Journal of IT in Asia, Vol. 6, 2016, 11 – 24

Improved Feature Selection Based on Mutual Information for Regression Tasks

[1]Muhammad A. Sulaiman and [2]Jane Labadin
[1,2]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
email: [1]muhalisu@gmail.com, [2]ljane@unimas.my

Abstract – *Mutual Information (MI) is an information theory concept often used in the recent time as a criterion for feature selection methods. This is due to its ability to capture both linear and non-linear dependency relationships between two variables. In theory, mutual information is formulated based on probability density functions (pdfs) or entropies of the two variables. In most machine learning applications, mutual information estimation is formulated for classification problems (that is data with labeled output). This study investigates the use of mutual information estimation as a feature selection criterion for regression tasks and introduces enhancement in selecting optimal feature subset based on previous works. Specifically, while focusing on regression tasks, it builds on the previous work in which a scientifically sound stopping criteria for feature selection greedy algorithms was proposed. Four real-world regression datasets were used in this study, three of the datasets are public obtained from UCI machine learning repository and the remaining one is a private well log dataset. Two Machine learning models namely multiple regression and artificial neural networks (ANN) were used to test the performance of IFSMIR. The results obtained has proved the effectiveness of the proposed method.*

Keywords: feature selection, filter, estimation, mutual information, machine learning.

1 Introduction

Feature Selection is an optimization method used to select optimal subsets of a full feature set. The selected optimal subset is expected to retain the relevant information in the full feature set. Feature selection algorithms rely on some certain criteria to score features or input variables, these criteria are divided into two categories and thereby forming the basis for the three types of feature selection models. The first category of criteria utilizes statistical and a probabilistic distribution of dataset attributes to measure the relevant of each input variable to the corresponding output variable. This category is refers to as filter models. The second category of criteria is search optimization algorithms used together with a particular learning machine model to find relevant feature subset based on the performance of the learning machine. This category comprises of wrapper models. And the third type of feature selection model is the hybrid models, which combined the power of both filter and wrapper to select feature subsets.

Mutual information is used as a criterion for feature selection method. It is a filter-based model that measures both linear and non-linear dependency relationships between two random variables, this property made it be a popular choice as feature selection criterion. Mutual information is formulated from probability density functions (pdfs) or entropies of two variables and their joint variable. However, despite that in theory mutual information formulation is suitable for a dataset with either discrete or continuous output variable, in practice its estimation is often assumed classification problems (that is a dataset with labeled output variable). The reason for this cannot be unconnected to the fact that it is not clear how or rather very difficult to estimate pdf or entropy of joint variable from a dataset with a continuous output variable. This study presents mutual information estimation formulated around the estimated entropy of a variable. In addition, this study provides an extension to feature selection greedy algorithms on how to select optimal feature subset when using filter model only.

The rest of the paper is organized as follows: Section 2 is a review of feature selection methods, information theory and how it relates to mutual information. Section 3 presents Machine learning models used in this study. Feature selection procedure based on MI estimation is presented in section 4. Section 5 presents experimental studies, results and discussion. And finally, the conclusion is presented in section 6.

2 Review

This section presents a review of feature selection methods from related works and how mutual information is formulated from entropy.

2.1 Overview of feature selection

There are three broadly categories of models for feature selection methods. The first category is the filter models, which are a group of models that utilize statistical and probabilistic distributions of dataset attributes in order to select feature subset from the input dataset. Hence, they select feature subset independent of any particular learning machine. This independent selection enables subsequent feature prediction by any learning machine. Feature Selection based on Mutual Information (MI) is an example of filter model. The second category is wrapper models, these are search optimization algorithms used together with a particular learning machine, these categories use performance of a learning machine in terms of accuracy of prediction to find the optimal feature subset. Genetic algorithm (GA) and Particle Swarm Optimization (PSO) algorithm are examples of wrapper models. The dependency on a particular learning machine to search the best feature subset makes wrappers have better predictive accuracy with relatively high computational overhead (Unler, Murat and Chinnam, 2011). However, there are high tendency that feature subset selected using wrapper based on a specific learning machine may fail woefully when used with a different learning machine. For example, feature subset generated based on GA and ANN may perform woefully when used to make a prediction with Support Vector Machine (SVM). The third category is the hybrid models, which combined the power of both filter and wrapper to select feature subsets.

Unler, Murat and Chinnam (2011) introduces hybrid filter-wrapper feature subset selection algorithm based on PSO with SVM. Mutual Information (MI) defined in terms of probability density functions of variables serves as filter algorithm while Particle Swarm Optimization (PSO) which is a population-based search optimization technique was used as the wrapper to find the best strongly relevant feature subset identified by MI and finally SVM is used for classification. Also, in another related work varied filter algorithms for managing uncertain data in data mining and machine learning application were proposed based on a mutual information (MI) criterion, a combination of ranking and expectation-maximization (EM) and Hilbert-Schmidt independence criterion (HSIC) as follow. Mutual information (MI) Criterion based on probability density function (pdf) of each data value was proposed (Doquire and Verleysen, 2011) and experimental results on 8 UCI machine learning repository have proved the effectiveness of this algorithm. MI was first evaluated between each feature of the training set and the output vector. The resulting MI scores for each feature are then used to rank the features. Song et al. (2012) used two different aging databases for the experiment, FG-NET containing 1002 face images of 82 persons with age ranging from 0 to 69. And Yamaha faces database containing 800 males, 800 females and 8000 images with ages ranging from 0 to 93. The Ranking model was built based on kernel trick and bilinear regression strategy, and the parameter learning technique was based on EM. Yan et al. (2007) introduces Hilbert-Schmidt Independence Criterion (HSIC) for feature selection. HSIC which is based on the covariance between variables mapped to produce kernel Hilbert spaces were employed for feature selection together with greedy backward elimination algorithm. Experimental results have shown that HSIC performed comparably to other state-of-the-art feature selectors such as SVM Recursive Elimination (RFE), RELIEF, L0 –norm SVM (L0) and R2W2.

In a nutshell, there are very many researches in the area of feature selection with mutual information as a selection criterion. This include Battiti (1994), Peng, Long and Ding (2005), Rossi et al. (2006), Evans (2008), Doquire and Verleysen (2011) and etc., all of which focus on classification tasks. However, there are only handful of works focus on regression tasks (Francois, et al., 2007; Verleysen, Rossi and Francoisi, 2009; Carmona et'al., 2011; Frenay, Doquirel and Verleysen, 2013).

