European Journal of Research Development and Sustainability (EJRDS)
Available Online at: https://www.scholarzest.com
Vol. 1 No. 3, November2020,
ISSN: 2660-5570

DESIGN AND DEVELOPMENT OF MISSING VALUES AND PREDICTION OF TIME SERIES DATA

Mukta Agarwal
Assistant Professor, Sabarmati University, Ahmedabad , Gujarat, India,
mukta09agarwal@gmail.com

Article history:

Received 26th October 2020
Accepted: 7th November 2020
Published: 21st November 2020

Abstract:

Data preprocessing plays a huge and essential capacity in the data mining measure. Data preprocessing is expected to improve the adequacy of a figuring. This paper bases on missing worth evaluation and estimate of time course of action data subject to the undeniable characteristics. Different figurings have been made to handle this issue, yet they have a couple of limitations. Most existing counts like KNNimpute (K-Nearest Neighbors attribution), BPCA (Bayesian Principal Component Analysis) and SVDimpute (Singular Value Decomposition credit) can't deal with the condition where a particular time point (portion) of the data is missing inside and out. This paper revolves around autoregressive-model-based missing worth appraisal technique (ARLSimpute) which is ground-breaking for the situation where a particular time point contains many missing characteristics or where the entire time point is missing. Data preprocessing yield is given to the commitment of the desire systems to be explicit direct conjecture and quadratic figure. These techniques are used to envision the future characteristics reliant on the chronicled values. The introduction of the figuring is assessed by execution estimations like precision and audit. Test results on certified datasets show that the proposed figuring is feasible and viable to reveal future time course of action data

Keywords: Temporal Databases, Auto-Regressive (AR) model, Prediction, time series analysis.

1.INTRODUCTION

Information mining is the way toward extricating or "mining" information from a lot of information. It permits clients to examine information from a wide range of measurements or points, sort it, and sum up the connections recognized. Information mining is worried about investigating enormous volumes of unstructured information to find intriguing normalities or connections, which thus lead to better comprehension of the fundamental cycles. Assortment of dataset is a troublesome assignment in information mining measure. Information preprocessing assumes a significant part in information mining task. Information preprocessing which incorporates information determination, trait choice, information cleaning, information joining, information rundown, information change and build a last dataset from a crude set. Information cleaning which includes missing worth assessment [1] and clamor expulsion. Information mining can be utilized in assortment of fields like climate forecast, securities exchange expectation [6], banking, extortion discovery, directed showcasing and logical information investigation. Fleeting information mining is the region of information mining which is characterized as extraction of information or data from the information base as for the time data.

For the instance of transient information mining, these undertakings might be assembled as follows: (I) expectation, (ii) arrangement, (iii) bunching, (iv) search and recovery and (v) design disclosure. Of the five classifications recorded over, the initial four have been explored widely in customary time arrangement investigation and example acknowledgment. Dissimilar to in search and recovery applications, in design revelation there is no particular question close by with which to look through the information base. The goal is basically to uncover all examples of interest. Transient information mining [8] uses fleeting information bases or time arrangement information bases. Transient information bases and time arrangement data sets both store time related information. A transient information base typically stores social information that incorporate time related characteristics. These trait may include a few timestamps, each having distinctive semantics. A period arrangement information base stores grouping of qualities or occasions got from rehashed estimations of time. The errand of time-arrangement expectation has to do with estimating [10] future estimations of the time arrangement dependent on its past examples. To do this, one necessities to construct a prescient model for the information. Essential objectives of information mining are expectation and depiction. Forecast utilizes existing factors in the information base to anticipate obscure or future estimations of interest, and portrayal centers around discovering patterns[11] depicting the information and the

resulting introduction for client understanding. The overall accentuation of both forecast and depiction vary regarding the application and the procedure.

Despite the fact that the term forecast alludes to both numeric and class mark expectation, in this paper we use it to allude fundamentally to numeric expectation. Numeric expectation is the errand of anticipating consistent or requested qualities for given information. Straight line Regression Analysis is a factual approach that is frequently utilized for numeric expectation. Relapse Analysis[12] can be utilized to demonstrate the connection between at least one free or indicator factors and a ward or reaction variable. By and large the estimations of the indicator factors are known. The reaction variable is the thing that we need to foresee.

The remainder of this paper is coordinated as follows. Area II portrays related work, segment III characterizes proposed strategy. Exploratory outcome and exhibitions of the proposed technique are accounted for in area IV and segment V covers end and future work.

