Relief for Regression with Missing Data in Variable Selection

By

Luquan Li

A thesis submitted to the
Graduate School-New Brunswick
Rutgers, The State University of New Jersey
In partial fulfillment of the requirements
For the degree of
Master of Science
Graduate Program in Operations Research
Written under the direction of
Prof. Myong K. Jeong
And approved by

________________________
________________________
________________________
________________________

New Brunswick, New Jersey
May 2014
ABSTRACT OF THE THESIS

RELIEF FOR REGRESSION WITH MISSING DATA IN VARIABLE SELECTION

By LUQUAN LI

Thesis Director: Prof. Myong K. Jeong

Variable selection is a significant pre-processing task for prediction in the field of data mining and machine learning, and involves the selection of a subset of relevant variables. Almost all researchers have faced the problem of missing data, which can occur due to nonresponse or loss of information. This thesis develops a new variable selection technique for dealing with missing data. Relief is an algorithm for estimating the quality of each variable and is applicable to categorical or continuous data. This thesis presents a new variable selection method, RM-Relief, by extending Relief to select the variables in a regression with missing data. RM-Relief weights all predictor variables by assigning bins for a response variable and estimating the conditional probability of unknown instances. Results on artificial and real-world datasets indicate that RM-Relief works well on regression problems with missing data.

Keywords: Variable Selection, Missing data, Relief, RM-Relief
# TABLE OF CONTENT

1. **INTRODUCTION** 4

2. **LITERATURE REVIEW** 4
   2.1 **MISSING DATA PATTERNS AND MISSING DATA MECHANISMS** 4
   2.2 **INFERENCE AND MISSING DATA** 4
      2.2.1 **COMPLETE CASE ANALYSIS** 5
      2.2.2 **AVAILABLE CASE ANALYSIS** 5
      2.2.3 **IMPUTATION BY REGRESSION** 6
      2.2.4 **MAXIMUM-LIKELIHOOD ESTIMATION** 7
   2.3 **APPROACHES FOR HANDLING MISSING DATA** 9
      2.3.1 **CASE DELETION** 9
      2.3.2 **MEAN IMPUTATION** 9
      2.3.3 **K NEAREST NEIGHBOR (KNN) IMPUTATION** 9
      2.3.4 **REGRESSION-BASED IMPUTATION** 9
      2.3.5 **EXPECTATION – MAXIMIZATION (EM) ALGORITHM** 9
   2.4 **RELIEF ALGORITHMS** 10
      2.4.1 **RELIEF ALGORITHM FOR CLASSIFICATION** 11
      2.4.2 **RELIEF ALGORITHM FOR REGRESSION** 12

3. **NEW VARIABLE SELECTION METHOD FOR REGRESSION WITH MISSING DATA** 13
   3.1 **RELIEF FOR REGRESSION WITH MISSING DATA (RM-RELIEF)** 13
   3.2 **STEPS OF RM-RELIEF** 15

4. **CASE STUDY** 17
   4.1 **CASE ONE** 17
      4.1.1 **COMPLETE CASE** 18
      4.1.2 **MISSING DATA (20% MISSING AT MCAR)** 20
   4.2 **CASE TWO** 25
      4.2.1 **COMPLETE DATA SET** 25
      4.2.2 **DATA SET WITH 5, 10, 15, AND 20% MISSING DATA** 27

5. **CONCLUSION** 29

**REFERENCE** 30
1. Introduction

Variable selection is a technique for selecting a subset of relevant variables from a large pool of available variables for use in model establishment. The goal of variable selection is to eliminate redundant or irrelevant variables, which provide no additional information or no useful information, and to choose the variables that have ample discriminatory information for classification/regression problems. Diminishing the number of variables into a sufficient minimum has many benefits, such as reducing data storage, shortening the training time, decreasing computational complexity, improving model performance, and enhancing generalization.

However, in real research, some of the data entries are lost or cannot be observed; these are referred to as missing data. The missing data problem represents the situation where no information is provided for some instances or variables; this happens in a broad range of applications. In this thesis, we propose a new variable selection algorithm named Relief for Regression with Missing Data (RM-Relief). RM-Relief uses approximate conditional probability and transfers a continuous response variable Y in a range to estimate the difference between missing values and observed data, and then conducts variable selection.

2. Literature Review

2.1 Missing Data Patterns and Missing Data Mechanisms
Many statistical methods are available today to analyze or explain datasets and a value can be unknown in a dataset for numerous reasons [Doquire and Verleusen, 2012]. For example, dysfunctional equipment or damaged sensor devices could fail to transmit data in industrial production, while in the medical area a doctor might not always be able to conduct all experiments on each patient. The missing data problem is especially common in the social science area, in fields like economics, sociology, or political science. On occasion, the researcher might fail to collect the data properly or might make errors in data entry. Furthermore, if the surveys are conducted to collect data, some participants may refuse to answer certain personal or sensitive questions, such as those involving political opinions, religions, private subjects, or financial income, which can then lead to missing data.

Little and Rubin [1987] defined the pattern of missing data, describing the observed data values, missing values, and the mechanism of missing data, as well as the relationship between values and missing data of variables in a dataset. In general, the pattern considers which values are missing, but the mechanism focuses on whether the missing data relate to the variables.

Let us use $X = [x_{ij}]$ to denote an $n \times K$ data matrix, where $x_{ij}$ is the value of variable $X_j$ for sample $i$. We then define $M = [m_{ij}]$ as the matrix of the missing data indicator. In matrix $M$, $m_{ij} = 1$ represents $x_{ij}$ where $X$ is missing; otherwise, $m_{ij} = 0$. The matrix $M$ represents the missing data pattern.

Missing data patterns are typically one of four types. We can use figures to illustrate these types as an easier way to visualize the structures of datasets with missing values.
In the following figures [Toutenburg and Heumann, 2002], the bars represent the number of observed values, which correspond to predictor variables $X_1, X_2, \ldots, X_n$ and response variable $Y$.

[Diagrams of missing data patterns]

The first figure shows the situation where only one variable has missing values while the others are complete. In the second figure, the variables have the pattern of monotone missing data. If $X_j$ is missing, then all variables $X_k$, with $k > j$ are also missing. In figure 3, the special missing data are such that when $X_i$ is observed, $X_j$ is missing, where $i \neq j$. No instance has both values of $X_i$ and $X_j$ at the same time. Figure 4 illustrates a common situation of missing data.

