A New Feature Selection Methodology for Environmental Modelling Support: The Case of Thessaloniki Air Quality

Nikos Katsifarakis, Kostas Karatzas

A new feature selection methodology for environmental modelling support: the case of Thessaloniki Air Quality

Nikos Katsifarakis and Kostas Karatzas

Aristotle University, Department of Mechanical Engineering, Informatics Systems and Applications – Environmental Informatics Research Group, Thessaloniki, Greece
{nikolakk, kkara}@auth.gr

Abstract. Environmental systems status is described via a (usually big) set of parameters. Therefore, relevant models employ a large feature space, thus making feature selection a necessity towards better modelling results. Many methods have been used in order to reduce the number of features, while safeguarding environmental model performance and resulting to low computational time. In this study, a new feature selection methodology is presented, making use of the Self Organizing Maps (SOM) method. SOM visualization values are used as a similarity measure between the parameter that is to be forecasted, and parameters of the feature space. The method leads to the smallest set of parameters that surpass a similarity threshold. Results obtained, for the case of Thessaloniki air quality forecasting, are comparable to what feature selection methods offer.

Keywords: Air quality, Feature selection, Self-Organizing Maps

1 Introduction

Environmental systems are complex, in terms of the parameters required to describe their status and spatiotemporal behavior. It is therefore expected that relevant models are complex as well, involving a large number of features. In the case of urban air quality systems [1] such parameters can be pollutant levels, meteorological conditions and any other feature that describes the impact that the atmospheric environment poses on human life [2]. Such features are commonly used as inputs to various data-driven environmental models [3], while some of them contain very little or no valuable information for the purposes of the model they are fed into. Therefore, it is necessary to select the most appropriate ones and thus reduce their number, using feature prioritization and selection methods [4]. In this paper, two well-established feature selection methods are used as a reference, while a new feature selection method based on Self Organizing Maps (SOM) is introduced. Methods are compared in terms of forecasting performance for a number of Computational Intelligence (CI) oriented models, for the case of air quality forecasting in Thessaloniki, Greece.

2 The case study

The application domain of this study is urban air quality in the area of Thessaloniki. It is the second largest city of Greece, characterized by high urbanization and a heavily used traffic network. Its wider area covers approximately 93 km^2 and has a population density of around 16000 inhabitants per km^2. Thessaloniki is located in the inner part of the Thermaikos gulf and has Hortiatis mountain and the Seich Sou forest to its north and north – east respectively, while the industrial zone is situated on its western part [5]. Regarding its air quality, it is characterized by very high levels of PM_{10} [6].

Available data include daily averages of air pollutant concentration levels (PM_{10}, CO, NO_2 and O_3 in $\mu g/m^3$), as monitored at seven stations located in different areas of the city, for the years 2000 to 2013 (a total of 17 features). In order to better evaluate the new proposed feature selection method, features with many missing values were omitted, this being the reason that the used dataset contains no meteorological parameters. A map of the city is presented in Figure 1, where the seven stations (Agia Sofia, Aristotle University, Kalamaria, Kordelio, Panorama, Sindos and Neochorouda) are marked.

