HCTL Open International Journal of Technology Innovations and Research (IJTIR)
http://ijtir.hctl.org
Volume 14, April 2015
e-ISSN: 2321-1814, ISBN (Print): 978-1-62951-946-3

A Comparative Study of Missing Value Imputation Methods on Time Series Data

Arumuga Nainar S.
sanainar@gmail.com

Abstract

Data mining is process of analyzing enormous sets of data and extracting meaningful information from data. Temporal data mining deals with data which has time information. In practical, the data collected may contain noisy, inconsistent data and in many cases the data may be missing. So one of the important step that need to be done in data mining is data pre-processing. Incomplete data may generate biased results and impact the accuracy of analysis. In order to rectify this it is important to predict the missing values based on other details in the dataset. The work focuses on predicting missing data using mean imputation, hotdecking and Inverse Distance Weighted Interpolation methods and compare the results of each of these methods. Machine learning methods applied to the imputed dataset will give better accuracy than that of the incomplete dataset

Keywords

Missing values, imputation, inverse distance weighted interpolation.

Introduction

Datasets are not complete often. There can be many reasons frequently characterized by their incompleteness. There are a number of reasons why values are missing in dataset. The reasons include ignoring values in dataset and the participants did not respond for questionnaires. Missing data is main problem while analyzing the data. Knowing the exact reason for missing data in advance is uncommon. Unavailability of full data hinders the process of decision making because it depends on complete dataset. Most decisions on business and scientific domain are based on data available during the process of such decision making. For example, decisions on business has dependency on complete information about sales and other data that are available. Simply ignoring missing data is not a good option as this may impact the result, for example, breakdown in automation systems. So it is important to do the decision making process based on completely available data.

There are different patterns of missing data as shown in Figure 1.1. The rows in this figure corresponds to observations and columns corresponds to variables. Figure 1.1(a) shows univariate pattern where data are missing for one variable, the variable B in Figure 1.1(a). In Monotone pattern the data are not available in many variables which is shown in figure 1.1(b). In Arbitrary pattern where the data are missing randomly as shown in Figure 1.1(c).

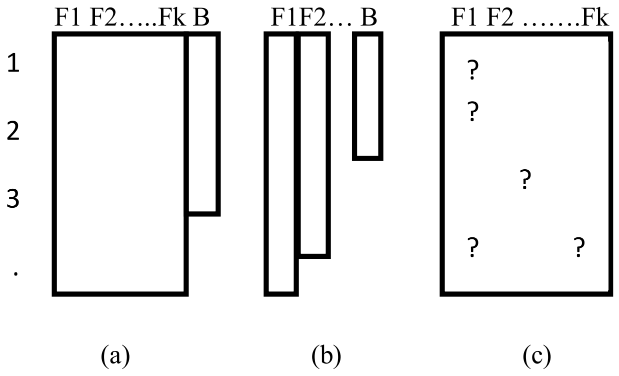

Figure 1.1

Missing data imputation deals with predicting missing data using various techniques. So that the dataset becomes complete and the results arrived from analysis is most accurate.

Related Work

There are number of methods exist for missing data imputation.

Listwise Deletion:

Listwise deletion is one of the simplest method to handle missing values. This is known as complete-case analysis [12]. In listwise deletion method, the software deletes the case which has missing values. Even if an attribute has less number of missing data, the entire collection will be deleted. This data reduction in listwise deletion may lead to decreased statistical power and the risk of missing important knowledge. Research community says this method of imputation is not mostly desirable. This method may be appropriate if the dataset is very large and the missing values are very minimal. So that the loss of statistical power is less. The advantage of listwise deletion method is the researcher does not need to spend time on conducting analysis on missing value imputation.

Mean Substitution:

In this method, the missing data of an attribute is found by calculating mean of total values of that attribute. For example, if average income of participants in analysis is \$1200.50, then \$1200.50 is populated in missing values for income in dataset. This method is appropriate if the dataset has minimum number of missing values. This method takes time in the beginning but it is decreased on further uses.

