Sobol Sensitivity: A Strategy for Feature Selection

Dmitry Efimov and Hana Sulieman

Abstract In this paper we propose a novel approach for feature selection in machine learning. The approach is based on the Sobol sensitivity analysis, a variance-based technique that determines the contribution of each feature and their interactions to the overall variance of the target variable. Similar to wrappers, Sobol sensitivity is a model-based approach that utilizes the trained model to evaluate feature importances. It uses the full feature set to train the model just as embedded methods do. Based on the trained model, it evaluates importance scores and, similar to filters, identifies the subset of important features with highest scores without retraining the model. The distinctive characteristic of the Sobol sensitivity approach is its computational efficiency compared to the existing feature selection algorithms. This is because importance scores for all individual features and subsets of features are calculated with the same trained model. We apply the proposed algorithm to a simulated data set and to four benchmark data sets used in machine learning literature. The results are compared to those obtained by two of the widely used feature selection algorithms and some computational aspects are also discussed.

Keywords Feature selection · Sobol index · Sensitivity analysis · Machine learning

Mathematics Subject Classification (2010): Primary 62P07 · Secondary 68U04

1 Introduction

The problem of variable (feature) selection in predictive modelling has received considerable attention during the past 10 years in both statistics and machine learning

D. Efimov · H. Sulieman (✉)
Department of Mathematics and Statistics, American University of Sharjah,
P.O. Box 26666, Sharjah, UAE
e-mail: hsulieman@aus.edu

D. Efimov
e-mail: defimov@aus.edu

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_4

literatures. The aim of feature selection is to identify the subset of predictor variables that provides a reliable and robust model for a given target variable. Feature selection plays a central role in many areas such as natural language processing, gene expression array studies, computational biology, image recognition, information retrieval, temporal modelling, consumer profile analysis and business data analytics. Curse of dimensionality in data collected in these and other areas and the increased level of noise in the associated features have motivated the development of various feature selection techniques. Feature selection is a key mechanism to reduce a large number of variables to relatively few.

In this article, a new approach for feature selection is proposed. The new approach is inspired by the popular Sobol sensitivity measure developed by I.M. Sobol in 1990 [18]. Sobol sensitivity measure is a variance-based sensitivity technique that decomposes the output (target) variable variance into summands of variances of the input variables (features) in an increasing dimensionality. It has been widely used in assessing global sensitivity of models in different fields such as environment, economics, engineering and many others. The approach is a model-based technique that utilizes the fitted model to compute the partial variances or variance contributions by each feature and their interactions to the overall variance of the target variable. It is shown to have lower computational cost than many existing feature selection algorithms but its effectiveness depends on the quality of the fitted model as it is the case for some popular algorithms.

The article is organized as follows. Section 2 provides a review of some existing methods on variable selection. Section 3 proposes Sobol sensitivity approach for variable selection and gives some theoretical foundation of the measure. It also discusses some computational aspects of the method. Section 4 applies the proposed Sobol sensitivity to several data sets used in machine learning literature and provides comparisons with some existing benchmark algorithms.

2 Literature Review

In this section we summarize some popular feature selection techniques, give some computational consideration and provide the motivation for the proposed new technique.

Feature selection techniques can be divided into three general frameworks: *wrappers*, *filters* and *embedded methods* [21, 25]. Wrappers evaluate the predictive power of subsets of features by retraining the model for different feature subsets. Filters evaluate the importance of features before the main prediction algorithm is trained. Embedded methods search for the optimal subset of features simultaneously with minimizing a loss function. What follows is a brief description of each framework:

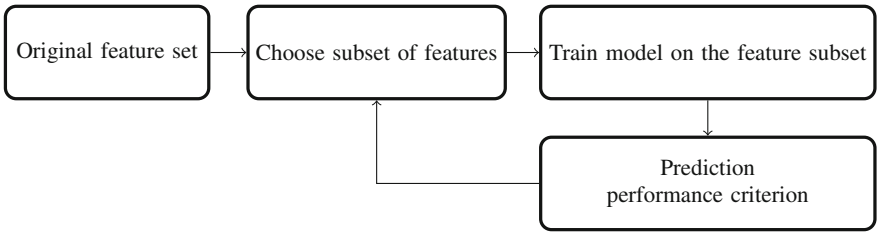

Fig. 1 Wrappers framework

1. Wrappers.
   Wrappers are model-based methods for feature selection and are considered to be the most effective and computationally intractable algorithms. Figure 1 shows the main principle of the wrapper methods' framework.
   Basically, to find the most relevant and informative subset of features, the prescribed model is trained for different subsets of features. The subset with the highest score on a particular prediction performance criterion is selected as best set of features. Because wrapper methods utilize the model algorithm they are considered more effective and hence more desired than filters and embedded methods.
   However, wrapper methods are generally criticized for their potential to overfit the training data and for their computational cost. Overfitting occurs when the complete data set is used for the training and the model becomes excessively complex to fit the data too precisely but still provides poor predictions when applied to new data set. This occurs when there is insufficient data to train and the data does not fully recover the concept learned. Several approaches have been proposed in the literature to overcome model overfitting:

   a. Cross-Validation (CV): in this approach the data set is split into training and validation sets. The model is trained on the training set and predictions are obtained on the validation set. A variety of techniques are developed to determine the fraction of data that should be used for training and that used for validation. These techniques include random sub-sampling, leave-one-out, K-fold and other CV sampling mechanisms.
   b. Probabilistic approach: based on information theory principles. The prediction accuracy of the algorithm is measured using various techniques such as Akaike Information Criteria (AIC, [3, 4]), Bayesian Information Criteria (BIC, [5]) and others.