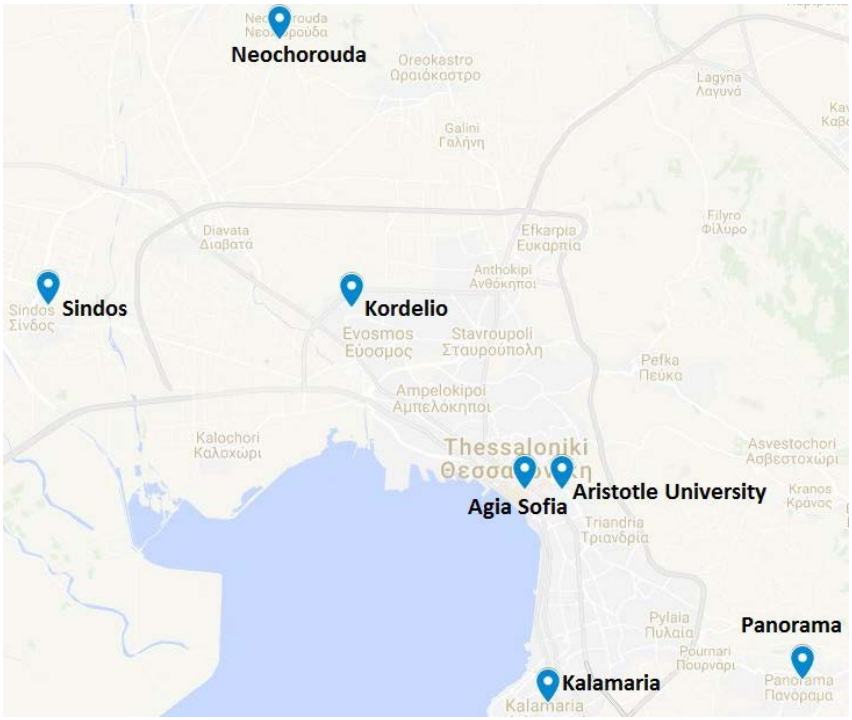

Fig. 1. Map of Thessaloniki, marking the seven air quality monitoring stations used in the current study.

3 The proposed methodology

With the aim of the study being the presentation and evaluation of a new feature selection methodology, two feature selection methods were rendered to be appropriate as reference methods, implemented in the WEKA computational environment [9]: (i) the Correlation - based feature selection (CfsSubsetEval) [7], and (ii) the ReliefFAttributeEval [8] The CfsSubsetEval method was chosen, as (a) it can be used in regression problems (i.e. the problem category where arithmetic values of feature(s) are forecasted), (b) it focuses on maximizing the forecasting ability of a model (our goal in this study), and (c) it employs features that demonstrate high predictive ability (i.e. lead to better forecasting statistics) and low intercorrelation [10]. The ReliefFAttributeEval method was chosen as it takes into account feature interrelationships by assigning a grade of relevance to each feature and then selecting those that are graded over a user given threshold [11].

The new feature selection method proposed makes use of the SOM method. SOMs are based on neural networks composed of a two-dimensional array of (initially) randomly weighted neurons [12]. All data points are passed through the neural network and are matched with a winning neuron, causing the network topology to adjust and eventually form clusters of similar attributes, while weights are updated to better fit into the process. The unified distance matrix (U-matrix) commonly used for SOM visualization, represents the Euclidean distance between neighboring neurons which is actually an expression of the relationship ("similarity") between neighboring neurons [13].

Here, the SOM Toolbox for Matlab [14] was used, as it offers a stable and commonly used implementation of the method. A typical example of the SOM representation, generated with the Matlab toolbox for the data used in this study is presented in Figure 2 that contains a SOM for each parameter, as well as the U - matrix.

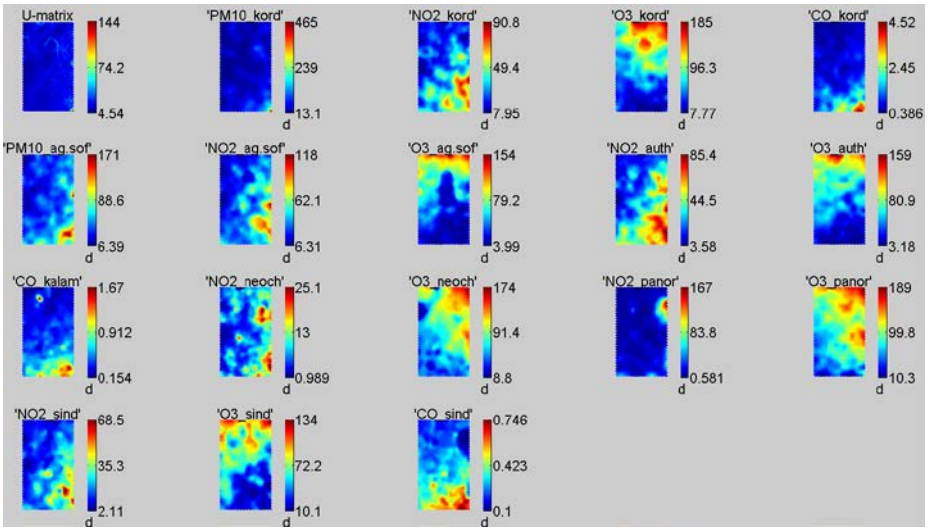

Fig. 2. A typical SOM representation of the case under study.

