A Supervised Feature Selection Approach Based on Global Sensitivity

Hana Sulieman and Ayman Alzaatreh

Abstract In this paper we propose a wrapper method for feature selection in supervised learning. It is based on the global sensitivity analysis; a variance-based technique that determines the contribution of each feature and their interactions to the overall variance of the target variable. First-order and total Sobol sensitivity indices are used for feature ranking. Feature selection based on global sensitivity is a wrapper method that utilizes the trained model to evaluate feature importance. It is characterized by its computational efficiency because both sensitivity indices are calculated using the same Monte Carlo integral. A publicly available data set in machine learning is used to demonstrate the application of the algorithm.

Hana Sulieman · Ayman Alzaatreh
Department of Mathematics and Statistics, American University of Sharjah,
P.O. Box 26666, Sharjah, United Arab Emirates
✉ hsulieman@aus.edu
✉ aalzatreh@aus.edu

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 5, No. 1, 2018

DOI 10.5445/KSP/1000087327/03
ISSN 2363-9881

1 Introduction

The purpose of feature selection, as a process for dimensionality reduction, is to identify the subset of features that provides the most reliable and robust learning performance. By removing noisy or redundant features, computational cost is reduced; higher accuracy and better interpretability are achieved. Feature selection plays a central role in many areas such as natural language processing, computational biology, image recognition, information retrieval, business analytics and many others.

In this article, we implement a sensitivity analysis approach to feature selection. Sensitivity analysis is concerned with understanding how the input variables (features) influence the changes of the output (target) variable. There exist two categories of techniques for sensitivity analysis: local sensitivity techniques and global sensitivity techniques. Local sensitivity techniques are mainly derivative-based techniques that measure the local impact of input variables on the output variable. On the other hand, global sensitivity techniques measure the impact on the output variable resulting from varying the inputs in their ranges of uncertainty. Over the last decade, global sensitivity analysis has gained acceptance among practitioners of model development. The proposed feature selection is a variance-based sensitivity technique that decomposes the target variable variance into summands of variances of the features in an increasing dimensionality. The approach utilizes the model training algorithm to compute feature sensitivities. Hence, the effectiveness of the resulting sensitivity measures depends largely on the precision of the trained model. The article is organized as follows. Section 2 presents a review of some existing methods for feature selection. Section 3 describes the global sensitivity analysis approach and gives some theoretical foundation of the method. Section 4 illustrates an application of the method and concludes with a discussion of some of its computational aspects.

2 Review of Feature Selection Methods

Based on the search strategy used, the existing methods of feature selection can be classified into three general frameworks (Guyon and Elisseeff, 2003; Janecek et al, 2008; Maio and Niu, 1981):

1. Filters,
2. wrappers, and
3. embedded methods.

Filters use the data itself to rank the features according to certain criteria and performs feature selection before the learning algorithm runs. Wrapper methods, on the other hand, evaluate the importance of features using the learning algorithm itself. Embedded methods search for the most discriminative subset of features simultaneously with the process of model construction. What follows is a brief description of each framework:

1. **Filters**
   Filters evaluate feature importance as a pre-processing operation to model construction as depicted in Figure 1. They use some information metric to calculate feature ranking from the data without direct input from the target. Popular information metrics include t-statistic, p-value, Pearson correlation coefficient, mutual information and other correlation measures. Computationally, filters are more efficient than wrappers as they require only the computation of n scores for n features. A drawback of filter methods is that they evaluate feature importance based on linear effects. Nonlinear effects of features are left undiscovered. The Markov blanket, a technique based on Bayesian networks, was developed to overcome this drawback of filter methods (Aliferis et al, 2010).

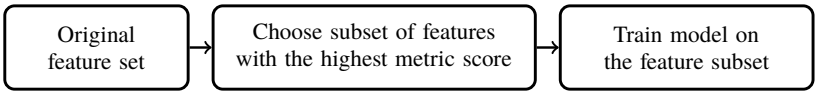

Figure 1: Filters framework.

2. **Wrappers**
   Wrapper methods are model-based methods for feature selection and are considered to be the most effective and computationally stable algorithms. Figure 2 shows the main principle of the wrapper methods' framework. Basically, to find the most relevant set of features, the intended model is trained for different subsets of features. The subset with the highest score on a particular performance criterion is selected as best set of features. Because wrapper methods utilize the learning algorithm they are considered more effective and hence more desirable than filters and embedded methods.

