of a geostatistical approach.

Key words Support Vector Machine, Machine Learning, geostatistics, geologic facies.

INTRODUCTION

Our knowledge of the spatial distribution of the physical properties of geologic formations is often uncertain because of ubiquitous heterogeneity and the scarcity and sparsity of information. 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 (e.g, hydraulic conductivity, K) as random processes that are characterized by multivariate probability density functions or, equivalently, by their joint ensemble moments. Whereas spatial moments of K are obtained by sampling K in physical space, its ensemble moments are defined in terms of samples collected in probability space. In reality only a single realisation of a geologic site exists. Therefore, 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 as input to a stochastic model of flow or contaminant transport. Ergodicity cannot be proved and requires a number of modeling assumptions.

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 (Morsey *et al.*, 2003), the theory of neural networks remains to a large extent empirical in this context. Tartakovsky and Wohlberg (2004) used a 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 is precisely the objective of this paper, where we demonstrate the potential of the SVM by means of a synthetic example.

PROBLEM SETTING

We consider a problem of reconstructing the spatial location of the boundary between two heterogeneous materials M_1 and M_2 from spatially distributed parameter data. The latter can consist of hydrodynamic dats (e.g., hydraulic conductivity), geophysical data (e.g., electric resistivity), and/or sedimentological data, collected at N selected locations $\mathbf{x}_i = (x_i, y_i)^{\mathrm{T}}$, where $i = 1, ..., N$ and $^{\mathrm{T}}$ is transpose. The first step in our facies delineation procedure is to analyse samples distributions with the goal of assigning an indicator function:

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

to each point where data are available. Let $\bar{\mathrm{I}}(\mathbf{x}, \alpha)$ be an estimate of a "true" indicator field $I(\mathbf{x})$, whose adjustable parameters α are consistent with, and determined from, the available data. Our objective is to construct an estimate that is as close to the true field as possible, i.e., to minimize the difference between the two, $||I-\bar{\mathrm{I}}||$.

SUPPORT VECTOR MACHINES

The theoretical foundation of SVM techniques relies on the definition of a bound for the expected risk, R_{exp}, which is provided by the maximum margin SVM (Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002). The simplest maximum margin SVM deals with linearly separable data collected from perfectly stratified geologic media, where different geologic facies are separated by planes (in three dimensions) or straight lines (in two dimensions). A description of the linear maximum SVM is provided by Tartakovsky and Wohlberg (2004) who applied the method to reconstruct a linear boundary between two materials in a two-dimensional domain on the basis of a few selected data points. In most practical problems, boundaries between geologic facies are significantly more complex than a straight line or a plane.

To account for this geometric complexity, one can generalize the linear maximum margin SVM by noting that data which cannot be separated by a straight line or plane in the two- or three-dimensional space of observation often become linearly separable (by a hyperplane) when projected onto another, usually higher-dimensional space. In this case, it can be shown (e.g., Wohlberg *et al.*, 2005) that the required indicator function is provided by the following decision function:

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{N} \gamma_i^* J_i \aleph(\mathbf{x},\mathbf{x}_i) + b^* \right) \tag{2}$$

Here $J(\mathbf{x}) = 2\, I(\mathbf{x}) - 1$ (so that $J = -1$ whenever $I = 0$ and $J = 0$ whenever $I = 1$), $J_i = J(\mathbf{x}_i)$, the kernel $\aleph(\mathbf{x}, \mathbf{x}') = F(\mathbf{x})\, F(\mathbf{x}')$, F being a mapping of the n-dimensional physical space onto a m-dimensional space (known as a feature space) in which linear SVM can be applied, γ_i^* ($i = 1, \ldots, N$) is defined as the solution of the dual optimisation problem

$$\underset{\gamma}{max} = \left\{ \sum_{i=1}^{N} \gamma_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_i \gamma_j J_i J_j \aleph(\mathbf{x}_i, \mathbf{x}_j) \right\} \tag{3}$$

and

$$b^* = -\frac{1}{2} \mathbf{a}^* (\mathbf{x}_+ + \mathbf{x}_-); \quad \mathbf{a}^* = \sum_{i=1}^{N} \gamma_i^* J_i F(\mathbf{x}_i) \tag{4}$$

where $\mathbf{x}_+$ and $\mathbf{x}_-$ denote arbitrary support vectors for which $J = 1$ and $J = -1$, respectively. We will consider the performance of the following set of expressions for the kernel $\aleph$ appearing in (2)

$$\aleph_{\mathrm{PLM}}(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + 1)^p; \qquad \aleph_{\mathrm{SIG}}(\mathbf{x}, \mathbf{x}') = \tanh(\rho\, \mathbf{x} \cdot \mathbf{x}' + \rho) \tag{5}$$

and the

$$\aleph_{\mathrm{ERB}}(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|/(2\sigma^2); \qquad \aleph_{\mathrm{GRB}}(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/(2\sigma^2) \tag{6}$$

where $\aleph_{\mathrm{PLM}}$, $\aleph_{\mathrm{SIG}}$, $\aleph_{\mathrm{ERB}}$, and $\aleph_{\mathrm{GRB}}$ denote a polynomial kernel of order p, a sigmoid kernel, an exponential radial basis kernel, and a Gaussian radial basis function kernel, respectively; here σ and ρ are fitting parameters.

