ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

A New Wrapper Method for Feature Subset Selection

Noelia Sánchez-Maroño[1] and Amparo Alonso-Betanzos[1] and Enrique Castillo[2] *

1- University of A Coruña- Department of Computer Science- LIDIA Lab.
Faculty of Informatics- 15071 A Coruña- SPAIN.

2- University of Cantabria. Civil Engineering School.
Department of Applied Mathematics and Computer Science.
39005 Santander- SPAIN

Abstract. ANOVA decomposition is used as the basis for the development of a new wrapper feature subset selection method, in which functional networks are used as the induction algorithm. The performance of the proposed method was tested against several artificial and real data sets. The results obtained are comparable, and even better, in some cases, to those accomplished by other well-known methods, being the proposed algorithm faster.

1 Introduction

Feature selection consists on selecting a subset of relevant features from a set, which remaining features will be ignored. Feature selection may not only help improve performance accuracy, but also results in better understanding and interpretation of the data. Given a set of n features and M samples $\mathbf{x} = \{x_{ij}; i = 1, \ldots, M; j = 1, \ldots, n\}$, feature subset selection methods find a subset $\mathbf{x}_s = \{x_{i1}, \ldots, x_{is}\}$, with $s < n$, that optimizes an objective function.

Feature subset selection requires a search strategy to select candidate subsets and an objective function to evaluate these candidates. Two different general approaches are commonly considered:

- Filter algorithms, in which case the selection method is used as a preprocessing that does not attempt to optimize directly the predictor (machine learning method) performance.
- Wrapper algorithms, in which the selection method optimizes directly the predictor performance.

In this paper, a new wrapper subset selection algorithm based on ANOVA decomposition is presented. Functional networks are used as the induction algorithm. The method has been tested against several benchmark and real data sets, and their results are presented and compared with those obtained by other filter and wrapper algorithms.

*This work has been partially funded by the Spanish Ministry of Science and Technology under project TIC-2003-00600 with FEDER funds, and by the Xunta de Galicia, under project PGIDIT04PXIC10502PN

2 Materials and Methods

2.1 The Sobol ANOVA Decomposition

According to Sobol, any square integrable function $f(x_1, \cdots, x_n)$ defined on the unit hypercube $[0,1]^n$ can be written as

$$y = f(x_1, \cdots, x_n) = f_0 + \sum_{\nu=1}^{2^n-1} f_\nu(\mathbf{x}_\nu), \tag{1}$$

where $\{\mathbf{x}_\nu | \nu = 1, 2, \cdots, 2^n - 1\}$ is the set of all possible subsets of the set $\{x_1, x_2, \cdots, x_n\}$. In the case $\nu = 0$, corresponding to the empty set, the function $f_\nu(\mathbf{x}_\nu)$ has no arguments, and it is assumed to be the constant f_0. The decomposition (1) is called ANOVA iff

$$\int_0^1 f_\nu(\mathbf{x}_\nu) dx_i = 0; \;\; \forall x_j \in \mathbf{x}_\nu \;\; i \neq j \;\; \forall \nu.$$

Then, the functions corresponding to the different summands are unique and orthogonal [1], i.e.:

$$\int_0^1 \int_0^1 \cdots \int_0^1 f_{\nu_1}(\mathbf{x}_{\nu_1}) f_{\nu_2}(\mathbf{x}_{\nu_2}) d\mathbf{x}_{\nu_1} d\mathbf{x}_{\nu_2} = 0; \; \forall \nu_1 \neq \nu_2.$$

Note that, since the above decomposition includes terms with all possible kinds of interactions among variables $x_1, x_2, \cdots, x_n$, it allows determining these interactions.