   As for computational cost, wrappers are deemed computationally expensive. For n features, the number of feature subsets is defined by $O(2^n)$, 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 value of n. The computational cost of wrapper methods can be reduced by using efficient search strategies to find the optimal subset of features.

One of the earliest attempt to improve the computational efficiency is due to Hocking and Leslie in [1]. Their method starts by fitting the full model and then features are eliminated based on the magnitudes of their t-statistics. The efficiency gain of the method lies in the fact that entire subsets of features are eliminated from further consideration when their reduced prediction error is greater than other subsets already evaluated. The method can assume independent features and works well when there is a small number of important features that dominate the target variable and can easily be identified.
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. Forward selection begins with no variables and progressively adds features until maximum reduction in prediction error is reached. The reverse of this strategy is the backward elimination which begins with full model and progressively removes features having smallest contributions. Once a feature is added in forward selection or eliminated in backward elimination, the operation can not be reversed. To overcome this drawback, stepwise selection is used. Stepwise selection starts by adding features until reaching some stopping criteria. Then the algorithm starts dropping features until reaching another stopping criteria and so on. While stepwise selection can reduce the computational cost of the best set of features it does not, however, guarantee the selection of the global optimal set.

2. Filters.
Filters evaluate feature importance as a pre-processing operation to model training as depicted in Fig. 2. The main difference between filters and wrappers is that filters do not use the training procedure to capture the relationship between features. Rather, 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. They are also more robust against overfitting than wrappers.
By using Pearson correlation, filters can only capture linear effects between features and target variable. The nonlinear effects are left undiscovered. A successful attempt to deal with nonlinear effects has been recently developed. Aliferis et al. [24] described Markov blanket technique that is based on Bayesian network. A Markov blanket of the target variable Y is defined as a minimal set of features on which all other features are conditioned so as they become independent of Y.

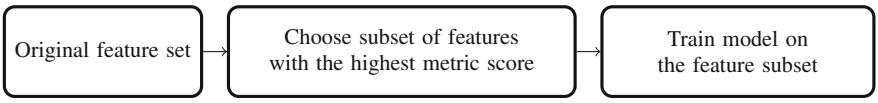

Fig. 2 Filters framework

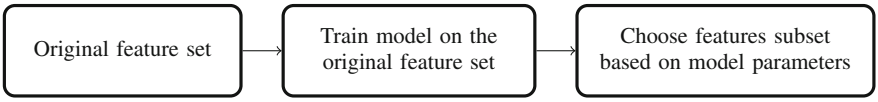

Fig. 3 Embedded methods framework

In terms of relevancy of selected features, Markov blanket is shown by Tsamardinos and Aliferis (2003) to provide the most relevant and optimal features in calibrated classifications in which the probability distribution of the target variable can be perfectly estimated with the smallest number of variables. The authors showed that neither filters nor wrappers are superior to one another in identifying the optimal features because filters lack universal optimality,i.e., independently of the classification algorithm and model-performance metric, and wrappers lack universal efficiency. Markov blanket technique does not suffer from these shortcomings.

3. Embedded methods.
   Embedded methods use training procedure to obtain feature rankings Fig. 3. The aim of embedded methods is two-fold: first, maximizing the prediction accuracy and second, minimizing the number of features in the predictive algorithm.
   Regularization methods such as Ridge Regression [2], Nonnegative Garrote [6], Least Absolute Selection and Shrinkage Operator (LASSO, [8]) are the most common forms of embedded methods. In these methods, the coefficients (weights) of the features are penalized by some regularization terms or forced to be exactly zero. Features with weights close to zero are then eliminated without compromising the prediction performance of the model. Analogy to filters, several developments have been achieved in embedded regularization methods. LASSO [15] and Elastic Net [11] are examples of methods that were developed to measure the importance of subsets of features. Boosted LASSO [13] and Smoothly Clipped Absolute Deviation (SCAD) [14] are examples of methods that use nonlinear regularization terms to produce more sparse and unbiased estimators of coefficients. Other embedded methods are based on decision tree algorithms. In this group of embedded methods, various decision trees are iteratively built using bootstrapping and for each tree, information gain (based on specific information entropy) is calculated for each feature. Features are then ranked based on the average information gain over all trees. Random Forest algorithm [7] is one popular example of decision tree methods. Louppe et al. in [17] provided comparative analysis of feature importance using various decision tree algorithms. Embedded methods can be disadvantaged for the fact that feature weights are often estimated iteratively not explicitly.

The reader is referred to the excellent reviews of feature selection methods found in [22, 23].

3 Sobol Sensitivity Approach

In this section we propose a new technique for feature selection in machine learning and provide some mathematical foundation of the algorithm. The proposed technique is based on variance decomposition principle of model output developed by Sobol (1990) [18] (in Russian) and Sobol (1993) [19] (in English). Sobol sensitivity analysis is intended to determine how much of the variability in model output is dependent upon each of the input variables, either upon a single variable or upon an interaction between different variables. The decomposition of the output variance in a Sobol sensitivity analysis employs the same principal as the classical analysis of variance in factorial experimental designs.

3.1 Theoretical Background

Let the function $f(\mathbf{x})$, where $\mathbf{x} = (x_1, ..., x_n)$ be defined on the unit n-dimensional cube

$$K^n = \{\mathbf{x} \mid 0 \leqslant x_i \leqslant 1, i = 1, ..., n\}.$$

Sobol's main idea is to decompose the function $f(\mathbf{x})$ into summands of increasing dimensionality, namely

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1 \leq i < j \leq n} f_{ij}(x_i, x_j) + ... + f_{12...n}(x_1, ..., x_n). \tag{1}$$

The decomposition in (1) holds true if f_0 is a constant and the integral of every summand over any of its variables is zero, i.e.

$$\int_0^1 f_{i_1...i_s}(x_{i_1}, ..., x_{i_s}) dx_{i_k} = 0, 1 \leqslant i_1 < .... < i_s \leq n, 1 \leqslant k \leqslant s, s = 1, 2, ...n.$$

For independent $x_1, ..., x_n$, all terms in Eq. (1) are orthogonal and f_0 can be calculated as:

$$f_0 = \int_{K^n} f(\mathbf{x}) d\mathbf{x} \tag{2}$$

which represents the average value (or expectation) of the function f. Sobol (1993) [19] showed that decomposition (1) is unique and all of its terms can easily be evaluated through multi-variable integrals.