2.2 Greedy methods of Feature Selection

There are three greedy methods for feature selection as presented in Battiti (1994), François et al. (2007), Liu, Liu and Zhang (2008). (1) The forward procedure involves selecting feature starting from the empty set until an optimal subset of the total set (say M sets) is obtained. While the forward procedure provides means for selecting features to an empty set and subsequently increases the feature subset, this procedure has no well-defined stopping criteria and chances are that it may lead to a suboptimal subset. (2) The backward procedure involves starting from the whole M sets and removing a feature from the whole set after each step until optimal subset is reached. This procedure is computationally expensive since it involves estimating mutual information for all M-dimensional variables. Also, like the forward procedure, it has no well-defined stopping criteria (3) The Forward-Backward procedure seems to be a better method since it involves starting from an empty set and select feature incrementally and occasional remove feature from the subset. This way it reduces the chances of having a suboptimal subset. Like the two procedures above it also lacked clear stopping criteria.

Lack of justifiable stopping criterion is identified as a limitation of greedy methods for feature selection (Verleysen, Rossi and Francoisi, 2009). Usually, the common way to stop the greedy forward feature selection is when estimated MI values started to decrease, this practice lacks justification. François et al. (2007) and Verleysen, Rossi and François

(2009) introduce the use of resampling and permutation to provide a statistically justifiable stopping criteria. However, in both cases, the method was used with a randomly generated dataset. Even though randomly generated artificial dataset was used by the previous studies to validate the effectiveness of the algorithm, the use of real-world dataset is required to test its effectiveness. More so, despite the progress made in the previous studies to improve the greedy methods for feature selection, chances are that they yielded feature subsets that are far from optimal, optimal or almost optimal (Pudil, Novovicova and Kittler, 1994; Verleysen, Rossi and François, 2009; Sulaiman and Labadin, 2015**b**).

So, the main contribution of this study is to establish a methodology for generating optimal feature subset for filter method by extending the previous work proposed by Verleysen, Rossi and François (2009).

2.3 Entropy and Mutual Information (MI)

The entropy of a random variable *X* is a concept in information theory that measures the uncertainty associated with *X*. While Mutual information (MI), another information theory concept quantitatively measures the amount of dependent information two random variables have about each other. Unlike correlation coefficient that measures linear dependence only, mutual information measures both linear and nonlinear dependence between variables, a property that made it a popular choice for feature selection (Doquire and Verleysen, 2011; François, et al., 2007; Battiti, 1994; Peng, Long and Ding, 2005; Rossi, et al., 2006).

Considering a pair of continuous random variables *X* and *Y*, the joint probability density function of *X* and *Y* is expressed as:

$$P_{X,Y}(x, y) = P_Y(y|x)P_x(x) \tag{1}$$

In a similar way, the joint differential entropy of *X* and *Y* is expressed as:

$$h(X, Y) = h(X) + h(Y|X) \tag{2}$$

where h(*Y*|*X*) is known as the conditional differential entropy of *Y* given *X*. In a word the equation 2 is expressed as the uncertainty about *X* and *Y* is equal to the uncertainty about *X* plus the uncertainty about *Y* given *X*. This can equally be said in the other way round as the uncertainty about *X* and *Y* is equal to the uncertainty about *Y* plus the uncertainty about *X* given *Y* as in equation 3:

$$h(X, Y) = h(Y) + h(X|Y) \tag{3}$$

Meanwhile, entropy of a random variable *X* is expressed as

$$h(X) = -\int f_X(x) \log f_X(x) d_x \tag{4}$$

To consider a learning system where the applications of a continuous random variable *X* to the input of the system, produces a continuous random variable *Y* at the output of the system. If by definition the differential entropy *h(X)* is the uncertainty about the system input *X* before the observation of the system output *Y*, while the conditional entropy *h(X|Y)* is the uncertainty about the system input after the output *Y* is observed. Then the difference, *h(X) – h(X|Y)*, is the uncertainty about the system input *X* that is determined by observing the system output *Y*. this is refers to mutual information between the system input *X* and the system output *Y* which is denoted as *I(X;Y)*. Thus:

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{X,Y}(x,y) \log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right) d_x d_y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{X|Y}(x|y)p_Y(y) \log\left(\frac{p_{X|Y}(x|y)}{p_Y(y)}\right) d_x d_y \end{aligned} \tag{5}$$

The first part of equation (5) can be expressed as equation (6) since *MI* is symmetric.

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(X) + h(Y) - h(X, Y) \end{aligned} \tag{6}$$

3 Machine Learning Model

The focus of this section is on supervised learning models, in which a learning algorithm is provided with both dependent and independent variables. That is, the learning algorithm has the right answer in advance and it will learn its parameters or hypothesis based on the available observations or examples. In general, supervised learning models are used for solving two kinds of prediction problems namely (1) Regression problems which involve training a model to solve or predict a continuous value output and (2) Classification problems which involve training a model to solve or classify a discrete value output.

It is strongly believed that feature selection helps in improving the predicting capabilities of machine learning models since it reduces the dimension of a dataset by selecting variable that are relevant to the predicting attributes. Other economic benefits of feature selection to learning models includes (1) building a concise model that avoid overfitting and generalized better, (2) improves the accuracy of prediction due to a reduction in estimation errors and (3) reduces burdens on data collection & computational complexities. So, the proper way to test the effectiveness of any feature selection is to see the relative performance of a choosing learning model based on a selected optimal feature subset and compare with its performance based on full feature set.

In this study, the overall performance of a learning model is computed using a global metric named Root Mean Squared as in equation 7.

$$RMSE = \sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N} \tag{7}$$

RMSE metric is sensitive to high errors and its lower value results in better predictive models.

3.1 Regression learning model

Linear regression tried to fit a straight line "h" in a training set. The accuracy of the fit depends on the model parameter. In figure 1, the blue line is the model line which is plotted against the actual data in red "*". The red line from the actual data points to the model line is the model error, and it indicates the difference between the models predicted value and the actual training value. The idea here is to choose the model parameter θ_0 and θ_1 (as the intercept on the y-axis and the slope respectively) so that $h_\theta(x)$ is closed to y for training examples (x, y). The goal is to find model parameters that will minimize the model sum of squared errors in equation 8:

Figure 1: Model Representation

$$\text{Cost function: } J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)} - y^{(i)})^2$$

$$\text{Goal: } \min_{\theta_0, \theta_1} imize \; J(\theta_0, \theta_1) \tag{8}$$

3.2 Gradient descent

This is the algorithm used to minimize different functions. In this study gradient descent is used to minimize the cost function for both Regression and Artificial Neural Network models (see table 1). The general idea is to start by picking a random combination of the parameters *(Θ_0, …, Θ_n)* and see what the cost function is, then search for the very next combination that will minimize the cost function until it reaches the local minimal.