The main advantage of this decomposition is that there are closed or explicit formulas to obtain the different summands or components of $f(x_1, \cdots, x_n)$. These expressions were given by Sobol in [1], allowing that the $f(x_1, \cdots, x_n)$ function can always be written as the sum of the 2^n orthogonal summands:

$$f(x_1, \cdots, x_n) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{i=1}^{n-1} \sum_{i<j}^{n} f_{ij}(x_i, x_j) + \cdots + f_{12\cdots n}(x_1, x_2, \cdots, x_n).$$

Since if $f(x_1, \cdots, x_n)$ is square integrable, then all $f_\nu(\mathbf{x}_\nu); \; \nu = 1, 2, \cdots, 2^n - 1$ also are square integrable, squaring $f(x_1, \cdots, x_n)$ and integrating over $(0,1)^n$ one gets

$$\int_0^1 \int_0^1 \cdots \int_0^1 f^2(x_1, \cdots, x_n) dx_1 dx_2 \cdots dx_n - f_0^2 = \sum_{\nu=1}^{2^n-1} \int_0^1 f_\nu^2(\mathbf{x}_\nu) d\mathbf{x}_\nu,$$

and calling D to the left part of this equation and D_ν to each summand in the right part, it results

$$D = \sum_{\nu=1}^{2^n-1} D_\nu.$$

If $(x_1, x_2, \cdots, x_n)$ is a uniform random variable in the unit hypercube, then the constant D is its variance. With this, the following set of global sensitivity indices, adding up to one, can be defined

$$S_\nu = \frac{D_\nu}{D}; \;\; \nu = 1, 2, 3, \cdots, 2^n - 1.$$

Therefore, the variance of the initial function can be obtained by summing up the variance of the components, and this allows assigning global sensitivity indices, adding to one, to the different functional components.

2.2 A basic description of functional networks

Functional networks are a generalization of neural networks that combine both knowledge about the structure of the problem, to determine the architecture of the network, and data, to estimate the unknown functional neurons. A functional network consists of: a) several layers of storing units; b) one or several layers of processing units that evaluate a set of input values and delivers a set of output values and c) a set of directed links, that indicate only the direction of the information flow and do not contain parameters.

In functional networks, the activation functions are unknown functions from a given family, i.e., polynomial, trigonometric, etc., to be estimated during the learning process. In addition, functional networks allow connecting neuron outputs, forcing them to be coincident. For a more detailed study of functional networks consult [2].

3 The proposed wrapper feature subset selection method

A method for learning the functional components of $f(x_1, \cdots, x_n)$ from data and calculating global sensitivity indices that will allow us to select the appropriate features is presented.

The idea consists of approximating each functional component $f_\nu(\mathbf{x}_\nu)$ in (1), using some set $\{h^*_{\nu 1}(\mathbf{x}_\nu), h^*_{\nu 2}(\mathbf{x}_\nu), \cdots, h^*_{\nu k^*_\nu}(\mathbf{x}_\nu)\}$ of simple basic functions (polynomial, Fourier series, etc.), i.e.:

$$f_\nu(\mathbf{x}_\nu) \approx \sum_{j=1}^{k^*_\nu} c^*_{\nu j} h^*_{\nu j}(\mathbf{x}_\nu), \tag{2}$$

where $c^*_{\nu j}$ are real constants. Then, an orthogonalization and orthonormalization processes are carried out so that the functional components become:

$$f_\nu(\mathbf{x}_\nu) \approx \sum_{j=1}^{k_\nu} c_{\nu j} p_{\nu j}(\mathbf{x}_\nu).$$

Notice that not only the functions h^* have changed but also the set of real constants, $c_{\nu j}$. These modified parameters, $c_{\nu j}$, will be estimated by solving

a minimization problem and they will allow to calculate the global sensitivity indices. These indices will indicate which variables or relations between variables are relevant. The different steps of the method are briefly described in the following subsections.

3.1 Sobol ANOVA decomposition algorithm

Input. A data set with M samples and n input variables $\{x_1, x_2, \ldots, x_n\}$ and one output variable y.
Output. One approximation to the Sobol ANOVA decomposition of function $f(x_1, x_2, \cdots, x_n)$ and the global sensitivity indices.
Step 1: Select a set of approximating functions.
Each functional component $f_\nu(\mathbf{x}_\nu)$ in (1) is estimated by (2) where $c^*_{\nu j}$ are real constants and $h^*_{\nu j}$ are simple basic functions that are elected in this step of the algorithm.

