Multitask Learning of Environmental Spatial Data

Mikhail Kanevski
Institute of Geomatics and Analysis of Risk, University of Lausanne
(Mikhail.Kanevski@unil.ch)

Abstract: The present research deals with an application of artificial neural networks for multitask learning from spatial environmental data. The real case study (sediments contamination of Geneva Lake) consists of 8 pollutants. There are different relationships between these variables, from linear correlations to strong nonlinear dependencies. The main idea is to construct a subsets of pollutants which can be efficiently modeled together within the multitask framework. The proposed two-step approach is based on: 1) the criterion of nonlinear predictability of each variable “k” by analyzing all possible models composed from the rest of the variables by using a General Regression Neural Network (GRNN) as a model; 2) a multitask learning of the best model using multilayer perceptron and spatial predictions. The results of the study are analyzed using both machine learning and geostatistical tools.

Keywords: Machine learning algorithms; multitask learning; environmental multivariate data; geostatistics

1 INTRODUCTION

Multitask learning (MTL) is an important recent trend in machine learning research. Loosely speaking the idea is that learning together several related (to be defined) tasks is more efficient, easier and gives rise to more parsimonious models. In a nonlinear modelling framework, e.g. multilayer perceptron, it means that several tasks (outputs) are learned together by sharing neurons in a hidden layer(s). One of the first detailed researches on this problem was published by Caruana (1997). If the tasks, are not related, for example, output variables are independent, the resulting network is equivalent to two non-interacting ones. There are many variants of MTL and recently the concept was generalized to kernel machines, Evgeni et al. (2005), and to Gaussian Processes, Bonilla and Williams (2009).

The important problem of “relatedness” of the tasks to be learned can be considered in a different way: from quite simple, like the tasks sharing the same covariance structures to more advanced approach by using SVM-like optimization procedures. The simple difference between single task and multitask learning is illustrated in Figure 1.

In the present study we apply multitask learning concept to model and to predict the spatial pollution of sediments in Geneva Lake. There are eight output (dependent) variables (pollutants) measured at 200 points in the Lake. The general problem is the following: in order to make a better prediction of the variable “k”, which other variables should be added/used to improve the quality of the analysis (better generalization and smaller uncertainty)?
We accept the hypothesis that the input space (independent variables) is fixed for all tasks, but not necessarily known a priori. Input space can include not only the geographical coordinates, but also bathymetry, distance to pollution sources, currents etc., which are potentially important for the predictions. It can be quite complex and high dimensional, like it was observed for wind speed modelling in Foresti et al. (2011). In a more general setting the input spaces can be different.

Let us have a look at the output space of the problem. In order to select a group of related variables there are many approaches proposed in machine learning: linear measures like correlations, nonlinear ones like mutual information, clustering of variables according to some similarity criteria, for example, by using self-organizing maps, see Kanevski et al. (2009). The models like PCA can be used to detect corresponding linear relationships between variables. In general, the relationships between variables are non-linear, see Figure 2, where the scatterplot matrix of real data is given. Therefore, the nonlinear modelling algorithm – GRNN, is applied in the present study. GRNN is a nonparametric, nonlinear and efficient tool for multidimensional regression problems. Let us remind, that in case of classification problems an analogue of GRNN for categorical data – Probabilistic Neural Networks can be used within the framework of the proposed methodology. The description of PNN and their applications can be found in Kanevski et al. (2009).

In the present research in order to select the most “related” (in the sense of MTL) group of variables a criterion based on the predictability of output variable of interest “k” by using all combinations of other outputs is considered. For example, in order to select the group for Ni pollutant, GRNN considers as an input space all possible combinations of other output variables in order to predict one GRNN output - Ni. The number of possible models is \(2^{(n-1)} - 1\), where \(n\) is the dimension of the input space of GRNN: the number of GRNN dependent variables constructed by using other pollutants. For \(n=8\) it equals to 127 models. The combination giving the minimum of cross-validation error is then chosen as the best. The procedure does not depend explicitly on the structure of the original input space (independent variables) because at this step only the dependent variables (pollutants) are used. Thus, the “relatedness” of the tasks is quantified via the criterion of nonlinear predictability. From another point of view, it looks like a problem of feature selection – the selection of the group of the most relevant features, see, for example, Kanevski and Timonin (2010), Foresti et al. (2011).

![Figure 1. Principles of single task (left) and multitask – right (5 tasks) learning using multilayer perceptrons](image-url)
In the following sections the methodology and corresponding case study along with some representative results are given.

Figure 2. Scatterplot matrix of eight output variables: Cd, Hg, Zn, Cu, Ti, Cr, V, Ni.

Table 1. Left – “relatedness” between dependent variables estimated by the best multiple linear regression model. Variables with significant beta (p=0.05) are marked. Right – “relatedness” between dependent variables estimated by GRNN.