SYNTHETIC EXAMPLE: RESULTS AND COMMENTS

To demonstrate the applicability of SVMs to subsurface imaging, and to elucidate its relative advantages with respect to a geostatistical approach, we reconstruct, from a few data points selected at random from a uniform distribution, the boundaries between two heterogeneous geologic facies in a synthetic porous medium shown in Figure 1a. This synthetic example was constructed by generating two autocorrelated, weakly stationary, and normally distributed processes, representing two distinct spatial distributions of log hydraulic conductivity $Y = \ln K$ with the ensemble means of -0.1 and 7.0. Both distributions have unit variance and Gaussian autocorrelation with unit correlation scale, and are mutually uncorrelated. None of these features is essential for the implementation of our approach. The composite porous medium in Figure 1a is constructed by setting an arbitrary shape of the internal boundary between the two materials and by assigning values of log-conductivity to cells in the domain. Assigning an indicator function with a threshold value of 4.0 to each element on the grid results in Figure 1b.

We use an SVM (Gunn, 1998) to reconstruct the boundary between the two geologic facies in Fig. 1b from a few (randomly) selected data points. Sampling densities ranging from 0.25% (9 data points) to 20% (720 data points) were considered. Figure 2 compares the performance of the SVMs whose kernels are given by the polynomial (PLM), exponential radial basis (ERB), Gaussian radial basis (GRB), and sigmoid (SIG) functions. For each sampling density, we randomly generated 20 realizations of the locations of data points and counted the number of elements on the grid that were misclassified by the SVMs. The error in Fig. 2 represents the average (over 20 realizations) number of misclassified elements. One can see that the kernels given by the exponential radial basis (ERB) function provide the best performance. Wohlberg *et al.* (2005) noted that, for sampling density exceeding 2%, the performance of the SVM is relatively insensitive to the choice of the fitting parameter σ in (6).

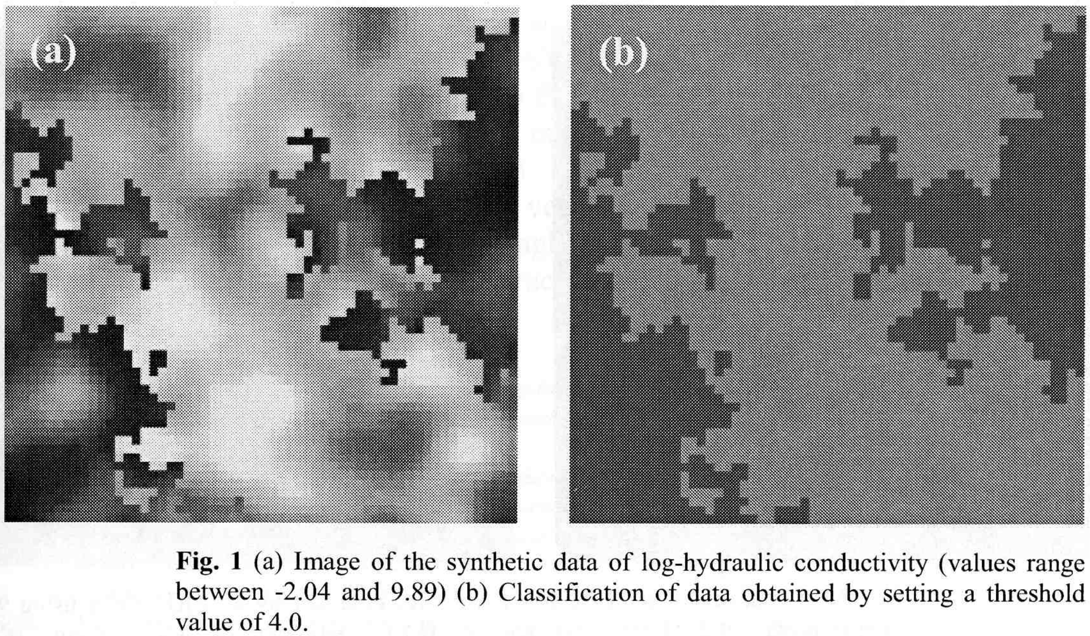

Fig. 1 (a) Image of the synthetic data of log-hydraulic conductivity (values range between -2.04 and 9.89) (b) Classification of data obtained by setting a threshold value of 4.0.