2. RELATED WORK

Troyanskaya et al. [1] sum up two attribution techniques, specifically k-Nearest Neighbors ascription (KNNimpute) and Singular Value Decomposition attribution (SVDimpute), where the previous is demonstrated to be beated by the last from the natural perspective. The benefits of KNN ascription are: (I) k-closest neighbor can foresee both subjective characteristics (the most regular incentive among the k closest neighbors) and quantitative traits (the mean among the k closest neighbors). (ii) It doesn't need to make a prescient model for each property with missing information. The disadvantages of KNN ascription are the decision of the separation capacity and it look through all the dataset searching for the most comparable occasions. To defeat this issue ARLSimpute was presented.

Little and Rubin [2] acquaints mean attribution strategy with discover missing qualities. The disadvantages of mean ascription are (I) Sample size is overestimated, (ii) fluctuation is thought little of, (iii) connection is adversely one-sided, and (iv) the circulation of new qualities is an off base portrayal of the populace esteems in light of the fact that the state of the dissemination is contorted by adding values equivalent to the mean. Supplanting all missing records with a solitary worth will collapse the fluctuation and misleadingly blow up the essentialness of any factual tests dependent on it. The Fixed-Rank Approximation Algorithm (FRAA) proposed by Friedland et al. [3] does the assessment of all missing passages in the dataset. The consequences of FRAA will like the mean attribution technique. This strategy will work in all circumstances, however their ascription results are extremely poor. ARLSimpute is utilized to tackle above issues.

The LLSimpute(Local Least Squares Imputation) calculation [4] utilizes the KNN cycle to choose the most corresponded qualities and afterward predicts the missing worth utilizing the least squares plan for the local quality and the non-missing passages. It functions admirably yet the time unpredictability is higher. Due to above disservices in [5] they examined an autoregressive-model-based missing worth assessment technique that considers the dynamic property of microarray transient information and the neighborhood similitude structures in the information. This strategy is particularly powerful for the circumstance where a specific time point contains many missing qualities or where the whole time point is absent.

Zhang et al. [6], present another calculation to be specific DIAL (Dynamic Interdimension Association rules for Local - scale climate forecast) to find likely relations between the exceptional change inclination and the serious climate. Refreshing climate dataset is a troublesome assignment. Past work on anticipating future information predominantly utilize fluffy strategies [7] or information mining procedures [8] to extricate highlights from the preparation informational index and play out the forecast undertakings on continuous arrangement. In any case, to accomplish exact outcomes that are obtuse toward the developing information, these techniques as a rule require preparing the indicators on the genuine information at significant expense.

To anticipate the future time arrangement vales utilizing grouping or preparing the neural organization [10] , anyway cause an exceptionally high update cost for either mining fluffy principles or preparing boundaries in various models. Accordingly, they are not appropriate to effective web based handling in the stream climate, which requires low expectation and preparing costs. To defeat this downsides Xiang et al [12] proposed three methodologies in particular polynomial, Discrete Fourier Transform (DFT) and probabilistic, to foresee the obscure qualities that have not shown up at the framework and answer similitude questions dependent on the anticipated information. They likewise applied proficient lists, that is, a multidimensional hash list and a B+-tree, to encourage the expectation and closeness search on future time arrangement, individually.

3. PROPOSED METHOD

This segment investigates missing worth assessment and the strategy to anticipate time arrangement information. It centers around autoregressive-model-based missing worth assessment strategy (ARLSimpute) which is successful for the circumstance where a specific time point contains many missing qualities or where the whole time point is absent. Information pre-handling yield is given to the contribution of the expectation strategies in particular direct forecast and quadratic forecast. These procedures are utilized to foresee the future qualities dependent on the authentic qualities. The flowchart of the proposed technique is appeared in figure 1

