ORTHOGONAL VARIANCE DECOMPOSITION BASED FEATURE SELECTION

FIRUZ KAMALOV

ABSTRACT. Existing feature selection methods fail to properly account for interactions between features when evaluating feature subsets. In this paper, we attempt to remedy this issue by using orthogonal variance decomposition to evaluate features. The orthogonality of the decomposition allows us to directly calculate the total contribution of a feature to the output variance. Thus we obtain an efficient algorithm for feature evaluation which takes into account interactions among features. Numerical experiments demonstrate that our method accurately identifies relevant features and improves the accuracy of numerical models.

1. INTRODUCTION

Feature selection is an important preprocessing step in machine learning tasks involving large datasets. Identifying and selecting the most relevant features enhances interpretability of the results and reduces computational cost. Feature selection methods can be broadly divided into three categories: filter methods, wrapper methods, and embedded methods. Filter methods use independent measures such as information gain or $\chi^2$-statistic to evaluate the importance of a feature subset. Wrapper methods evaluate a feature subset based on the accuracy of the learning model built with the given feature subset. Embedded methods perform feature evaluation as part of the model building process, with lasso regression being the canonical example of such approach. Among the three types of selection methods filter algorithms are the fastest to execute. However, most of the existing filter methods ignore the interactions between feature variables. For instance, a basic filter method would rank features based on the information gain between an individual feature and the target variable. However, it is possible that a pair of features, that individually have low information gain with respect to the target variable, may have a very high information gain if considered jointly. The canonical example where feature interactions are particularly prominent is the XOR problem.

Date: Mar 5, 2019.
Key words and phrases. feature selection; variance decomposition; Sobol decomposition; sensitivity index; total sensitivity index; wrapper methods.
There have been various attempts in the literature to address the effects of feature interactions. The authors of the maximum relevance minimum redundancy method [33] propose to incrementally add features to the optimal feature subset by maximizing the joint information gain between the feature subset and the target variable while minimizing the average information gain between the pairs of feature variables. However, this method and other similar approaches take only a partial account of feature interactions and thus cannot guarantee an optimal solution. In addition, existing heuristics for exhaustive feature selection are time consuming.

In this paper we propose a feature selection method that under certain conditions allows for an efficient and complete feature evaluation. Our method is based on orthogonal decomposition of the variance of the model output [21, 24]. In particular, the variance of the target variable $Y$ can be decomposed as

$$V(Y) = \sum_i V_i + \sum_{i,j} V_{ij} + \cdots + V_{12,.,k},$$

where each term $V_{i1,i_2,.,i_s}$ represents the contribution to the variance of $Y$ stemming from feature interactions in subset $\{X_{i_1}, X_{i_2}, ...X_{i_s}\}$. Based on the variance decomposition we calculate the Total Sensitivity Index (TSI) of a feature which takes into account the interactions of a feature with all other features in determining its effect on the output variance. The TSI can thus be used to evaluate the importance of a feature with respect to the model output. Our approach is similar to wrapper methods as it is trained on a particular learning model. However, it is more efficient than wrapper methods such as RFE in that the model being used needs to be trained only once. Thereafter all the remaining calculations are performed directly on the data using appropriate Monte Carlo integrals.

In Section 2, we give a brief overview of existing feature selection methods. In Section 3, we discuss decomposition of the output variance into its different components based on feature subsets. We give the necessary mathematical background and a Monte-Carlo method to perform required integral calculations. In Section 4, we present the results of numerical experiments that were performed using the proposed method. The results show that our method correctly selects relevant features and improves accuracy of learning models.