Table 1: Gradient Descent Algorithm

Repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1), \text{ for (j=0 and j=1)}$$

}

α is the learning rate and it controls the step taking by the algorithm when looking for a local minimum. And both Θ_0 *and* Θ_1 must be updated simultaneously.

3.3 Normal Equation Learning Model

Normal equation minimized the cost function for theta (Θ) by solving the equation 9.

$$\frac{\partial}{\partial \theta j} J(\theta_j) = 0 \tag{9}$$

when the derivative of J is equal to zero, it is where the cost function is minimized. Thus, normal equation model, unlike gradient descent, works for linear model only, it requires no learning rate and solve in 'one shot'.

3.4 Artificial Neural Network

ANN was developed by simulating human neurons or network of neurons in the brain. Neurons are cells with cell body and dendrites which are the input wires connected to the cell body and receive signals from body receptors. In addition to these two, neurons have output wire known as axon which sends signals or informative messages to other neurons. Thus, at a primitive level, a neuron is a computational unit that receives a number of inputs (electric pulses) through its input wires (dendrites), does some computation in the cell body and sends output through its axon to other neurons in the brain.

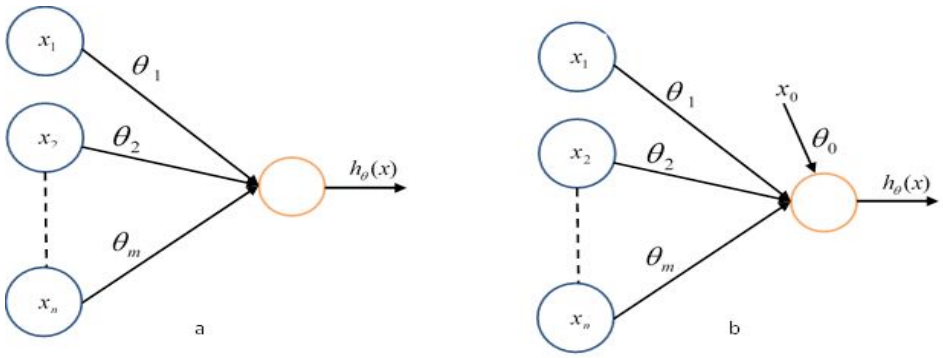


Figure 2: Neuron model: (a) without bias, (b) with bias input x_0 *and weight* θ_0 explicitly shown.

Figure 2 depicted ANN implementation of a simple model of a neuron; the empty circle (in orange) plays a role analogous to the body of neuron, each arrow linking an input xi to the neuron, has attached a parameter, or weight, θ_i, the diagram represents the computation of a sigmoid fed with the dot product of x and θ in the form of:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{10}$$

where x and θ are the parameter vectors:

$$x = \begin{bmatrix} x_0 \\ x_1 \\ . \\ . \\ x_m \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ . \\ . \\ \theta_m \end{bmatrix}$$

With this brief description and illustration, ANN is defined as a collection of different artificial neurons working together to form a predictor or classifier.

4 Proposed Feature Selection Procedure Based on Mutual Information

This section presents *MI* estimator as formulated by Kraskov, Stogbauer and Grasssberger (2004; 2008). Then followed by the improving feature selection by mutual information as well as introduction of stopping criterion for greedy procedure for feature selection as proposed by Verleysen, Rossi and François (2009). And lastly, the proposed enhancement on the feature selection method by *MI*.

The *MI* estimation presented in this study is based on the equation 6, which define *MI* in terms of entropy of *X, Y* and joint differential entropy of *X* and *Y*.

4.1 Mutual Information Estimation

The approach presented here is based on k-nearest neighbor consistent entropy estimator which was first formalized by Kozachenko-Leonenko (1987). And it is built upon, by Kraskov, Stogbauer and Grassberger (2004; 2008) to formulate *MI* estimator Kozachenko and Leonenko (1987) idealized the differential entropy estimator which can be estimated from data as presented in equation 11:

$$\hat{h}(x) = -\psi(K) + \psi(N) + \log c_D + \frac{D}{N}\sum_{n=1}^{N} \log \varepsilon(n, K) \tag{11}$$

where K is the parameter of the estimator, ψ function given by equation 12, N the number of samples in the dataset, D is the dimensionality of X, c_D the volume of unitary ball in a D-dimensional space and $\varepsilon(n, K)$ is twice the distance from a point x_n to its K-th neighbor.

$$\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)} = \frac{d}{dz}\ln\Gamma(z) \tag{12}$$

Based on equation 11, Kraskov, Stogbauer and Grassberger (2004) intuitively formulate mutual information estimator as:

$$MI(X,Y) = \psi(K) + \psi(N) - \frac{1}{N}\sum_{n=1}^{N}(\psi(\tau_x(n)) + \psi(\tau_y(n))) \tag{13}$$

where $\tau_x(n)$ *and* $\tau_y(n)$ are points in X and Y respectively whose distance from x_n *and* y_n respectively is strictly less than ϵ_n. ϵ_n is an infinite norm between x_n and its k-nearest neighbor expressed as:

$$\epsilon_n = \| z_n - z_{k(n)} \|_\infty = max(\| x_n - x_{k(n)} \|, \| y_n - y_{k(n)} \|) \tag{14}$$

K in equation 13 is known as a smoothing parameter which is to be choosing with care. It determines how effective is the equation 13 in estimating mutual information between X and Y. Secondly, like the traditional mutual information estimator, equation 13 is suitable for a low dimensional dataset. Since the accuracy of the estimator decreases as the dimension of selected future increases.

4.2 Improving feature selection based on MI and proposed stopping criterion for greedy procedure

This section provides details of stopping criteria and how to cope with bias/variance associated with the *MI* estimation as introduced by Verleysen, Rossi and François (2009).

- Smoothing Parameter

An empirical study by Battiti (1994), Kraskov, Stogbauer and Grassberger, (2004), Liu, Liu and Zhang (2008) have shown that choosing a value of smoothing parameter have an influence in bias/variance problem. A small value of K leads the estimator to have small bias and high variance, while a large value of K, makes the estimator have small variance and high bias. However, using student's t-test quantity in equation 15 whose parameters are obtained from resampling procedure as presented in François, et al. (2007) and Gringarten (2012) provides a means of choosing K that balanced for these two problems.

$$t_K = \frac{\mu_K - \mu_{K,\rho}}{\sqrt{\sigma_K^2 + \sigma_{K,\rho}^2}} \tag{15}$$

Detail flowchart for selecting smoothing parameter 'K' is presented in figure 3, $\widehat{I_K}(X;Y)$ & $\widehat{I_K}(X;p(Y))$ are two empirical distributions obtained from the cross-validation and permutation operations. ρ denote permutation operation and the subscript K emphasized the sensitivity of K in *MI* estimation. A good choice of K is where both empirical variance and mean of $\widehat{I_K}(X;Y)$ & $\widehat{I_K}(X;p(Y))$ are small. The difference between the two distributions can be measured based on equation 15, where μ_K & $\mu_{K,\rho}$ and σ_K & $\sigma_{K,\rho}$ are empirical means and standard deviations of $\widehat{I_K}(X;Y)$ and $\widehat{I_K}(X;p(Y))$ respectively.

