Analysis of Missing Value Estimation Algoithms for Data Farming

Mohd. Shahnawaz and Kanak Saxena

Department of Computer Applications

Samrat Ashok Technological Institute, Vidisha M.P., India

shahnawaznbd@gmail.com, kanak.saxena@gmail.com

Abstract: *In this paper we compare various statistical method of estimation of missing data values. Missing data estimation is a part of data farming. Data Farming is a process to grow the data & provides a more comprehensive understanding of the possible outcomes, and offers the opportunity to discover outliers, surprises. Many times data mining task use existing data collected for various other purposes, such as daily transactional data, monitoring & control data. Sometimes, the data set might be missing some values, to estimate these missing values various statistical methods exist in the literature. In this paper a comparison among these methods is given with implementation & comparative results on the real life data set. This research work will be helpful to understand the effect of missing values on the mining process.*

Keywords: *Data Farming, Missing Data, Error Factor, Least Square Method.*

I. Introduction

Adequate Data is required for Decisions making on the basis of knowledge extracted by the data mining process, data collection is a crucial process many times data is not adequate for the mining. In that case data reduction, selection and data farming techniques are applied to get adequate data [1]. After getting the adequate data one can apply the mining algorithms to extract more accurate & useful information compare to the former data.

Methodologies and tools are needed for determining the most appropriate data at an acceptable cost. The estimation of missing data values is helpful to achieve better results. Missing value estimation algorithms are based on statistical, numerical analysis, Regression & curve fitting.

II. Missing Data Analysis

Any times data set contain missing values of some attribute, due to unavailability, clerical mistakes or any other reason these values was not filled at the time of data entry. While data mining is going to process, these missing values mutate the results proportion to the missing data. To avoid this problem, various algorithms are exist for estimation of missing values and put these estimated values in the data set to fill the vacuums. In this paper various algorithms are analyzed on the basis of a newly developed performance parameters.

III Performance Measure

To analyze various missing values estimation algorithms a common performance measure proposed in this work is error factor. Let (x_1, x_2, x_3, ..., x_n) are the actual missing values and Let (x'_1, x'_2, x'_3, …, x'_n) are the estimated values then error factor is defined as

$$\text{Error Factor} = \sqrt{\frac{\sum_{i=1}^{n}(xi - x'i)^2}{n}}$$

As low as the value of the error factor for a missing value estimation algorithm as much better is the corresponding algorithm. Another performance measure is the time consume in the estimation process.

IV. Definition & System Model

In the proposed model we analyze the various missing values estimation algorithms like fill mean, fill median, fill mode, estimation by least square regression etc. As statistics allows us to estimate the values not calculate exactly. So to analyze the accuracy of various algorithms we take a standard data set any real life data set can be used for analyzing, and estimate missing values of that data set by various algorithms and then calculate the error factor, and calculate the time consume for each algorithm. On the basis of

these two performance measure, we analyze the various algorithms. An implementation is done for this purpose in java with the database in MySql. To properly understand the proposed work we should defined the following statistical terms.

A. Mean

It is the arithmetical average of the values of a vector. Mathematically Let (x_1, x_2, x_3, x_4 ... x_n) are values in a vector, arithmetical mean is defined as

$$Mean = \frac{\sum_{i=1}^{n} x_i}{n}$$

B. Median

This the term which is in the middle while the vector has values in the sorted order. Mathematically Let (x1, x_2, x_3, x_4 ... x_n) are values in increasing order in a vector, Median is defined as,

$Median = Value\ of\ the\ term \left\lfloor \frac{n}{2} \right\rfloor + 1$ If n is odd then

$Median = value\ of\ the\ term \left[\frac{n}{2}\right] + \left[\frac{n}{2} + 1\right] / 2$ If n is even then

Mode: It is the values of the term which is appear most of the times in a vector, Hence it is the most frequent term.

C. Variance

The variance is a term used to measure of how far a set of numbers are spread out from each other. how far the numbers lie from the mean (expected value). The variance is a parameter describing in part either the actual probability distribution of an observed population of numbers, in the simplest cases this estimate can be the sample variance Let (x_1 , x_2 , x_3 , x_4 … x_n) is a vector of values , M is the mean then variance can be defined as.

$$Variance = \frac{(M - x_1)^2 + (M - x_2)^2 + (M - x_3)^2 \ldots + (M - x_n)^2}{n}$$

D. Standard Deiation

The Standard Deviation or root mean square deviation is also a measure of how spreads out numbers are. It denoted by symbol is **σ** (the greek letter sigma), It can be calculate as

$$Standard\ Deviation\ (\boldsymbol{\sigma}) = \sqrt{Variance}$$

E. Least Square Method

The least-squares method was first described by Carl Friedrich Gauss around 1794 [10]. Least squares correspond to the maximum likelihood criterion if the experimental errors have a normal distribution and can also be derived as a method of moment's estimator.

