Efficiency Comparison of Data Mining Techniques for Missing-Value Imputation

Jarumon Nookhong and Nutthapat Kaewrattanapat
Suan Sunandha Rajabhat University, Bangkok, 10300, Thailand
Email: {jarumon.no, nutthapat.ke}@ssru.ac.th

***Abstract*—This research proposes to compare the efficiency data mining techniques for missing-value imputation by Naïve Bayesian, KNN, Linear Regression, Decision Tree and Rule Based Classifier (PART).There is adjusting parameters different set. The data was collected by data set of Mushroom Classification (Discrete data), Glass Type Classification (Continuous Data) and the Balance Scale data (Ordinal Data) from UCI Machine Learning Repository. The data was analyzed and compared the efficiency for each technique by comparing their performance in minimizing the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The result is found that the complete discrete data was well imputed by Decision Tree, but this technique needs enough rules to minimizing an error. For continuous data, it was well imputed by K-Nearest Neighbor. The last Naïve Bayes was good for the discrete data and hidden ordinal scale data.**

***Index Terms*—missing value, imputation, data mining, errorsd.**

I. Introduction

In obtaining and data storage, they are significantly important to an analyze or a quantitative research to obtain accurate research results to be utilized; however, data storage may have bug or some data is missing. [1], [2] For solving such problems, it can be solved by disposing data record that has problem; however, a lacking of data for utilizing in data analysis may occur. Therefore, in order to be able to bring the data to be utilized, it is necessary to update data utilizing the main technique, that is estimation of approximated similar data or imputation. [3]

As for the estimation to bring the data for imputation in missing data has several techniques. This research choose 5 data mining techniques, namely 1) Naïve Bayes 2) K-Nearest Neighbor: KNN) 3) Linear Regression 4) Decision Tree and 5) Rule Based Classifier to compare efficiency in imputation of missing data. In this research, an adjustment of parameter in each technique of data mining to obtain different models was conducted and utilizing of Root Mean Square Error: RMSE and Mean Absolute Error: MAE to show the comparison of efficacy of missing data imputation.

Manuscript received December 9, 2014; revised February 10, 2015.

II. Objectives

To study suitability in utilizing of data mining in imputing miss data having different format and bringing input parameter of mining data analysis, such as Naïve Bayes, KNN, Linear Regression, Decision Tree and Rule Based Classifier.

III. Delimitation of Research

Data utilized in the research was data set brought from UC Irvine Machine Learning Repository: UCI. The 3 data sets were used, namely mushroom data set (poisonous and non-poisonous), glass identification data set and balance scale data set, all of 3 data sets have different data specification and record number.

This research determined value of missing attribute by choosing with randomizing technique in each data set for only 1 attribute used in imputing missing value and testing errors.

Type of missing data would be in Missing Completely Random: MCAR.

IV. Related Literature

A. Naïve Bayes

Naïve Bayes' Technique forecasts with principle of classification by applying Bayes Theorem which is a supervised learning; the exercise must have answer keys meaning that type or class of data that wants to create learning to build concepts of such class [4] is suitable with the cases of large number of example sets and attribute of independent example sets [5] which probability of data sets to be Ci class for data having n attribute X = {A1, A2,...An} or P(Ci |A1,...,An} from Bayes' Rule; the results will be as the equation (1)

$$P(C_i \mid A_1, ..., A_n) = \frac{\prod_{j=1}^{n} P(A_1 \mid C_i) P(C_i)}{P(A_1, ..., A_n)} \quad (1)$$

By finding P(Ci| A1,...,An) from equation for ever group I, attribute at j; the obtained value will be brought to compare. The group having highest probability is an answer.

doi: 10.12720/jiii.3.4.305-309

B. K-Nearest Neighbor: KNN

K-Nearest Neighbor: KNN or nearest range cluster is unsophisiticated technique, applied to good irregular data in scattering formation. A principle of KNN is to extend boundary for finding members by measuring distance, applied in several fields of work, namely Impute Missing Values [6] and etc. as KNN will examine certain factors, such as a number of nearby members, distance measurement and Normalization.

The K parameter is a number of interested neighbors which, normally, would be self-determined. distance measurement could be conducted by several techniques, namely Euclidian Distance, Manhattan Distance, Mahalanobis Distance and etc.; these techniques are correlation neighbors that should apply normalization data. As for Euclidian Distance, it is not necessary to conduct normalization data.