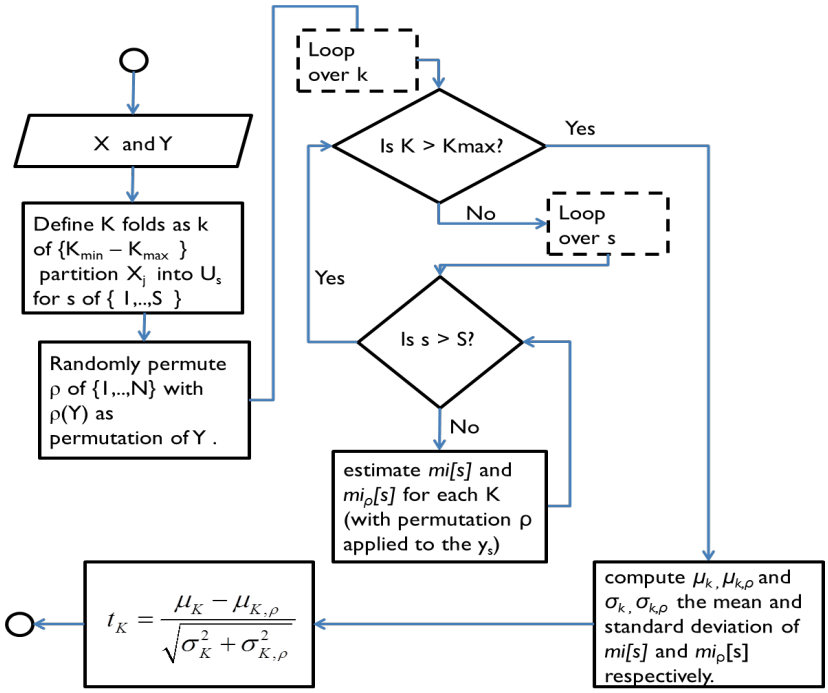


Figure 3: Algorithm for selecting K (smoothing parameter)

- Stopping Criterion

The method described here is the one presented in Verleysen, Rossi and François (2009) which is also based on resampling and random permutation operation.

The idea is to build two resampling distributions for $MI(S \cup X_{st}, Y)$ and $MI(S \cup \rho(X_{st}), Y)$, where $\rho(X_{st})$ is a randomized X_{st}. $\rho(X_{st})$ is generated independently from Y through permutation operation and X_{st} is the candidate variable. Figure 4 is the flowchart on how to stop greedy forward algorithm for feature selection.

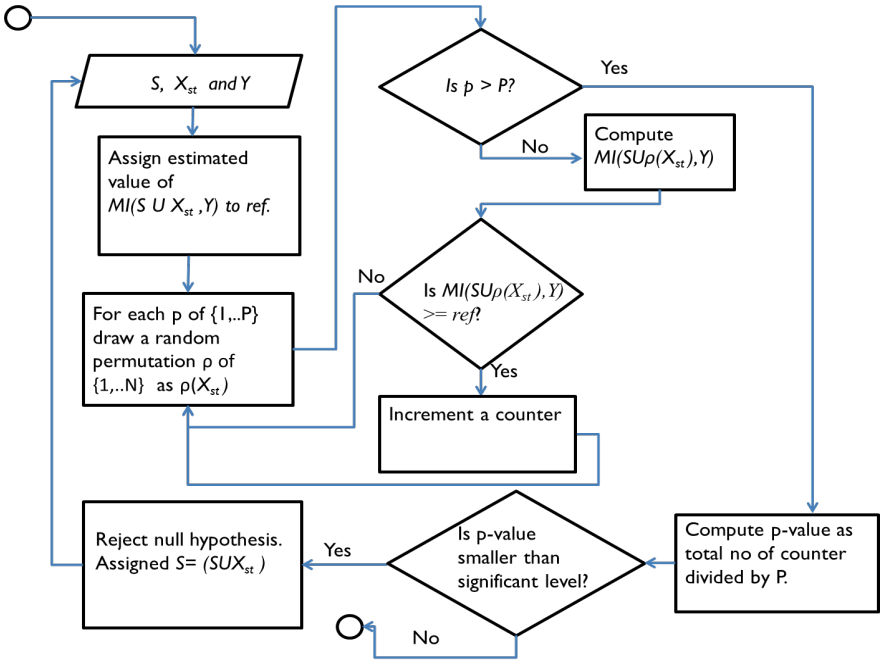


Figure 4: Algorithm for Stopping Criteria

4.3 Enhancement on Feature Selection Procedure based on Mutual Information

Despite that the resampling and permutation methods were effective in selecting feature subsets, investigation on this method (Sulaiman and Labadin, 2015**b**) suggest that this method needs further enhancement in order to select reliable optimal feature subsets. In practice, the resampling and permutation methods result in inconsistent suboptimal feature subsets each time the procedure is run. The suggested enhancement is presented in this section.

- Methodology for selecting optimal feature selection

After running the simulation for several times, and each run produces different sets of feature subsets. The procedure presented in figure 5 is used to consider a feature into optimal sets. For each feature, the number of times it appears in the total sets of feature subsets generated is counted. Based on a set threshold, an optimal feature subset can be regarded as Fine-grained feature subset (FGFS) selection containing strongly relevant features only or Coarse-grained feature subset (CGFS) selection containing both strongly and weakly relevant features. The intuition here is that if for example in 10 runs, feature 'A' appears 8 (or above) times in a total of 10 different feature subsets generated in 10 runs, this could be a good indication of how strong affinity is feature 'A' to the compared targeted variable.