Each SOM visualizes the areas where the corresponding variables have high or low values (blue colors here correspond to low values and red to high), and the U – matrix indicates the distance between neighboring tiles (neurons). Thus, similarities between maps are indications of interrelationships between the corresponding parameters. For example, the maps of Ozone in the area of Agia Sofia (O3_ag.sof) and Ozone in the area of the Aristotle University of Thessaloniki (O3_auth) present with high values in their upper parts and low values in the lower parts, thus they look topologically similar. This indicates that there is strong relationship between these two features.

For each of the 17 parameters - features, the SOM method generates a topographic map presenting the weight values for the neurons in each of the SOM nodes. These values are stored in an M*N matrix, where M is the number on nodes that the aforementioned maps contain, and N is the number of the features. In this way, the values at each column represent the relevant weight ("importance") of each feature for the nodes within the SOM.

After one of the features is set as the parameter of interest (forecasting goal), the method aims at identifying the features that maximize the forecasting ability of the mode(s) to be used. For this purpose, we make use of the values of the M*N matrix *W* of the SOM weights. As the method is based on the weights of the existing features, we introduce a number of N random additional "features", so that the weight matrix doubles its columns becoming an M * 2N matrix. This is done in order to enlarge the population of candidate features to be selected by introducing features that are not related whatsoever with the problem under investigation, thus acting as indicators of "noise" and therefore be used as a selection threshold criterion as explained next.The weights matrix is then normalized, so that its values become (real) numbers between 0 and 1. In the next step, each feature column N is compared with that of the – parameter of interest, element by element, to determine how many values are either very similar, or complementary (i.e. have a sum close to 1), according to **Eq 1**. An arbitrary parameter α is introduced, to quantitatively express the relationship between values being compared, α being a positive real number close to zero.

$$|Tar(j) - Feat_i(j)| \leq \alpha \quad \textbf{or} \quad |Tar(j) + Feat_i(j) - 1| \leq \alpha \tag{1}$$

Here *Tar(j)* is the j-th element of the target – parameter column and $Feat_i(j)$ is the j–th element of the column of the i-th feature. Thus, the initial M*2N matrix is transformed to a new one, *S*, with each column representing the same feature as the initial matrix, but with each element being equal to either one (when the original map's corresponding element is very similar or very complementary to the one of the target – parameter), or to zero otherwise.

With this new matrix, the feature that has the highest sum of values ("ones") is selected as the first feature of the current selection *CL*, and then the rest of the features are ranked accordingly. Then, a new ranking takes place, in which the remaining features are ranked again on the basis of Eq. 1, according to the amount of elements equal to "ones" they present, for element locations that the current selection demonstrates "zeros". Again, the feature that ranks first is selected, thus replacing the "zeros" of the current selection at the places where the newly selected feature had "ones". In this way, features are "completing" each other in order to be part of the population of the selected features, in an effort to maximize the amount of ones with the least possible features. This procedure also makes sure that there will be no redundancy, as simi-

lar features will have “ones” in similar places, therefore if one of them is selected, the rest will rank poorly in the next ranking round. The overall algorithm is presented in Table 1.

Table 1. The new feature selection algorithm based on SOM

```
Algorithm SOM_Feature_Selection

! N: the initial number of features
! M: the number of data rows (time stamps) in the data matrix
! Tar: the target parameter to be forecasted

  Get W;                    ! this is the SOM weights matrix
  CL := [];                 ! initialize current selection
  W := [W, rand(size(W))]; ! introduce N new random features
  NW := Normalized(W);      ! normalize the SOM weights matrix
  S := zeros(size(NW));
  S(|Tar(j) – Feati(j)|  ≤ α  or  |Tar(j) + Feati(j) - 1|  ≤ α) := 1;
  S2 := AddAllValuesPerLine(S) – ones(M,1);
  maxsim := (M - Sum(S2 == 0)) / M;
  R := Rank(features(feature != Tar)) ;
                            !Ranks according to Sum(S(:,i)), i = 1, …, 2N
   CL := [CL, R(1)];
   begin
      R := Rank(features(feature != Tar, feature not in CL)) ;
                            !Ranks according to
                            !Sum(S(:,i)), i ∈[1, 2N] and S(:,CL(i)) == 0

      CL := [CL, R(1)];
   until  Rank(1) > N or  Sum(CL(CL == 0)) == 0
   if CL(end) > N then
      CL := CL(1:end-1);
   end if
end
```

This process terminates once the maximum amount of ones is reached, or when one of the random “features” is selected, as this will indicate that the rest of the features offer little to no useful information for the parameter of interest. In this way, a set of features that describes the biggest part or the SOM concerning the parameter of interest is determined.

