ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Feature selection in environmental data mining combining Simulated Annealing and Extreme Learning Machine

Michael Leuenberger and Mikhail Kanevski

University of Lausanne - Institute of Earth Surface Dynamics (IDYST)
UNIL-Mouline, 1015 Lausanne - Switzerland

Abstract. Due to the large amount and complexity of data available in geosciences, machine learning nowadays plays an important role in environmental data mining. In many real data cases, we face the need to design input space with the most relevant features. Because the main goal is to understand and find relationships between phenomena and features, feature selection is preferred to feature transformation or extraction. To deal with the high-dimensional space of environmental data, a wrapper method based on Extreme Learning Machine and global optimization algorithm (Simulated Annealing) is proposed. This paper investigates the whole methodology and shows promising results for environmental data feature selection and modelling.

1 Introduction

Environmental science is a field in constant development. Because environmental phenomena lie in high dimensional spaces (e.g. for natural hazards: $d \approx 10 - 100$), it is challenging to reach the real dimension where the phenomena under study can be understood, explained and predicted [6]. Moreover, in most real data cases the relationships between features and phenomena are non-linear. Keeping in mind that these relationships involve not only one but several features, the main goal is to select relevant subsets of features according to their potential non-linear ability to explain or predict environmental phenomena.

There are a lot of methods in wrapper, filter and embedded methodologies [3][4][8]. On the one hand filter methods are faster but do not necessarily take into account the combinations of various features simultaneously (a feature can be irrelevant alone but may be relevant with other features together). On the other hand wrapper methods allow complex associations of features but suffer from the curse of dimensionality when considering all possible combinations of features.

To address this challenge, this paper proposes a new methodology based on combining Extreme Learning Machine (ELM)[5] and Simulated Annealing (SAN)[7] algorithms. ELM has showed good capability for merging methods [1] and SAN remains a good optimization algorithm despite the fact that it can perform faster by combining with a genetic algorithm [2]. The principal advantages of this new method are the following: (1) ELM allows the quick evaluation of the non-linear potential of subsets of features, (2) SAN allows the optimal subset of features to be reached without using an exhaustive search.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

The use of ELM instead of the more robust and accurate OP-ELM [10] resides in the fact that current version of OP-ELM cancel out the wrapper ability to detect irrelevant feature. The methodology is described in Section 2. Section 3 presents the results using real and simulated data, and Section 4 concludes the paper.

2 Method

2.1 Extreme Learning Machine

The ELM algorithm follows the structure of a single-hidden layer feedforward neural network (SLFN)[5]. For a given labelled training set $Z_{trn} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^n, y_i \in \mathbb{R}\}_{i=1}^N$ and for a number of hidden nodes $\tilde{N}$, it computes the output matrix $N \times \tilde{N}$ of the hidden layer:

$$H_{ij} \quad = \quad g(\mathbf{x}_i \cdot \mathbf{w}_j + b_j)$$

where $\mathbf{w}_j$ (the vector of weights connecting the input layer with the j^{th} neuron) and b_j (the bias of the j^{th} neuron) are randomly generated. Then, the vector $\boldsymbol{\beta}$ (connecting the hidden layer with the output layer) is estimated using the Moore-Penrose generalized inverse of the matrix H:

$$\hat{\boldsymbol{\beta}} \quad = \quad H^{\dagger}\mathbf{y}$$

Once all weights of the network are known, new data can be evaluated and error assessed using a hold-out validation set. Extremely fast, the only parameter that requires tuning is the number of hidden node $\tilde{N}$. See in Section 3.2 how to deal with this parameter in order to preserve the computational time.

2.2 Simulated Annealing

SAN is a metaheuristic algorithm for optimization problems inspired by the field of metallurgy. Initialized with a high temperature parameter, it performs a global random search from neighbour to neighbour. In a second stage, temperature decreases progressively and the search becomes local. Based on the following Metropolis criterion [9], it has the capability to accept bad solutions according to the level of the current temperature T.

Let θ_{cur} and θ_{new} respectively be the current and new states of the research, and f the function to minimize. If $\Delta f = f(\theta_{cur}) - f(\theta_{new}) \leq 0$ the new state θ_{new} is accepted, else θ_{new} is accepted with a probability:

$$P = \exp(-\Delta f / T)$$

In a theoretical way, the ability to accept bad solutions allows us to find the global minimum of any kind of problem. In a practical way, it cannot guarantee finding the optimal solution but it can approach it. The success of this convergence lies in a good parametrization of the initial temperature and in the annealing process.

2.3 Feature Selection Methodology

Let n be the number of features available and $\Theta = \{\theta \mid \theta = \{0,1\}^n\}$ the set of the whole combination of features, where θ_i indicates if we consider feature i or not. The goal is to find $\theta^* \in \Theta$ that minimizes the cost function f defined as follows:

$$\begin{aligned} f(\theta) &= MSE(\mathbf{y}_{val}, \hat{\mathbf{y}}_{val}) \\ \text{where,} \qquad \hat{\mathbf{y}}_{val} &= ELM(\theta, \tilde{N}, Z_{trn}, Z_{val}) \end{aligned}$$

Z_{trn} and Z_{val} correspond to two separate training and validation sets, and $\tilde{N}$ is the number of hidden nodes. Without loss of generality, $\tilde{N}$ can be defined a priori (see experimental part 3).