A missing data pattern offers a first impression of missing data, but if we want to study the relationship between observed values and missing values, the mechanism should be taken into consideration.
The missing data mechanism is based on conditional distribution of $M$, given $X$, as $f(M|X, \phi)$, where $\phi$ represents unknown parameters. Little and Rubin [1987] defined the data matrix $X = (X_{\text{obs}}, X_{\text{mis}})$, which means the data mechanism is established on the basis of the density function $f(M|X_{\text{obs}}, X_{\text{mis}}, \phi)$.

### 2.2 Inference and Missing Data

Using the concept of missing values, here we divide the response variable $Y = (y_1, ..., y_n)$ into two parts, $y_{\text{obs}}$ and $y_{\text{mis}}$ [Rao, C. R. and Toutenburg, H, 1999]. $y_{\text{obs}}$ represents all $y_i$ with complete data $X_i$:

$$y_{\text{obs}} = X_{\text{obs}} \beta + \epsilon_{\text{obs}}$$

$y_{\text{mis}}$ contains the $X_j$ with partially or fully missing values:

$$y_{\text{mis}} = X_{\text{mis}} \beta + \epsilon_{\text{mis}}$$

Therefore, $y_{\text{obs}}, y_{\text{mis}}$, and $X_{\text{obs}}$ are the datasets with complete values, and $X_{\text{mis}}$ is the dataset with partially or fully missing data.

We then combine the above two equations into one mixed model:

$$
\begin{pmatrix}
  y_{\text{obs}} \\
  y_{\text{mis}}
\end{pmatrix}
= 
\begin{pmatrix}
  X_{\text{obs}} \\
  X_{\text{mis}}
\end{pmatrix}
\beta + 
\begin{pmatrix}
  \epsilon_{\text{obs}} \\
  \epsilon_{\text{mis}}
\end{pmatrix}
$$

Due to the unobserved elements in $X_{\text{mis}}$, the not operational estimator $\hat{\beta}(X_*)$ is given by the mixed estimator of $\beta$ in the model [Rao and Toutenburg, 1999]:

$$
\hat{\beta}(X_*) = (X_*'X_* + X_*'X_{\text{c}})^{-1}X_*'y_* + X_*'y_*
$$

The estimator $\hat{\beta}(X_*)$ is unbiased for $\beta$ and has the covariance matrix:
\[ V(\beta(X_*)) = \sigma^2(X'_cX_c + X'_*X_*)^{-1} \]

### 2.2.1 Complete Case Analysis

The easiest way to conduct analysis is to use the completely observed data matrix \(X_{obs}\) only and discarding the elements with missing values contained in \(X_{mis}\), which is called classical LSE [Toutenburg and Heumann, 2002]. The corresponding estimator of \(\beta\) is:

\[ b_{obs} = X'_{obs}X_{obs}^{-1}X'_{obs}y_{obs} \]

For simplification, we let \(S_{obs} = X'_{obs}X_{obs}\) and \(S_{obs} = X'_{obs}X_{obs}\).

The complete case analysis excludes all elements for which any of the inputs are missing. However, this method may lose the information of observed values in \(X_{mis}\) for variable selection, when compared with estimating the missing values by first applying some imputation algorithms.

### 2.2.2 Available Case Analysis

The available case analysis approaches the problem by basing different analyses on distinguished subsets of data. Suppose the dataset \((X_1, ..., X_k, y)\) has a joint distribution with mean \(\mu = (\mu_1, ..., \mu_K, \mu_y)\) and covariance matrix \(\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \sigma_{yy} \end{pmatrix}\).

The \(\beta\) can be estimated by \(\hat{\beta} = \Sigma_{yx}\Sigma_{xx}^{-1}\).

Haitovsky [1968] conducted a Monte-Carlo experiment to investigate the performance between complete case analysis and available case analysis by adopting
various patterns of missing data, and arrived at the conclusion that the complete analysis $b_{obs}$ is better than the $\beta$ of the other methods in most situations.

2.2.3 Imputation by Regression

Here, we introduce two regression methods for imputation: imputation by Zero-Order-Regression (ZOR) and imputation by First-Order-Regression (FOR). Zero-Order-Regression, also called mean imputation, substitutes the missing value $x_{ij}$ from sample $i$ and variable $j$, with the mean $\bar{x}_{ij}$ of the observed values in $X_j$. Therefore, the matrix $X_{mis}$ with missing values is changed into a complete data matrix $X_{(1)}$ [Toutenburg and Heumann, 2002]. The mixed regression model then becomes:

$$
\begin{pmatrix}
  y_{obs} \\
  y_{mis}
\end{pmatrix} =
\begin{pmatrix}
  X_{obs} \\
  X_{(1)}
\end{pmatrix} \beta +
\begin{pmatrix}
  \varepsilon_{obs} \\
  \varepsilon_{(1)}
\end{pmatrix}
$$

ZOR also has some disadvantages; for example, the observed data might not have enough information to represent the missing data as the mean, which could reduce the variance of the imputed variable by shifting the possible extreme values into the mean. This always happens in time series trending and with a growth or decrease in the curve for $X_j$.

First-Order-Regression is a method to estimate missing elements of $X_{mis}$. The regression model is built as follows:

$$
x_{ij} = \theta_{0j} + \sum_{\mu=1, \mu \neq j}^{p} x_{i\mu} \theta_{\mu j} + \mu_{ij}, \ i \not\in \Phi = \bigcup_{j=1}^{p} \Phi_j
$$

where $i$ does not belong to $\Phi$, the union of $\Phi_j$, which is the index of missing values in $X_j$. The missing value $x_{ij}$ is estimated by:
\[ x_{ij} = \theta_{0j} + \sum_{\mu=1, \mu \neq j}^{p} x_{ij \mu} \beta_{\mu j}, \quad (i \in \Phi_j) \]

### 2.2.4 Maximum-Likelihood Estimation

Maximum-likelihood estimation is a method of estimating the parameters. Assume that the errors are based on normal distribution:

\[ \varepsilon_{obs} \sim N(0, \sigma^2 I_m), \quad \varepsilon_{mis} \sim N(0, \sigma^2 I_n) \]

The likelihood can then be factorized with one component for the observed data and one for the missing data [Little and Rubin, 1987]. The fully unobserved matrix \( X_{mis} \) can be estimated from the model:

\[
\begin{pmatrix}
Y_{obs} \\
Y_{mis}
\end{pmatrix} =
\begin{pmatrix}
X_{obs} \\
X_{mis}
\end{pmatrix} \beta +
\begin{pmatrix}
\varepsilon_{obs} \\
\varepsilon_{mis}
\end{pmatrix}, \quad \begin{pmatrix}
\varepsilon_{obs} \\
\varepsilon_{mis}
\end{pmatrix} \sim N(0, \sigma^2 I)
\]

First, we construct the maximum likelihood function, disregarding the missing data. Suppose there are \( n \) independent observations with \( k \) variables \( (x_{i1}, x_{i2}, \ldots, x_{ik}) \) with no missing data. The likelihood function is [Allison, 2012]:

\[
L = \prod_{i=1}^{n} f_i(x_{i1}, x_{i2}, \ldots, x_{ik}; \theta)
\]

where the function \( f_i \) is the joint probability of observation \( i \) and \( \theta \) is the parameter to be estimated.