4 Results and discussion

The parameter chosen to be forecasted as a test for the proposed methodology, was Ozone from the Panorama station (a typical inhabited urban area). The algorithms used for developing the forecasting models were Linear Regression (LR) [15] and Multilayer Perceptron (MLP) Artificial Neural Networks [16], since they are very commonly used in regression problems and have led to high AQ forecasting accuracy for the same geographic area [5]. Computations were again performed in WEKA [9].

Both LR and MLP were used for the cases of (a) the complete set of features, (b) the sets of features suggested by the CfsSubsetEval and ReliefFAttributeEval methods,

and (c) the features resulting from the suggested methodology, for each value from 0.005 to 0.5, with a step of 0.005, for the parameter α. It should be noted that model development and training was made via the 10-fold cross validation method [17].

The comparison of the CI models' performance with the set of features presented in each case, as well as these of the CfsSubsetEval and ReliefFAtttributeEval methods are presented in Figures 3, 4 and 5.

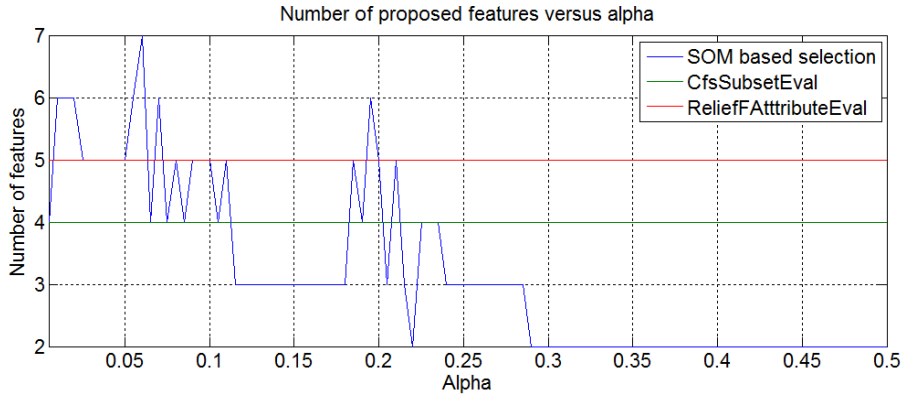

Fig. 3. Number of features of each feature set proposed by the SOM – based methodology, compared with these of the CfsSubsetEval and ReliefFAtttributeEval methods.

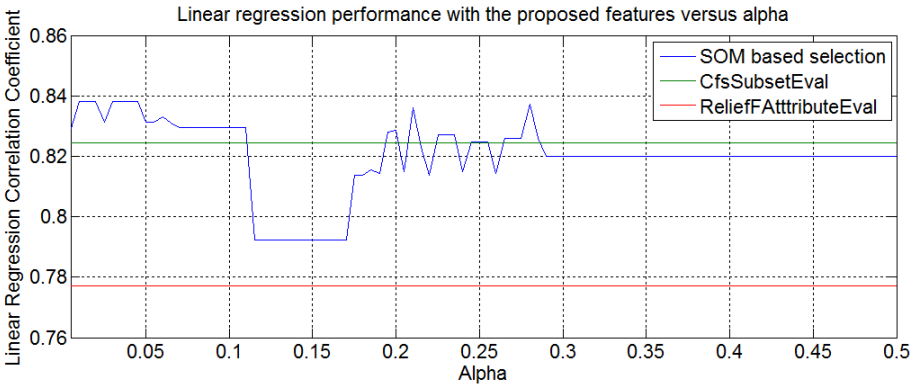

Fig. 4. Correlation coefficient of Linear Regression with each feature set proposed by the SOM – based methodology, compared with these of the CfsSubsetEval and ReliefFAtttributeEval methods.