2. Related Work

Feature selection algorithms can be broadly divided into three subsets: filter, wrapper and embedded methods. Filter methods use an independent metric - usually based on information gain [30] or $\chi^2$-statistic [13] - to measure the importance of a feature subset. This is done by measuring the
strength of the relationship between the features and the target class. The main advantage of filter methods is their relative robustness to changes in the learning model. In other words, a feature subset chosen with a filter method would perform similarly under different learning models. Wrapper methods evaluate the importance of a feature subset in the context of a particular model. The classical wrapper method uses a fixed classification (or regression) model to calculate its accuracy using a feature subset. The accuracy results determine the worth of the subset. The advantage of the wrapper method is that it is fine tuned to a particular model and thus gives better results than the filter approach on the particular chosen model. However, feature subsets that perform well on the original model do not perform as well on other models. In addition, wrapper methods are computationally intensive as the model must be trained for every subset evaluation. Embedded methods perform feature selection as part of the model building process as in the case of the lasso regression [29].

The simplest approach to filter features is to evaluate the importance of each feature using a statistic that measures the relationship between a feature and the target variable and then select the features with the top scores. However, this approach does not take into account interactions between the features. It is quite possible that a pair of features that are individually irrelevant can have a high predictive power when joined together as in the case of XOR problem. In [19, 32], the authors propose to evaluate features based on both their relevancy with respect to the target variable and their redundancy with respect to other features. In other words, the features that are highly correlated with the target variable but uncorrelated with other feature variables are given the preference. Authors use information gain as the measure of relevancy/redundancy though any other evaluation criteria would work as well. The authors in [14] use information gain and $\chi^2$-statistic jointly to evaluate features. In particular, the two metrics are combined into a single vector whose magnitude determines the importance of a feature. Combining the two metrics allows for lower variance in feature scores. This approach was shown to significantly reduce the number of features without hurting the accuracy of the model. In particular, the method was used effectively in selecting relevant factors in autism classification [28].

There also exist filter methods that don’t use the classical metrics from statistics or information theory. In [5], for instance, the authors propose to evaluate feature subsets based on the inconsistency measure whereby a subset is deemed inconsistent if there at least two instances with the same features labels that belong to different classes. The authors also discuss different search methods for finding the optimal subset based on the inconsistency measure.
Wrapper methods use specific models to evaluate features. One of the popular wrapper algorithms is Recursive Feature Elimination (RFE). In this method, features are recursively dropped from the initial set based on their weights in the model. For instance, we can use OLS to first fit a hyperplane to the (normalized) data. Then we drop the feature with the lowest coefficient in the hyperplane equation. We again fit a hyperplane to our data but now using the reduced feature set. This procedure is repeated until some stopping criteria is achieved. RFE has implementations with many estimators including SVM [8], Random Forests [9], and MLP [31]. Although wrapper methods generally perform well their biggest drawback is high computational cost.

In [6], the authors use Sobol’s method [24, 25] of decomposing output variance to evaluate features. In this approach, the variance of the target variable is deconstructed into a sum of variances based on feature subsets. In particular, the authors use first order sensitivity index to evaluate each feature

\[ S_i = \frac{\text{Var}(E[Y|X_i])}{\text{V}(Y)}. \]

In [15], the authors use total sensitivity index based on variance decomposition to evaluate features. Total sensitivity index takes a more complete account of feature interactions and ultimately yields better results. In this paper, we pursue a similar approach and use total sensitivity analysis to evaluate features.

3. Orthogonal Variance Decomposition

In this section, we will go through the steps of decomposing model output variance into orthogonal components based on feature subsets. We will discuss how to use variance decomposition to evaluate features. In the end, we will describe a method for calculating the required integrals using a Monte-Carlo approach. Our presentation follows that of Saltelli [21] with further details provided in [11, 24, 25].