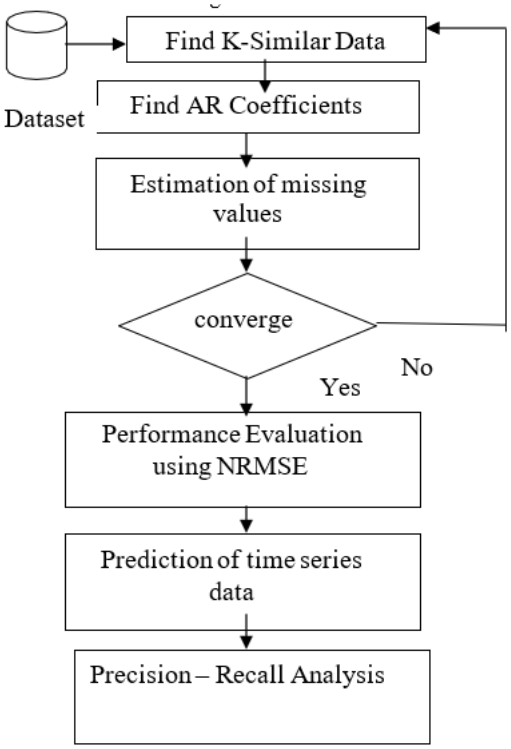

Fig .1 Flowchart of the Proposed System

3.1 Estimation of AR Coefficients

In the info dataset all the missing qualities are instated by setting them to zero. The AR model in framework structure can be portrayed as follows an AR model of request p, aj is the AR Coefficients and ‹j is the is a clamor grouping that we accept to be typically dispersed, with zero mean. The recipe can be modified as The forward–in reverse straight expectation technique is utilized rather than forward or in reverse forecast simply because this calculation builds the quantity of conditions to decide the coefficients. We accept that the emphatically related figures have similar AR coefficients. Connection between's information can be found by utilizing the measure lift. In the event that lift esteem is equivalent to one there is no relationship between's information. In the event that the worth is more noteworthy than one the information are emphatically connected else it is adversely related. This strategy has been demonstrated to be powerful in improving the precision of the assessed recurrence. With the blend of k reports, we attempt to discover the mutually displayed AR coefficients utilizing a most un-square arrangement dependent on SVD [4]. In this manner one can improve the solidness and exactness of the assessed AR coefficients [5] by focusing the little solitary qualities. The co-communicated information's are distinguished dependent on Euclidean separation, which has been demonstrated to beat other likeness measures. yj = Yj aj + ‹j (3.1)
Let S = {y1, . . . , yt, . . . , yn } be a stationary time series

3.2 Estimation of Missing Data

Let us assume that (y1,y2,….ys) are the observed data and
{x1, . . . , xm} are the missing data. Estimation of missing data in matrix form is given by
e=Az (3.3)
where z is a column vector that consists of the observed data y and the missing data x, and A is a Toeplitz matrix whose column number is n and row number is n – p . Matrix A can be Written as
If we separate the observed data from missing data and split A in the block matrix , the equation can be written as
e=Bx+Cy (3.5)
whereB = [B1 , B2 · · ·] and C = [C1, C2 · · ·] are block sub matrices of A corresponding to the respective locations of observed data y and missing data x. Finally the missing data can be calculated from B# (pseudo inverse of B). The corresponding equation is given by
x= -B# Cy (3.6)

3.3 Performance Measure of Missing value estimation

Normalized RMS Error (NRMSE) is used to measure the performance of missing value estimation method, it can be calculated as

$$\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}[E(i,j)-Y(i,j)]^2}$$

where Y is the true value, E is the estimated value, and m and n are the total number of rows and columns, respectively.

A. Prediction Techniques

Prediction technique uses H historical values(x1, . . . , xH)to predict Ot consecutive values in the future. Without loss of generality, let x1 be the value at time stamp 1, x2 at time stamp 2, and so on. In linear prediction, which assumes that all the H+Ot values can be approximated by a single line in the form

x=a.t + b (3.8)

where t is the time stamp, x is the estimated value, and parameters a and b characterize these H+Ot data. Where a and b is given by

a = 12 . i / H(H+1)(H-1)(3.9)

b = 6 . i / H(1-H) (3.10)

$$\sum_{i=1}^{H}(i-(H+1)/2).x$$
$$\sum_{i=1}^{H}(i-(2H+1)/3).x$$

Similarly, for the quadratic prediction, we approximate the values by a quadratic curve in the form

x= a.t2+b.t+c (3.11)

where a, b, and c are parameters that characterize the data.

The task of time-series prediction has to do with forecasting (typically) future values of the time series based on its past

4. EXPERIMENTAL RESULT

Missing qualities are assessed for stock dataset, UK measurements dataset, deals dataset and climate dataset utilizing Auto Regressive (AR) Model. The example dataset and the yield of AR model when request = 3 and request =4 are appeared in figure 2- 4. This calculation is utilized where the circumstance where a specific segment contains many missing qualities, and in any event, when esteems in a whole segment are absent. This attribution technique considers the dynamic conduct of the microarray time arrangement information where every perception may rely upon earlier ones.