Functional networks can be used for an adequate selection of families of functions. Several networks can be trained using different families of functions; then, the family with a better performance results is elected to be the $h^*_{\nu j}$ functions.
Step 2: Impose the orthogonality constraint for the functional components.

A new set of approximating functions, $\{h_{\nu 1}(\mathbf{x}_\nu), \cdots, h_{\nu k_\nu}(\mathbf{x}_\nu)\}$, is obtained when the following orthogonality constraint is imposed

$$\int_0^1 \sum_{j=1}^{k^*_\nu} c^*_{\nu j} h^*_{\nu j}(\mathbf{x}_\nu) dx_{i_r} = 0; \quad \forall x_{i_k} \in \mathbf{x}_\nu; \forall \nu.$$

Step 3: Orthonormalize the basic functions of each functional components.

A Gram matrix $\mathbf{G}$ with elements g_{ij} defined by:

$$g_{ij} = \int_0^1 \int_0^1 \cdots \int_0^1 h_{\nu_i}(\mathbf{x}_{\nu_i}) h_{\nu_j}(\mathbf{x}_{\nu_j}) d\mathbf{x}$$

is calculated. Then, the eigenvectors of matrix $\mathbf{G}$ are the columns of a matrix $\mathbf{Q}$ which is used for calculating the new basic functions as:

$$(p^*_{\nu 1}(\mathbf{x}_\nu), p^*_{\nu 2}(\mathbf{x}_\nu), \cdots, p^*_{\nu k_\nu}(\mathbf{x}_\nu)) = \mathbf{Q} \ \ (h_{\nu 1}(\mathbf{x}_\nu), h_{\nu 2}(\mathbf{x}_\nu), \cdots, h_{\nu k_\nu}(\mathbf{x}_\nu)).$$

Finally, these functions $p^*_{\nu k_\nu}$ are normalized obtaining the functions $p_{\nu k_\nu}$.
Step 4: Learn the coefficients by least squares.

The coefficients $c_{\nu j}; \nu = 1, 2, \cdots, 2^{n-1}; j = 1, 2, \cdots, k_\nu$ are obtained by solving the following minimization problem:

$$\underset{\substack{f_0, c_{\nu j}; \nu = 1, \cdots, 2^n - 1 \\ j = 1, \cdots, k_\nu}}{\text{Minimize}} Q = \sum_{k=1}^{m} \epsilon_k^2 = \sum_{k=1}^{m} \left(y_k - f_0 - \sum_{\nu=1}^{2^n-1} \sum_{j=1}^{k_\nu} c_{\nu j} p_{\nu j}(\mathbf{x}_{\nu k}) \right)^2 .$$

Step 5: Obtain the global sensitivity indices.

Since the resulting basic functions have already been orthonormalized, the global sensitivity indices (importance factors) are the squares of the coefficients, i.e.:

$$S_\nu = \sum_{j=1}^{k_\nu} c_{\nu j}^2; \;\; \nu = 1, 2, \cdots, 2^n - 1.$$

Step 6: Return solution.

The list of coefficients $c_{\nu j}; \nu = 1, 2, \cdots, 2^n - 1; \; j = 1, 2, \cdots, k_\nu$, the set of basic functions $\{p_{\nu 1}(\mathbf{x}_\nu), p_{\nu 2}(\mathbf{x}_\nu), \cdots, p_{\nu k_\nu}(\mathbf{x}_\nu) | \nu = 1, 2, \cdots, 2^n\}$, and the sensitivity indices $\{S_\nu | \nu = 1, 2, \cdots, 2^n\}$ are returned.