This finding is encouraging, since the optimal choice of σ is nontrivial. Figures 3a, b show the geologic facies reconstructed by an ERB SVM with $\sigma = 1.0$ from 9 and 180 sample points, respectively. The locations of sample points are indicated by the lighter shades.

The comparison of these reconstructions with the true field in Figure 1a, b shows that even very sparse sampling might be sufficient for the SVMs to capture general trends in the spatial arrangement of geologic facies. However, the performance of the

SVMs on such sparse data sets is highly sensitive to the locations of data points. As the sampling density increases, the SVMs capture finer features of the spatial arrangement of geologic facies, and their performance is less dependent on a sampling realization. Finally, we compare the accuracy of the facies reconstruction by means of the SVM with that obtained by the geostatistical approach (GSA) proposed by Ritzi et al. (1998). It is important to note that this and other geostatistical approaches to facies delineation assume that the relative volumes occupied by the two materials obtained from a sample are representative of the whole field.

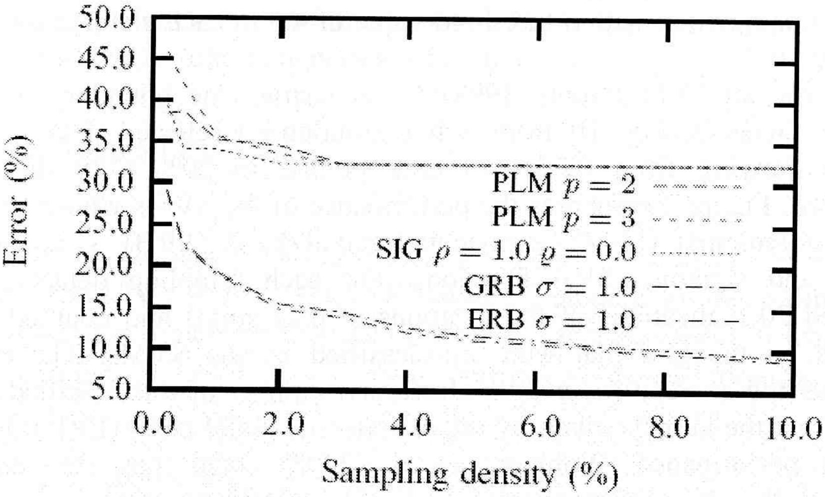

Fig. 2 Error rates corresponding to the SVMs with polynomial (PLM), exponential radial basis (ERB), and Gaussian radial basis (GRB) kernels.

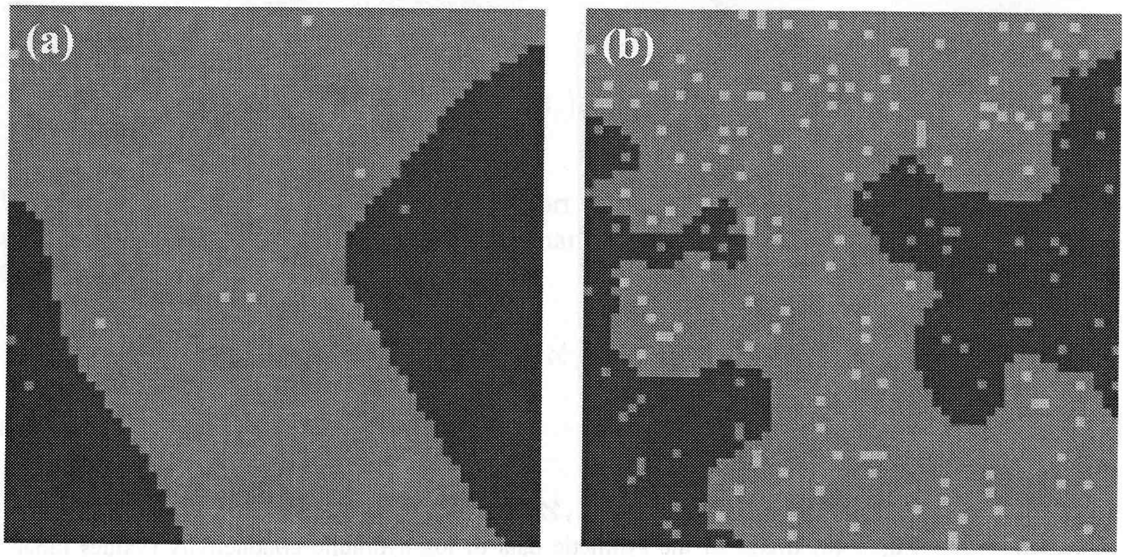

Fig. 3 (a) Classification of data in Figure 1a, obtained by an ERB SVM using 9 sample points (0.25% sampling density); (b) Classification of data in Figure 1a, obtained by an ERB SVM using 180 sample points (5% sampling density).

