Characterization of Hydrogeologic Systems with Machine Learning Algorithms and Geostatistical Models

M. Kanevski[1], L. Bolshov[2], E. Savelieva[3], A. Pozdnukhov[4], V. Timonin[5], S. Chernov[6]

Abstract: The detailed description of hydrogeologic structures is of great importance for groundwater flow modeling. The main parameters used in the flow equations (permeability, conductivity and others) are strongly dependent on the type of hydrogeologic units in the formation (stone, limestone, sand, fine sand, clay, silt, etc). One of the first problems of hydrogeologic structure modeling is to detect the zones of presence and/or absence for the whole set of hydrogeologic units in the region under study. This task is a classification problem and is considered in the present report by using machine learning algorithms (artificial neural networks of different architectures and Support Vector Machines) and geostatistical models including stochastic simulations. Hanford site hydrogeological data are used as a real case study.

The task of hydrogeologic data characterization can be considered as a mixture of classification and regression/simulation problems. Scarceness of data, multiscale variability, and heterogeneity of underground media give rise to high uncertainty in site characterization. At present, geostatistical simulation models are widely applied to develop probabilistic models for media and are used as an integrated part of flow models. In the present report the main attention is paid to the so-called data driven models – artificial neural networks (Multilayer Perceptrons and Probabilistic Neural Networks) and Support Vector Machines – for the classification part of the entire problem (Haykin 1999). The results are compared with a geostatistical approach to classification including stochastic simulation. 2D binary (2 class) and multi-class tasks are considered.

Adaptive Methods for Spatial Data Classification. *Probabilistic Neural Networks (PNN)* are supervised neural networks widely used for classification problems. The main idea is to construct the classifier using a Bayesian optimal or maximum a posteriori decision rule:

$$C(x) = \{c_1, c_2, ..., c_K\} = \underset{c_i}{argmax}\, P(c_i)p(x \mid c_i) \qquad i = 1,2,...,K\,, \qquad (1)$$

[1] Professor, Institute of Nuclear Safety (IBRAE) of Russian Academy of Sciences, B. Tulskaya 52, 113191 Moscow, Russia. m_kanevski@ibrae.ac.ru (corresponding author)

[2] Professor, Corresponding member of Russian Academy of Sciences, Institute of Nuclear Safety (IBRAE), Russian Academy of Sciences, B. Tulskaya 52, 113191 Moscow, Russia

[3] Senior Researcher, Institute of Nuclear Safety (IBRAE), Russian Academy of Sciences

[4] Researcher, Institute of Nuclear Safety (IBRAE), Russian Academy of Sciences

[5] Researcher, Institute of Nuclear Safety (IBRAE), Russian Academy of Sciences

[6] Senior Researcher, Institute of Nuclear Safety (IBRAE), Russian Academy of Sciences

where $P(c_i)$ is a prior probability of class c_i and $p(\mathbf{x}/c_i)$ is the density of class conditional distribution for all $\boldsymbol{x}$. $P(c_1)=P(c_2)$ when there is no additional information. The relative number of members of a class can be used as a priori information. The density $p(\mathbf{x}/c_i)$ is estimated by a non-parametric kernel method with a classical 2-dimensional Gaussian kernel function $W(X)$:

$$p.d.f(X)=\frac{1}{N}\sum_{n=1}^{N}W(X-X_n)=\frac{1}{2\pi\sigma^2 N}\sum_{n=1}^{N}e^{\frac{-\|X-X_n\|^2}{2\sigma^2}} \quad ,$$

where N is a number of samples in the training data set and $\sigma>0$ is the scaling parameter (kernel bandwidth) adjusted during the training session. The training procedure is to find the optimal (minimum mean square error criterion) kernel bandwidth. Cross-validation and jackknife procedures were used in the present study.