Hotdecking:

This method identifies a participant in dataset who has full data and has same characteristics as a participant who has missing values and that participant's data is used to populate missing data. A matrix is used to find what are the attributes (for example, attributes B and C) that are most correlated to the attributes that have missing values (attribute A, for example). Then the values are sorted by one of the correlated attributes (B or C) in ascending order. From the sorted values, the values for attribute A are populated by value of preceding participant. This results in missing data are populated using value from a sample that is similar on correlated attribute. For example, the participant A who do not have income data is given value from his precedent A whose income is $1800.

Hotdecking is appropriate when the attribute which is used for sorting is correlated with attribute which has missing data. Also this method is suitable when the dataset is large so that the correlated attribute can be identified easily. The advantage of this method is the standard deviation of attribute with imputed values better estimates the standard deviation of attribute without imputation. The limitation of hotdecking method is it is hard to implement. The programming need more time.

Regression Imputation:

In this method, more than one predictor of attribute that has missing data are found using a matrix. The predictor attributes are chosen and treated as independent attributes in regression equation. The attribute with missing value is treated as dependent attribute. Samples with full data for predictor attributes generate regression equation. The equation is used to find missing data. In the repeated process, data of missing attribute are substituted and dependent attribute is predicted using all cases. These steps are iterated until the difference between imputed values is minimal between consequent steps. The values from final iteration are used for imputing missing data.

Inverse Distance Weighted Interpolation:

Inverse distance weighted interpolation implements the approach that data points that are near to others are more similar compare to the data points that are far away. To find missing value, IDW uses observed values in the prediction area. The observed values nearest to the prediction area have more influence than those are away. IDW gives larger weights to data points nearest to the prediction area, and the weights decrease as a function of distance.

Weights are inversely proportional to the distance raised to the power value p. When the distance increases, the weights decrease rapidly. The value of p determines the rate at which the weights decrease. When p = 0, the weight is same and no decrease with distance and the prediction is mean value of all values in the prediction area. When p increases, the weights decrease rapidly. Analyst uses p=1 or greater. The value 2 is used by default for p.

When distance increases, the observed values will have less relationship to the value in prediction area. The points at very large distance can be excluded to speed up the calculation. So, it is common to specify neighborhood to limit the number of observations for calculation.

Method

The ozone level detection dataset is selected to study and compare result of different imputation methods. The dataset is chosen for demonstration purpose from UCI machine learning repository. The data were collected from year 1998 to 2004 at the cities Houston, Galveston and Brazoria in Texas, Unites States. The dataset has temperature and wind speed details measured at various time. The imputation methods mean imputation, hot decking and inverse distance weighted interpolation are applied to find missing values in the dataset and compared the results based on Root Mean Square Error (RMSE). RMSE is mostly used measure of differences between values estimated by model and actual observed values.

To find missing value using mean imputation method, the observed values for attribute which has missing values are taken for the current year, the mean is calculated and it is substituted on the missing value.

To find missing value using hotdecking method, the record which has complete data and has same characteristics as the record which has missing value is identified and value from that record is taken and replaced in missing value. In this dataset, the Date attribute is chosen as dependent variable. To find out missing value for a particular day, the value for the same day in another year was chosen as predicted value.

In IDW method, first the neighborhood is identified using k-nearest neighbor algorithm. Here the neighborhood is chosen as 31 days. From the data points available in the neighborhood the missing value is identified using IDW formula assuming the power = 2.

$$P(x) = \frac{\sum_{i=1}^{n} P(i)/d(x,i)\hat{}k}{\sum_{i=1}^{n} 1/d(x,i)\hat{}k} \qquad (3.1)$$

P(i) represents the value observed at point i

d(x,i) is the distance between the point x and the point i

n represents the total number of known nearest points for point i

k is a positive power function.