C. Linear Regression

Linear Regression is a study regarding how independent variables affects dependent variables or independent variables affects Y value varied in what model. The relationship could be explained by Regression Model [7] as equation (2)

$$Y = a + bX \tag{2}$$

As Y = Dependent variables
X = Independent variables
a = Constant
b = Slope

D. Decision Tree

Decision Tree is a data mining technique which is a learning model by categorizing data in the sample group into subgroups, using attributes of data as categorizing tools. The decision tree obtained from learning demonstrates attributes of data that is a determiner of answers and demonstrates how important or different of the attributes, helping users to analyze data and decide more accurately. A process starts with selecting attributes to be mode, which are the attributes when dividing samples into subgroups making most members in each subgroups have answers in the most same answers. Measurement of an ability in categorizing of gain in each mode is able to conduct by relying on Information Theory that bring entropy to be an indicator of disorder in data.

E. Rule Based Classifier

Rule Based Classifier is a technique for categorizing data record by using "if...then..." which the rule brought to create model will be present in connected model, such as R = (r1 V r2 ... V rk) as R is groups of rule and ri is a rule using in dividing or each rule [4]

F. Root Mean Square Error: RMSE

Root Mean Square Error or Square Root of Mean Standard Deviation is a technique measuring error from value forecast from model with occurred actual value; if RMSE has low value, it means that the model can forecast value nearly actual value; therefore, if this value equals zero, it means that there is no error in this model. RMSE could be calculated as equation (3).

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}(Y_i - \hat{Y}_i)^2}{n-1}} \tag{3}$$

As:

Y_i = Approximation from data value model from forecasting

$\hat{Y}_i$ = Actual value of actual data obtained from calculation

n = Number of sample size using in model estimation

G. Mean Absolute Error: MAE

Mean Absolute Error: MAE is a technique to measure difference value between actual value and forecast value from model. If MAE has low value, it means that the model can forecast nearly actual value; therefore, if this value equals zero, it means that there is no error in this model. MAE could be calculated as equation (4).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left| f_i - y_i \right| \tag{4}$$

f_i = Approximation from data value model from forecasting

y_i = Actual value of actual data obtained from calculation

n = Number of sample size using in model estimation

V. Related Research

Narong et al [8] have conducted a study and comparison regarding a technique of missing-value forecasting with statistical method, namely mean, correlation coefficient analysis, weighted correlation coefficient analysis and discriminant analysis by measuring efficacy by MMRE. It is found that the discriminant analysis can forecast missing value to be the most nearly actual data suitable for data having relationship with each other and clear scattering.

Kairung and Payunk [6] have conducted a study regarding replacing of missing value by weighted nearest technique of micro array by KNNFSW Impute and compared efficacy with Row Average KNN. The test demonstrated the better efficacy in term of Normalized Root Mean Squared Error (NRMSE).

VI. Research Methodology

A. Data Preprocessing

Comparison between efficacy of data mining technique to impute missing data in 5 techniques would use such data from UC L (UC Irvine Machine Learning Repository, http//:archive.ics.uci.edu/ml/index.html), total 3 sets of case study, namely mushroom data set consisted of 23 attributes, 8124 records consisted of Nominal Scale, Glass

Identification Data Set consisted of 11 attributes and 214 records comprise continuous data and nominal scale and Balance Scale Data Set consisted of 5 attributes and 625 records consisted of Ordinal Data and Class Label, all of data sets are different in property of data and number of records.

Before bringing data to analyze, it is necessary to convert data format to be suitable for analysis in each technique (Data Preprocessing), namely Linear Regression Analysis has to use numerical data only and etc., which the conversion of data format and attribute selection of missing data will be present in Table I

TABLE I: DATA PREPROCESSING

Data Sets	Missing Attribute	% Training Set	% Test Set	Real Data	Numeric Modeling	Nominal Modeling
Mushroom	Odor	80%	20%	a,l ,c,y,f, m,n, p,s	1,2,3,4,5, 6,7, 8,9	Real Data
Glass Identification	Class Label	80%	20%	1,2,3, 5,6,7	Real Data	ONE=1 TWO=2 THREE =3 ONE=1 TWO=2 THREE =3 FIVE
Balance Scale	Right-Weight	80%	20%	1,2,3, 4,5	Real Data	ONE=1 TWO=2 THREE =3 FOUR= 4 FIVE=5