Suppose that target variable \( Y \) is a function of a set of feature variables \( \{X_1, X_2, \ldots, X_k\} \), i.e., \( Y = f(X_1, X_2, \ldots, X_k) \). Assume that the features \( \{X_i\} \) are independently and uniformly distributed over the interval \([0, 1]\). We define

\[ f_0 = E[Y] \]
\[ f_i(x) = E[Y|X_i = x] - f_0 \]
\[ f_{ij}(x, y) = E[Y|X_i = x, X_j = y] - f_i(x) - f_j(y) - f_0 \]

and similarly for higher orders. For convenience, we drop the function arguments from the notation. Then it is not hard check that we obtain the
following functional decomposition

\begin{equation}
    f = f_0 + \sum_i f_i + \sum_{i,j} f_{ij} + \ldots + f_{12..k}
\end{equation}

**Example 1.** Let us demonstrate Eq. (3) in the case of two features

\[ f = f_0 + f_1 + f_2 + f_{12} = f_0 + f_1 + f_2 + (E[Y|X_1, X_2] - f_1 - f_2 - f_0) = E[Y|X_1, X_2] = Y \]

**Remark 2.** Our assumptions that the features are independent and uniformly distributed are highly optimistic. In practice, features have various non uniform distributions and are not independent of one another. In order to properly deal with these, more complex, scenarios one can employ other techniques for sensitivity analysis as described in [22]. However, as it will be demonstrated by our numerical experiments, even under the ”naive” assumption of i.i.d. our feature selection algorithm performs very well.

The functional decomposition in Eq. (3) is orthogonal in the sense that

\begin{equation}
    \int_0^1 f_{i_1i_2..i_s} \, dx_j = 0
\end{equation}

for every subset \( \{X_{i_1}, X_{i_2}, \ldots X_{i_s}\} \) and \( j \in \{i_1, i_2, \ldots, i_s\} \).

**Example 3.** Let us examine Eq. (4) in the case of two features. Since

\[ \int_0^1 E[Y|X_i] \, dx_i = E[Y], \]

then we have

\[ \int_0^1 f_1 \, dx_1 = \int_0^1 (E[Y|X_1] - f_0) \, dx_1 = 0. \]

Likewise, since

\[ \int_0^1 E[Y|X_i, X_j] \, dx_i = E[Y|X_j] \] and \( \int_0^1 E[Y|X_j] \, dx_i = E[Y|X_j] \), then

\[ \int_0^1 f_{ij} \, dx_i = \int_0^1 (E[Y|X_i, X_j] - f_i - f_j - f_0) \, dx_i \]

\[ = \int_0^1 (E[Y|X_i, X_j] - E[Y|X_i] + f_0 - E[Y|X_j] + f_0 - f_0) \, dx_i \]

\[ = \int_0^1 (E[Y|X_i, X_j] - f_0 + f_0 - E[Y|X_j]) \, dx_i = 0. \]

Another, useful property of the terms of the functional decomposition in Eq. (3) is

\begin{equation}
    E[f_{i_1i_2..i_s}] = 0,
\end{equation}
which follows directly from Eq. (4).

To obtain the variance decomposition, we square and integrate the two sides of Eq. (3)

\[ \int_{[0,1]^k} f^2 \, d\mathbf{X} = \int_{[0,1]^k} \left( f_0 + \sum_i f_i + \sum_{i,j} f_{ij} + \ldots + f_{12..k} \right)^2 \, d\mathbf{X}. \]

Using Eq. (4), we can eliminate most of the cross multiplied terms on the right-hand side. Then the final result can be written in the form

\[ \int_{[0,1]^k} f^2 \, d\mathbf{X} - f_0^2 = \int_{[0,1]^k} \left( \sum_i f_i^2 + \sum_{i,j} f_{ij}^2 + \ldots + f_{12..k}^2 \right) \, d\mathbf{X}. \]

Note that the left-hand side of the Eq. (6) represents the variance of the model output \( V(Y) \). Also note that by Eq. (5) we get \( \text{Var}(f_{i_1 \ldots i_s}) = \int f_{i_1 \ldots i_s}^2 \, d\mathbf{X} \). We now obtain Eq. (1) that was stated in the introduction of the paper

\[ V(Y) = \sum_i V_i + \sum_{i,j} V_{ij} + \ldots + V_{12..k}. \]

