ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

GeoKernels: Modeling of Spatial Data on GeoManifolds

Alexei Pozdnoukhov[1] and Mikhail Kanevski[1] *

1- Institute of Geomatics and Analysis of Risk - University of Lausanne
IGAR, UNIL, Amphipole, CH-1015, Lausanne - Switzerland

Abstract. This paper presents a review of methodology for semi-supervised modeling with kernel methods, when the manifold assumption is guaranteed to be satisfied. It concerns environmental data modeling on natural manifolds, such as complex topographies of the mountainous regions, where environmental processes are highly influenced by the relief. These relations, possibly regionalized and non-linear, can be modeled from data with machine learning using the digital elevation models in semi-supervised kernel methods. The range of the tools and methodological issues discussed in the study includes feature selection and semi-supervised Support Vector algorithms. The real case study devoted to data-driven modeling of meteorological fields illustrates the discussed approach.

1 Introduction

The problem of using unlabeled data is of increasing attention in Machine Learning. By unlabeled data, we mean those data samples which consist of the input values only, while the desired output value is unknown. The methods making use jointly of labeled and unlabeled data are called *semi-supervised.* When predictions have to be made to given unlabeled locations only, this particular situation is called *transductive learning*. Most real-life learning problems are actually semi-supervised, which gives rise to the developments of large-scale semi-supervised methods nowadays.

The information one obtains from the unlabeled part of the dataset can be of different nature. A common approach is to consider the *manifold assumption*. This implies that data actually belong to some lower dimensional manifold in high dimensional input space. A large body of literature is devoted to the exploration of such an approach; see [1] and references therein.

The ratio between the amount of available labeled and unlabeled data will always be in favor of the last. The unlimited amounts of unlabelled data may be available. Hence, the methodology of semi-supervised data-driven modeling is as important as the algorithms themselves, since the described situation influence every modeling stage starting from data splitting up to results comparison.

In this paper, the highlighted problems are discussed, mainly referring to a real-life problem of geospatial data modeling. In this field, the semi-supervised learning finds elegant applications. Moreover, the manifold assumption often can be

* The research was supported by Swiss National Science Foundation project “GeoKernels: Kernel-Based Methods for Geo- and Environmental Sciences” (project N° 200021-113944).

considered to be guaranteed and the amount of data to model a manifold is almost exhaustive.

2 Predictive Learning From GeoSpatial Data

Automatic environmental monitoring networks are becoming one of the main and easily accessible sources of information related to environmental and climatic research, including natural hazard risk assessment and renewable resources estimation. Following the recent advances in wireless sensor networks technologies, the spatially distributed information can be gathered at more local scales. Data-driven modeling and assimilation becomes one of the most important aspects of environmental and climate modeling nowadays [2, 3]. However, the data come noisy, contain outliers, missing values and gaps. Given that computational resources required to run a physical model in a region of complex topography or urban zone in the real time are hard to provide, the data-driven methods become more and more important for the processing and modeling of these data. Particularly, the task of topo-climatic mapping - spatial predictions of climatic and meteorological variables - can be efficiently approached in data-driven manner.

This paper presents an integrated data-driven methodology of semi-supervised data modeling on the manifolds for topo-climatic mapping with modern machine learning methods. It is introduced using the real case studies, providing the description of the modeling steps, the associated problems, approaches, empirical experience and discussions.

2.1 Monitoring Networks

Many machine learning modelers ignore the origin of data, assuming them to be i.i.d. However, it is essential to keep in mind how the data were sampled, or, in other words, how the monitoring network was organized. The purpose of a monitoring network is to detect and to understand/model spatio-temporal phenomena (natural or artificial) via the observations at a finite number of points in space.

Generally, the data coming from environmental monitoring networks are not i.i.d. There are numerous reasons for that such as clustered (measurements taken in clusters due to natural geomorphological bounds or administrative borders, for example) or preferential sampling (the density of measurements in some regions is higher). The effective dimension of the clustered network is lower then its spatial resolution and can be characterized by dimensional resolution, i.e. their ability to detect D-dimensional phenomena in a D-dimensional Euclidean space. Clustered monitoring networks cause the biased estimates of statistical moments and model hyper-parameters. Even simple mean estimation is not valid if the monitoring network is clustered. Clustering also affects the splitting of data into training, testing and validation subsets. The strategies to overcome these difficulties can be found in [2].

The particularly important feature of spatial data modeling is that even non complex and non clustered monitoring networks often induce low-dimensional manifolds in the registered high-dimensional data.

2.2 Learning on GeoManifolds

The environmental processes captured by modern monitoring networks can not be generally explained in 2D spatial coordinates. The simplest example is the modeling of mean air temperature in mountains. Generally, average temperature linearly decreases with altitude thus can be easily modeled in 3D space of coordinates-elevation. But actually, the domain of this process is the 2D manifold (mountainous terrain) embedded into 3D space.