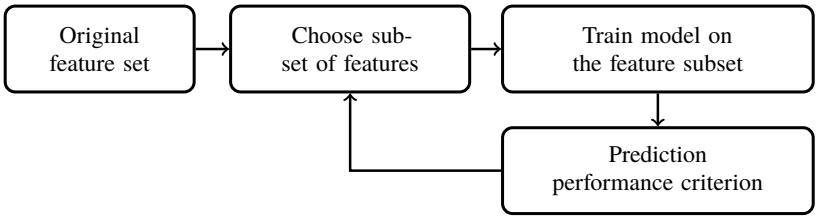

Figure 2: Wrappers framework.

   However, wrapper methods are generally criticized for their excessive computational cost. They require training and evaluation of the performance of the learning algorithm for every selected subset of features. For n features, the number of feature subsets is equal to $O(2^n)$, i.e., the computational needs of wrappers exponentially increase with the number of features in the model. This makes the search for all possible subsets of features impractical for even moderate values of n. Sequential search methods such as forward selection, backward elimination and stepwise regression became popular techniques used to overcome some of the computational demands of wrappers. Other efficient search strategies have also been developed to find the desired optimal subset of features (Hocking and Leslie, 1967).

3. **Embedded methods**
   Embedded methods combine the characteristics of filter and wrapper methods. They are implemented by learners that have built-in search procedures for the optimal subset of features (Figure 3). Embedded methods aim at maximizing the performance of the learning algorithm while minimizing the number of features in the algorithm.

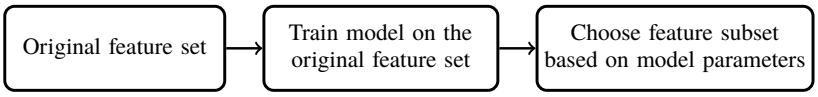

Figure 3: Embedded methods framework.

Regularization regression methods such as Ridge Regression (Hoerl, 1970), Least Absolute Selection and Shrinkage Operator (LASSO, Tibshirani, 1996) are the most popular forms of embedded methods. Regularized decision tree algorithms, such as random forests (Breiman, 2001), are other examples of embedded methods.

For more comprehensive reviews of feature selection methods, the reader is referred to Guyon and Elisseeff (2003) and Clarke et al (2009).

3 Global Sensitivity Indices

In this section we propose a variance-based feature selection approach for supervised learning in which the variance of the target variables is decomposed into contributions of individual or subsets of features. The technique which was first developed by Sobol (1990, in Russian) and Sobol (1993, in English) relies on the theory of ANOVA decomposition (Efron and Stein, 1981; Scheffé, 1959).

For a model function $Y = f(\mathbf{X})$, the importance of input variables (features) $\mathbf{X} = (X_1, \ldots, X_n)$ is quantified in terms of the reduction in the variance of Y when $\mathbf{X}$ are fixed at some value. For a particular X_i, this is expressed as

$$S_i = \frac{V_i}{V(Y)} = \frac{V_{X_i}[E_{\mathbf{X}_{-i}}(Y|X_i)]}{V(Y)}, \qquad i = 1, 2, \ldots, n, \tag{1}$$

and the joint (interaction) effect is given by

$$S_{ij} = \frac{V_{ij}}{V(Y)} = \frac{V_{X_i,X_j}[E_{\mathbf{X}_{-ij}}(Y|X_i, X_j)] - V_i - V_j}{V(Y)}, \quad i, j = 1, 2, \ldots, n,\ i \neq j, \tag{2}$$

where $\mathbf{X}_{-i}$ is the set of all variables except X_i and $E(.)$ and $V(.)$ represent the expected value and variance. S_i and S_{ij} represent the first order and interaction effects for X_i and (X_i, X_j), respectively. These effects measure the reduced

portion of the output (target) variable uncertainty caused by the input X_i and its interactions when the true value of X_i is known. Higher order interaction effects can be defined in a similar fashion. Another popular variance measure of X_i is the total effect index defined by:

$$TS_i = \frac{TV_i}{V(Y)} = \frac{E_{\mathbf{X}_{-i}}[V_{X_i}(Y|\mathbf{X}_{-i})]}{V(Y)} = 1 - \frac{V_{\mathbf{X}_{-i}}[E_{X_i}(Y|\mathbf{X}_{-i})]}{V(Y)}. \tag{3}$$

TS_i is the ratio of the remaining uncertainty of the output to the unconditional variance $V(Y)$ when the true values of all inputs except X_i are known. It measures the total effect comprising of first order, interaction and higher order effects involving X_i, i.e.,

$$TS_i = S_i + \sum_{j:j\neq i}^{n} S_{ij} + \ \ldots \ + S_{1\ldots i\ldots n}. \tag{4}$$