If there are \( m \) observations with complete data and the remaining \( n - m \) observations have missing values in \( x_1 \) and \( x_2 \), then the likelihood function with entire data set becomes:
\begin{align*}
L &= \prod_{i=1}^m f_i(x_{i1}, x_{i2}, \ldots, x_{ik}; \theta) \prod_{i=m+1}^n f^*_i(x_{i3}, x_{i4}, \ldots, x_{ik}; \theta)
\end{align*}

Obtaining the maximum likelihood estimates of \( \theta \) can maximize this likelihood.

The logarithm of the ML function is:

\[ \ln L(\beta, \sigma^2, X_*) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} (\sigma^2) - \frac{1}{2\sigma^2} (y_c - X_c\beta, y_* - X_*\beta)'(y_c - X_c\beta) \]

After differentiating the parameters \( \beta, \sigma^2, X_* \) and letting \( \frac{\partial \ln L}{\partial \beta} = 0, \frac{\partial \ln L}{\partial \sigma^2} = 0, \frac{\partial \ln L}{\partial X_*} = 0 \), we obtain the solution from the complete sub model:

\[ \hat{\beta} = b_{obs} = S_{obs}^{-1}X_{obs}'y_{obs} \]

\[ \sigma^2 = \frac{1}{m} (y_{obs} - X_{obs}b_{obs})'(y_{obs} - X_{obs}b_{obs}) \]

The Maximum likelihood estimate \( \hat{X}_{mis} \) from the equation is:

\[ y_{mis} = \hat{X}_{mis}b_{mis} \]

If the model is one-dimensional, the solution is unique:

\[ \hat{X}_{mis} = \frac{y_{mis}}{b_{mis}} \]

If the model is multi-dimensional, we use the \( \hat{X}_{mis} \) replace \( X_{mis} \) in the mixed regression model. We then estimate \( \beta \) by:

\[ \hat{\beta}(\hat{X}_{mis}) = (S_{obs} + \hat{X}_{mis}'\hat{X}_{mis})^{-1}(X_{obs}'y_{obs} + \hat{X}_{mis}'X_{mis}) \]

\[ = (S_{obs} + \hat{X}_{mis}'\hat{X}_{mis})^{-1}(S_{obs}\beta + X_{obs}'e_{obs} + \hat{X}_{mis}'\hat{X}_{mis}\beta) \]

\[ + \hat{X}_{mis}'\hat{X}_{mis}S_{obs}^{-1}X_{obs}'e_{obs} \]

\[ = \beta + (S_{obs} + \hat{X}_{mis}'\hat{X}_{mis})^{-1}(S_{obs} + \hat{X}_{mis}'\hat{X}_{mis})S_{obs}^{-1}X_{obs}'e_{obs} \]

\[ = \beta + S_{obs}^{-1}X_{obs}'e_{obs} \]

\[ = b_{obs} \]
Therefore, we use the estimated $\hat{X}_{mis}$ to replace the missing values in $X_{mis}$, and we conduct the regression analysis using the estimator $\hat{\beta}(\hat{X}_{mis}) = b_{obs}$.

2.3 Approaches for handling missing data

2.3.1 Case deletion

Delete the entire entry if data are missing in any variables. This strategy might lead to information loss, especially if a large proportion of data is missing in dataset.

2.3.2 Mean Imputation

Replace the missing value with the mean of the observed data for corresponding variable. However, this method will reduce the variance of imputed variables by shifting the possible extreme values into the mean.

2.3.3 K nearest neighbor (kNN) Imputation

Estimate the missing value by taking the average of the K-nearest neighbors. The kNN strategy has some limitations, such as a high proportion of missing data.

2.3.4 Regression-based Imputation

Estimate a regression model by predicting the missing variable based on the samples with complete information. Then compute the predicted value by regression coefficients, and replace the missing data with predicted scores. This approach has some problems, as it could underestimate standard errors by undervaluing the variance in $X$, which can add some random errors to the response value from the regression model.

2.3.5 Expectation – Maximization (EM) Algorithm
The EM algorithm estimates the covariance matrix, imputes missing data, and preserves the relationship with other variables. This algorithm has two steps: expectation and maximization. Expectation is the separation of the data into two parts: missing and non-missing, and then establishing initial values for parameters. The parameters are used by the EM algorithm to compute the predicted scores for the missing values. Maximization is use to predict the scores, maximize the likelihood function, and obtain new parameter estimations. This process is repeated until convergence is reached. The EM algorithm is an interactive procedure, which uses other variables to impute the missing value (Expectation) and then inspects whether it is the most likely value (Maximization).

2.4 Relief Algorithms

Notation:

\( W[V] \) : Weight for variable \( V \)

\( m \) : Number of random selected instances \( R \)

\( H \) : Hit, the nearest neighbor from the same class

\( M \) : Miss, the nearest neighbor from a different class

\( v \) : The number of variables

\( diff(V, I_1, I_2) \) : The difference between instance \( I_1 \) and \( I_2 \) in variable \( V \)

\( N_{dc} \) : Difference between the response values of \( R_i \) and \( I_j \)

\( N_{dv}[V] \) : Difference between the predictor values of \( R_i \) and \( I_j \) for variable \( V \)

\( N_{dc&dv}[V] \) : Difference between the response values and predictor values of \( R_i \) and \( I_j \) for variable \( V \)

\( k \) : The number of nearest neighbors from \( R_i \)
$d(i,j)$: The distance between instances $R_i$ and $I_j$

2.4.1 Relief algorithm for classification

The original Relief algorithm [Kira and Rendell, 1992] estimated the quality of variables according to how well their values distinguished between instances that were near to each other [Robnik-Sikonja and Kononenko, 2003]. The ReliefF algorithm [Kononenko, 1994] can be applied as a robust method for classification problems with multi-class and missing data.