With increasing dimensionality of the input space by means of related information such as digital elevation models, satellite/aerial remote sensing images, Geographical Information Systems, a large amount of *dependent* inputs is being added. This interdependence induce the lower-dimensional manifolds in the original input space, similarly to the temperature-elevation example above.

Semi-supervised manifold learning becomes particularly useful in approaching environmental modeling in data-driven manner. In this application area, the *manifold assumption* is actually guaranteed to be satisfied.

3 Semi-Supervised Kernel Methods

A semi-positive definite function *K(x,x')* which satisfies Mercer conditions is called a kernel. This implies that it corresponds to a dot product in some space (Reproducing Kernel Hilbert Space, RKHS), sometimes referred to as a feature space. Generally, given a (linear) algorithm, which includes data samples in the form of dot products only, one can obtain a (non-linear) kernel version of it by substituting the dot products with kernel functions [4]. The general model is a kernel expansion:

$$f(x,\alpha) = \sum_{i=1}^{N} \alpha_i K(x, x_i) + b \qquad (1)$$

The choice of the kernel function is an open issue. Using some typical kernels like Gaussian RBF, one takes into account some knowledge like distance-based similarity of the samples.

The non-parametric data-dependent kernels which reflect the inner geometry of the data are of particular interest for the manifold learning. The general idea of predictive manifold learning with kernels is to incorporate the geometry induced by the manifold. For distance-based kernels, the simplest approach is to use geodesic distances induced by the manifolds, which can be computed using Dijkstra algorithm [5], given enough unlabelled data.

Another way is to enforce the model to be smooth at the manifold by using a regularization properties of the kernel [6]. Given the original kernel function *K(x,x')*, not necessarily distance-based, and a set of data samples (both labeled and unlabelled) the modified kernel is given by

$$K(x,x') = K(x,x') - k_x^T (\mathrm{I}{+}\Omega\mathrm{K})^{-1} \Omega k_{x'}, \qquad (2)$$

where K is the complete kernel matrix of *K*(.,.), *k* is its column and I is identity matrix. The choice of matrix Ω implements the smoothness assumption with respect to the geometrical structure of the data. It can be achieved by taking Ω=γL, L being the graph Laplacian of the data and γ is a regularization parameter. This kernels were implemented for regression estimation in [7].

For N labeled and M unlabelled data samples, and V is the number of vertices in the manifold modeling graph, the first method can be implemented in $O(N+M+Vlog(V))$ or $O(N+M+V^2))$ computations, while the second one includes matrix inversion hence requires at least $O((N+M)^2)$. Though, the latter computation can potentially be speed up since the low number of vertices leads to a sparse structure of the graph Laplacian L.

3.1 Feature Selection

Large number of input variables brings both information and noise. It is important that feature selection in the semi-supervised learning scheme is actually a manifold construction: while adding or eliminating the features, the geometry of the input space and the manifold is influenced. Currently, there are no exhaustive studies concerning this issue.

The method which will be used in the case study below is a fully supervised recursive feature elimination (originated as pruning in neural networks) developed for Support Vector Machine [8].

3.2 Model Selection

In the semi-supervised setting, model selection is a non-trivial task. Both labeled and unlabelled data influence the model through a number of corresponding hyper-parameters. Concerning the labeled data, the usual cross-validation may be applied. However, due to the nature of the setting, every unlabelled sample acts as a user-defined parameter of the algorithm. This issue is not yet explored.

In the application below, the unlabelled dataset is fixed and just the hyper-parameters (the number of neighbors in the manifold-modeling graph and a smoothness parameter) are tuned by 10-fold cross-validation.

4 Case Study: Topo-Climatic Mapping

The task of topo-climatic mapping - spatial predictions of climatic and meteorological variables at local scales - is important for the use in decision support systems and natural hazard risk assessment.

The data used in this study comes from the Swiss Federal Office for meteorology and climatology (www.meteoswiss.ch). Generally, 107 labelled samples (the number of weather monitoring stations) and 650000 unlabelled samples (digital elevation model) were available for the study. The data are illustrated in Fig. 1.

Monitoring network is rather homogeneous, however the sampling is done at preferably low elevations and flat terrain. Two problems are considered below: temperature inversion mapping (short-time phenomena with highly nonlinear relations between relief and temperature) and mean annual wind speed mapping in complex mountain relief of Swiss Alps. The conventional tools such as geostatistics [2, 3] can not produce a reasonable result in this study since the variogram (spatial covariance function) can not be modelled.