Multilayer Perceptron (MLP) is the most widely used type of feedforward artificial neural network. Each neuron of a previous layer is connected through a weight w_i to each neuron of the following layer. The learning procedure for MLP classifier deals with minimization of empirical risk function with respect to the weight parameters W:

$$R(W)=\sum_{i=1}^{N}\left[s(g(X_i,W))-y_i\right]^2 ,$$

where $s(t)$ is a logistic sigmoid, $g(t)$ is the parameterization of current MLP output, $\boldsymbol{X}$ is a training data set of size N. For MLP classifier the data set $\boldsymbol{X}$ presents the categorical data – classes. MLP is an efficient and universal nonlinear tool for pattern classification and regression problems.

Support Vector Machines (SVM) approach is based on a Statistical Learning Theory. Statistical Learning Theory is a general mathematical framework for estimating dependencies from empirical finite data sets. The basic idea of SVMs is to determine a classifier or regression machine which minimizes the empirical risk (the training set error) and the confidence interval that corresponds to the generalization or validation error. The SVM provides non-linear classification (or regression) by mapping the input space into high dimensional feature spaces where special types of hyper-planes with maximal margins (giving rise to good generalizations – low errors on validation data sets) are constructed. SVMs are focusing on the marginal data (support vectors - SV) and not on statistics such as means and variances. Only data points close to the classification decision boundaries are important for the solution of the problem. Essentially the method is non-linear, robust and does not depend on the dimension of input space. Recently SVM have been successfully applied in different fields of data analysis and modeling, including spatial environmental data (Kanevski et al 2000).

Geostatistical classifications were performed using an indicator kriging and stochastic simulations. Geostatistical methods use the spatial correlation structure during analysis and modeling. Spatial correlation structure is described by the indicator variograms: $\gamma(\boldsymbol{h})=1/2Var(I(x)-I(x+\boldsymbol{h}))$, where $\boldsymbol{h}$ is a vector separating pairs of points. Experimental variograms are modeled by fitting to theoretical models. Indicator kriging is a best linear unbiased predictor of indicators. In order to estimate the uncertainty of classification, a geostatistical simulations based on indicators – sequential indicator simulations, and simulated annealing algorithms were applied. Simulations provide many equally probable realizations with the same first and second (geo)statistical moments. In the case of simulated annealing the results of machine learning and geostatistical indicator classifications were used as the initial images as well.

The probabilistic nature of classification methods allows to estimate the uncertainty of a classification result. The uncertainty can be signified by probability mapping – presenting for each location the probability to belong to a class. Defining a classification threshold – the minimum probability value for which the location is ascribed to a certain class – we can obtain zones of uncertainty. They can be defined as zones where the values of probability to belong to a class are lower than the threshold for all classes. Using the simulation approach we obtain the uncertainties of classification (fluctuation of the classification result on the borders).

Hanford site hydrogeologic data were considered as a real case study. The Hanford site hydrogeologic structure presents a layered structure of unconsolidated to semi-consolidated sediment with a zonal distribution of different hydrogeologic units. Unit 4 (U4 - Ringold upper mud formation) is used for the current presentation. U4 was selected as it is an important layer for groundwater modeling. The low permeability of the mud formation significantly changes the direction of water flow. U4 occurs in the two major areas at the Hanford site separated by an area where it was eroded. The classification task was to find the erosional borders of the areas where U4 occurs. The uncertainty of the obtained borders is of a great importance for flow modeling.

The work was partly supported by collaboration under Memorandum of Understanding between the U.S. Department of Energy and the Russian Academy of Sciences, Implementing Arrangement 2, Annex B (“Uncertainty Assessment and Advanced Modeling of Mathematical Geology in Development of Inverse Flow and Transport Models”) and INTAS grants 99-00099, 97-31726.

References

Haykin S (1999) *Neural Networks. A comprehensive foundation*. Prentice Hall, Upper Saddle River, New Jersey, 842 p.

Kanevski M., Pozdnukhov A., Canu S., Maignan M. (2000) “Advanced Spatial Data Analysis and Modeling with Support Vector Machines”. IDIAP Research Reports, RR-00-031. www.idiap.ch, 18 p.