MLRMUD: A MULTI LINEAR REGRESSION APPROACH FOR MISSING VALUES PREDICTION WITH UNKNOWN DEPENDENT VARIABLE

By

Ahmed Karama Mahboab Alhebshi

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE
in
Computer Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2019

MLRMUD: A MULTI LINEAR REGRESSION APPROACH FOR MISSING VALUES PREDICTION WITH UNKNOWN DEPENDENT VARIABLE

By

Ahmed Karama Mahboab Alhebshi

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Computer Engineering

Under the Supervision of

Prof. Dr. Samir I. Shaheen
Professor
Computer Engineering Department,
Faculty of Engineering, Cairo University

Prof. Dr. Amir F. Atiya
Professor
Computer Engineering Department,
Faculty of Engineering,Cairo University

Dr. Mona F. Ahmed
Assistant Professor
Computer Engineering Department,
Faculty of Engineering,Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2019

MLRMUD: A MULTI LINEAR REGRESSION APPROACH FOR MISSING VALUES PREDICTION WITH UNKNOWN DEPENDENT VARIABLE

By

Ahmed Karama Mahboab Alhebshi

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Computer Engineering

Approved by the
Examining Committee

_____

Prof. Dr. Samir I. Shaheen Thesis Main Advisor

_____

Prof. Dr. Ihab Elsayed Talkhan Internal Examiner

_____

Prof. Dr. Reda Abd-Alwahab Ahmed External Examiner

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2019

Engineer's Name: Ahmed Karama Mahboab Alhebshi
Date of Birth: 20/5/1985
Nationality: Yemeni
E-mail: engcmp505@gmail.com
Phone: 01091655230
Address: Giza-Main Mahattah street

Registration Date: 1/10/2014
Awarding Date: …./…./……..
Degree: Master of Science
Department: Computer Engineering

Supervisors:

Prof. Dr. Samir I. Shaheen
Prof. Dr. Amir F. Atiya
Dr. Mona F. Ahmed

Examiners:

Prof. Dr. Samir I. Shaheen (Thesis Main Advisor)
Prof. Dr. Ihab Elsayed Talkhan (Internal Examiner)
Prof. Dr. Reda Abd-Alwahab (External Examiner)
Professor in faculty of computers and information
Cairo university

Title of Thesis:

MLRMUD: A Multi Linear Regression Approach for Missing Values Prediction with Unknown Dependent Variable

Key Words:

missing values; splitting algorithm; dependent variable; multi linear regression; regression coefficients

Summary:

The Missing Value problem (MV) is the problem of predicting the missing value in the data set while achieving accurate values. An Additional attribute has been imposed on the missing value problem which is an unknown dependent variable.
In this work, a new approach, MLRMUD, based on Multiple Linear Regression is used to predict Missing values for a data set with an Unknown Dependent variable if complete rows are at least 20%. If they are less than that the Mean method is used to fill some rows until the complete rows reach 20%, after that MLRMUD can be applied normally. This approach is composed of three algorithms; splitting algorithm, dependent variable selection algorithm and multi linear regression algorithm.
MLRMUD is compared to other counterparts in the literature where it was proved that it outperforms them all in the accuracy of missing values computation determined in terms of the Root Mean Square Error (RMSE) and Mean Standard Error (MSE). A method to determine the unknown dependent variable from the training set is proposed. The results show that the proposed method can successfully select the dependent variable with an accuracy of 83% overall the data sets examined

Disclaimer

I hereby declare that this thesis is my own original work and that no part of it has been submitted for a degree qualification at any other university or institute.

I further declare that I have appropriately acknowledged all sources used and have cited them in the references section.

Name: Ahmed Karama Mahboab Alhebshi Date:10/7/2019

Signature:

Dedication

I'd like to dedicate this thesis to my wife and my family for supporting me during my work

Acknowledgments

I would have never gone through this work without the honest help and support of many people in my life to whom I dedicate this section.

First of all, I would like to express my gratitude to **Allah,** who is the reason of my strength and who always raises us over limits beyond our expectations, and who can do everything immeasurably far beyond our thoughts.

I would like to thank my supervisor **Dr. Mona Farouk** for suggesting the research path and for her consistent guidance and advice. He has assisted me in different aspects of my work and has been very keen on my progress in this thesis.

I would like to thank my supervisor **Prof. Samir I. Shaheen and Prof. Amir F. Atiya** for their continuous efforts and concern. They has provided me with important scientific information and aided me throughout my thesis.

I would like to thank my **beloved family** for always finding them by my side, and for their sincere encouragement throughout my work and throughout my entire l

Table of Contents

List of Tables

List of Figures

Nomenclature

Table I.1: List of Acronyms

Acronym	Definition
AKE	Applying K-nearest neighbor Estimation
BGA	Bayesian Genetic Algorithm
Bi	Regression Coefficients
CART	Multiple Classification And Regression Tree
CD	Complete Data
CDKDV	Complete Data with Known Dependent Variable
CDUDV	Complete Data with Unknown Dependent Variable
FPCM-SVRGA	Fuzzy Possibilistic C Means optimized with Support Vector Regression and Genetic Algorithm
MAR	Missing at Random
MCAR	Missing Completely at Random
MD	Missing Data
MLR	Multi Linear Regression
MLRMUD	Multi Linear Regression for Missing Value and Unknown Dependent Variable
MSE	Mean Standard Error
MV	Missing Value
MVPUDV	Missing Value Problem with Unknown Dependent Value
RMSE	Root Mean Square Error
TS	Training Set
TSz	Test Set
UDV	Unknown Dependent Variable
X_i	Independent Variables
Y	Dependent Variable