The underlying theory of the above variance-based measures is related to the decomposition of the function $f(\mathbf{X})$ (Sobol, 1993):

$$\begin{aligned} f(\mathbf{x}) &= f_0 + \sum_{i=1}^{n} f_i(X_i) + \sum_{1\leq i<j\leq n} f_{ij}(X_i, X_j) + \ \ldots \ + f_{12\ldots n}(X_1, \ldots, X_n) \\ &= f_0 + \sum_{s=1}^{2^n-1} f_{I_s}(\mathbf{X}_{I_s}), \end{aligned} \tag{5}$$

$$I = 1 \leq i_1 < \ldots < i_s \leq n, \quad s = 1, 2, \ldots, 2^n - 1, \tag{6}$$

where

$$f_0 = E(Y) \tag{7}$$

$$f_i(X_i) = E_{\mathbf{X}_{-i}}(Y|X_i) - f_0 \tag{8}$$

$$f_{ij}(X_i, X_j) = E_{\mathbf{X}_{-ij}}(Y|X_i, X_j) - f_i - f_j - f_0 \tag{9}$$

and so on.

The decomposition in Equation (5) holds true if the model function $f(\mathbf{x})$ is square-integrable and defined over the n-dimensional unit hypercube:

$$K^n = \{\mathbf{X} \mid 0 \leq X_i \leq 1, i = 1, \ldots, n\} \tag{10}$$

and the integral of every summand over any of its variables is zero, i.e.

$$\int_0^1 f_{I_s}(\mathbf{X}_{I_s})\, \mathrm{d}X_k = 0, \quad k \subset I, \tag{11}$$

where I_s is a subset of input variables. For statistically independent $X_1, \ldots, X_n$, the terms in Equation (5) are mutually orthogonal and the decomposition is unique. Additionally, with independent inputs, the unconditional variance $V(Y)$ can be decomposed in the same way as the model function, i.e.,

$$V(Y) = \sum_{i=1}^{n} V_i + \sum_{1 \leq i < j \leq n} V_{ij} + \ldots + V_{12\ldots n} \tag{12}$$

with

$$\begin{aligned} V_{i_1 \ldots i_s} &= V(\mathbf{f}(X_{i_1}, \ldots, X_{i_s})) \\ &= V[E(\mathbf{f}|X_{i_1}, \ldots, X_{i_s})], \quad 1 \leq i_1 < \ldots < i_s \leq n, \end{aligned} \tag{13}$$

where the conditional expectation is taken over all X_j not in I_s and the variance is computed over the range of possible values of I_s, $s = 1, 2, \ldots, 2^n - 1$. Dividing both sides of Equation (12) by $V(Y)$, we obtain:

$$1 = \sum_{i=1}^{n} S_i + \sum_{1 \leq i < j \leq n} S_{ij} + \ldots . + S_{12\ldots n} \tag{14}$$

With the identity in Equation (14) and the definitions of S_i and TS_i in Equations (1) and (3), it is easy to verify that, if interaction effects exist,

$$\sum_{i=1}^{n} S_i < 1 \tag{15}$$

and that the difference

$$1 - \sum_{i=1}^{n} S_i \tag{16}$$

is an indicator of the presence of interaction effects among features. It can also be concluded that

$$\sum_{i=1}^{n} TS_i \geq 1. \tag{17}$$

Saltelli et al (2010) suggested that for independent $X_1, \ldots, X_n$, one can avoid to explicitly include the probability distribution function of each X_i in the construction of f_i, $f_{ij}, \ldots$, etc.Therefore and without loss of generality, all input variables can be conceived as defined in K^n and the mapping from K^n to the actual probability distribution of X_i is intended to be part of the definition of f_i.

3.1 Monte Carlo Estimation

We construct in this section a Monte Carlo algorithm for the numerical estimation of Sobol sensitivity indices. The Monte Carlo procedure allows the estimation of both sets of indices S_i and TS_i using a single set of random samples generated from the assumed probability distributions of $\mathbf{X}$. The general framework for the numerical algorithm is as follows:

1. For a given design matrix $\mathbf{D}$ of size $M \times n$ and a vector of output (target) values $\mathbf{y}$ of size $M \times 1$, use an appropriate learning algorithm to train the model and obtain predictions $f(\mathbf{D})$. Here M is the number of realizations of features and of the target variable and n is the number of features.

2. For the assumed probability distributions, generate two independent samples $\mathbf{A}$ and $\mathbf{B}$ of $\mathbf{X} = (X_1, \ldots, X_n)$ each of size $N \times n$. N is the number of simulated observations of $\mathbf{X}$, $N \geq M$.