2 METHODOLOGY

The procedure of using the predictability criterion for the quantification of the “relatedness” was performed for all pollutants and the corresponding information is summarized in a matrix form (matrix of “relatedness”). The matrix of “relatedness” based on linear multivariate analysis (linear multivariate regression) is shown in Table 1 (left). The modelled variables are given at the left of the matrix while the most linearly related variables (nonzero betas with a p-value = 0.05) are given in
the corresponding rows. The matrix of nonlinear “relatedness” estimated by the GRNN is also shown in Table 1 (right). The selected models (group of variables) can be then analysed at the second step by using any multivariate model from machine learning algorithms.

Thus, the generic methodology proposed consists of two fundamental phases:

1. pre-processing of multivariate data and preparing of multitask “related” groups for each environmental variable to be studied,
2. multitask learning, testing/validation of the models and spatial predictions.

The first phase is based on GRNN. General Regression Neural Networks were introduced by Specht (1991). Below a short description of GRNN is given following Kanevski and Timonin (2010).

Let us consider data measurements $Z_n (n=1,…,N)$ in $m$-dimensional space $(x_1,…,x_m)$. According to GRNN, the prediction at unknown point $Z(x)$ is defined by the following formula

$$Z(x_1,…,x_m) = \sum_{n=1}^{N} \frac{Z_n \exp\left(-\frac{\sum_{i=1}^{m} D^2(x_i,x_{i,n})}{2\sigma_i^2}\right)}{\sum_{n=1}^{N} \exp\left(-\frac{\sum_{i=1}^{m} D^2(x_i,x_{i,n})}{2\sigma_i^2}\right)}$$

where $D^2(x_i,x_{i,n})$ is a distance between point of prediction and data, $\sigma_i$ is a kernel bandwidth corresponding to the $i$-th input dimension ($i=1,…,m$).

It is supposed that interaction between input variables is negligible; the procedure of feature selection and not a feature extraction is applied. Nevertheless, in a more general setting a full covariance matrix (Mahalanobis distance) or some data whitening procedures can be considered. The only unknown parameters in GRNN model are the kernel bandwidths ($\sigma_i$). These parameters are tuned either by splitting of data into training and validation subsets and trying to minimize the validation error by choosing corresponding bandwidths or (if there are not too many data) by using the leave-one-out criterion (called the cross-validation in geostatistics). Gradient search from a good starting point, for example estimated by isotropic kernel model, is another fast and efficient solution to minimize the cross-validation (or validation) error.

Let us remind some useful properties of GRNN for data processing: 1) for an isotopic kernel and the cross-validation error the minimisation solution is unique; 2) when there are no spatial structures, i.e. only spatially not correlated noise, there is no minimum on a cross-validation curve and the kernel bandwidth tend to infinity (i.e. it is much larger than the region of the study or maximal distance between measurements). This is also true when only some coordinates are very noisy and useless for the prediction (if $k$-input variable is a noise, then $\sigma_k \rightarrow \infty$). In such cases

$$\frac{D^2(x_k,x_{k,n})}{2\sigma_k^2} \ll 1 \quad \text{it follows that:} \quad \exp\left(-\frac{D^2(x_k,x_{k,n})}{2\sigma_k^2}\right) \approx 1$$

and the corresponding coordinate “$k$” disappears (does not contribute to the solution). This phenomenon was observed in Specht and Romsdahl (1994).
Recently the same idea was applied for the feature selection tasks in Gheyas and Smith (2010). More details on theory, training and applications of GRNN for spatial environmental data, can be found in Kanevski et al. (2009).

GRNN can also be used to analyse the residuals after the modelling by GRNN or other algorithms, like multilayer perceptrons, support vector machines etc. taking into account useful properties of GRNN mentioned above. GRNN is a good discriminatory (pattern/noise) tool.

A straightforward and direct approach in using isotropic GRNN for the selection of good group of variables/features is to consider all possible combinations of the models \(2^{(n-1)}-1\), and to select the best one, i.e. the one, which minimizes the leave-one-out error. This approach was considered in Kanevski and Timonin (2010) and Gheyas and Smith (2010), where the traditional feature selection task was considered - the selection of the best combination of the most relevant inputs.

When the number of variables is large, let us say >20, this approach is practically not possible and different approximate algorithms can be used, e.g. genetic algorithms, simulated annealing, to make the problem feasible, see Gheyas and Smith (2010).

For the second phase a multilayer perceptron (MLP) will be used in the present study. Now, an input and output spaces have to be constructed. The output space was prepared during the first phase: the most related groups of dependent variables were selected by GRNN. The simplest input space – geographical space composed of only longitude and latitude is considered in this paper. This space is typical in geostatistical studies of pollution mapping. Moreover, working in a geographical space permits also to use the geostatistical tools – variography, to control and to quantify the quality of MLP estimates, like it was proposed in Kanevski and Maignan (2004).

