International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-03, Issue-10, Jan 2018

Processing of Incomplete Data Sets: Prediction of Missing Values by using Multiple Linear Regression

Mr. Mahesh B. Shelke, Mr.Yogesh R Nagargoje

Asst Prof. Department of Computer Science, Everest College of Engg. & Tech., Aurangabad, M.S., India.

Mahesh_shelke21@hotmail.com, yogeshvcet1@gmail.com

Abstract: **A large number of Data Sets are available, but they are incomplete in nature so that they cannot be used for real applications. These incomplete data sets are produced due to various reasons like system failure, privacy of data, incomplete input, time delay in system, lack of available resources. As common examples Weather data sets, sensor image data sets, Measurement Reflection in Agriculture Data Sets. For these cases we use conceptual data reconstruction by using statistical models of multiple linear Regressions to Predict Missing values in Data Sets.**

Keywords: Data Mining, Incomplete Data Sets, Missing Values, Pre-Processing, Unsupervised Filter.

I. INTRODUCTION

In Data Mining, Data Retrieval System, Text Mining, Web Mining these different techniques needs different data sets in different format and all these techniques work with assumption that available data are Complete in nature. When Different data sets are produced they also includes noise in data, which are unnecessarily available in data sets and because of noise in data sets useful attributes also lose their meaning, so that before using these noisy data sets the must be preprocessed so that useful attributes can be retrieved and used for further processing of data sets. Many Data Mining algorithms are used for preprocessing of data which removes noise from data sets, redundancy in data sets, which makes data sets useful for processing of knowledge from data sets. We note that any missing data mechanism would rely on the fact that the attributes in a data sets are not independent of one another, that there is some predictive value from one attribute to another [1]. If the attribute values in available data sets are uncorrelated, then the output of preprocessor will lead to incorrect state of data sets. In this paper, we discuss the technique of reconstruction for correlated data sets attributes and predicting missing data along the linear directions such as Incremental Model.

II. CONTRIBUTIONS OF THIS PAPER

This paper discuss the technique of statistical reconstruction of incomplete data sets using multiple linear regression and which uses the correlation of data sets attribute to predict the missing values which helps to produce complete data sets in nature. This paper is theoretical and generalized algorithm approach to predict missing values by using multiple regressions Model in weka software it updates the weka software with preprocessing unsupervised filter.

III. WORKING OF LINEAR REGRESSION

Linear Regression is most commonly used technique for determining relation between a scalar dependent variable y and one or more explanatory variables denoted X. The case of explanatory variable is called simple linear regression. For more than one explanatory variable is called as Multiple Linear Regression [2].Linear Regression Model has many Practical uses [3].

1. To describe the linear dependence of one variable on another.
2. To predict values of one variable from values of another, for which more data are available.
3. To correct for the linear dependence of one variable on another, in order to clarify other features of its variability.

Assumptions behind Linear Regression Model:

The assumptions that must be met for linear regression to be valid depend on the purposes for which it will be used. Any application of linear regression makes two assumptions:

1. The data used in fitting the model are representative of the population.
2. The true underlying relationship between X and Y is linear.[3]

Steps in constructing good regression models:

1. Plot and examine the data.
2. If necessary, transform the X and Y variables so that the relationship between X and Y is linear.
3. Calculate the linear regression statistics.
4. Examine the slope and Intercept
5. Determine the Residuals.
6. Check results, both visually and statistically.[3]

Derivations of Linear Regression [3]:

Set of n points (Xi, Yi) on a scatter Plot. Find the best Fit-line,

$$Y_i = W_0 + W_1X + W_2X.$$

Such that the sum of Squared errors in Y, $\sum(\overline{Y}_i\text{-}Y)^2$ Where, W_i are the weights. Weights are calculated from the training data sets.[3]