Because of the orthogonality of the $\mathbf{x}$-space, the total variance D of $f(\mathbf{x})$ can also be partitioned in the same way as the original function, i.e.,

$$D = \sum_{i=1}^{n} D_i + \sum_{1 \le i < j \le n} D_{ij} + ... + D_{12...n} \tag{3}$$

where

$$D = \int_{K^n} f^2(\mathbf{x}) d\mathbf{x} - f_0^2$$

and

$$D_{i_1...i_s} = \int_0^1 ... \int_0^1 f^2_{i_1...i_s}(x_{i_1}, ..., x_{i_s}) dx_{i_1} ... dx_{i_s} \qquad \begin{array}{l} 1 \leqslant i_1 < .... < i_s \leq n, \\ s = 1, 2, ...n \end{array}$$

$D_{i_1...i_s}$ is the partial variance attributed to $x_{i_1}, ..., x_{i_s}$ defined by the variance of the conditional expectation of $f(\mathbf{x})$ conditioned on $x_{i_1}, ..., x_{i_s}$, namely,

$$D_{i_1...i_s} = Var[E(f|x_{i_1}, ..., x_{i_s})]$$

where the conditional expectation is taken over all x_j not in $\{i_1, ..., i_s\}$ and variance is computed over the range of possible values of $x_{i_1}, ..., x_{i_s}$.
The usefulness of $D_{i_1...i_s}$ as a measure of sensitivity is easy to grasp. Influential $x_{i_1}, ..., x_{i_s}$ control f significantly and so $E(f|x_{i_1}, ..., x_{i_s})$ will mimic f. In this case the total variance in f will be matched by the variability in $E(f|x_{i_1}, ..., x_{i_s})$ as $x_{i_1}, ..., x_{i_s}$ vary making $D_{i_1...i_s}$ large compared to the total variance D.

Sobol in [20] proposed the following indices to measure sensitivity of the function with respect to $x_{i_1}, ..., x_{i_s}$:

$$S_{i_1...i_s} = \frac{D_{i_1...i_s}}{D}, 1 \leqslant i_1 < .... < i_s \leqslant n, s = 1, 2, ...n \tag{4}$$

with $\sum S_{i_1...i_s} = 1$.

For $s = 1$, the sensitivity measure $S_{i_1} = S_i$ is called *first-order sensitivity index* which measures the fractional contribution of the individual variable x_i to the total variance of f. For $s = 2$, S_{ij} is called the *second-order sensitivity index* which measures the portion of the variability in f due to the interaction of x_i and x_j and so on. Total sensitivity index, defined as the sum of all sensitivity indices involving x_i up to the n-th order, i.e.,

$$TS_i = S_i + \sum_{j:j \neq i}^{n} S_{ij} + ... + S_{1....i...n} \tag{5}$$

was also proposed to quantify the overall effect of x_i on the model output.

Decomposition (1) or (3) has long history and was given in its general form by Efron and Stein [9]. More concisely, one can think of $f(\mathbf{x})$ as some statistics defined on the independent variables $x_1, ..., x_n$, then $f(\mathbf{x})$ may be decomposed into

a grand mean $f_0 = E[f(\mathbf{x})]$; i-th main effect $f_i(x_i) = E[f(\mathbf{x}|x_i)] - f_0$; ij-th interaction $f_{ij}(x_i, x_j) = E[f(\mathbf{x}|x_i, x_j)] - E[f(\mathbf{x}|x_i)] - E[f(\mathbf{x}|x_j)] + f_0$ and so on. Given these definitions, decomposition (1) follows immediately. The f_i's, f_{ij}'s,....are known in factorial experimental design as ANOVA-HDMR, where HDMR stands for High-Dimensional Model Representation.
For example, when $n = 2, f(\mathbf{x})$ can be decomposed into:

$$\begin{aligned} f(x_1, x_2) &= f_0 + f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2) = \\ &= f_0 + E[f(\mathbf{x}|x_1)] - f_0 + E[f(\mathbf{x}|x_2)] - f_0 + \\ &+ E[f(\mathbf{x}|x_1, x_2)] - E[f(\mathbf{x}|x_1)] - E[f(\mathbf{x}|x_2)] + f_0. \end{aligned}$$

The individual terms of decomposition (1) can easily be shown to have a zero mean. For example $E[f_i(x_i)] = E[E[f(\mathbf{x}|x_i)] - f_0] = E[f(\mathbf{x})] - f_0 = 0$. Decomposition (1) terms can also be shown to be mutually uncorrelated, implying that the unconditional total variance of $f(\mathbf{x})$, D, can simply be expressed as a sum of variances of these uncorrelated terms giving the variance decomposition (3) where

$$\begin{aligned} D_i &= Var(f_i(x_i)) = Var(E[f(\mathbf{x}|x_i)]) \\ D_{ij} &= Var(f_{ij}(x_i, x_j)) = Var(E[f(\mathbf{x}|x_i, x_j)]) + Var(E[f(\mathbf{x}|x_i)]) + \\ &+ Var(E[f(\mathbf{x}|x_j)]) \end{aligned}$$

and so on. It is noted that decomposition (3) is similar to the classical ANOVA decomposition without the residual error term.