This assumption is usually difficult to validate a priori. Figure 4 compares the performance of the GSA and the SVM with the ERB kernel and σ = 1.0 both averaged over 20 trials for each of sampling densities. When the sampling density is large enough, both methods perform equally well, with the SVM being slightly more accurate than the GSA. Two factors, however, argue strongly in favor of SVMs. First, they perform relatively well even on highly sparse data sets (see the boundary reconstruction from 9 sampling points in Figure 4), on which GSA fails. Second, SVMs are highly automated, while GSAs require manual data analysis to construct spatial variograms.

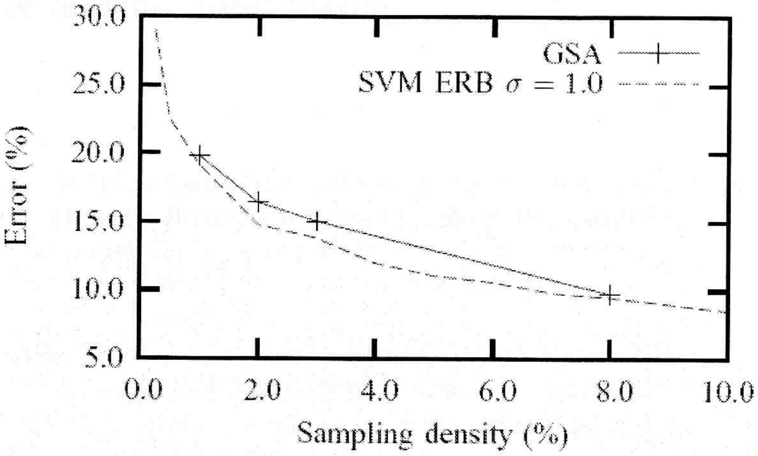

Fig. 4 Error rates corresponding to GSA and SVM approaches.

CONCLUSIONS

We explored the potential of the statistical learning theory in general, and support vector machines (SVMs) in particular, to delineate geologic facies from limited data. This was accomplished (i) by reconstructing, from a few data points, a synthetic randomly generated porous medium consisting of two heterogeneous materials; and (ii) by comparing the performance of the SVMs with that of the geostatistical approach (Ritzi *et al.*, 1994). Our analysis leads to the following major conclusions: (a) for any sampling densities the SVMs slightly outperforms the geostatistical approach in reconstructing the boundaries between two geologic facies, while significantly reducing the computational time; (b) For very low sampling densities, which make the inference of statistical parameters meaningless, the geostatistical approach fails, while the SVMs still offer a reasonable reconstruction of the boundaries.

REFERENCES

Bishop, C. M., (1995), *Neural networks for pattern recognition.* Oxford: Oxford University Press.

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

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

Deutsch, C. V., and A. G. Journel, (1992), *Geostatistical Software Library and User's Guide*. New York: Oxford University Press.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 Assessm.*, 17, 306 – 318.

Gunn, S. R., (1998), Support vector machines for classification and regression, University of Southampton, Southampton, U.K., Technical Report, School of Electronics and Computer Science. [Online]. Available: http://www.ecs.soton.ac.uk/~srg/publications/pdf/SVM.pdf.

Ritzi, R. W., Jayne, D. F., Zahradnik, A. J., Field, A. A., and Fogg, G. E., (1994). Geostatistical modeling of heterogeneity in glaciofluvial, buried-valley aquifers, *Ground Water* **32**, 666–674.

Schölkopf, B., and A. J. Smola, (2002), *Learning with Kernels*. Cambridge, MA, USA: The MIT Press.

Tartakovsky, D. M., and B. E. Wohlberg, (2004), Delineation of geologic facies with statistical learning theory, *Geophys. Res. Lett.*, **31**(18), L18502, 2004, doi:10.1029/2004GL020864.

Vapnik, V. N., (1998), *Statistical Learning Theory*. New York: John Wiley & Sons, Inc., 1998.

Wohlberg, B., Tartakovsky, D. M., and A. Guadagnini, (2005), Subsurface imaging with support vector machines, under review in *IEEE Trans. Geosci. Remote Sens.*