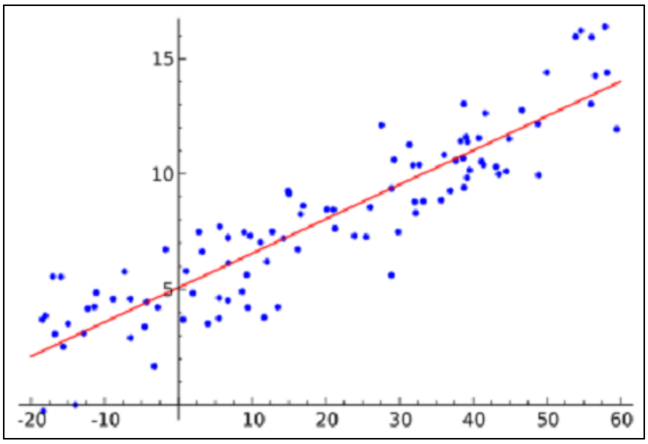

Figure 1 : Simple Linear Regression [2].

The linear regression model is used to adjust slope and intercept values to find the best fitting line

Between the coordinates of scatter plot. This can be achieved by minimizing sum of squares of the distances. The line determined by minimizing the sum of squares is most likely to be correct.

IV. RECONSTRUCTION ALGORITHM

1. Take Data Set as Input (which contains missing values in attributes).
2. Repeat the Steps Until all Attributes cannot be accessed in Particular Array. (Initially array is empty).It starts accessing data from first instance.
3. And check all Missing values and store these values into another array variable.
4. Take mean of each attribute that is $(\overline{X}, \overline{Y})$.
5. Calculate the co-efficient (Wi) by using Mean square error method.
6. To find the best fitting line by using straight line equations as:

   $$Y_i=W_0+W_1X+W_2X.$$
7. Retrieve the missing values and predict the missing value of X.
8. After completing first Instance repeat these steps for all Instances in Data Set. And Find all Missing values from data set.

   Note:

The linear regression does not analyses the values are linear or not. It assumes that all values are linear. And it works only for Numeric Values.

Following Example Shows how MLR works for Prediction of Missing values. Example: Suppose Data Set Contains data of Employees who works for certain organization with year of Experience and corresponding Salary in Thousands.

Experience (Year)	Salary (In Thousands)
1	20
3	30
3	36
6	43
8	57
9	64
10	**??**
11	59
13	72
16	83
21	90

Table 1: Employee Data Set With Missing Values.

Following Graph Shows the linear values of data set Employee and Salary having missing Attributes.

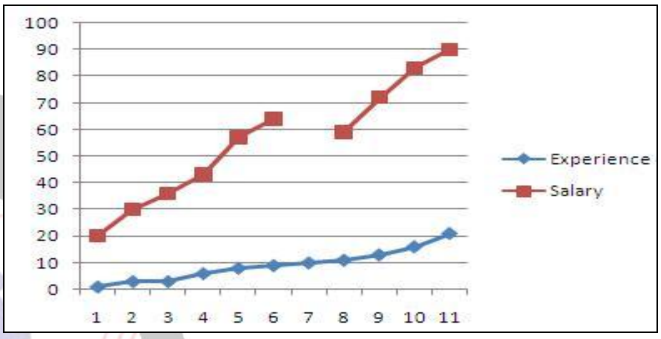

Figure 2: Graphical Representation of Data Set.

After processing this data set with multiple regression had produced the result of missing value of data set that is value for experience 10 is calculated as 58.57 which is approximately correct for this data set and the efficiency of this algorithm depend on the attributes missing in data set, If less values of data set are missing then the efficiency of prediction of these values is greater than the values which are more missing in data set. So in above example only one value had been missing so that it has calculated the value more efficiently. And following graph shows the predicted value for above data set that is value of experience 10 to value of salary as:

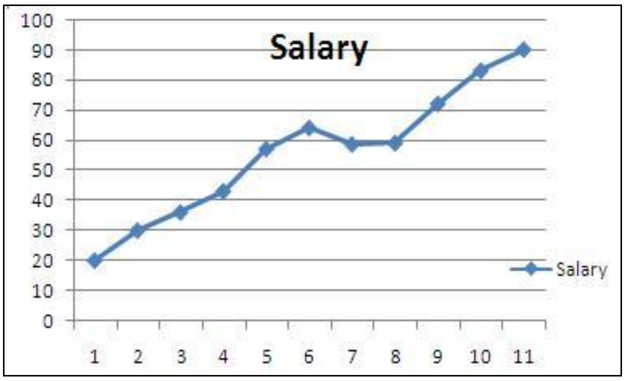

Figure 3: Complete Data Set after Prediction of Missing value.