The ReliefF method first selects a random instance $R_i$, and then chooses $k$ nearest hits $H_j$ with same class of $R_i$ and $k$ nearest misses $M_j(C)$ for all other classes $C$. Then, for variable $V$, we measure all differences between $R_i$ and $H_j$ or $M_j(C)$. If instance $R_i$ and $H_j$ have large differences in variable $V$, this represents a case where variable $V$ separates the two instances within the same class. This is not a desirable quality for variable selection; therefore, we give it a negative weight. On the contrary, if instance $R_i$ and $M_j(C)$ have large differences in variable $V$, this means that variable $V$ separates the two instances into different classes, which is a good quality, so we give it a positive weight. We select $m$ (a user-defined parameter) random instances; this means that each weight $W[V]$ is updated $m$ times. Finally, we rank the variables by weight in descending order.

The difference between two instances $I_i$ and $I_j$ in ReliefF is calculated by:

$$diff(V, I_1, I_2) = \frac{|value(V, I_1) - value(V, I_2)|}{\max(V) - \min(V)}$$
2.4.2 Relief algorithm for regression

In a regression problem, the response variable $Y$ is continuous, so hits $H$ and misses $M$ cannot be applied in ReliefF. Therefore, the differences between response values, disregarding classes and differences between the predictor values, are both taken into consideration.

Relief first selects a random instance $R_i$, and then chooses $k$ nearest instances $I_j$ of $R_i$. For each $I_j$, the closer instance should have a greater influence on the difference between $R_i$ and $I_j$. Therefore, the distance
\[ d(i,j) = \frac{d_1(i,j)}{\sum_{l=1}^{k} d_1(i,l)}, \quad d_1(i,j) = e^{-\left(\frac{\text{rank}(R_i)}{\sigma}\right)^2} \]
is multiplied for each difference to control the influence. $N_{dc}$ measures the difference between response value of $R_i$ and $I_j$, $N_{dv}[V]$ determines the difference between the predictor value of $R_i$ and $I_j$ for variable $V$, $N_{dc\&dv}[V]$ takes into account the multiplication of difference between the response value $\text{diff}(y, R_i, I_j)$ and the difference between the predictor value for variable $V$ $\text{diff}(V, R_i, I_j)$, and $N_{dc\&dv}[V]$ measures the consistence of $\text{diff}(y, R_i, I_j)$ and $\text{diff}(V, R_i, I_j)$. If $N_{dc\&dv}[V]$ is small, the difference between $\text{diff}(y, R_i, I_j)$ and $\text{diff}(V, R_i, I_j)$ is large; otherwise, if $N_{dc\&dv}[V]$ is large, the difference between $\text{diff}(y, R_i, I_j)$ and $\text{diff}(V, R_i, I_j)$ is small. Therefore, $N_{dc\&dv}[V]$ has a positive influence on the weight of variable $V$, while $N_{dc}$ and $N_{dv}[V]$ have a negative impact.
3. New Variable Selection Method for Regression with Missing Data

Our research addresses the variable selection problem with missing data. Currently, two main different approaches are taken to resolve this issue, as shown in Figure 5.

![Figure 5: Approaches to variable selection](image)

In this section, we present a new variable selection Relief method for regression with missing data (RM-Relief) by modifying Relief to a regression (RRelief) algorithm.

3.1 Relief for Regression with Missing Data (RM-Relief)

To deal with missing data problem, the ReliefF improves the calculation of the \( \text{diff} \) function. ReliefF calculates the probability that two given instances have different values for given attributes conditioned over class value [Robnik-Sikonja and Konoenko, 2003]:

For example, if one instance, such as \( I_1 \), is missing, the \( \text{diff} \) function changes to

\[
\text{diff}_M(V, I_1, I_2) = 1 - P(\text{value}(V, I_2)|\text{class}(I_1))
\]

where \( \text{value}(V, I_2) \) means the value of instance \( I_2 \) in variable \( V \), \( \text{class}(I_1) \) represents the class of instance \( I_1 \), and \( P(\text{value}(V, I_2)|\text{class}(I_1)) \) is the probability of
\( \text{value}(V, I_2) \) in the same class of instance \( I_1 \). Conditional probability is approximated with relative frequencies from the training set.

However, in a regression problem, the response variable \( Y \) is continuous. In order to obtain the probability in an interval, we need to use a definite integral. For example, the probability of \( x \) in range \([a, b]\) is 
\[
P_r[a \leq x \leq b] = \int_a^b f_X(x)dx,
\]
where \( f_X(x) \) is the probability density function. Hence, the formula to obtain the probability in discrete classes cannot be used in a continuous situation.

Instead of the concepts of “Hits” and “Misses,” we introduce the “bin” concept for Relief in regression with missing data. In RM-Relief, a “bin (range)” is assigned for each instance, so that the probability can be calculated based on the definite integral of the probability density function. The continuous response variable \( Y \) is divided into \( Q \) bins.

Therefore, the decision regarding the width of each bin is a significant problem for RM-Relief. In this thesis, we decide the bin width by a user-defined number \( Q \). The width of the bins also needs to be appropriate: if it is too small, some bins might be empty and the probability could be also small, whereas if it is too large, the difference between the values of response variables will be difficult to determine. We set a similar value of \( Y \) into one bin. The probability \( P(\text{value}(V, I_2)|\text{class}(I_1)) \) changes to 
\( P(\text{value}(V, I_2)|\text{bin}(I_1)) \).
$Q$ is the number of bins in the variable selection in the regression problem, which is a user-defined parameter. The width of each bin is $idth = \frac{Max(Y) - Min(Y)}{Q}$, where $Y$ is the response variable. Therefore, the range for each instance is $Y \pm \frac{W_i}{2}$.

The $Q$ number of bins method has the benefit that calculating the width is easy, but it has the disadvantage that some bins may contain small or even a zero number of instances, when the number of bins is in excess, the dataset has large variance, or outlier points exist.