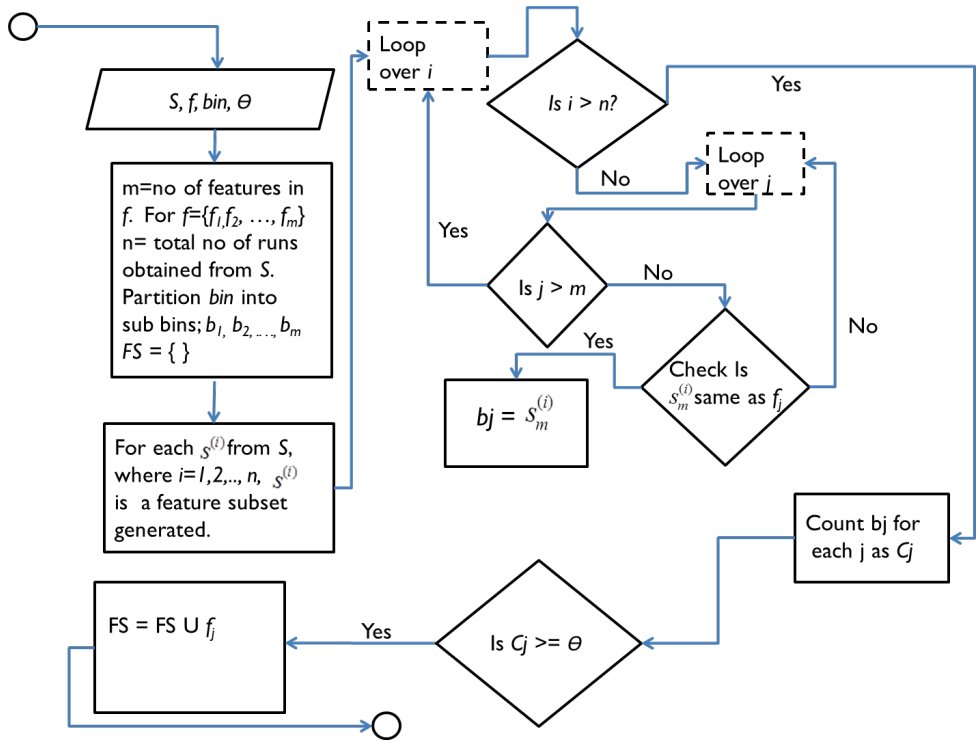


Figure 5: Algorithm for selecting optimal feature subset

S, f, bin and Θ in figure 5 are the input. Where *S* contains a total set of all feature subsets, *f* is a set of feature labels, *bin* is used to collect and count the number of times each feature appears in total set in *S and Θ* is set as a threshold for considering a feature into an optimal subset based on its count. And *FS* is the output containing the selected optimal feature subset.

5 Experimental Studies

This section comprises of the description of the various dataset used in this work, experimental setups, simulation, results obtained and discussion.

5.1 Datasets

The focus of this work is to investigate feature selection based on MI estimator for regression task, in this regards three regression datasets were obtained from UCI machine learning repository and one private regression dataset. The four datasets are used for investigation in this study.

5.1.1 Forest fire dataset: this dataset is described as very difficult regression task and is publically available for research. Cortez and Moris (2007) provide the detail description of the dataset. It has 517 instances, 12 input features/variables, and 1 output or target variable with no missing attribute value. Many of the attributes may be correlated, as such suitable for testing feature selection methods.

5.1.2 Concrete dataset: concrete dataset is a public dataset obtained from UCI machine learning repository. The target attribute is the concrete compressive strength, which is a quantitative variable. And it is responsible for categorizing the dataset as regression task. The dataset has 1,030 instances and 8 quantitative input features namely cement (kg), blast furnace (kg), water (kg), superplasticizer (kg), coarse aggregate (kg), fine aggregate (kg) and age (day). Detail description of the concrete dataset is available in Yeh (1998), in which Yeh (1998) modeled concrete compressive strength using Artificial neural network and concluded that strength model based on ANN is more accurate than a model based on regression analysis.

5.1.3 Communities dataset: the data comprises of socio-economic data from1990 US Census. No previous published results from the dataset; however, Redmond and Baveja (2002) investigate a related dataset and provided the detail description of the dataset. Community's dataset is multivariate with real type of attributes. It has 1994 instances, 127 input variables, and the target attribute is "the total number of violent crimes per 100k population". However, 122 input variables are predictable why 5 are unpredictable. And due to a large number of missing data further reduction of the input dataset to 99 was made. Many variables are included to make the dataset suitable for testing feature selection methods, but none of the variables is unrelated to the crime. And all predictable variables are normalized into the decimal range of 0.00 – 1.00. Like forest fire and concrete data sets, communities' dataset is publically available at UCI machine learning repository.

5.1.4 Well log dataset: It is private data collected from wells of a Middle Eastern region. The dataset consists of 12 well logs (feature set) and permeability as the target core log. The dataset consists of 880 instances. The well log data variables used for this study are DEPTH (depth), MSFL (Micro spherically Focused Log), DT (Sonic travel time), NPHI (Neutron porosity), PHIT (Total porosity), RHOB (Bulk density), SWT (Water saturation), CALI (Caliper log), CT (Electric conductivity), DRHO (Density), GR (Gamma Ray Log) and RT (Deep Resistivity), and permeability which is the output reservoir attribute.

5.2 Experimental setups

Each of the four datasets is divided into training, validation and test sets in the ration of 3:1:1 respectively. The k nearest neighbor obtained for each feature is in accordance with the procedure described in the previous section. K value was used with the training data to select relevant features by applying the forward procedure and stop according to the proposed stopping algorithm. A significant level of 0.01 is used with the estimated p-value for the null hypothesis. Also, in this study 80% of the total sets of feature subsets generated is set as threshold for Fine-grained feature set (FGFS) selection while 70% of the total sets of feature subsets generated is considered as threshold for Coarse-grained feature set (CGFS) selection. The threshold values were set empirically in accordance with the proposed procedure for selecting optimal feature subset.

After successful generation of optimal feature subsets for each dataset, the generated features subsets are tested for a performance alongside the full feature sets using two machine learning techniques namely multiple regression and artificial neural networks models. Two forms of multivariate regression models were implemented, the first form used gradient descent to minimizes the cost function. And the second form used the normal equation to minimizes cost function in one shot. Finally, each model with multiple input variables and single layer continuous output variables are implemented. In addition, one hidden layer is used for the artificial neural network architecture.