We can also view the variance decomposition in Eq. (1) from a slightly different angle:

\[ V(Y) = \sum_i V_i' + \sum_{i,j} V_{ij}' + \ldots + V_{12..k}'. \]

where \( V_i' = \text{Var}(E[Y|X_i]) \), \( V_{ij}' = \text{Var}(E[Y|X_i, X_j]) - V_i' - V_j' \), and similarly for higher orders. In the context of Equation (7), \( V_i' \) represents the variance of \( Y \) due solely to the feature \( X_i \), \( V_{ij}' \) represents the variance due to interaction between features \( X_i \) and \( X_j \), and similarly for higher orders.

**Example 4.** Let us show that Eq. (1) and Eq. (7) are equivalent in the case of two features. First, it is easy to see that

\[ V_i = \text{Var}(E[Y|X_i] - f_0) = \text{Var}(E[Y|X_i]) = V_i'. \]
Next,

\[ V_{ij} = \text{Var}(E[Y|X_i, X_j] - f_i - f_j - f_0) \]

\[ = \int_0^1 \int_0^1 (E[Y|X_i, X_j] - f_i - f_j - f_0)^2 \, dx_i \, dx_j \]

\[ = \int_0^1 \int_0^1 (E[Y|X_i, X_j]^2 + f_i^2 + f_j^2 + f_0^2) \, dx_i \, dx_j \]

\[ - \int_0^1 \int_0^1 2E[Y|X_i, X_j](f_i + f_j + f_0) \, dx_i \, dx_j \]

\[ = \int_0^1 \int_0^1 (E[Y|X_i, X_j]^2 - f_i^2 - f_j^2 - f_0^2) \, dx_i \, dx_j \]

\[ = \text{Var}(E[Y|X_i, X_j]) - V'_{i} - V'_{j} = V'_{ij}. \]

We used the fact that \( \int f_k f_l \, dx = 0 \) and \( \int E[Y|X_i, X_j] f_k \, dx = f_k^2 \) in our calculations.

Based on Eq. (7) we define the first order sensitivity index of feature \( X_i \) by

\[ S_i = \frac{V'_i}{V(Y)}, \]

which measures the contribution of \( X_i \) alone to the output variance. We can use \( S_i \) as a simple tool to perform feature evaluation and selection. In fact, Efimov and Sulieman [6] used this approach to design their feature selection method. A more comprehensive metric to evaluate features would be the total sensitivity index defined as the sum of all variance terms of Eq. (7) that contain contributions of feature \( X_i \):

\[ S_{Ti} = \frac{V'_i + \sum_j V'_{ij} + \sum_{j,k} V'_{ijk} + \ldots + V'_{12\ldots i\ldots k}}{V(Y)}. \]

To simplify the expression in Eq. (9) we can use the following useful identity:

\[ V'_i + \sum_j V'_{ij} + \sum_{j,k} V'_{ijk} + \ldots + V'_{12\ldots i\ldots k} = V(Y) - \text{Var}(E[Y|X_{\sim i}]), \]

where \( X_{\sim i} \) is the vector of all features except \( X_i \). Let us illustrate the identity in Eq. (10) with an example based on three features.

**Example 5.** Suppose that we have \( Y = f(X_1, X_2, X_3) \). Let us verify that

\[ V'_1 + V'_{12} + V'_{13} + V'_{123} = V(Y) - \text{Var}(E[Y|X_{\sim 1}]), \]

where \( X_{\sim 1} \) is the vector of all features except \( X_1 \). Indeed, we know that \( \text{Var}(E[Y|X_{\sim 1}]) = V'_{23} + V'_2 + V'_3 \). Since \( V(Y) \) is the full sum of partial variances the result follows.
Using Eq. (10) we can rewrite the definition of the total sensitivity index as

\[ S_{T_i} = 1 - \frac{\text{Var}(E[Y|X_{\sim i}])}{V(Y)}, \]