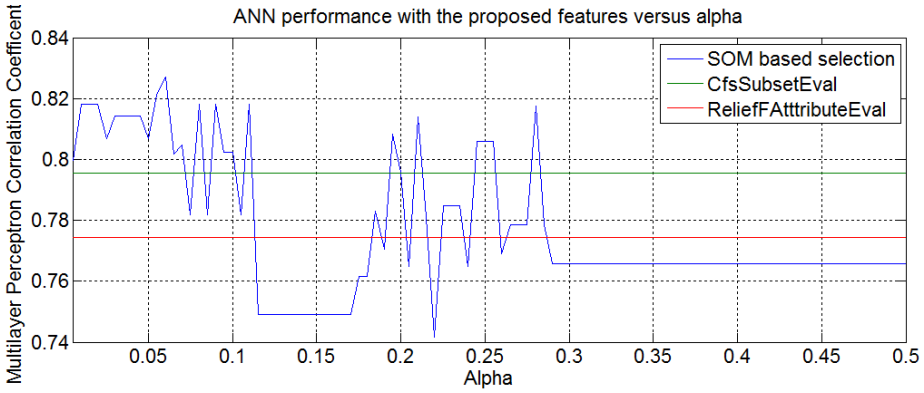

Fig. 5. Correlation coefficient of Multilayer Perceptron with each feature set proposed by the SOM – based methodology, compared with these of the CfsSubsetEval and ReliefFAttributeEval methods.

Results indicate that even a slight change in α can lead in different sets of features being selected. In more detail, for low values of α, (from 0.005 to 0.11), all the feature sets resulting from the new proposed method lead to better LR performance in comparison to the feature sets from both the reference methods (with 0.8382 being the highest value for correlation coefficient, 1.7% higher than the CfsSubsetEval method and 7.85% higher than ReliefFAtttributeEval method), while regarding MLP, they all outperform the ReliefFAtttributeEval method, and most outperform the CfsSubsetEval method as well (with 0.8332 being the highest value for correlation coefficient, 3.92% higher than the CfsSubsetEval method and 6.75% higher than ReliefFAttributeEval method). Values of 0.115 to 0.17 for α lead to the feature set with the poorest performance of all the sets offered by the new proposed method. Values of α between 0.175 and 0.295 lead to different sets that perform close to the set offered by the CfsSubsetEval method with LR and somewhere between the two reference methods with MLP. Finally, for α varying between 0.3 and 0.5, the feature set offered by the new proposed method does not change, and performs close to the set offered by the CfsSubsetEval method with LR and worse than the two reference methods with MLP, however, with only two features, as opposed to the four of the CfsSubsetEval method and the five of the ReliefFAtttributeEval method.

The most popular features among all the offered feature sets are Ozone from the Kordelio station, which is also selected by the CfsSubsetEval method, and Ozone from the Neochorouda station, which is selected by both the reference methods as well. This selection seems very plausible, as it makes sense for the same pollutant from the other stations to be more correlated to our target – parameter than other atmospheric quality parameters.

Overall, the CI model's performance is comparable to this of the sets presented by the CfsSubsetEval and the ReliefFAtttributeEval methods for each α. Table 2 presents the mean value and standard deviation of the CI models' performance for every α between 0.005 and 0.3 (as after 0.3 the results remain the same).

Table 2. Mean value and standard deviation of the CI models' performance for Ozone for the Panorama Station, with the sets of features from the suggested methodology, for α varying-between 0.005 and 0.3.

	Mean	**Standard Deviation**
Linear Regression	0.8202	0.0156
MLP	0.7848	0.0262

Table 2 indicates that even a poor choice of α will outperform the ReliefFAt-tributeEval method, and perform close to the CfsSubsetEval method. The low standard deviation of the results, despite the number of different set of features offered, shows that the concept of complementarity of features with the aim of maximum similarity with the target – parameter will always lead to acceptable results for the case under investigation. It is also apparent that there are values of α that lead to sets of features which outperform both the ReliefFAtttributeEval and the CfsSubsetEval methods, as well some that lead to good performance with less features.

For reasons of further investigation, the same tests were also run for a different pollutant and monitoring station, namely NO_2 for the Sindos station. The mean value and standard deviation of both models' performance is presented in Table 3. It should be mentioned that in this case results did not change for any α greater than 0.27. Again low values for α lead to feature sets with better performance, at times superior to the two reference methods, with the maximum correlation coefficient for the LR model reaching 0.74 (for a=0.11) while for the MLP model it was 0.67 (again for a=0.11). Reference methods on the other hand led to the feature sets that correspond to the modelling results presented in Table 4, demonstrating better performance in comparison to the mean performance indicators of Table 3.

Table 3. Mean value and standard deviation of the CI models' performance for NO_2 for the Sindos Station, with the sets of features from the suggested methodology, for every α between 0.005 and 0.27.