The method of least squares is a standard approach to the approximate solution of over determined systems, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the errors made in solving every single equation. The most important application is in data fitting. The best fit in the least-squares sense minimizes the sum of squared residuals, a residual being the difference between an observed value and the fitted value provided by a model. In case of simple linear regression analysis a single variable is used to predict another variable on the assumption of linear relationship between the given variables [9]. Regression models involve the following variables:

- The unknown parameters denoted as B.
- The independent variables, X_i.
- The dependent variable, Y_i.
- The unknown parameter denoted A.
- The unpredictable random element called as residual error e_i

In various fields of application, different terminologies are used in place of dependent and independent variables.

A regression model relates *Y* to a function of X and B.

$$Y = f(X, B)$$
$$Y_i = A + BX_i + e_i$$

A: it is the Y intercept, B: it is a constant denoting slope of the regression line

If the two constant A & B are known the accuracy of our prediction Y is depends on the magnitude of the values of e_i.

V Proposed Model

A. Function: Proposed model is a integration of 5 subtask described as below

Data Load: in this phase data stored in the database is fetched through JDBC/ODBC connectivity and stored in a multidimensional array. All further processing will be done on these data stored in these arrays.

Data Missing: in this phase a fixed (user input) percentage of the data will be missed artificially from random location i.e. a random row and random column location. Information of the missed data will be stored separately for further calculation & references. Concept of sparse matrix is used for storing the information of the missing data; due to sparse matrix this required low memory space for storing the missing data.

Estimation: in this phase estimation of the missing values is done by various algorithms like

- ✓ **Fill Mean:** in this method, missing values are filled by the arithmetic mean of the respective columns.
- ✓ **Fill Median:** in this method, missing values are filled by median of the respective columns.
- ✓ **Fill Mode:** in this method, missing values are filled by mode of the respective columns.
- ✓ **Least Square Regression**: In this method the missing values are estimated by the least square method of regression.

Statistics Calculation : in this method calculation of the other statistical terms like Mean, variance, standard deviation are calculated, these terms are useful in the analysis of the result with different aspects.

Performance Measurement: Performance measure for each estimation algorithm is calculated in this phase, in terms of the error factor as describe earlier. Time consumption is also measure in this phase.

VI. Result Analysis

Table 1 shows the results in the form of various statistics term like mean, variance & standard deviation, this table contains the value of these parameter for each column of the dataset. In this work we take 4 column and 60 rows. Graphical representation of the result is also described in the figure 3.

Parameter	Col 1	Col 2	Col 3	Col 4
Mean	16	26	40	560
Variance	3.22	31.83	96.74	766.4
Standard Deviation	1.79	5.64	9.83	27.68

Table 1: Statistical term for each column of dataset

Percentage of Missing Data	ERROR FACTOR			
	Fill Mean	Fill Median	Fill Mode	Least Square Method
1 % Miss	4	2	2	3.8
2 % Miss	3.97	2.34	5.24	2.29
4 % Miss	18.72	8	12.67	17.45
5 % Miss	15.29	11.86	17.48	26.29
8 % Miss	35.5	14.23	51.86	34.77
10 % Miss	48.94	13.96	67.75	71.15

Table 2: Error Factor on various data missing rate

Table 2 shows the results in the form of error factor for various missing value estimation algorithms like fill mean, fill median, fill mode & least square regression method. Each algorithm analyzed on different missing data rate i.e. 1%, 2%, 4%, 5%, 8%, 10%. Implementation of the proposed model is flexible to analyze missing value estimation algorithms on any data missing rate. It is quite clear that as the rate of missing data increase error factor is also increase proportionally. Graphical representation of the result is also described in the figure 1 & figure 2.