Our final task is to calculate \( \text{Var}(E[Y|X_{\sim i}]) \). There are exist various estimators for \( \text{Var}(E[Y|X_{\sim i}]) \) [11,12,26]. In this paper, we choose to follow the approach of Homma and Saltelli [11]. Let \( A \) and \( B \) be a pair of independent sampling matrices. Let \( j \) and \( i \) denote row and column indexes respectively. Define \( A^{(i)}_B \) to be matrix \( A \), where its \( i \)th column replaced with the \( i \)th column of \( B \). Then our estimator is

\[ \text{Var}(E[Y|X_{\sim i}]) = \frac{1}{n} \sum_{j=1}^{n} f(A)_{ij} f(A^{(i)}_B)_{ij} - f_0^2 \]

4. Numerical Experiments

In this section, we will apply the total sensitivity (\( S_{T_i} \)) index to evaluate and select features. We will first test our approach on simulated data where the relevant features are known and show that our method correctly identifies the right features. Then we apply our method to real world datasets and show that our method can reduce the number of features used in a model without losing its accuracy.

The details of the feature evaluation process using the total sensitivity index are given below

Algorithm

1. Let \( D \) be a dataset consisting of \( n \) rows, \( k \) features, and a target variable. Train a model \( f \) (SVR, RF, NN, etc) on \( D \).

2. Discard the target variable so that \( D \) consists of only feature variables. Calculate the average \( f_0 = \frac{1}{n} \sum_{j=1}^{n} f(D)_{ij} \) and total variance \( V(\hat{Y}) = \frac{1}{n} \sum_{j=1}^{n} f(D)_{ij}^2 - f_0^2 \).

3. Shuffle the rows of \( D \) and split it into 2 halves: \( A \) and \( B \).

4. For each feature \( X_i \), create matrix \( A^{(i)}_B \) by replacing the \( i \)th column of \( A \) with the \( i \)th column of \( B \). Then calculate the corresponding partial variance using Eq. (12).
5. For each feature $X_i$, calculate the corresponding index $S_{Ti}$ using Eq. (11).

We begin our experiments with dataset generated using the Friedman function.

*Example 6.* In this example, we generate 20 independent features $X_i$ using the uniform distribution over the interval $[0, 1]$ and the normal distribution with mean 0.5 and standard deviation 0.25. We use the first 5 features to calculate the target variable according the Friedman function $[2, 7]$:

\[
y(x) = 10 \sin(\pi x_0 x_1) + 20(x_2 - 0.5)^2 + 10x_3 + 5x_4 + \sigma \cdot N(0, 1),
\]

where $\sigma$ represents the amount of noise added to the data. The remaining features are thus redundant. We begin by calculating the sensitivity index using the original Friedman function as described in Equation (13). As can be seen from Figure 6, both $S_i$ and $S_{Ti}$ based approaches perform well when the noise level is zero. In fact, even when the features are generated using the normal distribution - as opposed to the uniform distribution as required by the theory - the sensitivity indices of the relevant features are higher than that of the redundant features. When the noise level is increased to $\sigma = 1$ the $S_{Ti}$ continues to classify correctly the relevant features while $S_i$ struggles with normally distributed features. When the noise level is increased further to $\sigma = 2$ $S_{Ti}$ continues to outperform $S_i$ although it can no longer classify confidently the relevant features under the normal distribution.

Next we train a Random Forrest (RF) regressor (with 10 estimators) on the data and use it to calculate the sensitivity index. As can be seen from Figure 6, the $S_{Ti}$ values are significantly higher for the relevant features. Somewhat surprisingly, the $S_{Ti}$ values are even higher for the normally distributed data which indicates that our method can perform well even when the theoretical assumptions on the data, i.e. i.i.d uniform distribution, are not strictly satisfied. We can also see from the last subplot that using $S_{Ti}$ we can identify the relevant features even under high levels of noise.