2010 Apr	117.4	110.7	94.3
2010 May	118.0	109.1	94.8
2010 Jun	118.7	109.2	95.2
2010 Jul	118.1	109.9	94.7
2010 Aug	117.7	109.8	93.6
2010 Sep	118.4	111.2	93.1
2010 Oct	119.1	112.0	92.3
2010 Nov	120.3	112.4	91.2
2010 Dec	119.3	109.3	89.2
2011 Jan	120.4	112.5	92.2

EAQW RSI:Predominantly food stores (val sa):All Business Index
Seasonally adjusted
2006 = 100
Industry: 52.11/52.2
Updated on 16/ 2/2011
EARA RSI:Textiles, clothing & footwear (val sa):All Business Index
Seasonally adjusted
2006 = 100
Industry: 52.41_3

Fig. 2 Sample Dataset

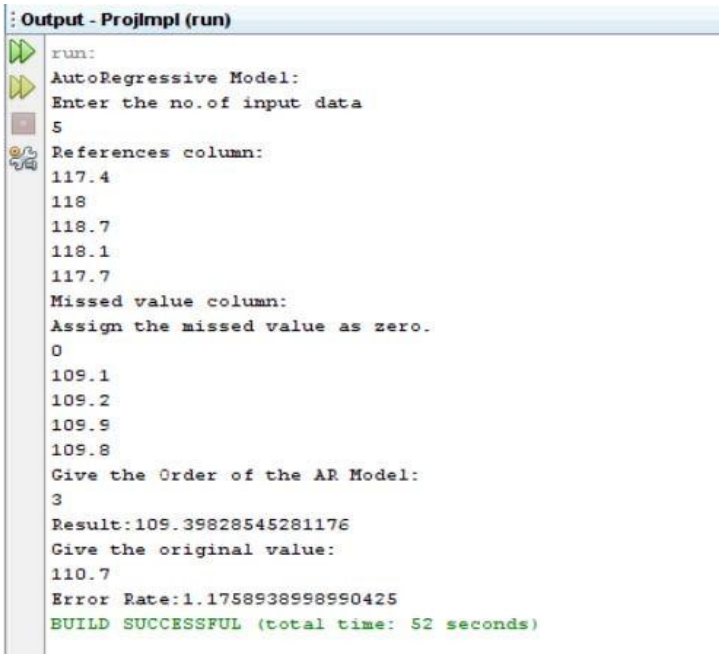

Fig .3 Output of AR Model when order =3

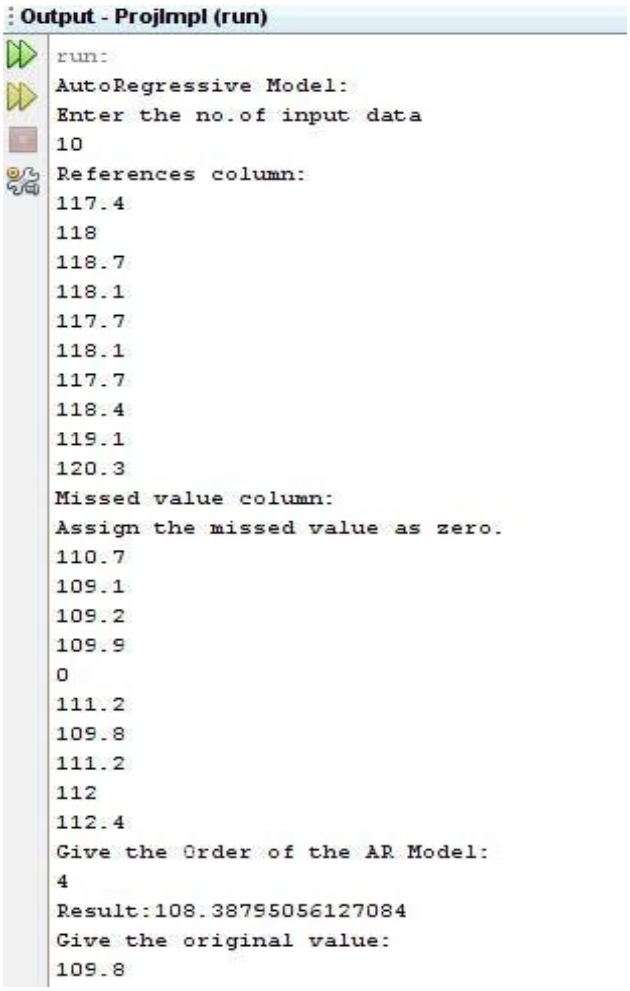

Fig .4 Output of AR Model when order =4

The performance of the AR model is measured by Normalized RMS(Root Mean Square) Error (NRMSE). This was shown in figure 5.Error rate is increased if we increase the order p. Comparison of NRMSE result when order =3 and order =4 are shown in table 1 and 2.

Fig. 5 NRMSE output

FOR AR MODEL WHEN ORDER=3

Techniques	Missing values in %		
	10%	15%	20%
ARL Simpute	0.0791	0.0983	0.1697
KNN Impute	0.2264	0.2983	0.3283

TABLE 2
NRMSE RESULT FOR AR MODEL WHEN ORDER=4

Techniques	Missing values in %		
	10%	15%	20%
ARL Simpute	0.1103	0.1514	0.2513
KNN Impute	0.2424	0.3010	0.3367

NRMSE result of ARLSimpute is compared with KNNimpute when order=3 and order =4. This was shown in figure 6 and 7. The graph shows that the error rate of KNNimpute is higher than ARLSimpute and the error rate is increased if we increase the order p.

5. CONCLUSION AND FUTURE WORK

We have created Autoregressive strategy to discover missing qualities and expectation strategies to gauge the future qualities dependent on chronicled values. Autoregressive strategy is utilized where the circumstance where a specific segment contains many missing qualities, and in any event, when esteems in a whole segment are missing and it is found to shows serious outcomes when contrasted with different procedures, for example, KNN attribute technique, Row normal technique, and mean ascription strategy. Trial results on genuine datasets exhibit that the proposed calculation is compelling