As Multiple regressions working as preprocessor in this paper then why we are calling it as an preprocessor in weka Software? As we know that filtering in weka works as preprocessor and it includes two major categories in it as follows:

1) Unsupervised Filtering
2) Supervised Filtering

And Unsupervised Filters works for the data sets which has having missing attribute values in data sets so that we are including multiple linear regression as an Unsupervised Filter and after applying MLR algorithm it will generate the data set with prediction of all missing values from data set which results complete data set as an Output. After implementing multiple linear regressions in weka as preprocessor tools and Unsupervised Filter (Dataset contains missing values so that it is incomplete in nature.) it produces following results and also generates the output which is complete data set and which can produce more useful statistical output for further processing of dataset and efficiency of these filters depend on how many attributes are missing in data if it contains large missing values its efficiency will be less but if dataset contains less dataset values are missing then efficiency will be high.

Following Table shows Results After applying regression filter.

Original dataset	Dataset with Five missing value	Output after applying filter
@relation glass @attribute Id integer @attribute RI real @attribute Na real @attribute Mg real @attribute Al real @attribute Si real @attribute K real @attribute Ca real @data 1,1.52101,13.64,4.49,1.10,71.78,0.06,8.75 2,1.51761,13.89,3.60,1.36,72.73,0.48,7.83 3,1.51618,13.53,3.55,1.54,72.99,0.39,7.78 4,1.51766,13.21,3.69,1.29,72.61,0.57,8.22 5,1.51742,13.27,3.62,1.24,73.08,0.55,8.07	@relation glass10%m @attribute Id integer @attribute RI real @attribute Na real @attribute Mg real @attribute Al real @attribute Si real @attribute K real @attribute Ca real @data 1,1.52101,13.64,4.49,1.10,**?**,0.06,8.75 2,1.51761,**?**,3.60,1.36,72.73,0.48,7.83 3,1.51618,13.53,**?**,1.54,72.99,0.39,7.78 4,1.51766,13.21,3.69,1.29,72.61,0.57,**?** 5,1.51742,13.27,3.62,1.24,73.08,**?**,8.07	@relation:glass10- @attribute Id integer @attribute RI real @attribute Na real @attribute Mg real @attribute Al real @attribute Si real @attribute K real @attribute Ca real @data 1,1.0,1.52101,13.64,4.49,1.1,**73.7391**,0.06,8.75 2,2.0,1.51761,**13.2551**,3.6,1.36,72.73,0.48,7.83 3,3.0,1.51618,13.53,**3.29470**,1.54,72.99,0.39,7.78 4,4.0,1.51766,13.21,3.69,1.29,72.61,0.57,**7.4854** 5,5.0,1.51742,13.27,3.62,1.24,73.08,**0.53974**,8.07

Table 2 : Data Set after Applying Unsupervised Filter

V. Conclusion

In this paper we have discussed the approach to find missing values of data sets by using technique of Multiple Regressions in which mathematical model helps us to find those corresponding values, this paper includes theoretical and reconstruction algorithm for this technique in data mining which will help us to reconstruct incomplete data set and produce complete data set before processing these data sets by using various data Mining algorithm.

Acknowledgements

The completion of my paper, “Processing of Incomplete Data Sets: Prediction of Missing Values by using Multiple Regression” Would not have been so easy without the guidance of various individuals, who gave an appropriate direction for our efforts.

REFERENCES

[1] Srinivasan Parthasarathy and Charu C. Aggarwal, On the use of conceptual Reconstruction for Mining Massively Incomplete Data Sets, IEEE PP.1512-1521,2003.

[2] http://en.wikipedia.org/wiki/Linear_regression

[3] http://seismo.berkeley.edu/~kirchner/eps_120/EPSToolkits.html

[4] C.C. Aggarwal, On the Effects of Dimensionality Reduction on High Dimensional Similarity Search, 2001.

[5] Data Mining: Concepts and Techniques, 2nd^{ed} by Jiawei Han and Micheline Kamber.