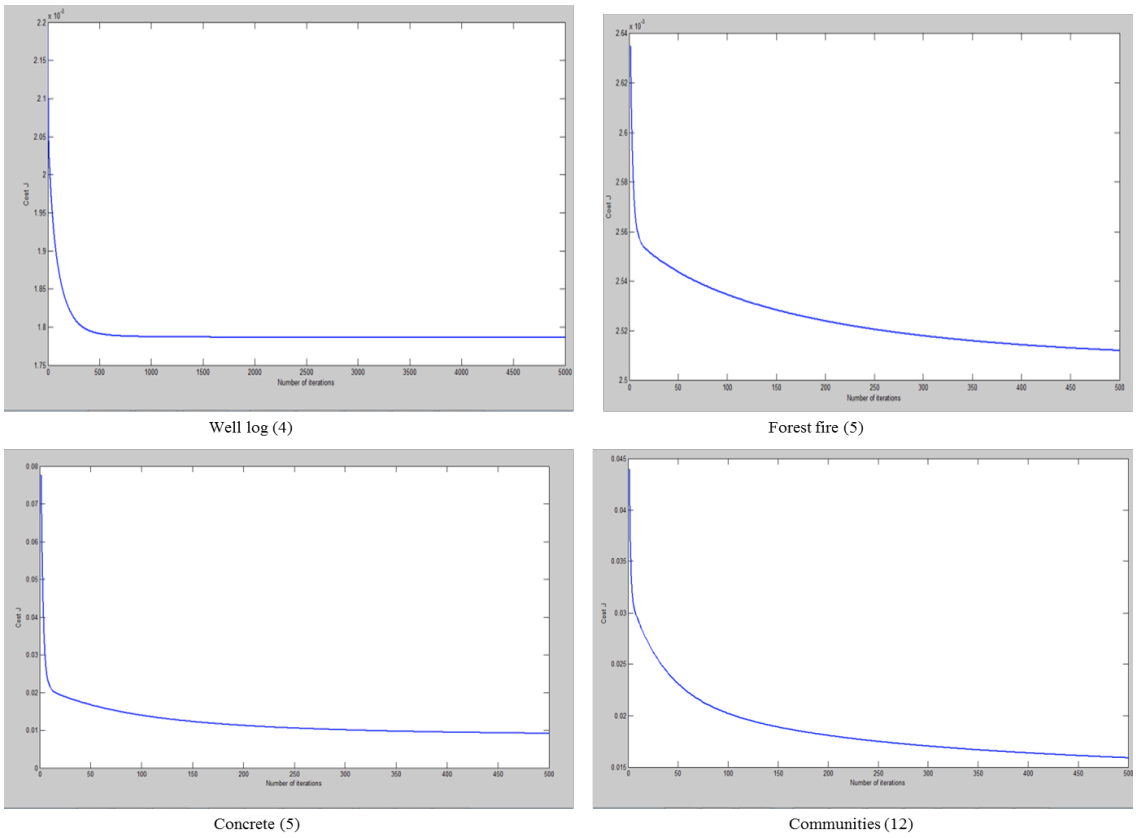
5.3 Results and discussion

Before training the models, both artificial neural networks and multiple regression models were evaluated to ensure that the learning process during training stage descent as expected. Figure 6 to figure 8 present gradient descent evaluations for regression models. Figure 6 is the FGFS selection in which numbers in the bracket represents the number of features in a selected feature subset by the *IFSMIR* method. Figure 7 is the CGFS selection and figure 8 is the full feature sets. For concrete dataset, the CGFS selection is the same as full feature sets. In all cases, the cost function is minimized.

Table 1 presents the number of features in an optimal feature subset selected by *IFSMIR* method and the *k* value. From the table 1, FGFS selection is more restricted compared to CGFS selection. Full feature set in table1 is the original feature sets for each data set before applying *IFSMIR* method to generate FGFS and CGFS. A close examination of CGFS selection, one can easily see that is a loose form of FGFS selection and it consists of both strongly and weakly relevant features.

Table 1: Selected Fine-grained and Coarse-grained feature subset selections using IFSMIR method.

Dataset	Full feature set	Fine-grained Feature Set (FGFS)	Coarse-grained Feature Set (CGFS)	K value
Well log	12	4	8	4
Forest Fire	12	5	10	1
Concrete	8	5	8	1
Communities	99	12	29	3



Well log (4) Forest fire (5)

Concrete (5) Communities (12)

Figure 6: Gradient Descent for Fine-grained feature subsets (FGFS). *X*-axis is the number of iterations and *Y*-axis is the value of Cost J

Table 2 to table 4 present the results of 10 runs based on *IFSMIR* method for three datasets namely, well log, fire forest and concrete. The "***Feature Subset Category***" refers to the list of features or variables in each dataset and the "***No of occurrences of the feature subset***" is the number of times each feature is selected to a suboptimal set in the total runs.

Table 2: Results of 10 runs for well log dataset using IFSMIR method.

Selected Feature Subsets												
Feature Subset Category	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***
No of occurrences of the feature Subset	***8***	***6***	***5***	***6***	***5***	***8***	***6***	***8***	***10***	***4***	***4***	***6***

