An Optimum Model for the Retrieval of Missing Values for Data Cleansing using Regression Analysis

Deepshikha Aggarwal
Associate Professor
Jagan Institute of Management Studies, Delhi

V. B. Aggarwal, Ph.D
PhD (USA), Dean (IT)
Jagan Institute of Management Studies, Delhi

ABSTRACT
An important aspect of the data mining is the pre-processing of the data. Pre-processing of the data is important because real world data is susceptible to inconsistencies, noise and missing values. Such a data cannot be used in data mining as that would produce highly inadequate results .There are basically two methods through which we can remove the problem of the missing values the first one is to ignore the data set with the missing value the second one is to predict those values. Prediction can be made based on assuming the continuity of the data or giving them some suitable value based on previous knowledge .In this paper our focus is on providing an adequate method to fill those missing values by predicting a suitable value by comparing and choosing a suitable regression method based on both the statistical and the subjective analysis of the graph from the various known regression method.

General Terms
Data Cleaning, Data warehousing, Regression techniques for data cleansing.

Keywords
Data Quality, Missing Values, Data Cleaning, Regression, Linear, Quadratic, Exponential, Gaussian, Prediction, RMSE.

1. INTRODUCTION
The Data in various applications may be dirty due to noise or due to missing values as we are extracting the information from a number of locations or may be from heterogeneous formats [1]. The data may be missing due to equipment malfunctioning, due to non-understanding of data, certain data values not captured at time of entry etc. Real data may be error onus due to huge size. The error in the data may lead to wrong information which may results into wrong decision [2]. So we have to clean the noisy data. The missing data may be correct by various ways, we can have mean value of all the existing values to fill missing value, we can ignore that tuple, we can use any global value to replace missing value or we can use the regression method to predict the missing values.

There are various reasons for predicting missing values and not discarding. Firstly however the data may be continuous or random discarding the data may lead to shrinkage of data [3]. Secondly there might be several important information that might be in conjugation with missing value that might get lost[4] .There are several ways to predict the values the main problem lies in deciding about choosing the method there are statistics that can help us quantify in choosing the values but again those statistics alone cannot be taken as absolute but when we combine visualization i.e. the subjective view of our data and the statistics it help in better visualizing the method from which we need to choose .

2. LITERATURE REVIEW
A wide variety of work has been done in the field of data cleaning and different researchers have presented different methods and algorithms. We have studied the literature to gather information about the existing methods of data cleansing.

In a research paper titled "Robust and Efficient Fuzzy Match for Online Data Cleaning" [5], the researchers have developed a robust and efficient fuzzy match algorithm, applicable across a wide variety of domains. They have defined the fuzzy match similarity (fms) function for comparing tuples. , the similarity between an input tuple and a reference tuple is the cost of transforming the former into the latter—the less the cost, the higher the similarity

The research [6] explains a method where a user selects two or three most important fields and ranks them based on their power to uniquely identify records. The elements in the selected fields are tokenized, resulting in a table of tokens. The two most uniquely identifying fields of the table are used as two different main sort keys on the table of "tokens" to produce two sorted token tables Token records in close neighborhoods are compared for a match, and warehouse ID. (WID) is generated for each record.

Authors in [7] explain two levels of domain independence, Domain independence at attribute level and at record level. The positional algorithm is an enhancement to the recursive algorithm which tokenizes the two strings (or fields) being compared into words. If one of the strings has only one word-token and the other string has more than one. Word-token then, breaks up the word-token in the string with one token into its constituent atomic characters, then, break up the word-token in the string with one token into its constituent atomic characters

In the TB-Cleaner algorithm [8] the Names and important fields like Date of Birth and Address are converted into tokens (short format) for the purpose of comparison Duplicate Detection, Elimination and WID generation: Using each of the 2 token tables, identify all pairs of records as duplicates if they are (1) perfect match because their similarity match count (smc) is 1.0or (2) near perfect match because their SMC is between0.67 and 0.99.

The authors [9] explain that Transitivity rule states that if A is equivalent to B and B is equivalent to C then A is equivalent to C. Transitivity rule is used for similar-duplicate detection. Proposed Domain-Independent Transitivity (DIT) algorithm for similar-duplicate detection is totally domain independent, and it does not require any rule, or user involvement. The proposed algorithm also explores the idea to make the duplicate detection algorithms faster and more efficient for real-life data. A domain-independent transitivity rule is also proposed to reduce the negative aspect of the transitivity rule for domain-independent cases.

3. METHODOLOGY
Regression is a statistical process for estimating the relationships among variables. It can model relation between dependent and independent variables Regression can be used

in variety of application one such application that we are going to discuss is the problem of cleaning data that has been corrupted. Sometimes there is a data loss and those values are either rendered zero or are null values. Such type of data might result in error while computation. We therefore need a method that based on known data can accurately predict the unknown value. There are a number of regression methods that can predict that value our task is aimed at comparing that method for effectiveness.

3.1 Linear Regression

Linear regression is a type for regression model that is used to model the relationship between two variables x and y where y is a scalar dependent variable and x is an explanatory or independent variable.

Suppose there are n data points (x_i, y_i) where i =1…n

We need to find the best fit equation that can give us the relation between x and y

f(x) =b + a*x (equation 1)

The function Y is a dependent variable whose value depends on the variable x.

Best fit is a line that minimizes the sum of squared residuals of the linear regression model. This type of model can be found through least fit squares.

$err = \sum(d_i)^2 = (y_1-f(x_1))^2+\ldots..+(y_i-f(x_i))^2$ (equation 2)

Now we can substitute equation 2 in equation 1

$$err = \sum_{i=1}^{n}(d_i)^2 = \sum_{i=1}^{n}(y_i-(a+b*x))^2$$

$$\frac{\partial err}{\partial a} = -2\sum_{i=1}^{n}(yi-(a+b*x)) = 0$$

$$\frac{\partial err}{\partial b} = -2\sum_{i=1}^{n}(yi-(a+b*x))*x = 0$$

$a\sum x_i^2 + b\sum x_i = \sum x_i y_i$

$a\sum x_i + b*n = \sum y_i$

Using linear regression to predict the missing value:

Step 1 :

Choose the type of function that you want to model .

Firstly we choose a linear model of the form

Y = a*x+b ;

Step 2 :

Code the problem in Matlab and find the values of the constants a and b

Step 3 :

The Final equation obtained is found to be

F(x) = 34.87*x+11000

Predicted F(100) =14487

Actual F(100)=14395.07

Root Mean Square Error : 410.9

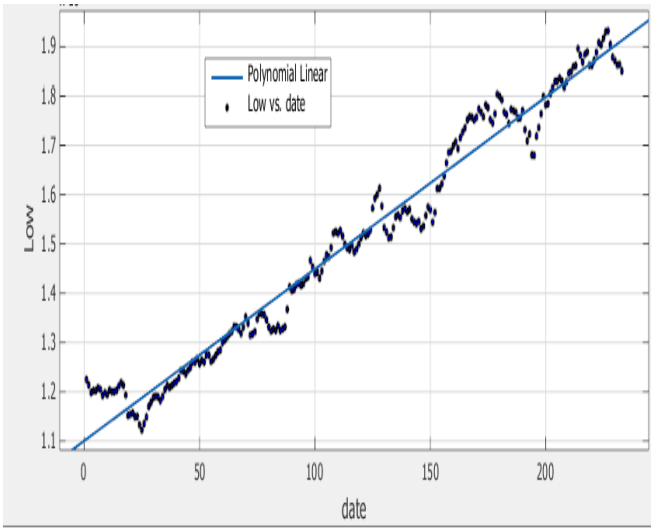

Fig1 Linear Regression method to predict Missing value

3.2 Quadratic Regression

Consider the polynomial equation of the form

$f(x)=a + b*x + c*x^2$ equation(3)

To Pick the coefficient that best fits the curve use the least square approach

The general expression for any error using least square error approach is :

$err = \sum(di)^2 = (y_1-f(x_1))^2+\ldots..+((yi-f(xi))^2$ equation(4)

now substitute the form of our equation 3 into the general equation 4

$err = \sum(di)^2 = \sum_{i=1}^{n}(yi-(a+b*x+c*x^2)$ equation (5)

Where n is the number of data points given, i is the current data point being summed. The polynomial being Quadratic,

find the best line = minimize the error (squared distance) between line and data points

Working with Quadratic Regression to predict missing values:

Step 1:

Choose the type of function that you want to model .

Firstly we choose a linear model of the form

$Y = a+b*x+c*x^2$

Step 2 :

Code the problem in Matlab and find the values of the constants a and b

Step 3 :

The Final equation obtained is found to be

$F(x) = 0.02798*x^x + 28.32*x + 11260$

Predicted F(100) =14371.81

Actual F(100)=14395.07

Root Mean Square Error : 395.7

3.3 Gaussian Regression

Start with the Exponential function

Y = a*e $e^{\frac{-(x-b)}{c^2}}$

Take logarithm of both sides

logY = log(a*e $e^{\frac{-(x-b)}{c^2}}$)

The properties of logarithm gives

LogY = log a - $\frac{x-b}{c^2}$ * log e

This expresses log Y as a linear function of x with,

Slope m = - $\frac{x-b}{c^2}$

Intercept c = log a

To obtain a best fit exponential curve of the form

Find the regression line for the data (x, log Y)

Step 1 :

Choose the type of function that you want to model .

Firstly we choose an exponential model of the form

Y = a*e $e^{\frac{-(x-b)}{c^2}}$

Step 2 :

Code the problem in Matlab and find the values of the constants a and b

Step 3

The Final modelled equation is found to be

Y = a* $e^{\frac{-(x-b)}{c^2}}$

a = 4.575e+04

b = 1073

c = 904.7

Y = 45750* $e^{\frac{-(x-1073)}{904.7^2}}$

Predicted F(100) =14389.56

Actual F(100)=14395.07

Root Mean Square Error : 395.2

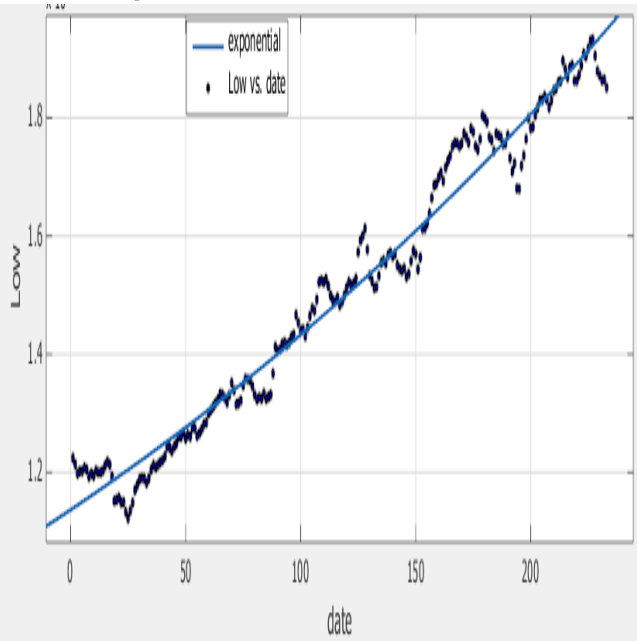

Fig 2 Quadratic Regression Method to predict missing value

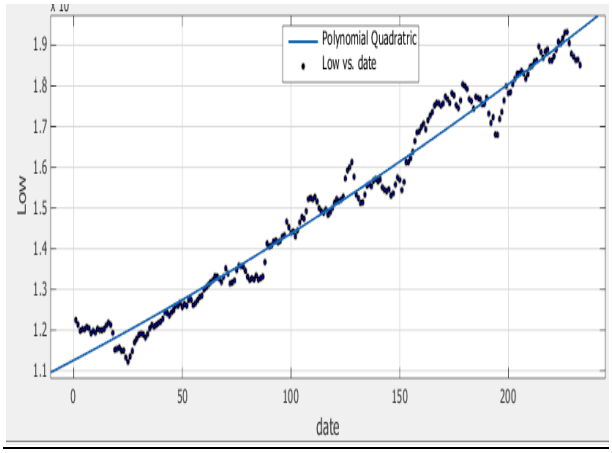

Fig 3 Exponential Regression method to Predict Missing Value

3.4 Exponential Regression

Start with the Exponential function

Y = $a*e^{x*b}$

Take logarithm of both sides

logY = log($a*e^{x*b)}$

The properties of logarithm gives

LogY = loga + x*b*log e

This expresses log Y as a linear function of x with,

Slope m = b*log e

Intercept c = log a

To obtain a best fit exponential curve of the form

Y = $a*e^{b*x}$

Find the regression line for the data (x, log Y)

Step 1 :

Choose the type of function that you want to model .

Firstly we choose an exponential model of the form

Y = $a*e^{x*b}$

Step 2 :

Code the problem in Matlab and find the values of the constants a and b

Step 3 :

The Final equation obtained is found to be

F(x) = a*exp(b*x)

a = 1.137e+04

b = 0.002314

F(x)=11370exp(.002314*x)

Predicted F(100) =14330.33

Actual F(100)=14395.07

Root Mean Square Error : 401.2

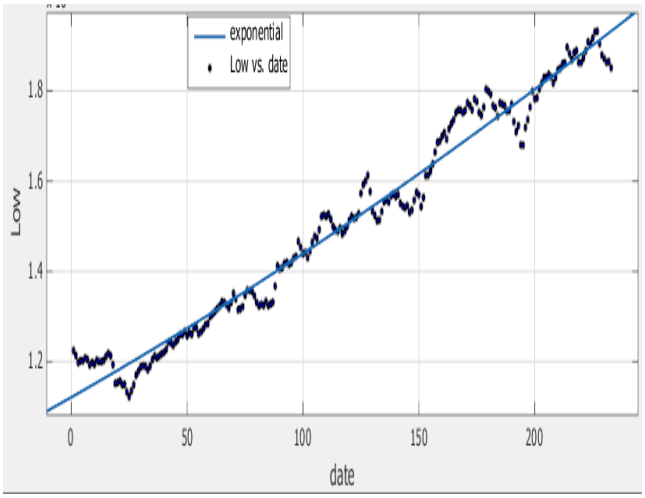

Fig4 Gaussian Regression method to predict missing value

3.5 Fourier Series Regression

Measured Values can be approximated through the periodic function . The procedure for this is leading to development of Fourier Series . The element of the series are sine and cosine function.

The Fourier Series is

F(x) = a+ bcos (w*x) + csin(w*x)

With Fourier coefficients b and c and

w = $\frac{2\pi}{kT}$

This is the period with T=a1-a2 where a1 is the initial interval and a2 final interval .

The coefficient b and c satisfy the least square fit condition . The Coefficient is calculated as follows:

b = $\frac{2}{\pi}\int_{a1}^{a2} f(x)\cos(w*x)*dx$

c = $\frac{2}{\pi}\int_{a1}^{a2} f(x)\sin(w*x)*dx$

Fourier coefficients

The Fourier coefficients a and b are here computed numerically using the trapezoidal method for the numerical integration. The accuracy can be improved by increasing the number of measurements is increased in the interval . Final simplification would lead to:

$$\int_a^b f(x)dx = (b-a)\left[\frac{f(a)+f(b)}{2}\right]$$

where f(x) denotes the function being integrated.

Step 1 :

Choose the type of function that you want to model .

Firstly we choose a linear model of the form

Y = a*x+b ;

Step 2 : Code the problem in Matlab and find the values of the constants a and b

Step 3 :

The Final equation obtained is

f(x) = a + b*cos(x*w) + c*sin(x*w)

Predicted F(100) =14273

Actual F(100)=14395.07

Root Mean Square Error : 358.5

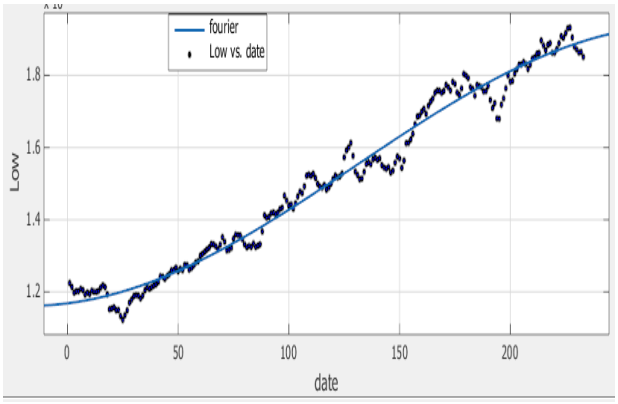

Fig 5 Fourier Series Regression method to predict missing values

3. RESULTS

After the coding was done in Matlab to find the missing values using different regression techniques, the following results were gathered.

Table1 (Analysis of Various methods to Predict Missing Values) *14395.07(actual missing value)

Sr. No	Method	Predicted Value	Root Mean Square Error
1	Linear Regression	14487	410.9
2	Quadratic Method	14371.81	395.7
3.	Exponential Method	14330.33	401.2
4.	Gaussian Method	14389.56	395.2
5.	Fourier Series Method	14273	358.5

The table shows the different regression techniques used along with the predicted value of missing data. We have calculated the root mean square error to check the accuracy of the predicted values.

4. CONCLUSION AND FUTURE SCOPE

Through this research we have given a comparison between the different regression techniques for the purpose of predicting missing values in data for the purpose of data cleansing. Comparing the above values for best fit through RMSE we find that the Fourier method is the best fit method according to both the visual as well as statistical parameter i.e. the root mean square error. But still the difference between the actual and the predicted value is larger than the Gaussian equation which has a considerably higher RMSE .We see that although the statistical parameter does not guarantee an accurate fit at low value but it gives a better fit for higher values Hence the statistical parameter can be used as a criteria

for discarding the poor fit values by choosing a suitable threshold through averaging of some of the best fit values. Based on the above observation we are in the process of designing an algorithm that can help us determine the suitable threshold value above which we can discard the predicted value. This research may be extended to test the discussed regression techniques on real time data and find the most accurate method for predicting missing values for the purpose of data cleansing and hence improving the quality of data in data warehouses.

5. REFERENCES

[1] "A Data Cleaning Method Based on Association Rules" by Weijie Wei, Mingwei Zhang, Bin Zhang, www.atlantis-press.com

[2] "Data Cleansing for Web Information Retrieval using Query Independent Features" by Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma- www.thuir.cn

[3] "An Extensive Framework for Data Cleaning " by Helena Galhardas, Daniela Florescu, Dennis Shasha, Eric Simon

[4] "A Token-Based Data Cleaning Technique for Data Warehouse" by Timothy E. Ohanekwu International Journal of Data Warehousing and Mining Volume 1

[5] Surajit Chaudhuri Kris Ganjam Venkatesh Ganti Rajeev Motwani, SIGMOD 2003, June 9-12, 2003, San DiegoCA. "Robust and Efficient Fuzzy Match for Online Data Cleaning"

[6] Christie I. Ezeife, Timothy E. Ohanekwu, University of Windsor, Canada, International Journal of Data Warehousing & Mining, 1(2), 1-22, April-June 2005 Research paper titled "Use of Smart Tokens in Cleaning Integrated Warehouse Data"

[7] Ajumobi Udechukwu, Christie Ezeife, Ken Barker Dept. of Computer Science, University of Calgary, Canada School of Computer Science, University of Windsor, Canada, 5th International Conference on Enterprise Information Systems (ICEIS) 2003, Research paper titled "INDEPENDENT DE-DUPLICATION IN DATA CLEANING"

[8] G.Siva Nageswara Rao, Dr.K.Krishna Murthy, Dr.B.V.Subba Rao, Dr.J.Rajendra Prasad, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 3, March 2012) research paper titled "Removing Inconsistencies and Errors from Original Data Sets through Data Cleansing"

[9] Kazi Shah Nawaz Ripon Department of Informatics, University of Oslo, Norway Computer Science and Engineering Discipline, Khulna University, Bangladesh Ashiqur Rahman and G.M. Atiqur Rahaman Computer Science and Engineering Discipline, Khulna University, Bangladesh, JOURNAL OF COMPUTERS, VOL. 5, NO. 12, DECEMBER 2010 research paper titled "A Domain-Independent Data Cleaning Algorithm for Detecting Similar-Duplicates"

[10] "The role of visualization in effective data cleaning" by Yu Qian, Kang Zhang – Proceedings of 2005 ACM symposium on applied computing

[11] "A Statistical Method for Integrating Data Cleaning and Imputation" by Chris Mayfield, Jennifer Neville, Sunil Prabahakar- Purdue University(Computer Science report-2009)

[12] "Data cleansing based on mathematical morphology" by Sheng Tang published in ICBBE 2008 The second International Conference-2008

[13] "A Domain Independent Data Cleaning Algorithm for detecting similar-duplicates" by Kazi Shah Nawaz Ripon, Ashquir Rahman and G.M. Atiqur Rahaman – Journal of Computer Vol 5, No.12,2010

[14] P.Pehwa "An Efficient Algorithm for Data Cleaning" www.igiglobal.com -2011.

[15] "Attribute Correction-Data cleaning using Association Rule and Clustering Methods" by R.KavithaKumar, Dr. RM. Chandrasekaran, IJDKP,Vol.1,No.2 March-2011.

[16] Random Forest Based Imbalanced Data Cleaning and Classification – Jie Gu –Lamda.nju.edu.cn

[17] Data Cleansing Based on Mathematical Morphology S.Tang-2008 –ieeeexplore.ieee.org. Bioinformatics and Biomedical Engineering , 2008 ICBBE 2008. The 2nd International conference.

[18] "An efficient Algorithm for Data Cleaning of Log File using File Extension" International journal of Computer Applications 48(8):13-18, June-2012 Surabhi Anand , Rinkle Rani Aggarwal.

[19] A New Efficient Data Cleansing Method – Li Zhao, Sung Sam Yuan, Sun Peng and Ling Tok Wang – ftp10.us.freebsd.org

[20] Computer Research and Development (ICCRD), 2011, 3rd International Conference.", Web log cleaning for mining of web usage patterns" –T.T.Aye.

[21] "Mass Data Cleaning Algorithm based on extended tree-like knowledge base" – Yan Cai-rong,SUN Gui-ning , GAO Nian-gao Computer Engineering and application 2009