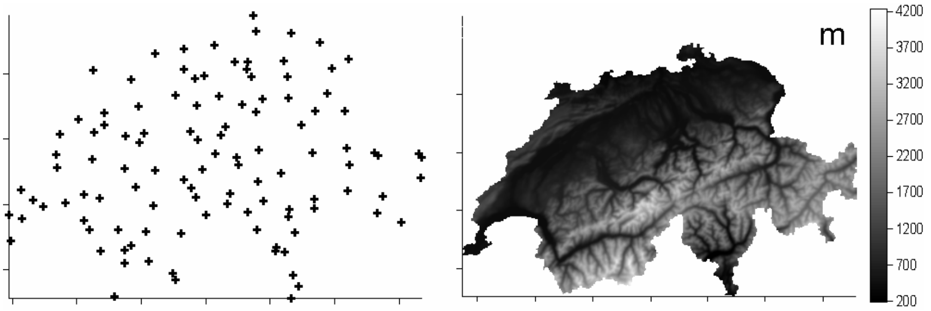

Fig. 1 Left: X and Y input features (spatial coordinates) of the training dataset coming from 107 monitoring stations (labeled data). Right: digital elevation model with 650000 samples which is used to create the manifold model for semi-supervised learning (unlabeled data).

4.1 **Modeling Scheme**

The complex topographies of the mountainous regions highly influence all the environmental processes. These relations, possibly regionalized and non-linear, can be modeled from data using the Digital Elevation Models. Numerous features such as first and second derivative forms (slope, aspect, curvatures, difference of elevation smoothed at different scales) were computed from high-resolution DEM. Recursive feature elimination was used to select the set of relevant features for a particular prediction problem, resulting in 7 features for temperature and 12 for wind modeling. Modeling was done with Support Vector Regression using leave-one-out cross-validation over 107 training samples to tune the parameters. Gaussian RBF kernel with geodesic distances on the manifold was used ("Geodesic SVR"). To speed-up the computations, the distances were computed on the low-resolution DEM. The low-resolution DEM was also used to apply the semi-supervised kernel modification according to (2). This model is denoted as "Manifold SVR" below.

Method	CV RMSE, wind speed, m/s	CV RMSE, temperature inversion, ^{o}C
SVR	3.2	3.4
3D SVR	1.9	2.2
Geodesic SVR	0.82	1.8
Manifold SVR	0.85	1.7

Table 1: Cross-validation RMSE of the compared methods.

The cross validation RMSE of the methods are shown in Table 1, and the prediction mapping results (the model predictions for temperatures and wind speed in space) in Fig. 2.

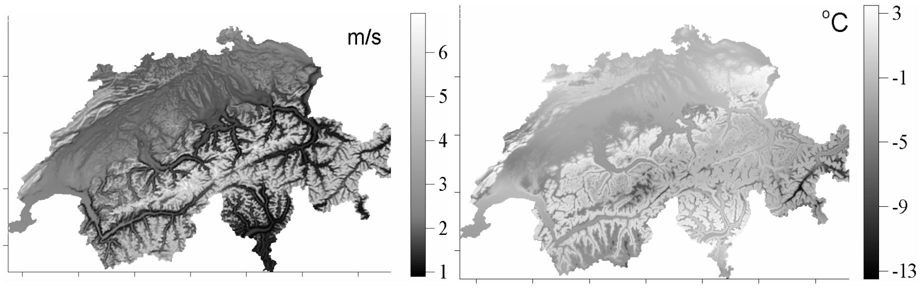

Fig. 2: Mean annual wind speed prediction (left) and temperature inversion modeling (right) with the semi-supervised SVR model on geomanifold.

5 Conclusions

The paper presented a methodology of data modeling with semi-supervised kernel methods in a unique situation when the manifold assumption is guaranteed to be satisfied. It concerned the application in the domain of spatial environmental data modeling. The low-dimensional manifolds appear in data when integrating the related spatially distributed information into the model.

The problems related to complex monitoring network structures were discussed. It was shown how semi-supervised kernel methods can be applied in this domain, starting from feature selection, model selection up to visualization of the results. Real case study concerning the topo-climatic mapping was considered. The described methodology of data-driven modeling of complex environmental processes with machine learning methods enabled to improve the modeling considerably. The computational issues concerned with large amount of unlabelled data, resolved here by using a representative subset of the latter, still have to be investigated.

References

[1] Chapelle O., Scholkopf B., and Zien A. (Eds.) Semi-Supervised Learning. (2006). MIT Press. 498 p.

[2] Kanevski M. (Ed.) (2008). Advanced Mapping of Environmental Data: Geostatistics, Machine Learning and Bayes-ian Maximum Entropy. ISTE, London, 2008.

[3] Kanevski M and Maignan M. (2004). Analysis and Modelling of Spatial Environemental Data. EPFL Press, 288 pp.

[4] Scholkopf B. and Smola A. (2002). Learning with kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press. 644 p.

[5] Dijkstra E. W. A note on two problems in connexion with graphs. In Numerische Mathematik, 1 (1959), S. 269–271.

[6] Sindhwani V., Niyogi P., Belkin M. Beyond the Point Cloud: from Transductive to Semi-supervised Learning. In Proc. of ICML'05, Bonn, Germany.

[7] Pozdnoukhov A., Bengio S., Semi-Supervised Kernel Methods for Regression Estimation. In proc. of ICASSP'06. Toulouse, France, 2006.

[8] Guyon I., Weston J., Barnhill S., Vapnik V., Gene selection for cancer classification using Support Vector Machines. Machine Learning, 46, pp. 389-422, 2002.