Applying this notation and using the simulated annealing algorithm, the proposed new feature selection algorithm is as follows:

Algorithm 1 SANELM

```
Require: Initialize $\theta_0 \in \Theta$ and $T_0$ the initial temperature
Generate a model with $ELM(\theta_0, \tilde{N}, Z_{trn}, Z_{val})$
Compute $f(\theta_0)$, and put $\theta_{cur} = \theta_0$
for $i = 1$ to $STOP$ do
    Compute $T_{new} = Ann(T_0, i)$
    Generate $\theta_{new}$ in the neighbourhood of $\theta_{cur}$
    Compute $f(\theta_{new})$ and $\Delta f = f(\theta_{cur}) - f(\theta_{new})$
    if $\Delta f \leq 0$ then
        Accept $\theta_{new}$: $\theta_{cur} \leftarrow \theta_{new}$
    else
        Generate $U$ uniformly in $[0,1]$, and compute $P = \exp(-\Delta f / T_{cur})$
        if $U \leq P$ then
            Accept $\theta_{new}$: $\theta_{cur} \leftarrow \theta_{new}$
        else
            Reject $\theta_{new}$
        end if
    end if
end for
```

For more details of the methodology, see section 3.2.

3 Data and Results

3.1 Data

The data used for this application come from 200 measurement points in Lake Geneva. Composed of 3 real input variables (i.e. X, Y and Z coordinates), 21 simulated variables were added to the database. These additional input variables are composed of 3 shuffled variables from the original X, Y and Z coordinates, and of 18 random variables following a uniform distribution. Finally, the database was composed of 21 input variables and 1 output variable which is the pollutant, Nickel.

The principal objective is to investigate the parameter of the SANELM for this particular database, important for environmental risk studies and to evaluate the robustness and the accuracy of such methodology according to the parameters. The expected result is to find the original features, that is the X, Y and Z coordinates.

3.2 Experimental setup

First of all, the whole database must be normalized in order to fit to the range $[0, 1]$ within which ELM works. Secondly, because of the need to assess the ELM model at each iteration of the SAN algorithm, the database must be split into two subsets. About 75 per cent of the data are allocated to the training set and the remaining 25 to the validation set.

Once the preprocessing task is completed, several SAN parameters have to be fitted. The first one is the annealing schedule $Ann(T_0, i)$. Written as a function of the initial temperature T_0 and the iteration index i, the schedule can take different forms. No preferential function exists, but as the optimization space Θ is discrete and not continuous, a basic schedule can be considered such as:

$$Ann(T_0, i) = \frac{T_0}{c \cdot i} \qquad \text{or} \qquad Ann(T_0, i) = \frac{T_0}{c \cdot \log(i)}$$

where c is the parameter of the schedule. In practice, since T_0 and c have to be parametrized, the most simple way is to fix $c = 1$ and to fit the parameter T_0 by trial and error.

Another important proceeding in the algorithm is the generation of a new state $\theta_{new} \in \Theta$ in the neighbourhood of the current state θ_{cur}. For this purpose, θ_{new} is defined as a neighbour of θ_{cur} if and only if the Hamming distance between the two is equal to 1 (i.e. θ_{new} and θ_{cur} differ in just one coefficient). This allows them to reach any state of the Θ space in at least n steps (where n is the number of input variable).

In order to complete the parameter setup, it remains to tune the number of hidden nodes $\tilde{N}$. In the first stage of the paper, an additional loop was added in the algorithm in order to compute $f(\theta_{new})$ with the optimal number of hidden nodes. Because this process is time consuming, an analysis of the distribution of the optimal number of node was carried out. It appears that this distribution shows the same range of optimal number of nodes for any kind of $\theta \in \Theta$. Furthermore, if we fix the number of nodes $\tilde{N}$ that is not necessary the optimal one for the desired best subset of features θ^*, it appears that

$$f(\theta^*) \leq f(\theta) \qquad \forall \theta \in \Theta$$

In other words, even if the model f is not perfect for a fixed number of hidden nodes $\tilde{N}$, it would be minimal for subset of relevant features.

3.3 Results

The first results show the stability of the methodology according to the choice of the number of hidden nodes $\tilde{N}$. For this purpose, 1000 subsets of features were generated randomly and all are evaluated with ELM for $\tilde{N} \in \{5, 10, 15, ..., 70\}$. In Figure 1 each dashed line correspond to one random subset of features and the solid line coincides with the best subset of features. Examining 1000 random subsets of features reveals that the range of the number of hidden nodes where they reach the minimum value of MSE is approximately $[15, 30]$.