3. For the i-th feature, construct a matrix A_B^i, consisting of all columns of A except the i-th column which is taken from matrix B.

4. Estimate V_i and TV_i , the numerators of S_i and TS_i in Equations (1) and (3), respectively, and $V(Y)$ by the following formulas (Saltelli et al, 2010):

$$\hat{V}_i = \frac{1}{N} \sum_{r=1}^{N} f(\mathbf{B})_r [f(\mathbf{A}_B^i)_r - f(\mathbf{A})_r] \tag{18}$$

$$\hat{TV}_i = \frac{1}{2N}\sum_{r=1}^{N}[f(\mathbf{A}_B^i)_r - f(\mathbf{A})_r]^2 \tag{19}$$

$$\hat{V}(Y) = \frac{1}{N}\sum_{r=1}^{N} f(\mathbf{A})_r f(\mathbf{B})_r - f_0^2 \tag{20}$$

where r is the r^{th} row of $\mathbf{A}$ and $\mathbf{B}$ and

$$f_0 = \frac{1}{N}\sum_{r=1}^{N} f(\mathbf{A})_r. \tag{21}$$

5. Use Equations (1) and (3) to obtain the estimated values for first-order and total effect indices, $\hat{S}_i$ and $\hat{TS}_i$.
6. For a given k, select the features with the highest k % first-order or total effect scores. k depends on the desired accuracy of the model.

The number of model evaluations required by the above estimation algorithm for the combined set of $\hat{S}_i$ and $\hat{TS}_i$ is $N(n + 2)$. A larger N produces more accurate estimates. Another important aspect of the computation is the random sampling of the two matrices A and B. The Monte Carlo method makes use of quasi-random (QR) points of the Sobol sequences (Bratley and Fox, 1988). QR sequences possess the desired uniformity property and optimally fill the unit hypercube K^n in a fashion that avoids inhomogeneity of selected points. Sobol sequences possess the additional favored property of being low discrepancy points and they perform well in existing QR methods. More details about the characteristics and generation of Sobol QR sequences can be found in Saltelli et al (2010).

4 Numerical Example

The above described computational algorithm of Sobol sensitivity indices is applied here to a published data set containing 100 independent variables (features) and one output (target) variable (source: `https://drive.google.com/file/d/0ByPBn4rtMQ5HaVFITnBObXdtVUU/view`). The features represent various profiles of certain stocks while the target variable represents a rise in the stock price (1) or a drop in stock price (−1).

The data set includes 3000 observations. The model was first trained by a Random Forest (RF) utilizing 67 % of the data set and the model was validated on the remaining 33 % of the observations. The reported accuracy of the validated model – measured by the Area Under Curve (AUC) – was 0.4523. Using the mean decrease in the Gini index measure for feature importance (Clarke et al, 2009), the top 20 features were selected and the model was retrained with these most important 20 features. The resulting validated prediction accuracy was 0.4767 (5.4 % increase in accuracy). We followed the same learning algorithm and applied instead the global sensitivity analysis to select the top 20 features. We used $N = 15000$ simulations. A filter method based on the point-biserial correlation coefficient of each feature with the binary target variable is added to the analysis. The results are shown in the following Figure 4 and Table 1.