If the relationship between $\mathbf{x}$ and the model output is additive linear, then a straightforward variance decomposition can be provided by regression analysis. It can be shown, in this case, that the first-order sensitivity index, S_i is equal to the squared standardized regression coefficients, i.e., $S_i = \beta_i^2$. That is, the β_i's give the fractional contribution of each predictor to the variance of the response variable. The effectiveness of β_i's as a measure of sensitivity, in this case, depends on the quality of the fitted model and the degree of linearity in the relationship between the response variable and predictors.

The definition of Sobol sensitivity index given in (3) can be extended to include group indices for subsets of variables and their joint sensitivity behaviour. Suppose the variables $x_1, ..., x_n$ are partitioned into r disjoint groups $\mathbf{x}^1, ..., \mathbf{x}^r$, $r < n$, then decomposition (1) can be expressed as:

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{r} f_i(\mathbf{x}^i) + \sum_{1 \leqslant i < j \leqslant r} f_{ij}(\mathbf{x}^i, \mathbf{x}^j) + ... + f_{1,2,.....,r}(\mathbf{x}^1, ..., \mathbf{x}^r).$$

For $r = 2$ for example, the variables $\mathbf{x}$ are partitioned into two groups $\mathbf{y}$ and $\mathbf{z}$, giving the following decomposition:

$$f(\mathbf{x}) = f_1(\mathbf{y}) + f_2(\mathbf{z}) + f_{12}(\mathbf{y}, \mathbf{z}).$$

The variances D_1 and D_2 for each of $\mathbf{y}$ and $\mathbf{z}$ are calculated as

$$D_1 = \int_0^1 \ldots \int_0^1 f_1^2(\mathbf{y}) d\mathbf{y}, \quad D_2 = \int_0^1 \ldots \int_0^1 f_2^2(\mathbf{z}) d\mathbf{z} \tag{6}$$

and

$$D = \int_0^1 \ldots \int_0^1 f^2(\mathbf{x}) d\mathbf{x} - f_0^2, \quad D_{12} = D - D_1 - D_2. \tag{7}$$

For this two-set decomposition, we define the following sensitivity index:

$$\begin{aligned} SI_{\mathbf{y}} &= \frac{1}{D}(D_1 + D_{12}) \\ SI_{\mathbf{z}} &= \frac{1}{D}(D_2 + D_{12}). \end{aligned} \tag{8}$$

In the next section, $SI_{\mathbf{y}}$ and $SI_{\mathbf{z}}$ will be referred to as Sobol Importance (SI) measure that will be used as the basis for feature selection mechanism.

In practice, variables are usually ranked based on the magnitude of their Sobol sensitivity indices, the higher the magnitude, the more influential respective variables are. Although no distinct cutoff value has been defined, the rather arbitrary value of 0.05 is frequently accepted for this type of analysis for distinguishing important from unimportant variables. It should be noted though that this value of 0.05 is primarily used for more complex models and it may be not stringent enough for relatively simple models that contain only few input variables.

3.2 Sobol Sensitivity for Machine Learning

3.2.1 General Framework

Let m be a number of samples in the dataset and n be a number of features (variables). Denote the set of feature indices as $J = \{1, \ldots, n\}$. For the purpose of feature selection in machine learning, we propose to partition the set of indices J into two subsets $J_1 = \{j_1, \ldots, j_s\}$ and $J_2 = \{j \in J \mid j \notin J_1\}$ and estimate the importance for the features from each group separately using Sobol sensitivity index given in Eq. (8). In many machine learning problem settings, splitting features into two groups is deemed sufficient for identifying important features. In the classical Sobol's sensitivity analysis, variables (features) are assumed independent and uniformly distributed over the interval [0, 1]. In our proposed analysis, we consider normally distributed features following the work of Arwade et al. (2010) [26] and continue to assume independent

features. The Monte-Carlo procedure [20] can be applied to evaluate the quantities from above (6) and (7). Assuming that X is an original design matrix we generate two new matrices Y and Z such that Y is obtained from X by random shuffling each column with index $j \in J_1$, Z is obtained by random shuffling each column with index $j \in J_2$. We denote $\mathbf{x}_i$, $\mathbf{y}_i$, $\mathbf{z}_i$ the *i-th* row of the matrices X, Y, and Z accordingly.

Based on Monte-Carlo exploration of the features' space, the quantities in (6) and (7) can be estimated as follows:

$$\begin{aligned} &f_0 = \tfrac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i), && D + f_0^2 \approx \tfrac{1}{m}\sum_{i=1}^{m} f^2(\mathbf{x}_i), \\ &D_1 + f_0^2 \approx \tfrac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i) f(\mathbf{z}_i), && D_2 + f_0^2 \approx \tfrac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i) f(\mathbf{y}_i), \\ &D_{12} = D - D_1 - D_2, \end{aligned} \tag{9}$$

where summation is taken by all dataset entries.

The suggested permutation procedure of the design matrix X allows the generation of random values from the assumed distribution of features for use in the Monte-Carlo algorithm. Sobol importance scores for the feature subsets J_1 and J_2 are calculated using (8).

3.2.2 Computational Aspects

The general framework described above gives rise to the following algorithm that calculates feature importance for subsets of features. The main output of this algorithm is the Sobol importance (SI) score for a given subset of features.