In the dataset, the value for variable WSR0 for the day 6/30/1998 is missing. Using IDW method, it is calculated as 1.76.

Experimental Results

The implementation of different methods of imputation is done using Java programming language. Table 4.1 shows missing values in the dataset represented by blank.

Date	WSR0	WSR1	WSR2	WSR4

1/1/1998	0.8	1.8	2.4	2
1/2/1998	2.8	3.2	3.3	3.3
1/3/1998	2.9	2.8	2.6	2.2
1/4/1998	4.7	3.8	3.7	2.9
1/5/1998	2.6		1.6	0.9
1/6/1998	3.1	3.5	3.3	1.6
1/7/1998	3.7	3.2	3.8	6
1/8/1998	2.25	2.9	3.4	4.7
1/9/1998	1	1.5	1.2	0.7
1/10/1998	0.9	0.6		0.6
1/11/1998	1.1	1.7	1.4	0.9
1/12/1998	3.7	4.2	3.1	2.3
1/13/1998	2.25	0.6	0.3	1.3
1/14/1998	1.3	1.3	1.6	1.4
1/15/1998	4.2	5.1	5.1	5.5
1/16/1998	0	0.2	0.1	0.7
1/17/1998	2.1	2.2	2.2	2.1
1/18/1998	2.25	2.3	1.3	1.6
1/19/1998	2.7	2		3.3
1/20/1998	0.3	0.6	1.1	1.7
1/21/1998	2.1	1.8	1.1	1.4
1/23/1998	3.4	3.5	3.8	3.1
1/24/1998	2.25	0.7	1.4	1.6
1/25/1998	1.2	0.7		0.3

Table 4.1

When applied, mean imputation method, the value predicted for the variable WSR0 for the date 6/30/1998 is 1.64. For the same variable and date, the hotdecking method predicted the value as 1.7. The IDW interpolation method estimated the value as 1.76 for the same variable and date. The Table 4.2 shows observed value versus estimated value using all the three imputation methods.

	Observed Value	Estimated Value
Mean imputation	2.3	1.64
Hotdecking	2.3	1.7
IDW interpolation	2.3	1.76

Table 4.2

The IDW method of estimation is compared with mean imputation and hotdecking methods. The performance of IDW method is compared with performance of mean imputation and hotdecking with respect to their corresponding RMSE values.

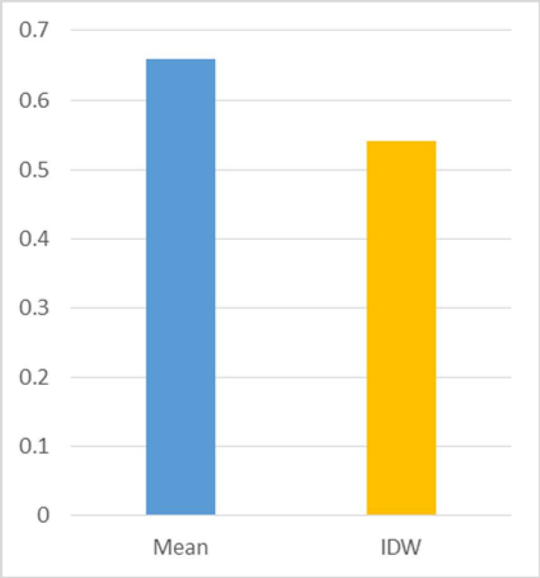

Figure 4.1 RMSE values for IDW and Mean Imputation

The chart in Figure 4.1 shows that the RMSE is less when IDW method is used when compared to mean imputation method. The chart in Figure 4.2 below shows that the RMSE is less when IDW method is used when compared hotdecking imputation method. Thus it proves to be a better method for missing value imputation.