Finally, we use a trained neural network (NN) model to calculate the sensitivity index. In our neural network model we used 2 hidden layers with 64 nodes in each layer. As shown in Figure 6, the $S_{Ti}$ values for uniformly distributed features are significantly higher than that of the redundant features. Note that $S_{Ti}$ produces better results than $S_i$ as was the case in the previous scenarios. However, when the noise level is increased the $S_{Ti}$ values for relevant features become very close to that of redundant features in the case of normally distributed data. Also note that the neural network model did not fit the data well as can be seen from the relatively high mean absolute error values. This might help to explain the fact that
Figure 1. $S_i$ and $S_{Ti}$ values based on the Friedman function under different levels of noise.

The neural networks approach underperformed relative to the previous two models.

**Example 7.** In this example, we generate a dataset consisting of 20 independent features $X_i$ with the first four features taken as the relevant variables. The relevant features are uniformly distributed over the following intervals

- $0 \leq X_0 \leq 100$
- $40\pi \leq X_1 \leq 560\pi$
- $0 \leq X_2 \leq 1$
- $1 \leq X_3 \leq 11$.

The remaining features are uniformly distributed over the interval $[0, 1]$. The target variable is calculated according another Friedman function [2, 7] :

$$y(x) = \sqrt{x_0^2 + (x_1 x_2 - \frac{x_1 x_3}{x_1 x_3})^2} + \sigma \cdot N(0, 1).$$

We apply various feature ranking methods to test if they can correctly identify the relevant features. To this end, we train Support Vector Machines (SVR) [23], RF, and NN models and use them to calculate the sensitivity indices. We also use the original function from Equation 15 to calculate the corresponding sensitivity indices. We benchmark the performance of the
Figure 2. \( S_i \) and \( S_{T_i} \) values based on the Random Forrest model under different levels of noise.

\( S_{T_i} \) based approach to the performance of RFE methods. In particular, we use the RFE algorithms based on SVR, Linear Regression (LR), and RF regressors. The results in Table 7 show that the \( S_{T_i} \) based approach produces the best results in correctly identifying the relevant features. In particular, \( S_{T_i}^{Friedman} \) and \( S_{T_i}^{SVR} \) rank the relevant features in top 4 even when we apply noise to the data. Similarly, \( S_{T_i}^{RF} \) and \( S_{T_i}^{NN} \) perform well with only a single feature (4) being ranked one spot below the top 4. By comparison RFE \( SVR \) and RFE \( LR \) rank only one feature correctly. We note that RFE \( RF \) does perform well, somewhat surprisingly, when noise is applied to the data.

Example 8. In this example, we generate a dataset consisting of 20 independent features \( X_i \) with the first four features taken as the relevant variables. The relevant features are uniformly distributed over the following intervals

- \( 0 \leq X_0 \leq 100 \)
- \( 40\pi \leq X_1 \leq 560\pi \)
- \( 0 \leq X_2 \leq 1 \)
- \( 1 \leq X_3 \leq 11 \).

The remaining features are uniformly distributed over the interval \([0, 1]\). The target variable is calculated according another Friedman function \([2, 7]\).
Figure 3. $S_i$ and $S_{Ti}$ values based on the Neural Network model under different levels of noise.