Algorithm 1 Sobol importance scores for feature subset

1: Let X be an $m \times n$ design matrix for the given dataset, y be a vector of outputs.
2: Train the model M on the original dataset X, y and obtain predictions p on the dataset X
3: Evaluate $f_0 = \dfrac{1}{m}\sum_{i=1}^{m} p_i$ and $D = \dfrac{1}{m}\sum_{i=1}^{m} p_i^2 - f_0^2$
4: Define a feature subset of interest

$$J_1 = \{j_1, \ldots, j_s\}$$

and complimentary feature subset

$$J_2 = \{j \in \{1, 2, \ldots, n\} \,|\, j \notin J_1\}$$

5: Create matrix Y from X by random shuffling columns with indices $j \in J_1$ and matrix Z from X by random shuffling columns with indices $j \in J_2$
6: Use model M with design matrix Z as an input to obtain D_1 using Eq. (9)
7: Use model M with design matrix Y as an input to obtain D_2 using Eq. (9)
8: Evaluate $D_{12} = D - D_1 - D_2$
9: Compute Sobol importance score for the subset J_1 using Eq. (8):

$$SI_{J_1} = \frac{1}{D}(D_1 + D_{12})$$

As mentioned earlier, the main challenge in feature selection algorithms is the high computational cost due to huge number of subsets that need to be investigated. With Sobol sensitivity approach, the importance of both individual features and subsets of features can be computed using the same Monte Carlo integral. The next algorithm utilizes this computational efficiency of the approach and calculates importance score based on the total sensitivity index given in Eq. (5) up to second-order interactions. In many application areas, second order interactions are deemed sufficient to capture the joint sensitivity behaviour of features. To calculate the Sobol importance score SI_i for individual features, the subset J_1 in Algorithm 1 is set to $J_1 = \{i\}, i = 1, 2, \ldots, n$ and for joint importance score SI_{ij}, $J_1 = \{i, j\}, j = 1, 2, \ldots, i-1, i+1, \ldots, n$.

Algorithm 2 Feature selection based on total second order Sobol importances

1: Initialize the $n \times n$ matrix $\mathbf{S}$ of zeros
2: **for** i=1 **to** n **do**
3: **for** j=i **to** n **do**
4: **if** j == i **then**
5: Using Algorithm 1 calculate the individual importance SI_i of feature with index i and assign it to the diagonal elements of $\mathbf{S}$, i.e. $\mathbf{S}_{ii} = SI_i$
6: **else**
7: Using Algorithm 1 calculate the importance SI_{ij} of features with indices i, j and assign it to $\mathbf{S}_{ij}$
8: Sort the features based on the total second order sensitivity indices given by:

$$TSI_i = \sum_{j=1}^{n} \mathbf{S}_{ij}$$

9: For a given k, select the features with the highest k-TSI scores. k depends on the desired accuracy of the model.

Algorithm 2 requires $\frac{n(n+1)}{2}$ score evaluations for n features. All these evaluations are completed using the same Monte Carlo integral.

Similar to wrappers, Sobol sensitivity is a model-based approach that utilizes the trained model to evaluate feature importances. While wrappers select a subset of features to train the model, Sobol sensitivity uses the full feature set to train the model just as embedded methods do. Based on the trained model, it evaluates importance scores and, similar to filters, it identifies the subset of important features with highest scores without retraining the model. As the case for filters and wrappers, the optimality and efficiency of the technique depend on the training algorithm (learner) and/or model-performance metric used. Sobol sensitivity assumes normally distributed and statistically independent features. These two distributional assumptions are popular in many feature selection algorithms. It can, however, consider other feature probability distributions and can be implemented for statistically dependent features. Because it is variance-based measure, it can be applied for linear and nonlinear relationships between target variable and features. In terms of computational needs,

Sobol sensitivity approach can be considered one of the most tractable techniques because importance scores for all feature subsets are computed using the same Monte Carlo integral.

4 Application and Comparisons

In this section, we apply the proposed feature selection techniques to several data sets known in machine learning community and presents comparative evaluations of the results obtained with results obtained using: Random Forest (RF) and Support Vector Machine Recursive Feature Elimination (SVM-RFE).

Example 1: *Effect of noise on Sobol importance*.

In this example we demonstrate the behavior of Sobol importance approach under different levels of noise in the model starting from zero-level noise model (100 % accurate predictions). We consider a model function given by Friedman in [28]:

$$f(x_1, x_2, x_3, x_4, x_5) = 10\sin(\pi x_1 x_2) + 20\left(x_3 - \frac{1}{2}\right)^2 + 10x_4 + 5x_5 + \\ +\sigma h(x_6, x_7, ..., x_{15}). \tag{10}$$

We generate the dataset with 1000 training examples and 15 features (drawn from the normal distribution). Only first 5 features are important. Figure 4 shows the results of Algorithm 2 for different values of σ. Using 0.05 as the cutoff value to declare importance, it is easily seen that the first 5 features continue to be the important features for $\sigma \leq 1$. Once σ is inflated beyond 1, more features exhibit themselves as important. However, for all σ values (except $\sigma = 2.5$, the overall ranking of features continues to agree with the first five features being the subset of features demonstrating highest importance scores.

Example 2: *Comparisons using simulated data*.

Friedman model function used in example 1 is used here to generate a data set with 1000 training examples and 15 features drawn from the normal distribution, only first 5 features are used to calculate the function in (10). The values of the function are used as values of the target variable y.

We calculate Sobol importances based on four different models: Neural Network (NN), Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting Trees model (XGB). We compare the results with importances obtained using Random Forest (RF) and Support Vector Machine Recursive Feature Elimination (SVM-RFE). Figure 5 depicts the findings. All algorithms correctly identify the first five features as the important set of features. One exception is observed in SVM-RFE where features 12 and 14 are identified as equally important. This wrong feature identification can be due to nonlinearity in the relationship between target variable and features for which SVM-RFE can not capture.