### 3.2 Steps of RM-Relief

Step 1: Decide the range of bins and set the bin (class) for each instance

User defines the number of $Q$ bins, then for each bin, the width is $Width = \frac{Max(Y) - Min(Y)}{Q}$. Therefore, the ranges for each bin are:

$Bin_1[min(Y), min(Y) + Width]$

$Bin_2[min(Y) + Width, min(Y) + 2Width]$

$...$

$Bin_Q[min(Y) + (Q - 1)Width, max(Y)]$

Insert a vector of the bin into the data set, and decide the bin for each instance. The illustration of a data set is as follows:
Table 1. Illustration of data set

<table>
<thead>
<tr>
<th>Instance/Variable</th>
<th>$V_1$</th>
<th>$V_2$</th>
<th>...</th>
<th>$Y$</th>
<th>Bin</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_1$</td>
<td>value($V_1$, $I_1$)</td>
<td>value($V_2$, $I_1$)</td>
<td>...</td>
<td>$Y_1$</td>
<td>Bin($I_1$)</td>
</tr>
<tr>
<td>$I_2$</td>
<td>value($V_1$, $I_2$)</td>
<td>value($V_2$, $I_2$)</td>
<td>...</td>
<td>$Y_2$</td>
<td>Bin($I_2$)</td>
</tr>
<tr>
<td>$I_3$</td>
<td>value($V_1$, $I_3$)</td>
<td>value($V_2$, $I_3$)</td>
<td>...</td>
<td>$Y_3$</td>
<td>Bin($I_3$)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Step 2: Convert the regression problem into a multi-class (i.e., Q class) classification problem

The regression problem has been changed into a classification problem with Q classes, where ReliefF can be applied. Therefore, the difference function between instance $I_i$ and $I_j$ for variable V with missing data becomes:

$$diff_M(V, I_i, I_j) = 1 - P(value(V, I_j)|bin(I_i))$$

where $I_i$ is missing here. A kernel probability density function is applied to estimate the probability of $value(V, I_j)$ in $bin(I_i)$.

Step 3: Adopt the idea of variable selection for a multi-class classification problem with missing data

In order to make sure that only one missing value exists in the $diff$ function, we randomly select the instance $R_i$ from the complete data set, and then find the k nearest hits $H_j$ and misses $M_j(C)$, which can be either observed or unknown.

The pseudo code of the RM-Relief algorithm is as follows:
Input: training data with a vector of each variable values \( x \), response variable \( y \), user-defined number of bins \( Q \)

Output: vector weight for all variables

1. Insert a vector Bin for dataset;
2. Calculate width of Bin and decide the bin for each instance;
3. Set all weights \( W[V] := 0 \);
4. For \( i := 1 \) to \( m \)
   a. Randomly select an instance \( R_i \);
   b. If \( R_i \) is unknown:
      i. Return to last step;
   c. Else \( R_i \) is observed:
      i. Find \( k \) nearest hits \( H_j \);
      ii. For each Bin \( B \neq Bin(R_i) \)
          i. Find \( k \) nearest misses \( M_j(B) \) from Bin B;
      iii. End;
   d. For \( V := 1 \) to \( v \) do
      i. \( W[V] := W[V] - \sum_{j=1}^{k} \frac{diff(V, R_i, M_j(B))}{mk} \)
      ii. End;
5. End;

4. Case Study

4.1 Case One

In this chapter, we compare complete cases with variable selection in forward selection using Mallows’s \( C_p \) criterion and RM-Relief. We then create random missing data of 5, 10, 15, and 20% under the MCAR missing mechanism. Finally, a comparison is made between variable selection after imputation (i.e., K-nearest neighbor, Expectation–Maximization (EM) and ILLsimpute) and RM-Relief.

Case one is an example of an application based on an artificial data set from Abt(1967) [Thompson, 1978]. There is one response variable \( Y \) and three predictor variables \( X_1, X_2, \) and \( X_3 \). A linear regression model \( y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i \) is applied for
variable selection in regression with and without missing data. First, we use the complete case to make the variable selection by forward selection and Mallows’s $C_p$ criterion, and by RRelief. In the second step, 20% of the data are erased randomly so that the data becomes MCAR. Three imputation methods: K-nearest neighbor, Iterated Local Least Squares Imputation (ILLsimpute) [Cai and Lin, 2005], and Expectation Maximization (EM) Algorithms [Dempster, Laird, and Rubin, 1977] are first used to impute the data set. The same variable selection methods are then applied. Finally, we use the newly developed algorithm, RM-Relief, to conduct variable selection in the regression with missing data. The MATLAB code for RM-Relief uses user-defined parameters K, Q, and m, which represent number of nearest neighbors, bin, and random selected instances, respectively. The data for case 1 are shown in Table 2.

Table 2. Data for case 1

<table>
<thead>
<tr>
<th>Y</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.69</td>
<td>4</td>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td>18.22</td>
<td>4</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>17.68</td>
<td>6</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>22.43</td>
<td>7</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>26.7</td>
<td>8</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>14.04</td>
<td>8</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>23.48</td>
<td>10</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>19.62</td>
<td>10</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>29.46</td>
<td>12</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>17.08</td>
<td>12</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>28.61</td>
<td>13</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>22.08</td>
<td>13</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>25.45</td>
<td>14</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>33.44</td>
<td>15</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>24.81</td>
<td>16</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

Suppose we select two out of three variables from this data set.

4.1.1 Complete case
4.1.1.1 Forward Selection with Mallows’s $C_p$

Step 1: Obtain the linear regression functions of $(\bar{Y}, X_1), (\bar{Y}, X_2), \text{and} (\bar{Y}, X_3)$. Calculate Mallows’s $C_p$ for each predicted function.

\[
\hat{Y} = 12.348 + 0.997X_1 \quad C_{p1} = \frac{260.08}{0.553} - 15 + 2 \times 2 = 459.31
\]

\[
\hat{Y} = 11.256 + 1.098X_3 \quad C_{p2} = \frac{205.90}{0.553} - 15 + 2 \times 2 = 361.33
\]

\[
\hat{Y} = 35.832 - 1.098X_3 \quad C_{p3} = \frac{74.40}{0.553} - 15 + 2 \times 2 = 123.54
\]

Note: in Mallows’s $C_p$, $C_p = \frac{SSE_p}{S^2} - N + 2P$. $SSE_p$ is the residual sum of squares for the model with $P$ regressors. $S^2$ is residual mean square after regression on the complete set of variables. $N$ is the sample size.

We select the variable $\{X_3\}$ with minimum Mallows’s $C_p$.