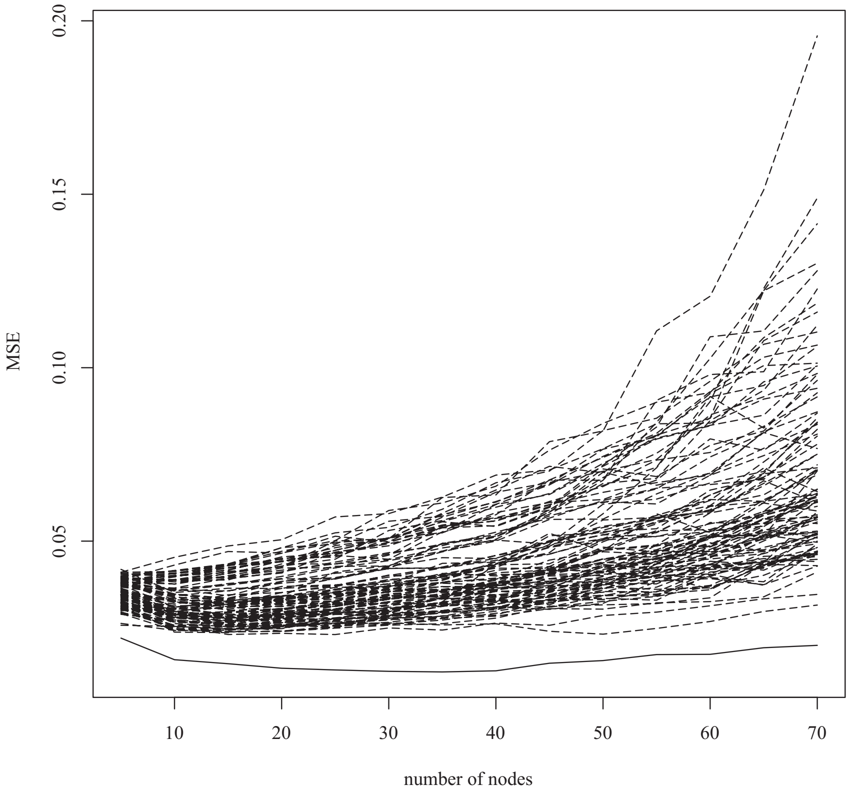

Fig. 1: Each dashed line correspond to one random subset of features and the solid line coincides with the best subset of features. The graph shows the MSE of the ELM for these different subsets of features according to the number of hidden nodes.

According to this first result, it is recommended that for each new problem the behaviour of $\tilde{N}$ is explored through randomly generative several subsets of features. By doing this, the range of the minimum number of nodes can be determined, and the SANELM algorithm can be performed using a fixed $\tilde{N}$ in that range.

By using the Lake Geneva database with the additional 18 irrelevant variables and with a fixed $\tilde{N} = 20$, the SANELM algorithm reaches the optimal subset of feature, that is the original X, Y and Z coordinates, in less than 4000 iterations. By comparison, the exhaustive search need $2^n - 1$ iterations (in this case more than 2 million) to evaluate all the possible combinations of features. The same results are obtained using different $\tilde{N} \in [15, 30]$.

4 Conclusion

This paper develops a combination of two algorithms, the Extreme Learning Machine as a wrapper method and the Simulated Annealing as an optimization algorithm. Analyses were performed in order to investigate the behaviour of both ELM and SAN parameters. As the optimization space is a discrete one, the annealing schedule of SAN can be standard. For the remaining T_0 and c parameters, trial and error are needed according to the complexity and dimensionality of the problem. For the unique ELM parameter $\tilde{N}$ (the number of hidden nodes), it has been shown that it is quite stable within the range determined by the problem. Therefore, $\tilde{N}$ can be fixed during the process and computational time can be reduced. In future research, this benefit will allow to investigate more complex phenomena in high dimensional space and multivariate data, as well as to perform a comprehensive comparison in computational time and accuracy with other feature selection algorithms.

References

[1] B. Frénay and M. Verleysen, Using SVMs with randomised feature spaces: an extreme learning approach. In M. Verleysen, editor, *proceedings of the* 18^{th} *European Symposium on Artificial Neural Networks* (ESANN 2010), d-side pub., pages 315-320, April 28-30, Bruges (Belgium), 2010.

[2] I. Gheyas and L. Smith, Feature subset selection in large dimensionality domains, *Pattern Recognition*, 43:5-13, 2010.

[3] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *The Journal of Machine Learning Research*, 3:1157-1182, 2003.

[4] I. Guyon, S. Gunn, M. Nikravesh and L.A. Zadeh. *Feature extraction: foundations and applications*, volume 207, Springer, 2006

[5] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew, Extreme learning machine: Theory and applications, *Neurocomputing*, 70(1-3):489-501, 2006.

[6] M. Kanevski, A. Pozdnoukhov and V. Timonin. *Machine Learning for Saptial Environmental Data*, EPFL Press, 2009.

[7] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi, Optimization by simulated annealing, *Science*, 220:671-680, 1983.

[8] J.A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*, Information Science and Statistics, Springer Verlag, 2007.

[9] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, Equation of State Caluculations by Fast Computing Machines, *The Journal of Chemical Physics*, 21:1087-1092, 1953.

[10] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten and A. Lendasse, OP-ELM Optimally Pruned Extreme Learning Machine, *IEEE Transactions on Neural Networks*, 21(1):158-162, 2010.