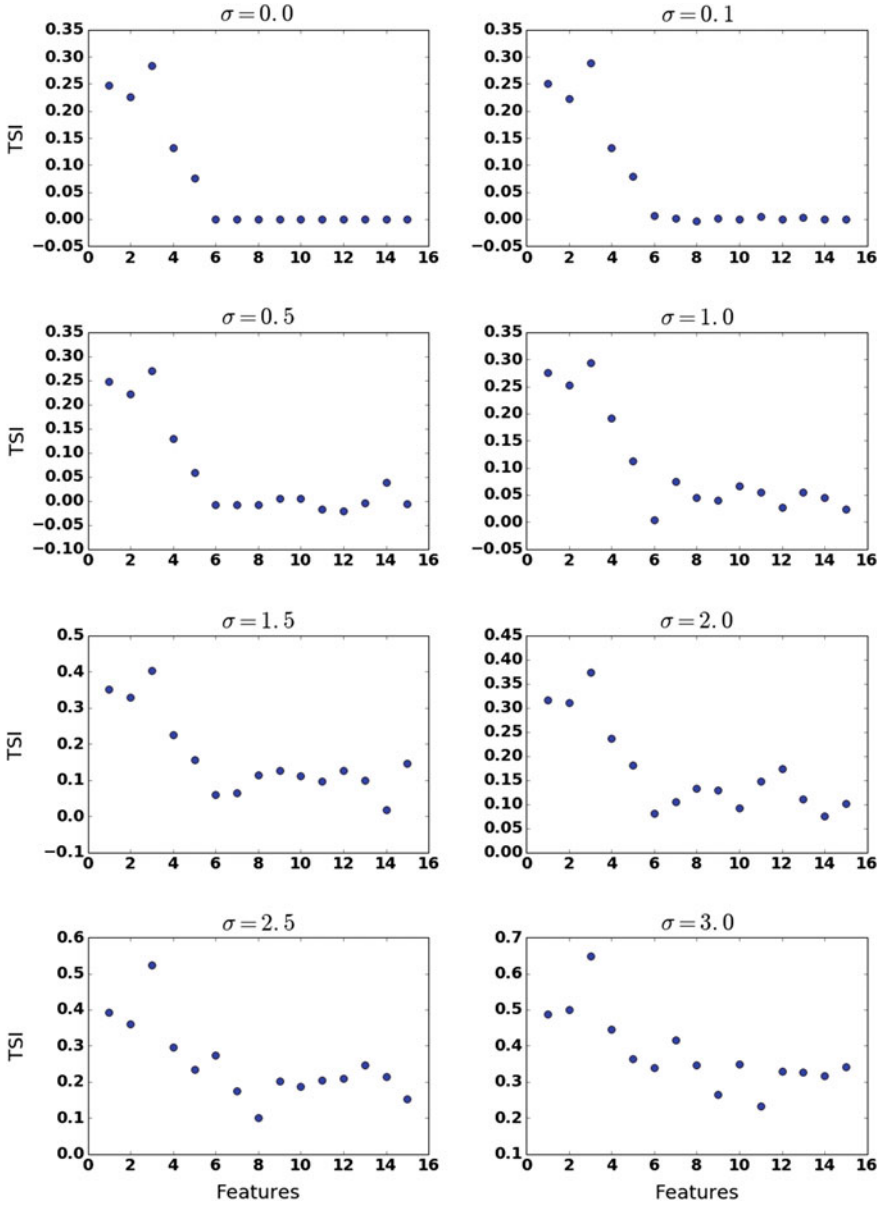

Fig. 4 Sobol importances for Friedman function with noise

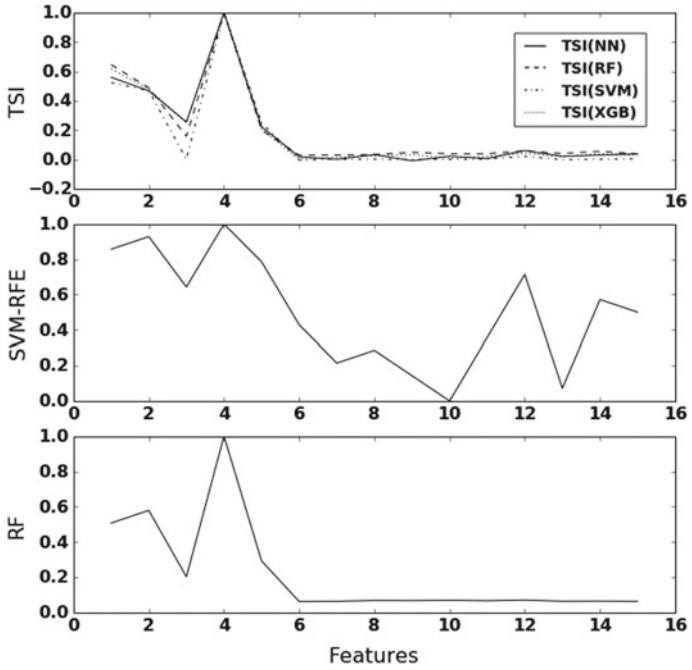

Fig. 5 Comparisons using simulated Friedman dataset

Example 3: *Comparisons using benchmark data sets*.

In this example, four benchmark data sets, described in Table 1, are used to evaluate the performance of Sobol sensitivity as compared to SVM-RFE and RF algorithms.

The model is trained using NN, RF and SVM methods and Sobol importance scores are calculated for all features in each data set using the three training methods. The results are compared to importance measures obtained using SVM-RFE and RF benchmark methods. For more meaningful comparisons, different threshold values for the number of important features selected from the benchmark importances. For example, threshold 10 means that we take top 10 important features from the benchmark algorithms. To compare the results we use Area Under the Curve (AUC) metric. The AUC calculates the overall differences in the feature rankings obtained by the benchmark method for a given threshold and those obtained from the Sobol importance scores. The general work flow is visualized in Fig. 6.