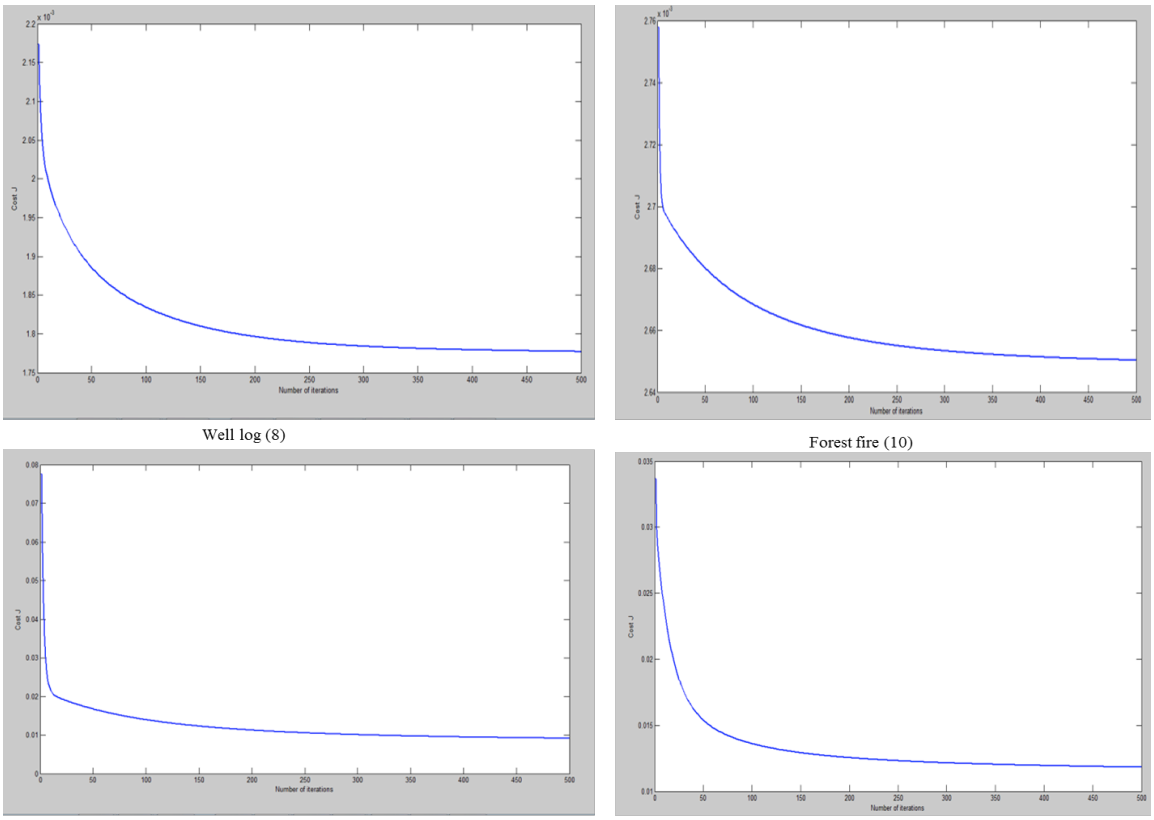


Figure 7: Gradient Descent for Coarse-grained feature subset (CGFS). *X*-axis is the number of iterations and *Y*- axis is the value of Cost J

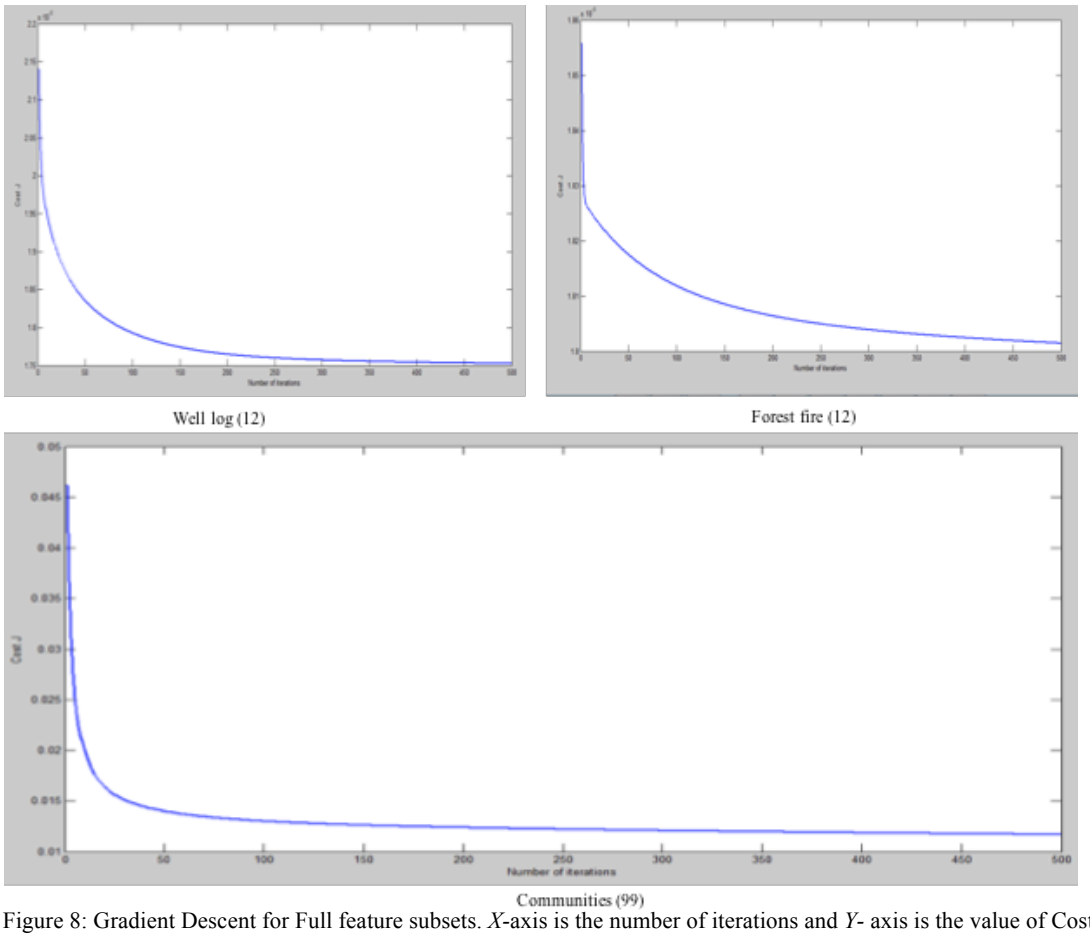


Figure 8: Gradient Descent for Full feature subsets. *X*-axis is the number of iterations and *Y*- axis is the value of Cost J

Table 3: Results of 10 runs for forest fire dataset using IFSMIR method.

Selected Feature Subsets												
Feature Subset Category	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***
No of occurrences of the feature Subset	***9***	***6***	***5***	***9***	***6***	***4***	***6***	***8***	***8***	***6***	***6***	***9***

Table 4: Results of 10 runs for concrete dataset using IFSMIR method.

Selected Feature Subsets								
Feature subset category	1	2	3	4	5	6	7	8
No of occurrences of the feature Subset	7	9	8	9	7	6	9	8

Table 5: Overall performance Metric for predictive models

Dataset	**Predictor model**	**RMSE for FGFS**	**RMSE for CGFS**	**RMSE for Full feature set**
Well log	Regression	0.1136	0.1131	0.1126
Well log	Normal Equation	0.1136	0.1207	0.1703
Well log	Neural Network	0.1133	0.19333	0.1946
Forest fire	Regression	0.0348	0.0232	0.0700
Forest fire	Normal Equation	0.0350	0.0233	0.0700
Forest fire	Neural Network	0.0343	0.0227	0.0691
Concrete	Regression	0.1682	0.1278	-
Concrete	Normal Equation	0.1616	0.1190	-
Concrete	Neural Network	0.1333	0.0929	-
Communities	Regression	0.2057	0.1745	0.1393
Communities	Normal Equation	0.1883	0.1785	0.1729
Communities	Neural Network	0.1910	0.1704	0.1663

In order to test the effectiveness of *IFSMIR* method in selecting optimal feature subsets. This study compares three categories of the feature sets based on Root mean squared error (RMSE) values of the predictive models. The three categories of the feature sets are the FGFS selection, the CGFS selection, and the full set feature set. Two machine learning model were used in this regard, the Regression model, and Artificial neural networks. The regression model is of two categories namely multivariate regression with gradient descent as cost minimization function (simply refers as Regression in table 5) and multivariate regression with the normal equation as cost minimization function (refers as Norma Equation in table 5). RMSE is a global metric which is sensitive to high errors and whose lower value indicates a better predictive model. Table 5 presents the performance evaluation of the IFSMIR method based on the RMSE values of predictive models.