Table 1. Rankings of the four relevant features (1, 2, 3, and 4) using various feature selection methods. The dataset used consists of 20 features with the target variable being generated via the Friedman function (Equation 15).

| Noise  | $\sigma = 0$ | $\sigma = 10$ |
|--------|--------------|---------------|
| Feature | 1 2 3 4      | 1 2 3 4       |
| $S_i$, Friedman | 20 2 1 19   | 19 2 1 5      |
| $S_{Ti}$, Friedman | 3 2 1 4   | 3 2 1 4      |
| $S_i$, SVR | 20 1 2 3   | 20 1 2 3     |
| $S_{Ti}$, SVR | 2 1 4 3   | 2 1 4 3     |
| $S_i$, RF | 13 2 1 19  | 20 2 1 11    |
| $S_{Ti}$, RF | 3 2 1 5  | 3 2 1 5      |
| $S_i$, NN | 2 1 20 4   | 19 1 20 9    |
| $S_{Ti}$, NN | 2 1 3 1   | 2 1 3 5     |
| RFE, SVR | 19 17 1 14 | 19 17 1 11 |
| RFE, LR | 17 18 1 19 | 18 19 1 20 |
| RFE, RF | 6 2 1 7   | 3 2 1 6      |
We apply various feature ranking methods to test if they can correctly identify the relevant features. To this end, we train Support Vector Machines (SVR) [23], RF, and NN models and use them to calculate the sensitivity indices. We also use the original function from Equation (15) to calculate the corresponding sensitivity indices. We benchmark the performance of the \( S_{Ti} \) based approach to the performance of RFE methods. In particular, we use the RFE algorithms based on SVR, Linear Regression (LR), and RF regressors. The results in Table 7 show that the \( S_{Ti} \) based approach produces the best results in correctly identifying the relevant features. In particular, \( S_{Ti,Friedman} \) and \( S_{Ti,SVR} \) rank the relevant features in top 4 even when we apply noise to the data. Similarly, \( S_{Ti,RF} \) and \( S_{Ti,NN} \) perform well with only a single feature (4) being ranked one spot below the top 4. By comparison RFE \( SVR \) and RFE \( LR \) rank only one feature correctly. We note that RFE \( RF \) does perform well, somewhat surprisingly, when noise is applied to the data.

\[
y(x) = \sqrt{x_0^2 + (x_1x_2 - \frac{1}{x_1x_3})^2} + \sigma \cdot N(0, 1).
\]
5. Conclusion

In this paper we discuss a new approach to evaluating features based on total sensitivity index (TSI). There are two main advantages to using TSI in feature evaluation. First, the TSI incorporates the effects of interactions between the features. Most of the modern feature selection methods do not fully consider the effects that other features have on the relationship between a feature and the target class. In this sense, TSI approach stands in a small crowd. Second advantage of TSI lies in its relative efficiency. Although it is not as efficient as a filter method, it is still much faster than most of other wrapper methods.

The experiments with artificially generated data (Friedman data set) where the relevant features were known showed that TSI is the highest for the relevant features. In other words, TSI effectively identified the important features in the data set. In particular, TSI computed using the neural networks model can effectively identify the relevant features even with high level of noise in the data. We also tested our approach to a real life data set (Communities and Crime). The experiments with this data set showed that TSI is very competitive with other modern feature selection models. In particular, the highest performance on the data set was achieved using the features selected via TSI-SVR.

REFERENCES

[1] G. Archer, A. Saltelli, and I. Sobol, "Sensitivity measures, anova-like techniques and the use of bootstrap," Journal of Statistical Computation and Simulation 58(1997), 99120.
[2] L. Breiman, Bagging predictors, Machine Learning 24 (1996) 123-140.
[3] L. Breiman, Random Forests, Machine Learning, 45(1), 5-32, 2001.
[4] K. Buza, "BlogFeedback Data Set ," UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science (2014).
[5] M. Dash and H. Liu, "Consistency-based search in feature selection," Artificial intelligence 151(1-2)(2003) 155-176.
[6] D. Efimov and H. Suliman, "Sobol Sensitivity: a strategy for Feature Selection," Springer Proceedings in Mathematics and Statistics 190(1)(2017) 57-76.
[7] J. Friedman, "Multivariate adaptive regression splines," The Annals of Statistics 19(1)(1991) 1-67.
[8] I. Guyon, J.Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Machine Learning 46(1):389, (2002).
[9] P. Granitto, et al, "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products," Chemometrics and Intelligent Laboratory Systems 83(2), (2006) 83-90.
[10] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," Procedia Computer Science 17 (2013), 26-32.
[11] T. Homma and A. Saltelli, "Importance measures in global sensitivity analysis of model output," *Reliability Engineering and System Safety* **52**(1) (1996), 1-17.