The resulting AUC values for the different models and the four datasets are plotted against different threshold values in Fig. 7. The Fig. 7 demonstrates that the different algorithms have succeeded in identifying comparable sets of important features. For example, the AUC between RF and Sobol RF feature rankings is higher than 0.9, which means that more than 90 % of features are common between the two algorithms.

Table 1 Datasets description

Set	Domain	Num. var.	Num. samples	Target	Data type	Ref.
SYLVA	Ecology	216	14394	Ponderosa pine	Continuous and discrete	WCCI 2006 Performance Prediction Challenge
HIVA	Drug discovery	1617	4229	Activity to AIDS HIV infection	Discrete (binary)	WCCI 2006 Performance Prediction Challenge
NOVA	Text	16969	1929	Separate politics from religion topics	Discrete (binary)	WCCI 2006 Performance Prediction Challenge
BANK	Financial	147	7063	Personal bankruptcy	Continuous and discrete	Foster and Stine [29]

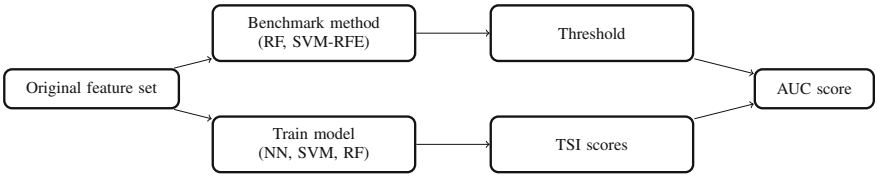

Fig. 6 Feature importances comparison framework

Furthermore, for each algorithm in Fig. 7, the top N% important features are selected and the model is trained by SVM algorithm on the selected subset of features. Table 2 presents the reduced model accuracy values expressed by the Root Mean Square Errors (RMSE). Reported in Table 2 also are the RMSE values for the trained model on all features in each data set.

Table 2 demonstrates that for all benchmark data sets, the reduced models give significantly more accurate predictions than that given by the full model. When top 10 % important features are used in the model, the Sobol approach gives better result than the benchmark algorithms SVM-RFE and RF. When including more than 10 % important features, SVM-RFE performs marginally better than Sobol approach in two of the data sets. In calculating variance contributions of features to the overall variability in the target variable, Sobol sensitivity approach can identify the small number of most important features more accurately than other methods. When larger number of features are desired in the model, the approach may fail to provide most accurate predictions due to the increased level of noise in the data. According to the analysis of Example 1, greater level of noise can distort the Sobol rankings of

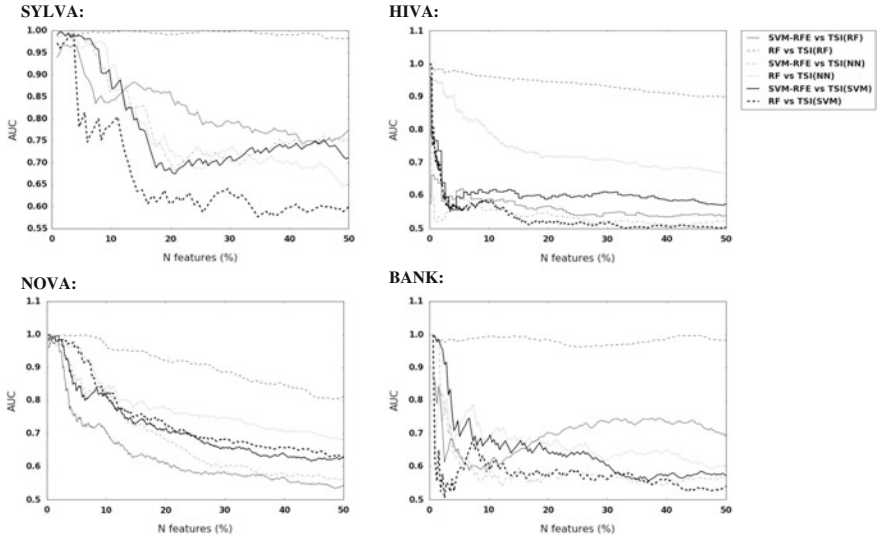

Fig. 7 Comparisons using benchmark data sets

important features. It may also imply that higher order interactions between features are needed for more accurate predictions in this situation.

The performance consistency of the three techniques measured by the standard deviation of model accuracy over N values was also examined. The standard deviation values, s, reported in Table 2 clearly exhibits comparable consistency between the three methods.

In summary, for the given data sets the total Sobol second order indices *TSI* produce the most accurate model predictions for the majority of cases. By varying the values of N, the *TSI* model accuracy shows small fluctuations (measured by s values) that are comparable to fluctuations observed by SVM-RFE or RF benchmark methods.

5 Conclusion

In this paper we implemented Sobol sensitivity analysis to select important features for the supervised data mining problem. We have proposed two algorithms for importance scoring: one algorithm to compute importance scores for the individual features and another one to compute importance scores for subsets of features. The main advantage of our proposed approach is lower computational cost and higher efficiency compared to many other existing algorithms. It can be applied for all types of relationships (linear or nonlinear) between target variable and features. A concern about the algorithm is that it estimates the importance of features with respect to the