TABLE II: PARAMETER TUNING OF NAÏVE BAYESIAN

Debug	DisplayModelInOldFormat	UseKernel Estimator	UseSupervised Discretization
True , False	True , False	True , False	True , False

TABLE III: PARAMETER TUNING OF K-NEAREST NEIGHBOR: KNN

KNN	DistanceFunction	DistanceWeigthing
K=1, K=3	EuclideanDistance, ManhattanDistance	No, 1/Distance, 1-Distance

TABLE IV: PARAMETER TUNING OF LINEAR REGRESSION

AttributeSelectionMethod	Debug	EliminateColinearAttribute
M5Method, GreedyMethod	True, False	True, False

TABLE V: TUNING OF DECISION TREE

Confidence Factor	Binary Split	ReducedError Pruning	UseLaplace
0.25, 0.50, 0.90	True, False	True, False	True, False

B. Parameter Tuning Design

In this research, the researchers tuned parameter value by WEKA for tuning because the parameter tuning would affect a change in model used in different imputation. The scope of parameter tuning would present in Table II to Table VI.

TABLE VI: PARAMETER TUNING OF RULE BASED CLASSIFIER

Confidence Factor	Binary Split	ReducedError Pruning	Unprune
0.25, 0.50, 0.90	True, False	True, False	True, False

C. Process of Data Mining Analysis

When the data for preprocessing is prepared and parameter data is finished preparing, the analysis is conducted in each technique of data mining according to the process as Fig. 1:

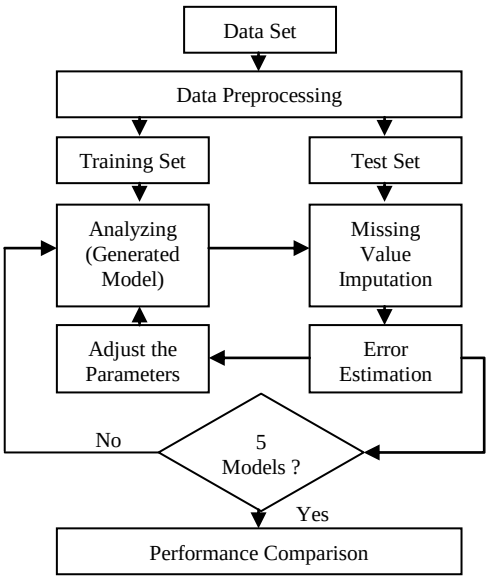

Figure. 1. Missing value of data in 5 techniques.

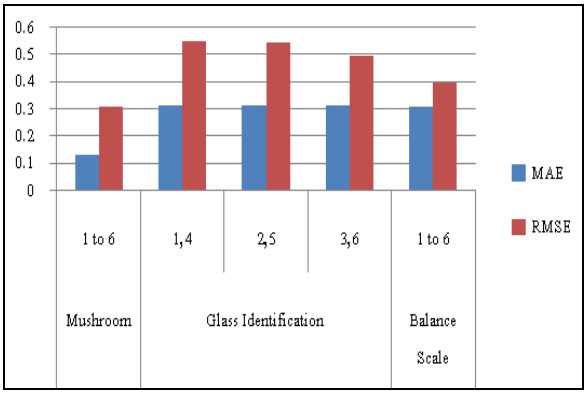

Figure. 2. Error value from imputation in missing value by naïve bayesian technique.

VII. RESULTS

A. The Results in Comparison of Errors

When the data which is ready to analyze is brought to use in each technique in order to obtain forecasting model utilized in missing value. Then, the comparison for errors

by MAE and RMSE is conducted. The models and error value is presented in Fig. 2 and Table VI.

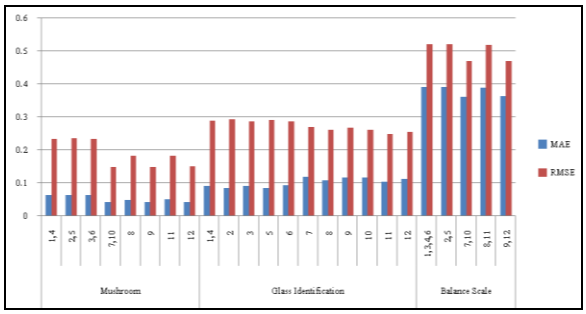

Figure. 3. Error value from imputation in missing value by k-nearest neighbor technique.