In fact, MLP can be replaced by any other nonlinear multivariate model, but it has some interesting for this research properties: after training the weights can be analysed to understand the results of MTL, an application of shuffled variables can be considered as a particular regularization procedure (noise injection), as it was demonstrated, for example, in Kanevski et. al. (2009). It is well known that MLP application for environmental tasks is not a trivial problem, see ASCE Tasks (2000). Many of the problems - model selection, training/overfitting, etc. can be solved if the level of noise in data is estimated independently. Then, only the structured information can be extracted from the data without fitting the noise. In the present paper – the variography is used. The variography is applied both for the original data and for the residuals (see Figure 5). The level of noise and small scale variability in data were modelled with a nugget effect in the variogram. Finally, the problem of MLP selection is simplified, e.g. by applying an early stopping criterion at the level of noise.

3 RESULTS

Let us consider a real case study and some results. In the present paper, for the brevity of presentation, the main results only for Ni are given. In order to test the efficiency of the approach and algorithm two spatially non-structured variables were additionally generated by shuffling the original data (shuffling = randomization, which preserves the global distribution but destroys the spatial patterns). Therefore, 10 output variables were considered (8 real - Cd, Hg, Zn, Cu, Ti, Cr, V, Ni; + 2 shuffled – Zn and Ni) which gives rise to the total number \((2^{(n-1)}-1) = 511\) of possible models. It is expected that the models with shuffled data will be neglected according to the cross-validation error criterion.
In order to compare the results of modelling at the second step of the analysis, the original 200 measurements were split into training (150) and validation (50) subsets. The splitting of data was carried out using spatial declustering procedures, which guarantees that all geographical regions are represented and the variability of data are preserved in both subsets.

The same procedure was performed also for MTL using the non-optimal models as well. Because of the different variability of pollutants all variables before modelling were projected from original min-max values to the interval [0,1]. It helps during the training procedures and in the interpretation of the results.

In Figure 3 the results for Ni (models of GRNN MTL pre-processing), are ranked according to the cross-validation error (the best model is left). Only the top ten models are shown. The dark colour codes the selected variables for a particular model/group, organised by columns. The models including the last two variables, shuffled Ni and Zn, were never selected. The same analysis was carried out for all eight pollutants and the results are summarized in Table 1 (right).

**Figure 3.** The structure of best (from left to right) models for Ni prediction. Important co-variables are the following: Cd, Zn, Cr, and V. Two variables at the bottom are shuffled data on Zn and Ni.

**Figure 4.** The structure of the best (from left to right) models for Zn prediction. For the best (left) model all co-variables with the exception of shuffled are important.
The ranking of models according to the cross-validation error for Zn pollutant is shown in Figure 4. It follows, that for Zn multitask learning all other pollutants should be taken into account.

After the detailed pre-analysis described above, a nonlinear multitask MLP model was developed and applied. The input layer of MLP consists of geographical coordinates, one or two hidden layers with a varying number of neurons selected by minimizing the validation error. Typical number of hidden neurons was between 7 and 15. The output layer includes the best group of variables selected by GRNN at the MTL pre-processing step. The best model developed was compared also with other models when the variables in MTL group were replaced either by shuffled ones or by non-optimal variables, for example for Ni instead of Zn, Hg was used. For the best MTL model the mean square error was always lower (10% - 30%).

Finally, the quality of the modelling procedure was also controlled by a geostatistical tool – variogram, see Kanevski and Maignan (2004). This test can be easily done in a low dimensional space. In Figure 5 the experimental variogram of raw Ni data and the theoretical variogram model are compared with the variogram of the training residuals of the best MTL model. For the clarity of interpretation only the isotropic variograms are shown. The variogram of the residuals demonstrates pure nugget effect – no spatial structures in the residuals. Moreover, the level of the nugget exactly corresponds to the nugget value in the theoretical model, which means that after modelling all structured information was correctly extracted and the residuals correspond to the noise level in raw data.

![Variogram](image)

Figure 5. The results of Ni mapping (top) and the variography of raw data and training residuals.

4 CONCLUSIONS

The paper presents a new two stage approach to multitask learning of spatial environmental multivariate data: 1) pre-processing of data for multitask learning and preparing of related groups of the variables; 2) analysis, testing and nonlinear multivariate modelling by using the best MTL group.
The approach is quite generic, does not depend on the structure of the input space and can be applied for a variety of environmental data mining tasks: spatial pollution data, environmental monitoring time series data, renewable resources assessments, natural hazards, etc. The “relatedness” of the variables is quantified using the nonlinear predictability criterion. For a reasonable number of dependent variables (≤ 20) the complete space of solutions can be explored in order to select the best group. When the number of variables is higher some empirical optimization procedures, like simulated annealing, genetic algorithms etc. can be applied.

The obtained results demonstrate the efficiency of the approach for the multitask learning of environmental phenomena. Moreover, the criterion of nonlinear predictability can be useful itself for better understanding of the relationships between the variables. The future research will deal with the generalization of the approach to spatio-temporal data and the application for other challenging case studies.

ACKNOWLEDGMENTS

The research was carried out within the framework of Swiss National Science Foundation projects 200020-121835 and 200021-126505. The author thanks to Dr. V. Timonin for many useful discussions on GRNN modelling. The constructive comments of the reviewers enabled to improve the paper.

REFERENCES