Step 2: Obtain the linear regression functions of $(\bar{Y}, X_1, X_3)$ and $(\bar{Y}, X_2, X_3)$. Calculate Mallows’s $C_p$ for each predicted function.

\[
\hat{Y} = 31.061 + 0.302X_1 - 0.796X_3 \quad C_{p13} = \frac{61.78}{0.553} - 15 + 2 \times 3 = 102.72
\]

\[
\hat{Y} = 30.737 + 0.296X_2 - 0.771X_3 \quad C_{p23} = \frac{64.84}{0.553} - 15 + 2 \times 3 = 108.25
\]

Therefore, we select two variables $\{X_1, X_3\}$ with the smallest $C_p = 102.72$. 
4.1.1.2 RRelief

For the RRelief algorithm, we randomly select instance \( R_i \) and keep updating the quality weight for each variable \( V \). The MATLAB statistical toolbox has an embedded function for RRelief\(^1\).

\[
[\text{Ranked}, \text{Weight}] = \text{relieff}(X, Y, K)
\]  

(31)

which computes ranks and weights of each variable \( V \) for input data matrix \( X \) and response vector \( Y \) using the RRelief for regression with \( K \) nearest neighbors. The term “Ranked” lists the importance of each variable, where “Ranked 1” means the most important variable. “Weight” is the attribute ranging in \([-1, 1]\), where the largest positive “weight” is the most significant variable.

In the current data set, the answer given by MATLAB is:

\[
>> \text{Ranked, Weight} = \text{relieff}(X, Y, 5)
\]

\[
>> \text{Ranked: 3 1 2}
\]

\[
>> \text{Weighted: 0.0183 0.0162 0.0201}
\]

Therefore, we select two variables \( \{X_1, X_3\} \), with first and second rank.

4.1.2 Missing data (20% missing at MCAR)

<table>
<thead>
<tr>
<th>( Y )</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.69</td>
<td>4</td>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td>18.22</td>
<td>4</td>
<td>NaN</td>
<td>17</td>
</tr>
</tbody>
</table>

\(^1\) http://www.mathworks.com/help/stats/relieff.html
The term NaN represents missing in MATLAB. The response variable Y is complete and the predictor variables $X_1$, $X_2$, and $X_3$ have three missing data (20% of 15 samples) for each. We then use the sum of least squares, $S = \sum_{i=1}^{M}(x_i - \hat{x}_i)^2$, to test the performance of the imputation method for each variable and the total. A smaller $S$ will provide a closer prediction of the imputation algorithm.

4.1.2.1 Missing data imputed by: K-nearest neighbor, ILLsimpute and EM Algorithm

4.1.2.1.1 K-nearest neighbor

The result of knnimpute (Note: the number in bracket is the original value):

Table 4. Imputation by knn

<table>
<thead>
<tr>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td>16(8)</td>
<td>4(12)</td>
<td>22.43(11)</td>
</tr>
<tr>
<td>10(10)</td>
<td>12(16)</td>
<td>28.61(11)</td>
</tr>
<tr>
<td>7(10)</td>
<td>16(8)</td>
<td>25.45(15)</td>
</tr>
</tbody>
</table>

Table 5. Sum of least squares for knn imputation

<table>
<thead>
<tr>
<th>$S_{knn-X_1}$</th>
<th>$S_{knn-X_2}$</th>
<th>$S_{knn-X_3}$</th>
<th>$S_{knn}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>73</td>
<td>144</td>
<td>549.95</td>
</tr>
</tbody>
</table>

Note: $S_{knn} = \sum_{i=1}^{\# of variables} S_{knn-xi}$
The result of 5-nnimpute:

Table 6. Imputation by 5nn

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td>17(8)</td>
<td>9.86(12)</td>
<td>24.7(11)</td>
<td></td>
</tr>
<tr>
<td>14.2(10)</td>
<td>15.4(16)</td>
<td>18.3(11)</td>
<td></td>
</tr>
<tr>
<td>11.8(10)</td>
<td>16.8(8)</td>
<td>16.6(15)</td>
<td></td>
</tr>
</tbody>
</table>

Table 7. Sum of least squares for 5nn imputation

<table>
<thead>
<tr>
<th></th>
<th>$S_{5nn\times 1}$</th>
<th>$S_{5nn\times 2}$</th>
<th>$S_{5nn\times 3}$</th>
<th>$S_{5nn}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>101.88</td>
<td>82.38</td>
<td>243.54</td>
<td>427.80</td>
</tr>
</tbody>
</table>

The imputed data after knn and 5nn are shown in Tables 4 and 6, and we can compare the imputed data and the original values. $S_{kmn}(S_{5nn})$ is the sum of least squares of variable $X_i$, which represents the sum of squares of errors for each variable after imputation.

4.1.2.1.2 ILLsimpute

The Iterated Local Least Squares Imputation method (ILLsimpute) was proposed by Zhipeng Cai[14], to estimate the missing values in the microarray. ILLsimpute is used to learn the similarity threshold at each iteration by using the observed values. The threshold is used to acquire a set of coherent genes for each target gene, which contain missing data. Therefore, coherent genes in linear regression functions represent the target gene. This algorithm terminates after certain iterations or when the imputation converges.
Table 8. Imputation by ILLsimpute

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7.4(8)</td>
<td>12(12)</td>
<td>14.2(11)</td>
</tr>
<tr>
<td></td>
<td>11.5(10)</td>
<td>15.3(16)</td>
<td>10.6(11)</td>
</tr>
<tr>
<td></td>
<td>11(10)</td>
<td>5.7(8)</td>
<td>22.4(15)</td>
</tr>
</tbody>
</table>

Table 9. Sum of least squares for ILLsimpute imputation

<table>
<thead>
<tr>
<th></th>
<th>S_{ILLsimpute-X1}</th>
<th>S_{ILLsimpute-X2}</th>
<th>S_{ILLsimpute-X3}</th>
<th>S_{ILLsimpute}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>3.61</td>
<td>5.78</td>
<td>65.16</td>
<td>74.55</td>
</tr>
</tbody>
</table>

4.1.2.1.3 EM Algorithm

The EM algorithm is an iterative method for finding the maximum likelihood of parameters, which depends on unobserved latent variables.