Table 2 Reduced datasets model accuracy

N (top N % features selected)	SVM-RFE	RF	TSI(RF)	TSI(NN)	TSI(SVM)
SYLVA: RMSE (all features) = 0.144					
10 %	0.103	0.09	0.09	**0.085**	0.091
20 %	0.104	**0.088**	0.094	0.096	0.095
30 %	0.104	0.1	**0.094**	0.102	0.099
40 %	0.106	0.113	0.109	**0.102**	**0.102**
50 %	0.117	0.118	0.117	0.113	**0.11**
s	0.005	0.012	0.01	0.009	0.006
HIVA: RMSE (all features) = 0.179					
10 %	**0.163**	0.165	0.166	0.168	**0.163**
20 %	**0.162**	0.168	0.168	0.168	0.165
30 %	**0.161**	0.175	0.173	0.173	0.168
40 %	**0.165**	0.177	0.18	0.175	0.17
50 %	**0.168**	0.181	0.182	0.176	0.17
s	0.003	0.006	0.006	0.003	0.003
BANK: RMSE (all features) = 0.252					
10 %	0.227	0.235	0.24	0.225	**0.213**
20 %	**0.21**	0.236	0.239	0.224	0.217
30 %	**0.213**	0.229	0.238	0.223	0.216
40 %	**0.216**	0.229	0.23	0.225	0.222
50 %	**0.217**	0.226	0.221	0.23	0.226
s	0.006	0.004	0.007	0.002	0.005
NOVA: RMSE (all features) = 0.242					
10 %	0.256	0.275	0.27	**0.24**	0.28
20 %	0.24	0.241	0.259	**0.221**	0.26
30 %	0.237	0.24	0.253	**0.219**	0.248
40 %	0.234	0.24	0.246	**0.217**	0.238
50 %	0.236	0.244	0.246	**0.218**	0.233
s	0.008	0.01	0.009	0.009	0.017

model objective function which means that if the modeling algorithm is not accurate or overfitting, the Sobol approach may give misleading feature importances. The authors intend to further investigate the robustness of the approach to the training algorithm. As Example 3 has shown, the accuracy of the reduced model is higher than the full feature set model. It is then possible to train the model on a subset of features identified by a pre-processing algorithm and use the reduced model to compute Sobol feature importances. In addition to increasing Sobol importance reliability and efficiency, using reduced model to calculate importances reduces the computational

cost of the method. Another possible approach to reduce the computational cost is using Kolmogorov representation theorem [27] in which the model objective function can be expressed in an additive form of sub-functions, each as a single-variable function. If the hypothesis of additive model is true, then the Sobol algorithm can be simplified so as to require $2n$ model evaluations only.

Acknowledgments The authors wish to acknowledge the support of the American University of Sharjah, United Arab Emirates.

References

1. Hocking, R.R., Leslie, R.N.: Selection of the best subset in regression analysis. Technometrics **9**, 531–540 (1967)
2. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)
3. Akaike, H.: Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika **60**(2), 255–265 (1973)
4. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **AC-19**, 6, 716–723 (1974)
5. Schwarz, G.: Estimating the dimension of a model. Ann. Statis. **6**(2), 461–464 (1978)
6. Breiman, L.: Better subset regression using the nonnegative garrote. Technometrics **37**(4), 373–384 (1995)
7. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
8. Tibshirani, R.J.: Regression shrinkage and selection via the lasso. J. Roy. Statis. Soc. **58**(1), 267–288 (1996)
9. Efron, B., Stein, C.: The Jackknife estimate of variance. Ann. Statis. **9**, 586–596 (1981)
10. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Statist. **32**(2), 407–499 (2004)
11. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Statis. Soc. **67**(2), 301–320 (2005)
12. Zou, H.: The adaptive lasso and its oracle properties. J. Am. Statis. Assoc. **101**, 476, 1418–1429 (2006)
13. Zhao, P., Yu, B.: Stagewise lasso. J. Mach. Learn. Res. **8**, 2701–2726 (2007)
14. Zhang, H.H., Ahn, J., Lin, X., Park, C.: Gene selection using support vector machines with non-convex penalty. Bioinformatics **22**(1), 88–95 (2006)
15. Yuan, M., Lin, Y.: On the non-negative garrotte estimator. J. Roy. Statis. Soc. **69**(2), 143–161 (2007)
16. Ishwaran, H.: Variable importance in binary regression trees and forests. Electron. J. Statis. **1**, 519–537 (2007)
17. Louppe, G., Wehenkel, L., Sutera, A., Geurts, P.: Understanding variable importances in forests of randomized trees. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.). Advances in Neural Information Processing Systems vol. 26, pp. 431–439. Curran Associates, Inc. (2013)
18. Sobol, I.M.: On sensitivity estimation for nonlinear mathematical models (in Russian). Matematicheskoe Modelirovanie **2**, 112–118 (1990)
19. Sobol, I.M.: Sensitivity estimates for nonlinear mathematical models. Math. Model. Comput. Exper. **1**(4), 407–414 (1993)
20. Sobol, I.M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. **55**, 271–280 (2001)

21. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**(1), 1157–1182 (2003)
22. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.: Feature Extraction: Foundations and Applications. Springer-Verlag, Berlin (2006)
23. Clarke, B., Fokoué, E., Zhang, H.H.: Principles and Theory for Data Mining and Machine Learning. Springer-Verlag, Berlin (2009)
24. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and Markov Blanket induction for causal discovery and feature selection for classification. J. Mach. Learn. Res. **11**, 171–234 (2010)
25. Janecek, A.G.K., Gansterer, W.N., Demel, M.A., Ecker, G.F.: On the relationship between feature selection and classification accuracy. In: JMLR: Workshop and Conference Proceedings, vol. 4, pp. 90–105 (2008)
26. Arwade, S.R., Moradi, M., Louhghalam, A.: Variance decomposition and global sensitivity for structural systems. Eng. Struct. **32**, 1–10 (2010)
27. Kolmogorov, A.N.: On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. Amer. Math. Soc. Transl. **17**, 369–373 (1961)
28. Friedman, J.: Multivariate adaptive regression splines. Ann. Statis. **19**(1), 1–67 (1991)
29. Foster, D.P., Stine, R.A.: Variable selection in data mining: building a predictive model for bankruptcy. J. Amer. Stat. Assoc. **99**, 303–313 (2004)