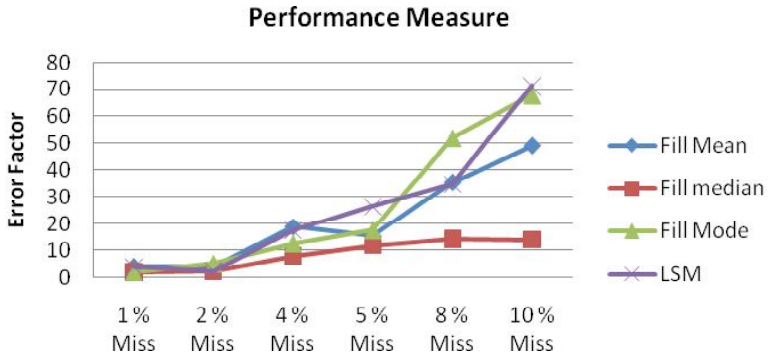

Figure 2: 2 D Representation of Error factor Measured for various data missing rate.

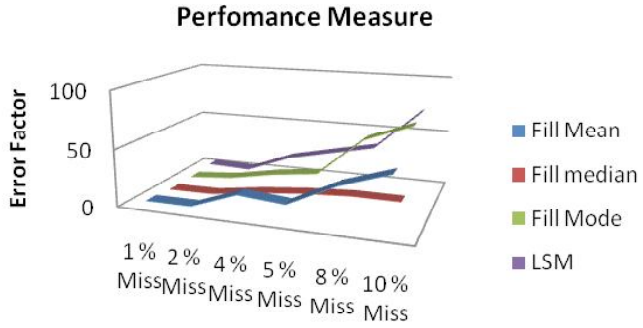

Figure 3: 3 D Representation of Error factor Measured for various data missing rate.

Figure 5: Statistical term for each column of data set

Figure 4: Running Snapshot of the Proposed Work

VII. Conclusion

Data is significant to mine the useful information from the dataset, accuracy of the mining results is highly depends on the data set available; to prepare the data for mining various preprocessing steps are executed before the actual mining. Work described in this paper is a step to analyze various data missing algorithm on the basis of newly developed performance measuring parameter known as error factor. As growing more data by using of available data is called as data farming, filling of missing values in the dataset is also a part of data farming. We can observe from the results that as low as the error factor as much better is the estimation of the missing values. As much as the percentage of missing data is increased the value of the error factor is also increased. in this work we obtain the error factor in six different (1%, 2%, 4%, 5%, 8%, 10%) percentage of missing data.

One another fact can be concluded from the results that if the data is distributed in a short range then the fill median method is the best for estimation of missing values. If data is distributed in a wide range then least square method of regression give the best result for estimating a missing value. Fill mean method provide an average results in each case. Time is also a measuring factor to analyze the various algorithms, here we calculate the time which shows that execution time is depend on the number of the data values (i.e. MxN, M is number of column and N is number of rows in the dataset) and percentage of missing data (directly effects the number of the

estimations). Fill Mean, Fill Median, Fill Mode algorithm are not depends on percentage of data missing while Least square regression depends on both.

References

[1] SRIVATSAN LAXMAN and P S SASTRY, A survey of temporal data mining, *S¯adhan¯a* Vol. 31, Part 2, April 2006, pp. 173–198. © Printed in India

[2] Gary E. Horne, Klaus-Peter Schwierz, DATA FARMING AROUND THE WORLD OVERVIEW, *Proceedings of the 2008 Winter Simulation Conference, S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.*

[3] Andrew Kusiak, Data Farming Methods for Temporal Data Mining, Intelligent Systems Laboratory, 2139 Seamans Center, The University of Iowa, Iowa City, Iowa 52242 - 1527

[4] C.S. Choo,E.C. Ng,Dave Ang, C.L. Chua, DATA FARMING IN SINGAPORE: A BRIEF HISTORY , *Proceedings of the 2008 Winter Simulation Conference S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.*

[5] Simulation Experiments and Efficient Design (SEED) centre for data farming (2009) http:// harvest.nps.edu.

[6] Gary Horn, Stephen Seichter, Karsten Haymann, Data Farming in Support of Military Decision Makers 2010. http:// harvest.nps.edu.

[7] S.S. Sastry , Introductoary Methods of Numerical Analysis, Prentice hall of India. ISBN 978-81-203-2761-0.

[8] Kishor S. Trivedi, Probability & Statistics with Reliability, Queing and Computer Science applications, Prentice hall of India. ISBN 978-81-203-0508-3.

[9] Han J, Kamber M 2001 Data Mining: Concepts and Techniques (San Fransisco, CA: Morgan Kauffman)

[10] Bretscher, Otto (1995). *Linear Algebra With Applications, 3rd ed.*. Upper Saddle River NJ: Prentice Hall.