In well log dataset, Regression model seems to perform better in the overall performance of learning models. While FGFS selection is the optimal feature set that gives a competitive performance for all the learning models based on low RMSE values. The performance of neural network agrees with Sulaiman and Labadin (2015**a**) in which the same dataset was used in a similar study.

In Forest fire dataset which was described as difficult regression task since its several attributes seem to be correlated (Cortez and Moris, 2007), Neural network performed all round better compared to the two other models and CGFS selection is the optimal feature set compared to the other two categories.

The same trend in forest fire dataset is seen in Concrete dataset with neural network performing all round better compared to the two other predictive models and CGFS selection is the optimal feature set as well. This result agrees with Yeh (1998) who concluded that a strength model based on ANN is more accurate than a model based on regression analysis. Moreover, it performed better in terms of low RMSE value of 0.0929 compared to Yeh (1998) best RMSE value of 0.814.

Though no records of previous prediction study using the communities' dataset, it appears that IFSMIR method is less effective in communities' dataset. The reason for this is not clear, however, it may be that the set thresholds for both

FGFS and CGFS selections respectively are not suitable for a dataset with large features or variables. This may be evidence in the number of features in feature subsets generated for each run, which is between 58 to 68, but the majority of them could not appear in up to 30% of the total sets of feature subsets, this suggests further studies in order to reach a conclusion. However, Frenay, Doquirel and Verleysen (2013) concluded that with regards to mutual information as a selection criterion for regression tasks, feature selection could give suboptimal results. Meanwhile, CGFS optimal feature set gives a competitive performance for all learning models after the full feature set, based on the low value of RMSE. And regression model performed much better with full feature set compared to the other two machine learning models.

6 Conclusions

This study focuses on feature selection based on mutual information for regression tasks. To the best of our knowledge, this is the first time several real-world datasets were used with the k-nearest neighbor mutual information estimator as feature selection criterion function. Previous works focus on the use of randomly generated relationship functions to validate the feature selection method.

Also, in this study a methodology for selecting optimal feature sets is proposed. This is to serve as an enhancement to the method of feature selection from previous work by Verleysen, Rossi and François (2009).

The idea of fine-grained feature set (FGFS) and coarse-grained feature set (CGFS) selections has been introduced. And based on the number of datasets and the learning models used in this study, ANN performs all round better based on the RMSE values as a global metric. Multivariate regression model with normal equation although runs fast in one shot without the need for training, performed least. ANN work well with both Forest fire and Concrete datasets using CGFS as optimal feature set, FGFS as an optimal feature set gives competitive results for Well log dataset with all the learning models. And multivariate regression with gradient descent gives best result based on full set feature set for Well log and communities datasets.

In future, the used of hybrid filter-wrapper methods will be considered, so as to compare the effectiveness of *IFSMIR* method with evolutionary search method.

References

Battiti, R. (1994). Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transaction on Neural Networks*, 5.

Cortez, P. & Morais, A. (2007). A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos & J. Machado (Eds.), *New Trends in Artificial Intelligence*, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, 512-523. Available at: http://www.dsi.uminho.pt/~pcortez/fires.pdf

Doquire, G. & Verleysen, M. (2011). Feature Selection with Mutual Information for Uncertain Data. *Springer Link Data Warehousing and Knowledge Discovery Lecture Notes in Computer Science*, 6862, 330-341.

Evans, D. (2008). A Computationally efficient estimator for mutual information. *Proc. R. Soc. A*, 464, 1203–1215.

François, D., Rossi, F., Wertz, V. & Verleysen, M. (2007). Resampling methods for parameter-free and robust feature selection with mutual information. *Neurocomputing,* 70, 7-9, 1276-1288.

Frénay, B., Doquire, G., & Verleysen, M. (2013). Is mutual information adequate for feature selection in regression? *Neural Networks Letter*, 48, 1-7.

Gringarten, E. (2012). Integrated uncertainty assessment – from seismic and well-logs to flow simulation. PARADIGM, SEG Las Vegas 2012 Annual Meeting.

Guyon, I. & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3, 1157-1182.

Kozachenko, L. F. & Leonenko, N. N. (1987). Sample estimate of entropy of a random vector. *Probl. Inf. Transm.*, 23, 95–101.

Kraskov, A., St¨ogbauer, H. & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69, 066138.

Latorre Carmona, P., Sotoca, J.M., Pla, F., Phoa, F.K.H., Bioucas Dias, J. (2011), Feature Selection in Regression Tasks Using Conditional Mutual Information. *Pattern Recognition and Imag e Analysis Volume 6669 of the series Lecture Notes in Computer Science*, 224-231.

Liu, H., Liu, L. & Zhang, H (2008). Feature Selection Using Mutual Information: An Experimental Study. *PRICAI 2008, LNAI 5351*, Springer-Verlag Berlin Heidelberg, 235–246.

Liu, H. & Yu, L (2005). Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17.

Peng, H., Long, F. & Ding, C. (2005). Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27.

Pudil, P., Novovicova, J. & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15, 1119-1125.

Redmond, M. A. & Baveja, A. (2002). A Data-Driven Software Tool for Enabling Cooperative Information Sharing Among Police Departments. *European Journal of Operational Research*, 141, 660-678.

Rossi, F., Lendasse, A., François, D., Wertz, V. & Verleysen, M. (2006). Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and Intelligent Laboratory Systems*, 80, 2, 215-226.

Song, L., Smola, A., Gretton, A., Bedo, J. & Borgwardt, K. (2012). Feature Selection via Dependence Maximization. *Journal of Machine Learning Research*, 13, 1393 – 1433.

Sulaiman, M. A. & Labadin, J. (2015a). Feature Selection Based on Mutual Information for Machine Learning Prediction of Petroleum reservoir properties. *The 9th International Conference on IT in Asia (CITA)*, 1-6, DOI: 10.1109/CITA.2015.7349827.

Sulaiman, M. A. & Labadin, J. (2015b). Feature Selection with Mutual Information for Regression Problems. *The 9th International Conference on IT in Asia (CITA),* 1-6. DOI: 10.1109/CITA.2015.7349826.

Unler, A., Murat, A. & Chinnam, R.B. (2011). Mr2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. *Elsevier Journal of Information Sciences,* 181, 4625 – 4641.

Verleysen, M., Rossi, F. & François, D. (2009). Advances in Feature Selection with Mutual Information, arXiv:0909.0635 [cs.LG].

Yan, S., Wang, H., Huang, T. S., Yang, Q. & Tang, X. (2007). Ranking with Uncertainty Labels. In *Proceedings of IEEE International Conference on Multimedia and Expo*.

Yu, L. and Liu, H (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proceeding of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC.