Regression model approach to predict missing values in the Excel sheet databases

Filling of your missing data is in your hand

Z. Mahesh Kumar
School of Computer Science & Engineering
VIT University
Vellore, India
mahesh.cse349@gmail.com

R. Manjula
School of Computer Science & Engineering
VIT University
Vellore, India
rmanjula@vit.ac.in

Abstract— **The most important stage of data mining is *pre-processing,* where we prepare the data for mining. Real-world data tends to be incomplete, noisy, and inconsistent and an important task when pre-processing the data is to fill in missing values, smooth out noise and correct inconsistencies. We can handle the missing values by ignoring data row, using global constant to fill miss missing value, using attribute mean to fill missing value, using attribute mean for all samples belonging to the same class, using most probable value to fill the missing value , and finally we can use the data mining algorithm to predict the value. We use Regression method for this prediction of missing values. This method is used to map a data item to a real valued prediction variable. All these operations can be done by using EXCEL sheet database also.**

KEYWORDS: Preprocessing, Missing values, Regression, Prediction.

1. INTRODUCTION

1.1 Overview and Problem definition:

Everyone doing analysis has some missing data, especially survey researchers, market researchers, database analysts, researchers and social scientists. Missing data are questions without answers or variables without observations. Even a small percent of missing data can cause serious problems with your analysis leading you to draw wrong conclusions.

Real-world databases are highly susceptible to noise, missing, and inconsistent data due to they are typically huge in size often in gigabytes or more. We have to preprocess the data in order to help improve to quality of data and so as to improve the efficiency and ease of mining access. There are number of data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse or a data cube. Data transformations, such as normalization, may be applied. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance.

Need of preprocessing data: The data you wish to analyze by data mining techniques are incomplete (lacking attribute values or certain attributes of interest), noisy (containing errors) and inconsistent. Incomplete data can occur in many reasons. Attribute values may not be available, not considering important at the time of entry. Missing data[12], particularly tuples with missing values for some attributes, may need to be inferred.

Data cleaning:

Real world data tend to be noisy, incomplete, and inconsistent. Data cleaning routines[9] attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data.

We concentrate mainly on filling of missing values by ignoring the data row completely, filling the missing values manually, use the global constant to fill the missing values, use the attribute mean for 1 column of data, same using to fill all columns of data, using most probable value to fill missing value (Regression algorithm).

In the regression method[12], a regression model is fitted for each variable with missing values. Based on the resulting model, a new regression model is then drawn and is used to impute the missing values for the variable. Since the data set has a several missing data patterns, the process is repeated sequentially for variables with missing values.

2. METHODOLOGY

We have an excel sheet that having missing values.

Importing data:

From the jdbc-odbc connection we import the excel sheet data into a *ResultSet.*

Filling missing values:

We have to fill those missing data cells with 6 possible ways.

1. Ignoring the data row completely
2. Filling missing values manually
3. Use a global constant to fill the missing values
4. Use the attribute mean to fill the missing value
5. Use the attribute mean for all samples belonging to the same class as the given tuple
6. Use the most probable value to fill the missing value (Predicting by Regression algorithm)

We use Microsoft Office Excel sheet[10] to have our data.

Regression Methodology:

Regression Definition:

A regression is a statistical analysis[3] assessing the association between two variables. It is used to find the relationship between two variables.

RegressionFormula:

RegressionEquation (y) = a + bx
slope ‘b’, Intercept ‘a’

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

or Intercept(a) = $\frac{\Sigma Y - b(\Sigma X)}{N}$

and correlation coefficient is 'r'
where
x and y are the variables.
b = the slope of the regression line
a = the intercept point of the regression line and the y axis.
N = Number of values or elements
X = First Score
Y = Second Score
ΣXY = Sum of the product of first and Second Scores
ΣX = Sum of First Scores
ΣY = Sum of Second Scores
ΣX^2 = Sum of square First Scores

Exporting data:

We create *HSSFWorkbook* [1[6][8]]in excel file and in that *HSSFSheet* [5][7][9] is created. We perform the row and cell operations on that sheet to export data. We export the data of modified into created excel sheet by *FileOutputStream[2]*.

Regression Example:
To find the Simple/Linear Regression of

X Values	Y Values
60	3.1
61	3.6
62	3.8
63	4
65	4.1

To find regression equation, we will first find slope, intercept and use it to form regression equation

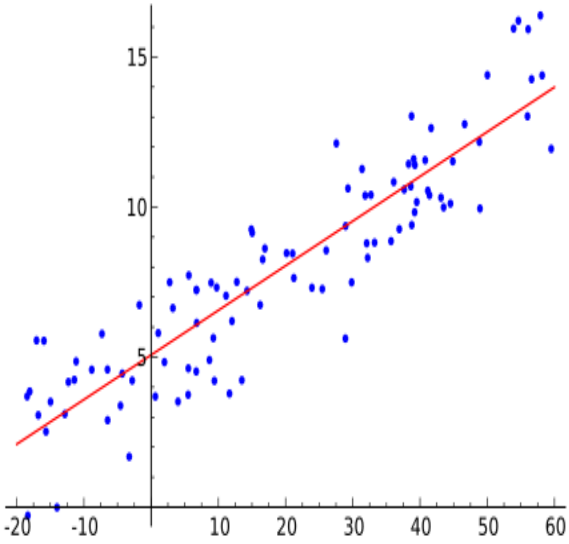

Fig 1: Linear regression model

Step 1: Count the number of values.

N = 5

Step 2: Find XY, X^2

See the below table

X Value	Y Value	X*Y	X*X
60	3.1	60 * 3.1 = 186	60 * 60 = 3600
61	3.6	61 * 3.6 = 219.6	61 * 61 = 3721
62	3.8	62 * 3.8 = 235.6	62 * 62 = 3844
63	4	63 * 4 = 252	63 * 63 = 3969
65	4.1	65 * 4.1 = 266.5	65 * 65 = 4225

Step 3: Find ΣX, ΣY, ΣXY, ΣX^2.

ΣX = 311

ΣY = 18.6

ΣXY = 1159.7

ΣX^2 = 19359

Step 4: Substitute in the above slope formula given.

Slope(b) = $\frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X2 - (\Sigma X)2}$

= ((5)*(1159.7)- (311)*(18.6))/((5)*(19359)-$(311)^2$)

= (5798.5 - 5784.6)/(96795 - 96721)

= 13.9/74

= 0.19

Step 5: Now, again substitute in the above intercept formula given.

Intercept(a) = $\frac{\Sigma Y - b(\Sigma X)}{N}$

= (18.6 - 0.19(311))/5

= (18.6 - 59.09)/5

= -40.49/5

= -8.098

Step 6: Then substitute these values in regression equation formula

Regression Equation(y) = $a + bx$

= -8.098 + 0.19x.

Suppose if we want to know the approximate y value for the variable x = 64. Then we can substitute the value in the above equation.

Regression Equation(y) = a + bx

= -8.098 + 0.19(64).

= -8.098 + 12.16

= 4.06

This example will guide you to find the relationship between two variables by calculating the Regression from the above steps.

3. IMPLEMENTATION AND RESULTS

Implementation steps:

1. First of all create JDBC-ODBC connection with excel sheet data (Data Sources→ JDBC-ODBC →system DSN→choose driver→set). Give the data source name as related to dataset.

2. Import the excel sheet data[4] by using JDBC-ODBC connection

3. Try different methods to fill the missing values as told earlier (6 methods)[12].
4. Export the same filled data into a new excel sheet.

Pseudo Code :

Start

1. Import all the packages which are belongs to excel data.
2. Create an output file to export our filled data.
3. Create Excel Sheet requirements i.e. Creating Workbook, sheet, fields etc.
4. Import the data of our excel sheet which having missing data and storing each and every column data into an array.
5. Check each and every row such that whether there is missing data or not.
If any missing data is found, we fill by 2 ways.
    (a) If it is the string value we manually fill that value
    (b) If it is integer value we can fill with any of the 6 methods as said earlier.
6. Linear Regression algorithm is applied for integer data in the 6^{th} step of filling missing values.
7. Finally export our new data that is modified i.e. missing values filled data, into new excel sheet.
End

Result & Conclusion:

Finally we got our modified excel sheet with filled data of missing values. And these are further used in statistical analysis and even more. Reduction in sample size also reduces the power of statistical significance testing. The most important advantages of these mean imputation methods. In this way we can find the missing values and fill it in the database without changing manually by using regression model.

References:

[1] Sundaram, Elango (2004-03-22), *Excelling in Excel with Java*, Java World.
[2] *POI homepage from October 2004*, Coyote Song, showing original explanations for naming.
[3] DeMarco, Jim (2008). "Excel's data import tools". *Pro Excel 2007 VBA*. Apress. p. 43 *ff*.ISBN 1590599578.
[4] Harts, Doug (2007). "Importing Access data into Excel 2007". *Microsoft Office 2007 Business Intelligence: Reporting, Analysis, and Measurement from the Desktop*. McGraw-Hill Professional.ISBN 0071494243.
[5] Harvey, Greg (2007). *Excel 2007 Workbook for Dummies* (2nd ed.). Wiley. p. 296 *ff*.ISBN 0470169370
[6] "XML Spreadsheet Reference". *Microsoft Excel 2002 Technical Articles*. MSDN. August 2001. Retrieved 2008-11-10.
[7] "OpenOffice.org's documentation of the Microsoft Excel File Format". 2008-08-02.
[8] Dodge, Mark; Stinson, Craig (2007). *Microsoft Office Excel 2007 inside out*. Microsoft Press. ISBN 073562321X.
[9] Adèr, H.J.(2008). "Chapter 13: Missing data". In Adèr, H.J., & Mellenbergh, G.J. (Eds.) (with contributions by Hand, D.J.), *Advising on Research Methods: A consultant's companion* (pp. 305-332). Huizen, The Netherlands: Johannes van Kessel Publishing. ISBN 9079418013
[10] Graham, J.W., Olchowski, A.E., and Gilreath, T.D. (2007) "How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory". *Preventative Science* 8 (3), 208-213 doi:10.1007/s11121-007-0070-9
[11] Zarate LE, Nogueira BM, Santos TRA, Song MAJ (2006). "Techniques for Missing Value Recovering in Imbalanced Databases: Application in a Marketing Database with Massive Missing Data".*IEEE International Conference on Systems, Man and Cybernetics, 2006. SMC '06.*. **3**. pp. 2658–64. doi:10.1109/ICSMC.2006.385265.

Authors:

Mr. Z.Mahesh Kumar received his B.Tech degree in Computer Science and Engineering in 2011 from Acharya Nagarjuna University, Guntur, Andhra Pradesh and pursuing M.Tech degree in Computer Science and Engineering in 2011-13 from VIT university, Vellore, Tamilnadu. His areas of interest are Web Technologies, Database management systems. As part of this paper, he is working on developing new shopping system based on web which depend and work on XML data source, XML-Web Services and DOM Parsing.

Prof. R.Manjula received her B.E in Computer Science & Engineering from University of Vishwesvaraya and Engineering, Bangalore, Karnataka State, India in 1992 and M.E in Software Engineering from Anna University, Tamil Nadu, India in 2001. She is now working as Associate Professor and also as Ph.d Candidate affiliated with School of Computing Science and Engineering at Vellore Institute of Technology, Vellore, India. Her area of specialization includes Software Process modeling, Software Metrics, Software Metrics, Software Testing and Metrics, XML-Web Services and Service Oriented Architecture

Regression Analysis in Excel. Explanation of Regression Mathematically. How to Perform Linear Regression in Excel? #1 â€“ Regression Tool Using Analysis ToolPak in Excel. #2 â€“ Regression Analysis Using Scatterplot with Trendline in Excel.Â It occurs because Yâ€™s predicted value will never be exactly the same as the actual value for a given X. We donâ€™t need to worry about this error term as some software do the calculation of this error term in the backend for you. Excel is one of that software. In that case, the equation becomesÂ In this case, we want to see the output on the same sheet. Therefore, given range accordingly.Â It gives values of coefficients that can be used to build the model for future predictions. Now our, regression equation for prediction becomes Another approach for filling in the missing data is to use the forecasted values of the missing data based on a regression model derived from the non-missing data. For the data in Figure 1, this results in the following. Figure 5 â€“ Regression imputation. This time we impute the values of the five missing cells by inserting the array formula =FORECAST(J13:J17,G6:G12,F6:F12) in the range K13:K17. Since this results in a perfect linear correlation between the math and science values for the last five data elements, it is not surprising that the correlation coefficient between math and science rises from .769 (cell F20) to .859 (cell J20). The standard deviation also falls. You can perform predictive modeling in Excel by following a few simple steps. In this article learn how to create a linear regression model in Excel.Â An old customer of yours named Aleksander walks in and we wish to predict the sales from him. We can simply plug in the number from the data in the linear regression model and we are good to go! Aleksander has an income of 40k and lives 2km away from the store. What is the estimated sales? The equation becomes: Here, our model has estimated that Mr. Aleksander would pay 4218 units to buy his new pair of shoes! Thatâ€™s the power of linear regression done simply in Microsoft Excel. End Notes. In this article, we learned how to build a linear regression model in Excel and how to interpret the results. How to restore missing data (nulls) in height feature based on existing dependancy (correlation) between these variables? To be more clear: Input and output variables have clear correlation. I guess that predicting missing values for excel is not a difficult procedure. But I need some directions how to implement it. excel regression correlation predict. Regression model approach to predictmissing values in the Excel sheet databases. Abstract. KEYWORDS.Â missing values, smooth out noise while identifying outliers and correct inconsistencies in the data. We concentrate mainly on filling of missing values by ignoring the data row completely, filling the missing. values manually, use the global constant to fill the missing values, use the attribute mean for 1 column of data, same using to fill all columns of data, using most probable value to fill missing value (Regression algorithm). In the regression method[12], a regression model is fitted for each variable with missing values. Based on the resulting model, a new regression model is then drawn...