Table 10. Imputation by the EM algorithm

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8.4(8)</td>
<td>11.5(12)</td>
<td>14.9(11)</td>
</tr>
<tr>
<td></td>
<td>11.3(10)</td>
<td>13.9(16)</td>
<td>8.9(11)</td>
</tr>
<tr>
<td></td>
<td>11.2(10)</td>
<td>8.5(8)</td>
<td>11.3(15)</td>
</tr>
</tbody>
</table>

Table 11. Sum of least squares for the EM algorithm imputation

<table>
<thead>
<tr>
<th></th>
<th>S_{EM Algorithms-X1}</th>
<th>S_{EM Algorithms-X2}</th>
<th>S_{EM Algorithms-X3}</th>
<th>S_{EM Algorithms}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>3.29</td>
<td>4.91</td>
<td>33.31</td>
<td>41.51</td>
</tr>
</tbody>
</table>

Finally, we compare these four imputation methods in Table 12. We can see that the EM algorithm and ILLsimpute hold the best imputation for the missing data, tested by least squares. In addition, $X_1$ is the best variable and can be estimated more precisely than can the other two variables.
Table 12. Comparison among the four imputation methods

<table>
<thead>
<tr>
<th>Imputation Methods/Least Square</th>
<th>$S_{x1}$</th>
<th>$S_{x2}$</th>
<th>$S_{x3}$</th>
<th>$S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>knn</td>
<td>73</td>
<td>144</td>
<td>549.95</td>
<td>766.95</td>
</tr>
<tr>
<td>5nn</td>
<td>101.88</td>
<td>82.38</td>
<td>243.54</td>
<td>427.80</td>
</tr>
<tr>
<td>ILLsimple</td>
<td>3.61</td>
<td>5.78</td>
<td>65.16</td>
<td>74.55</td>
</tr>
<tr>
<td>EM</td>
<td>3.29</td>
<td>4.91</td>
<td>33.31</td>
<td>41.51</td>
</tr>
</tbody>
</table>

4.1.2.2 Variable Selection

We then apply forward selection with Mallows’s $C_p$ and RRelief to the imputed dataset. We can see that the different imputation methods have a substantial impact on variable selection. The result is shown in Table 13.

Table 13. Selected variables with/without missing data

<table>
<thead>
<tr>
<th>Selection Method/Data Set</th>
<th>Complete Data</th>
<th>20% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>knn</td>
<td>5nn</td>
</tr>
<tr>
<td>FW with $C_p$</td>
<td>${X_3, X_1}$</td>
<td>${X_2, X_1}$</td>
</tr>
<tr>
<td>RRelief</td>
<td>${X_3, X_1}$</td>
<td>${X_2, X_3}$</td>
</tr>
</tbody>
</table>

4.1.2.3 RM-Relief

Finally, we apply Relief for Regression with Missing Data (RM-Relief) to this data set. We set 5 nearest neighbors, 3 bins, and 100 random selected instances to test the algorithm.
\[ Weight = rmrelief(X, Y, 5, 3, 100) \]

\[
\text{Weighted: } 1.0e+04 \times \\
-8.6666 & -4.6666 & -2.4762
\]

Therefore, the selected variables are \( \{X_2, X_3\} \).

4.2 Case Two

The second dataset is housing data, which concerns housing values in suburban Boston. The data set is from the UCI Machine Learning repository\(^2\), and we chose 8 variables and 200 complete instances for data analysis. We then randomly set 5, 10, 15, and 20% missing data to investigate RM-Relief.

4.2.1 Complete Data Set

Suppose we select the four best variables out of a total of eight.

4.2.1.1 Forward Selection with Mallows’s \( C_p \)

By applying forward selection, we calculate Mallows’s \( C_p \) in linear regression step by step, and then choose the variable of minimum \( C_p \). Table 16 shows the value of \( C_p \) for each predictor variable. For example, the first row is the \( C_p \) of the linear function \( (Y, CRIM), (Y, INDUS), \ldots, (Y, LSTAT) \). Because the minimum value is 96.31, we first choose the variable RM. In the second row, the values are the \( C_p \) of the linear function \( (Y, RM, CRIM), (Y, RM, INDUS), \ldots, (Y, RM, LSTAT) \), and then we choose

\(^2\) http://archive.ics.uci.edu/ml/datasets/Housing
PTR. Therefore, after four steps, the selected variables are \{RM, TAX, PTR, LSTAT\}

Table 15. Mallows’s $C_p$ in forward Selection

<table>
<thead>
<tr>
<th>CP</th>
<th>CRIM</th>
<th>INDUS</th>
<th>RM</th>
<th>AGE</th>
<th>DIS</th>
<th>TAX</th>
<th>PTR</th>
<th>LSTAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CP</td>
<td>862.98</td>
<td>791.61</td>
<td>96.31</td>
<td>778.32</td>
<td>883.90</td>
<td>862.17</td>
<td>774.28</td>
<td>332.77</td>
</tr>
<tr>
<td>CP (RM)</td>
<td>98.04</td>
<td>95.79</td>
<td>76.67</td>
<td>95.52</td>
<td>88.97</td>
<td>52.16</td>
<td>63.92</td>
<td></td>
</tr>
<tr>
<td>CP (RM+PTR)</td>
<td>48.60</td>
<td>49.76</td>
<td>34.86</td>
<td>54.16</td>
<td>39.44</td>
<td></td>
<td>24.45</td>
<td></td>
</tr>
<tr>
<td>CP (RM+PTR+LSTAT)</td>
<td>26.36</td>
<td>26.45</td>
<td>24.29</td>
<td>23.93</td>
<td>17.65</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.2.1.2 RM- Relief

Because the result of RM-Relief is based on the user-defined parameters $K$, $Q$, and $m$ (the number of nearest neighbors, bin, and random selected instances, respectively), we can check the differences in rank base for different values of $K$, $Q$, and $m$.

Table 16. Ranks of variables based on different K nearest neighbors

<table>
<thead>
<tr>
<th>K, Q, m</th>
<th>Rank of all variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5,5,5)</td>
<td>INDUS AGE LSTAT RM TAX PTR CRIM DIS</td>
</tr>
<tr>
<td>(10,10,10)</td>
<td>TAX RM PTR AGE LSTAT INDUS DIS CRIM</td>
</tr>
<tr>
<td>(15,15,15)</td>
<td>TAX AGE LSTAT RM INDUS DIS PTR CRIM</td>
</tr>
<tr>
<td>(20,20,20)</td>
<td>TAX AGE PTR INDUS LSTAT DIS RM CRIM</td>
</tr>
<tr>
<td>(25,25,25)</td>
<td>TAX LSTAT RM AGE INDUS PTR DIS CRIM</td>
</tr>
<tr>
<td>(30,30,30)</td>
<td>TAX AGE LSTAT PTR CRIM DIS INDUS RM</td>
</tr>
</tbody>
</table>

We take the number of appearances in each iteration into account, and select the
variables that appear the most frequently, which are \{AGE, TAX, LSTAT, RM\}. Unlike the forward selection with Mallow’s $C_p$, the variable AGE is selected instead of PTR. The difference caused by the different algorithms is due to the fact that RM-Relief tests the weight of each variable independently, but forward selection with Mallow’s $C_p$ checks the closeness of the fit between the test variable and all selected variables. AGE is the variable means proportion of the owner-occupied units built prior to 1940. As an independent variable, AGE is very close to the housing price, but may not work well with other variables. This is the major reason for the difference in the selected variables between forward selection with Mallow’s $C_p$ and Relief.

