Machine Learning Algorithmes for Spatial Data Analysis

GILARDI Nicolas ; November 25, 2002

Supervisor:Prof. Michel Maignan, Institut de Minéralogie et Géochimie

Environmental survey needs maps as decision making tools. These maps can be separated into two main categories: the prediction maps, where the problem is to estimate the measurement value of an unsampled
location with a minimum error, and the risk maps, where the problem is, for example, to give the confidence interval of a prediction (confidence maps), or to estimate, for an unsampled area, the probability of exceeding a critical value (probability maps). These maps are built using data sets of sample measurements. The number of these measurements is often small (a few hundreds) for various technical reasons, and the sampling locations might be inhomogeneous. This is producing various difficulties when analyzing the data set (noise, distorted distribution, non-stationarity, etc.), which make the analysis of spatial data a very difficult problem.
Geostatistics was designed to deal with this kind of problems. It is based on the modeling of the spatial correlation of data which it used to construct a minimum variance interpolator called Kriging. Many algorithms can then be constructed to solve different problems, like Ordinary Kriging (OK) for prediction, Indicator Kriging (IK) for probability maps and Sequential Gaussian Simulations (SGS) for any kind of risk maps.
The problem of Geostatistics is that it makes a lot of hypotheses on the distribution of the data set. Especially, it is supposed to be stationary, and each value is supposed to be drawn from a Normal distribution. When a data set does not respect these constraints, geostatisticians need to proceed to various data transformations which are reducing the hope of making a good analysis of the studied phenomenon. In addition, this cause the efficiency of geostatistical models to be very dependent of user's experience.
An interesting alternative to solve spatial data problems is to use Machine Learning (ML) algorithms. These methods are modeling the relationship between input information (for example sample location) and output
information (for example the measurement value) of the available data set. The parameters of the model are chosen automatically in order to optimize a quality criterion (mean squared error, likelihood, etc.). This phase is called the training procedure.
ML algorithms are interesting because they are not making any supposition about data distribution. Moreover, as the parameters are automatically fitted, they are less dependent of user's experience. However, they need a large number of examples to give reliable results and are not specifically adapted to spatial data analysis.
In this thesis, various ML algorithms are used to solve spatial data problems. Different implementations of Support Vector Regression (SVR) and Multi Layer Perceptron (MLP) are used to build prediction maps. The results are compared to various methods used during the Spatial Interpolation Comparison 1997 (SIC97), among which half are using geostatistical methods. Ridge Regression Confidence Machines (RRCM) are used to build confidence maps and compared to SGS. Support Vector Machines (SVM) and Conditional Gaussian Mixture
Models (CGMM) are used to draw probability maps and compared respectively to IK and SGS.

The conclusion of these experiments is first that ML algorithms proved to be able to manage spatial data sets. In addition, they performed generally better than geostatistical methods for drawing risk maps. This is mainly due to the lack of hypotheses on data distribution which allows ML algorithms to fit directly the real data distribution, while SGS needs a Gaussian transformation of data which might destroy some aspects of data distribution like multi-modality. On the other hand, ML algorithms did not convince for what concern the construction of prediction maps. Even if their results were usually as good as other methods, they appeared to be more difficult to handle than simple Geostatistics. The reason is first that Geostatistics has been designed specifically for prediction mapping, and that the focus made on spatial correlation is of a great advantage in such problems. In addition, when the studied data set is very complex, expert knowledge can be used more easily inside a geostatistical model (tuning by hand the spatial correlation model) than in a ML algorithm.
ML algorithms and Geostatistics have complementary properties for constructing efficient spatial data analysis tools. Some hybrid methods have been proposed, which proved to be efficient. However, a « real » learning algorithm which would be able to model and use the spatial correlation of data, might be an even more powerful approach.