[12] M.J.W. Jansen, ”Analysis of variance designs for model output,” *Computer Communications* **117** (1999), 3543.

[13] X. Jin, et al., ”Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles,” *International Workshop on Data Mining for Biomedical Applications* Springer, Berlin, Heidelberg, 2006.

[14] F. Kamalov and F. Thabtah, ”A feature selection method based on ranked vector scores of features for classification,” *Ann. Data Sci.* **4** (2017), 1-20.

[15] F. Kamalov. ”Sensitivity Analysis for Feature Selection.” 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018.

[16] Y. LeCun, L. Bottou, G. Orr and K. Miller, Efficient BackProp, *Neural Networks: Tricks of the Trade* 1998.

[17] Y. Li, C. Luo, and S. M. Chung, ”Text clustering with feature selection by using statistical data.” *IEEE Transactions on Knowledge and Data Engineering* **20**(5) (2008), 641-652.

[18] H. Liu and H. Motoda eds., *Computational methods of feature selection*, CRC Press, 2007.

[19] H. Peng, F. Long, and C. Ding, ”Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on pattern analysis and machine intelligence* **27**(8) (2005), 1226-1238.

[20] M. Redmond, ”Communities and Crime Data Set .,” UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science (July 2009).

[21] A. Saltelli, et al., ”Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index,” *Computer Physics Communications* **181**(2) (2010), 259-270.

[22] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, M. (2004). Sensitivity analysis in practice: a guide to assessing scientific models. John Wiley & Sons.

[23] A. Smola and B. Schlkopf, A Tutorial on Support Vector Regression, *Statistics and Computing Archive* **14** (3) (2004), 199-222.

[24] I.M. Sobol’, ”Sensitivity analysis for non-linear mathematical models,” *Mathematical Modelling and Computational Experiment* **1**(1993), 407414.

[25] I.M. Sobol’, ”Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates,” *Mathematics and Computers in Simulation* **55**(2001), 271280.

[26] I.M. Sobol, ”Global sensitivity analysis indices for the investigation of non-linear mathematical models,” *Matematicheskoe Modelirovanie* **19** (11) (2007) 2324 (in Russian).

[27] B. Tang, S. Kay, and H. He, ”Toward optimal feature selection in naive Bayes for text categorization,” *IEEE transactions on knowledge and data engineering* **28**(9) (2016), 2508-2521.

[28] F. Thabtah, F. Kamalov, and K. Rajab, ”A new computational intelligence approach to detect autistic features for autism screening,” *International Journal of Medical Informatics* **117** (2018), 112-124.

[29] R. Tibshirani. ”Regression shrinkage and selection via the lasso: a retrospective.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**.3 (2011): 273-282.
[30] J. Vergara and P. Estevez, ”A review of feature selection methods based on mutual information,” Neural Computing and Applications 24(1) (2014), 175186.

[31] J.B. Yang, et al, ”Feature selection for MLP neural network: The use of random permutation of probabilistic outputs,” IEEE Transactions on Neural Networks 20(12) (2009), 1911-1922.

[32] L. Yu and H. Liu, ”Efficient feature selection via analysis of relevance and redundancy,” Journal of machine learning research 5 (2004), 1205-1224.

[33] H. Peng, F. Long, and C. Ding. ”Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.” IEEE Transactions on pattern analysis and machine intelligence 27, no. 8 (2005): 1226-1238.

Canadian University Dubai, Dubai, UAE.
E-mail address: firuz@cud.ac.ae