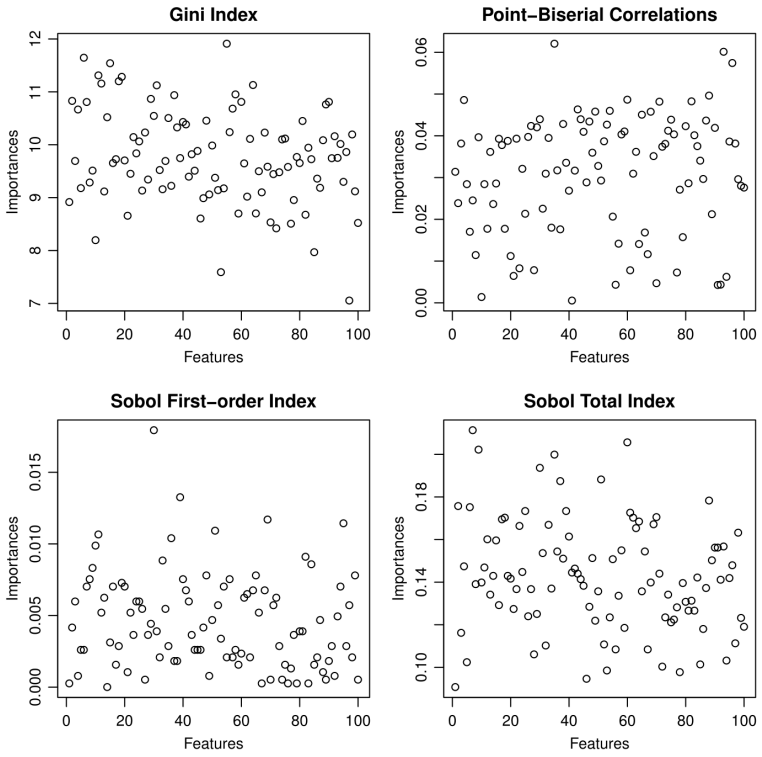

Figure 4: Scatter plots of feature importance values for the stock data based on four different measures.

Figure 4 depicts scatter plots of the feature relevance (importance) values computed based on mean decrease in the Gini index (as reported in the source article), point-biserial correlation coefficient and Sobol sensitivity indices versus feature index. A quick examination of the scatter plots reveals that the first-order Sobol sensitivity index S_i has the strongest discrimination ability of the most important features, followed by biserial correlation and total sensitivity index TS_i. The three measures possess more dispersed scatters in the upper range as compared to the Gini index.This is confirmed by the values of the coefficient of variation which are from highest to lowest: 73 % for S_i, 45 % for biserial correlation, 37 % for TS_i and 9 % for Gini index.

Table 1: Prediction accuracy of reduced model.

Selection Method	**Top 20 features AUC**	**% accuracy increase (baseline 0.4523)**
Mean Decrease Gini	0.4767	5.4
Biserial correlation	0.5170	14.3
First-order Sobol effect	0.5185	14.6
Total Sobol effect	0.5207	15.2

Table 1 clearly indicates that the Sobol sensitivity indices produce the highest predictive accuracy when the model is trained on the top 20 selected features.

5 Conclusion

In this paper a supervised feature selection approach is proposed based on the global sensitivity analysis. Feature importance is measured by the first-order and total Sobol sensitivity indices that quantify the contribution of individual features and their interactions to the overall variance of the target variable. A computational algorithm of the two sensitivity indices was described and applied to a publicly available data set. It is shown that Sobol sensitivity indices identify the most important features more distinctly than other existing feature selection techniques and provide a higher predictive accuracy.

The classical Sobol-based sensitivity analysis presented in this paper assumes statistically independent features. In subsequent work, the authors wish to extend the analysis to include dependent features and also compute sensitivity indices for subsets of features.

Acknowledgements The authors gratefully acknowledge the support of the American University of Sharjah, United Arab Emirates.

References

Aliferis C, Statnikov A, Tsamardinos I, Mani S, Koutsoukos X (2010) Local causal and Markov Blanket induction for causal discovery and feature selection for classification. Journal of Machine Learning Research 11:171–234.

Bratley P, Fox B (1988) Sobol's quasirandom sequence generator. Association for Computing Machinery Transactions on Mathematical Software 14:88–100.

Breiman L (2001) Random forests. Machine Learning 45(1):5–32, Kluwer Academic Publishers. DOI: 10.1023/A:1010933404324.

Clarke B, Fokoué E, Zhang H (2009) Principles and theory for data mining and machine learning, 1st edn. Springer, New York. DOI: 10.1007/978-0-387-98135-2.

Efron B, Stein C (1981) The Jackknife estimate of variance. The Annals of Statistics 9(3):586–596.

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. Journal of Machine Learning Research 3:1157–1182.

Hocking R, Leslie R (1967) Selection of the best subset in regression analysis. Technometrics 9:531–540.

Hoerl R A.E.and Kennard (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67.

Janecek A, Gansterer W, Demel M, Ecker G (2008) On the relationship between feature selection and classification accuracy. Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery at ECML/PKDD 2008, Proceedings of Machine Learning Research (PMLR) 4:90–105.

Maio J, Niu L (1981) A Survey on Feature Selection. DOI: 10.1016/j.procs.2016.07.111.

Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M, Tarantola S (2010) Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Computer Physics Communications 181(2):259–270. DOI: 10.1016/j.cpc.2009.09.018.

Scheffé H (1959) The analysis of variance. Wiley series in probability and mathematical statistics. Probability and mathematical statistics, John Wiley & Sons, New York.
Sobol I (1990) On sensitivity estimation for nonlinear mathematical models (in Russian). Matematicheskoe Modelirovanie 2:112–118.
Sobol I (1993) Sensitivity estimates for nonlinear mathematical models. Mathematical Modeling and Computational Experiment 1:407–414.
Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B (Methodological) 58(1):267–288.