	Mean	**Standard Deviation**
Linear Regression	0.7049	0.0173
MLP	0.6261	0.0258

Table 4. Performance of the CI models with the feature sets offered by the two reference feature selection methods, for NO_2 and for the Sindos station.

	CfsSubsetEval	**ReliefFAttributeEval**
Linear Regression	0.7317	0.6931
MLP	0.6564	0.6286

5 Conclusions

In the present paper a new methodology for feature selection in data-driven air quality forecasting problems is presented and compared with two existing and commonly used feature selection methods, namely CfsSubsetEval and ReliefFAttributeEval. The new method is based on SOMs thus making use of additional information concerning feature interrelationship and cross-influence. Two air quality forecasting models were used for testing the performance of the feature selection methods, one being a standard multivariate Linear Regression model and the other being an Artificial Neural Network of the Multi-Layer Perceptron type. Model results show that the new feature selection method is comparable to the reference ones, while being also able to outperform them producing results up to +7.85% for the correlation coefficient, provided that computational experiments take place in order to determine the values for the selection parameter *a* (alpha). As such, this approach seems promising, and is offered for further investigation, involving testing additional cases while also improving the estimation of the alpha parameter.

References

1. Gulia S., Nagendra S., Khare M., Khanna I. (2015), "Urban air quality management-A review", Atmospheric Pollution Research **6**(2), pp 286–304
2. Chen H., Goldberg MS., Villeneuve PJ. (2008). "A systematic review of the relation between long-term exposure to ambient air pollution and chronic diseases." Reviews on environmental health **23(**4), pp. 243–97
3. Araghinejad S.(2014),"Data-Driven Modeling: Using MATLAB® in Water Resources and Environmental Engineering", Water Science and Technology Library, Volume 67
4. Mesin et al. (2010), "A Feature Selection Method for Air Quality Forecasting", Artificial Neural Networks – ICANN 2010, Volume 6354 of the series Lecture Notes in Computer Science pp 489-494
5. Voukantsis D., Karatzas K., Kukkonen J., Räsänen T. Karppinen A. and Kolehmainen M. (2011), "Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki", Science of the Total Environment, 409, pp. 1266–1276 DOI:10.1016/j.scitotenv.2010.12.039
6. Moussiopoulos N., Vlachokostas Ch, Tsilingiridis G., Douros I., Hourdakis E., Naneris C., Sidiropoulos C. (2009), "Air quality status in Greater Thessaloniki Area and the emission reductions needed for attaining the EU air quality legislation", Science of the Total Environment. 2009 Feb pp. 1268-85
7. Hall, M. A. (1998), "Correlation-based feature subset selection for machine learning". PhD thesis, The University of Waikato, http://www.cs.waikato.ac.nz/~ml/publications/1999/99MH-Thesis.pdf
8. Kenji Kira, Larry A. Rendell: (1992), "A Practical Approach to Feature Selection". In: Ninth International Workshop on Machine Learning, pp. 249-256
9. Hall M.A., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H. (2009), "The WEKA Data Mining Software: An Update"; SIGKDD Explorations **11(**1). http://www.cs.waikato.ac.nz/ml/weka/
10. Yue C. et al (2010), "Correlation-Based Feature Selection and Regression", Volume 6297 of the series Lecture Notes in Computer Science pp. 25 – 35.
11. Arauzo-Azofra A., Benítez J.M., Castro J. L. (2004) "A Feature Set Measure Based on Re-

lief”
12. Kohonen, Teuvo (1982). "Self-Organized Formation of Topologically Correct Feature Maps". Biological Cybernetics **43** (1), pp. 59–69
13. Ultsch A. and Siemon H.P. (1990).” Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis” Proceedings of International Neural Networks Conference (INNC) (1990), pp. 305-308
14. http://www.cis.hut.fi/somtoolbox/
15. Rencher A.C., Christensen W.F. (2012), “Chapter 10, Multivariate regression – Section 10.1, Introduction”, Methods of Multivariate Analysis, Wiley Series in Probability and Statistics,709 (3rd ed.), John Wiley & Sons, p. 19
16. Rumelhart D.E., Hinton G.E., Williams R.J. (1986). “Learning Internal Representations by Error Propagation”. David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations. MIT Press
17. Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 2 (12), pp. 1137–1143