3.2 Selecting the proper variables

Once the ANOVA decomposition algorithm has been applied, a global sensitivity index is obtained for each variable and each combination of variables considered. Then, only the variables with a high index by its own or included in a combination of variables with a high index are considered. Therefore, a minimum limit has to be determined, in such a way that those variables above this minimum are considered relevant and those below are discarded. This limit depends on the problem being solved and it has to be defined for each one in terms of the global sensitivity indices obtained.

Besides the indices, an approximation for the function to be estimated is also obtained. This approximation may have a very good performance, in terms of accuracy. In this case, the method is used as a pure wrapper algorithm. If the performance obtained could be improved, the previous algorithm from step 4 on or the functional networks used in the step 1 of it can be employed as induction algorithm considering only the relevant variables and its combinations.

4 Results

The proposed method has been applied to several datasets used in previous studies [3]. All datasets except for Corral, introduced in [4], were obtained from the Irvine repository [5]. For all the cases, the set of basic functions chosen was the polynomial family considering only functions with degree two. Therefore, equation (1) is simplified and the first three steps of the methodology need to be applied only once.

The results obtained are shown in Table 1, datasets above the horizontal line are artificial and those below are real. For the artificial data, training and testing samples were selected such as it was done in [3]. Similarly, a ten-fold cross-validation was carried out for the real datasets. In these cases, the results shown are the mean and standard deviation of the ten test accuracies obtained, moreover, it has to be considered that different global sensitivity indexes are obtained for each fold, so a variable that is upper the limit in a fold can be below it in another fold. Therefore, all the variables that were included in at least two folds were considered. In [3], the authors compared the performance of

different wrapper and filter approaches. From them, in our table we have selected those with the best average results in real and artificial data sets. The results are comparable to those of the other methods. Although a comparison in CPU time consumption is not directly possible because of the different processors used, an extrapolation of the processing speed of both makes our method 10 times faster than the best of the others. As an example, each fold of the breast-cancer takes $0.15s$ in a Pentium 4 (260GHz).

Dataset (Features)	B.A.	ANOVA+FN		ID3-RLF		ID3-BFS	
		Sel	Acc	Sel	Acc	Sel	Acc
Corral (6)	56.25	4	100	5	100	4	100
Monk1 (7)	50.00	3	100	3	97.22	3	97.22
Monk2 (7)	67.13	4	67.59	4	63.90	3	64.35
Monk3 (7)	52.78	2	97.22	3	100	2	97.22
		ANOVA+FN		NB-RLF		NB-BFS	
Pima (8)	65.52	2	76.82 ± 4.16	1.2	64.57 ± 2.4	4.4	76.03 ± 1.6
Breast (7)	65.10	5	95.42 ± 3.07	5.7	95.14 ± 1.3	5.9	96.00 ± 0.6

Table 1: A comparison of the methodology proposed with the ID3 algorithm for the artificial data and Naive-Bayes(NB) for the real data. Both methods are applied with the Relived-F filter (RLF) and with the wrapper using backward best-first search with compound operators (BFS). B.A. stands for Baseline Accuracy and it is defined as the accuracy when predicting the majority class. Acc stands for test accuracy and Sel for the mean number of selected features.

5 Conclusions

A new method for feature subset selection based on ANOVA decomposition and functional networks has been developed and tested using several data sets. The results achieved show the adequacy of the approach, although better results could be obtained selecting appropriate families for each data set. The method has a complexity exponential to the number of initial features, but it has the advantage of given an interpretation in terms of variance to the selected subset that most of the wrapper methods can not do.

References

[1] I. M. Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55:271–280, 2001.

[2] E. Castillo, A. Cobo, J.M. Gutiérrez, and E. Pruneda. *Functional Networks with Applications*. Kluwer Academic Publishers, 1998.

[3] R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial Intelligence journal, special issue on relevance*, 97(1-2):273–324, 1997.

[4] G. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Machine Learning: Proceedings of the Eleventh International Conference*, pages 121–129, 1994.

[5] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.