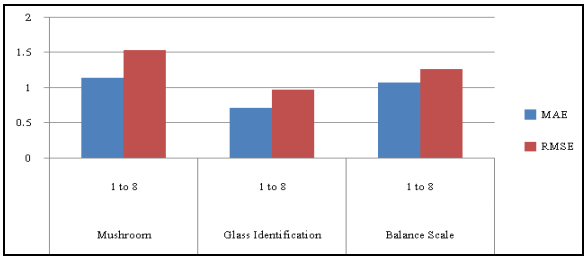

Figure. 4. Error value from imputation in missing value by linear regression technique.

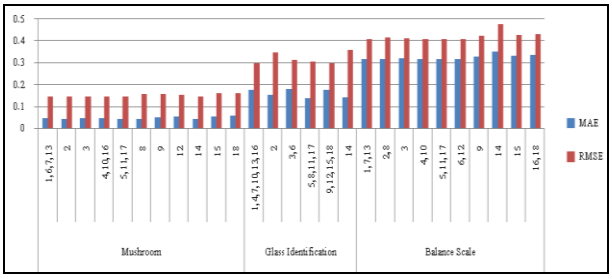

Figure. 5. Error value from imputation in missing value by decision tree.

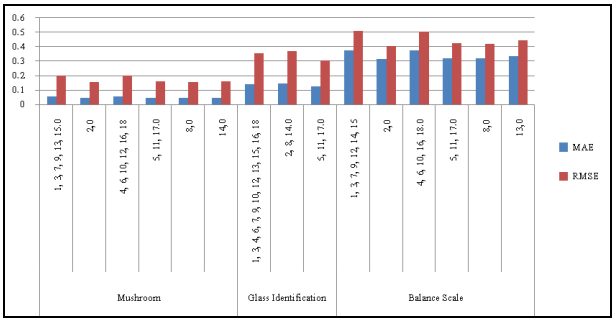

Figure. 6. Error value from imputation in missing value by rule based classifier.

B. *Summarize of Error Comparison*

In the comparison of data mining efficacy in order to impute missing value, error between forecasting data of imputation and actual data was used. As the results in comparison of errors are presented in Table VII

When the data in Table VII was brought to convert into comparison chart, it demonstrates a clear difference. The comparison chart is as Fig. 7:

TABLE VII: THE LOWEST ERRORS IN EACH MODELING

Data Set	Model	MAE	RMSE
Mushroom	Decision Tree	0.0405	0.1424
	KNN	0.0419	0.148
	Rule Based	0.0429	0.1518
	Na ïve Bayes	0.1306	0.3079
	Linear Regression	1.1299	1.5318
Glass Identification	KNN	0.0855	0.2918
	Rule Based	0.1236	0.3045
	Decision Tree	0.137	0.3041
	Na ïve Bayes	0.3109	0.5459
	Linear Regression	0.7094	0.9608
Balance Scale	Na ïve Bayes	0.3091	0.3971
	Decision Tree	0.3139	0.4051
	Rule Based	0.314	0.4032
	KNN	0.3612	0.4675
	Linear Regression	1.0674	1.2629

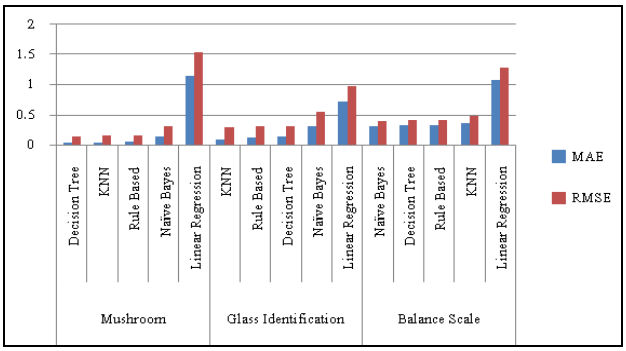

Figure. 7. Comparison for errors by mae and rmse in each mining data in order to impute missing value.