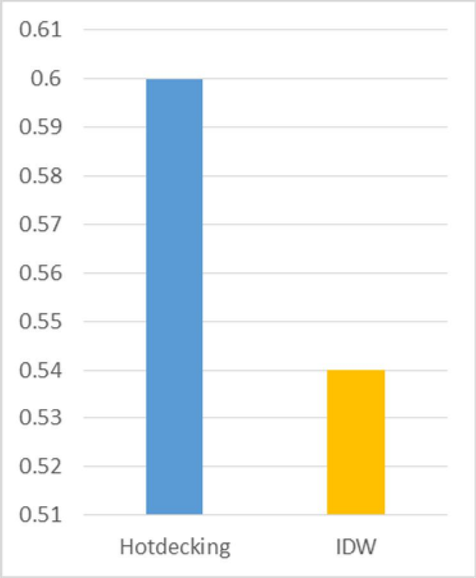

Figure 4.2 RMSE values for IDW and Hotdecking

Conclusion and Future Work

The work focuses on imputing missing values using different imputation methods such as mean imputation, hotdecking and inverse distance weighted interpolation. The experimental results show that the RMSE value calculated on observed value and estimated value is less in IDW method when compared to mean imputation and hotdecking methods.

As a future work the data set after imputation can be further used for analysis such as prediction and its performance can be compared with the analysis results of incomplete data set to determine its accuracy.

References

[1] Allison, P. D. (2002). Missing data (Sage University Paper Series on Quantitative Applications in the Social Sciences 07-136).Thousand Oaks, CA: Sage Publications.

[2] Figueredo, A. J., McKnight, P. E., McKnight. K. M., & Sidani, S. (2000). Multivariate modeling of missing data within and across assessment waves. Addiction, 95(Suppl. 3), S361-S380.

[3] Graham,J.W.,Taylor,B.J.,& Cumsille, P. E. (2001). Planned missing-data designs in analysis of change. In L. M. Collins & A. G. Sayer (Eds.), New methods for the analysis of change (pp. 335-353). Washington, DC: American Psychological Association.

[4] Hadley-Ives, E., Stiffman,A. R., Elze, D.,Johnson, S. D., & Dore, P. (2000). Measuring neighborhood and school environments: Perceptual and aggregate approaches. JoMrna/ of Human Behavior in the Social Enuironment, 3(1), 1-28.

[5] Little, R.J.A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). New York: John Wiley & Sons.McDonald, R. A.,Thurston, P.W, & Nelson, M. R. (2000). A Monte Carlo study of missing item methods. Organizational Research Methods, 3,71—92.

[6] Orme.J. G., & Reis,J. (1991). Multiple regression with missing data. Journal of Social Service Research, /5(l/2), 61-91.

[7] Pigott,T. D. (2001). A review of methods for missing data. Educational Research and Evaluation, 7, 353-383.

[8] Raaijmaken, Q.A.W. (1999). Effectiveness of different missing data treatments in surveys with Likert-type data: Introducing the relative mean substitution approach. Educational and Psychological Measurement, 59,725-748..

[9] Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons.

[10] Rubin, D. B. (1991). EM and beyond. Psychometrika, 56, 241-254.SAS Institute Inc. (2000). SAS language and procedures.Cary, NC: Author

[11] Schafer,J. L. (1998). NORM: Multiple imputation of multivariate continuous data under a normal mode. Version 2. [Computer software for Windows 95/98/NT]. Available at http://www.stat.psu.edu/~jls/misofrwa.html

[12] Schafer.J. L., & Graham, J.W. (2002). Missing data: Ournview of the state of the art. Psychological Methods, 7, 147-177.

[13] Streiner, D. L. (2002).The case of the missing data: Methods of dealing with dropouts and other research vagaries. Canadian Journal of Psychiatry, 47,68-75.

[14] Tabachnick, B. G., & Fidell, L. S. (1983). Cleaning up your act: Screening data prior to analysis. In B. G. Tabachnick & L. S. Fidell, Using multivariate statistics (pp 68-81). New York: Harper & Row.

[15] UCI Repository: UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/ Istanbul stock exchange Data Set.