and productive to uncover future time arrangement information. Our future work targets finding the missing qualities for multiple segments of missing information and anticipating the future qualities utilizing the calculations like Probabilistic , Group Probabilistic and Prediction utilizing Fuzzy rationale and hereditary calculations.

REFERENCES

1. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics, vol. 17, pp. 520–525,2001.
2. Little, R. J. and Rubin, D.B., Statistical Analysis with Missing Data, Second Edition. John Wiley and Sons, New York. 2002
3. S. Friedland, A. Niknejad, and L. Chihara, A simultaneous reconstruction of missing data in DNA microarrays, Linear Algebra Appl., vol. 416, pp. 8–28, 2006.

4. Y.Tao , , D. Papadias, and X. Lian, Reverse KNN Search in Arbitrary Dimensionality, Proc. 30th Int'l Conf. Very Large Data Bases (VLDB '04), 2004.
5. Miew Keen Choong, Maurice Charbit, and Hong Yan, Autoregressive-Model-Based Missing Value Estimation for DNA Microarray Time Series Data, IEEE Transactions On Information Technology In Biomedicine, VOL. 13, NO. 1, page no 131-138, JANUARY 2009.
6. Zhang, Weili Wu and Huang, Mining Dynamic Interdimension Association Rules for Local -scale Weather Prediction, Proceedings of the 28th Annual International Computer Software and Applications Conference, pp.200-204, 2004..
7. Vilalta and S. Ma, Predicting Rare Events in Temporal Domains, Proc. Int'l Conf. Data Mining (ICDM '02), 2002.
8. Abdullah Uz Tansel et al,Temporal Databases-Theory, Design and Implementation, Benjamin/Cummings publications, 1993.
9. Y. Tao, D. Papadias, X. Lian, and X. Xiao, Multidimensional Reverse kNN Search, VLDB J., 2005
10. S.Policker and A. Geva, A New Algorithm for Time Series Prediction by Temporal Fuzzy Clustering, Proc. 15th Int'l Conf.Pattern Recognition (ICPR '00), 2000.
11. N.Jovonovic et al, Foundations of Predictive Data Mining,IEEE Transactions on Knowledge and Data Engg, Vol. 1, No. 2,pp.439-450, July 2002.
12. Xiang Lian, Lei Chen, Efficient Similarity Search over Future Stream Time Series, IEEE Transactions On Knowledge And Data Engineering, VOL. 20, NO. 1,pp- 40-55, JANUARY 2008.