VIII. CONCLUSION

In the comparison of data mining efficacy for imputing missing value, the researchers chose 5 techniques available in WEKA software, namely Na ïve Bayes, KNN, Linear Regression,

Decision Tree, Rule Base Classifier (PART) to be applied in clearly different data, for example all discrete data based classifier of mushroom, continuous data classifier of glass and discrete and hidden-order data of balance scale. This research utilized errors for comparison, namely MAE and RMSE as they helped to know on difference between imputed data and actual data. From the research, it is found that when the mushroom data set was imputed with missing value by Decision Tree Technique, it is suitable with Complete Discrete Data and clustered in each group; however, this technique will give low errors results when there are a large number of rules in imputation of missing value. When glass identification data set is imputed with missing value and conducted errors resulting in KNN to be suitable for numerical data (Continuous Data) and necessary to tune neighborhood value to be low as the data clusters really near. When balance scale data set is is imputed with missing value and conducted errors resulting

in Na ïve Bayesian to be suitable in imputing missing value as the property of balance scale data comprised discrete data and ordinal scale. Moreover, it is found that all 3 data sets are not suitable for linear regression technique as forecasting model giving value as continuous number which are not batch data, making high rate in errors.

The matter that should be developed further is the case of each record having missing value more than 1 value, what suitable technique should be used to solve the problem and which attributes should be imputed before, which will give the lowest errors.

ACKNOWLEDGMENT

I would like to thank Institute for Research and Development Suan Sunandha Rajabhat University's fund in supporting this research study and sponsorship.

REFERENCES

[1] K. Kularbphettong and C. Tongsiri, "Mining educational data to analyze the student motivation behavior," *World Academy of Science, Engineering and Technology*, International Science Index 68, vol. 6, no. 8, pp. 1036-1040, 2012.

[2] N. Kaewrattanapat and W. Kunnu, "The automatic classification of thai news by similarity method," in *Proc. 2th International Symposium on Business and Social Sciences: ISBSS*, Osaka Japan, 2013.

[3] N. Kaewrattanapat, "Recruitment system based on SOAP and XML web services," in *Proc. Human Resource National Conference*, Chulalongkorn University, 2010.

[4] T. Bunyang, A. Roapichian, and N. Pongam, "Comparison of data classification in k nearest neighbor, na ïve bayesian, decision tree and rule based classifier," in *Proc. 7th National Conference on Computing and Information Technology*, pp. 19-23, 2011.

[5] K. Srion, "Forcastin of cause in electric current failure by utilizing data mining technique in electric distribution of regional electricity authority," Area 1, Central Region. Master of Engineering (Electrical Engineering), Faculty of Engineering, Kasetsart University, 2009.

[6] K. Heng-praphrom and P. Mesut, Selection of Property by Nearest Member Technique for Imputing Missing Value by Weighted Nearest Technique of Micro Array Data in NCCIT, pp. 504-509, 2008.

[7] S. Jusu and T. Jansutivechakul, "Comparison of book borrowing for nabon school's library by regression analysis and artificial neural network," in *Proc. 6th NCCIT*, 2010, pp. 110-115.

[8] N. Poti, S. Prakancharoen, D. Thammasiri, P. Ammaruekarat, and V. Nuipian, "A study and comparison of missing value estimation by statistical method" in *Proc. 5th National Conference on Computing and Information Technology*, 2009, pp. 1160-1165.

Jarumon Nookhong is Lecturer of Information Management, Faculty of Humanities and Social Sciences at Suan Sunandha Rajabhat University, Bangkok (SSRU), Thailand

She got her Bachelor degree of Information Science (Information Management) Walailak University. Also, She graduated her Master of Sciences (Management Information System) King Mongkut's University of Technology North Bangkok and Ph.d. candidate in Information and Communication Technology for Education from King Mongkut's University of Technology North Bangkok, Thailand. She's research focuses on Data Mining, Knowledge Management, Artificial Neural Networks, Green ICT, Cloud Computing and Green Cloud System.

Nutthapat Kaewrattanapat is Lecturer of Information Management, Faculty of Humanities and Social Sciences at Suan Sunandha Rajabhat University, Bangkok (SSRU), Thailand

Nutthapat Kaewrattanapat was born on 8 March 1983 in Thailand and received his B.S. degree with 1st class honors in Computer Science and M.S. degree in Management Information System and Ph.d. candidate in Information Technology (Royal Thai government scholarship) from King Mongkut's University of Technology North Bangkok, Thailand. He is currently a faculty member of Information Management Program, Suan Sunandha Rajabhat University since 2009. His research focuses on Information Technology, Computational Linguistic, Natural Language Processing and Data Mining.