4.2.2 Data set with 5, 10, 15, and 20% missing data

In order to test the accuracy of RM-Relief, we randomly eliminate 5, 10, 15, and 20% of the complete data set, so the data is MCAR. First, the data are imputed by knn and the EM algorithm. The RM-Relief is then applied for variable selection. Next, we use the RM-Relief to select the variables directly, and compare the results.

4.2.2.1 The data set was imputed by EM algorithms, and then relief is applied

In table 18, the first column shows the percentage of missing data, the imputation method, and the number of nearest neighbors. Each row represents the rank of all thirteen variables, and the top five are the selected variables. Compared with the selected four variables \{AGE, TAX, LSTAT, RM\} for RM-Relief for a complete case, the imputation by EM performs poorly, and successfully chooses the same two variables in the 5, 10, and 20% missing data cases, with one difference for the 15% missing data case. In contrast, the knn imputation shows one difference in the 5, 10, and 15% missing data cases, and two differences in the 20% missing data case.
Table 1. Rank of variables in Relief based on different percentages of missing data and imputation methods

<table>
<thead>
<tr>
<th>Imputation</th>
<th>Rank of all variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>5% EM (25)</td>
<td>AGE, TAX, CRIM, INDUS, LSTAT, RM, PTR, DIS</td>
</tr>
<tr>
<td>5% knn (25)</td>
<td>TAX, AGE, LSTAT, PTR, INDUS, DIS, RM, CRIM</td>
</tr>
<tr>
<td>10% EM (25)</td>
<td>AGE, TAX, PTR, INDUS, LSTAT, RM, DIS, CRIM</td>
</tr>
<tr>
<td>10% knn (25)</td>
<td>TAX, AGE, DIS, LSTAT, PTR, RM, INDUS, CRIM</td>
</tr>
<tr>
<td>15% EM (25)</td>
<td>AGE, TAX, INDUS, RM, DIS, LSTAT, PTR, CRIM</td>
</tr>
<tr>
<td>15% knn (25)</td>
<td>TAX, RM, AGE, INDUS, DIS, PTR, LSTAT, CRIM</td>
</tr>
<tr>
<td>20% EM (25)</td>
<td>AGE, TAX, PTR, CRIM, LSTAT, INDUS, RM, DIS</td>
</tr>
<tr>
<td>20% knn (25)</td>
<td>TAX, DIS, CRIM, AGE, LSTAT, PTR, INDUS, RM</td>
</tr>
</tbody>
</table>

4.2.2.2 RM-Relief

Table 19 shows the variables ranked by RM-Relief. The first column represents the percentage of missing data, nearest neighbor, bins, and iteration. One difference appears in the 5 and 10% missing data cases, and two differences in the 20% missing data case.

Table 18. Rank of variables in RM-Relief

<table>
<thead>
<tr>
<th>K, Q, m</th>
<th>Rank of all variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>5% (25, 25, 25)</td>
<td>TAX, LSTAT, AGE, INDUS, PTR, DIS, RM, CRIM</td>
</tr>
<tr>
<td>10% (25, 25, 25)</td>
<td>TAX, CRIM, AGE, LSTAT, PTR, DIS, RM, INDUS</td>
</tr>
<tr>
<td>15% (25, 25, 25)</td>
<td>TAX, LSTAT, INDUS, AGE, DIS, CRIM, RM, PTR</td>
</tr>
<tr>
<td>20% (25, 25, 25)</td>
<td>TAX, LSTAT, INDUS, AGE, DIS, CRIM, RM, PTR</td>
</tr>
</tbody>
</table>

The results of the case study with missing data, shown in Table 19, reveal the accuracy (compared with relief in regression in complete case) of each method. The RM-Relief has a good accuracy rate in variable selection with missing data. The imputation methods could lead to bias for the entire data set. RM-Relief does not
impute the data, but divides the response variable Y into bins and calculates the difference based on the probability of the missing value being in a particular bin.

Table 19. Accuracy rate

<table>
<thead>
<tr>
<th>Method/percentage of missing data</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>knn - RM-Relief</td>
<td>75%</td>
<td>75%</td>
<td>75%</td>
<td>50%</td>
</tr>
<tr>
<td>EM - RM-Relief</td>
<td>50%</td>
<td>50%</td>
<td>75%</td>
<td>50%</td>
</tr>
<tr>
<td>RM-Relief</td>
<td>75%</td>
<td>75%</td>
<td>75%</td>
<td>50%</td>
</tr>
</tbody>
</table>

5. Conclusion

This thesis proposes a new method for variable selection in regression with missing data. RM-Relief divides the continuous response variable Y into different bins, and then applies the conditional probability to acquire the differences between the missing values and the observed instance. The advantage of RM-Relief is that it extends the RRelief and ReliefF algorithms into regression problems with missing data, whereas the traditional Relief is only available for classification with missing data. The procedure is also based on methods for deciding bin width, conditional probability, and K nearest neighbors. All developments could also be applied to classification problems with missing data, where bins can be treated as classes.

Experimental results on the first artificial data set indicate that RM-Relief is more effective at making variable selections, when compared with forward selection, Mallows’s $C_p$, and Relief after imputation. The same variables are selected by these three methods. Working with a real-world data set – Boston housing prices – the accuracy rate was used to test the performance among RM-Relief, knn- RM-Relief,
and EM- RM-Relief. The results for RM-Relief were good for low missing data percentages (i.e., 5, 10, and 15%), but decreased for 20% missing data.

Reference


[5] Columbia University, Missing-data imputation, Ch 25


[19] Matteo Chemometrics and Francesca Grisoni, Variable selection methods: an introduction, Molecular Descriptors
