Imputation for the Analysis of Missing Values and Prediction of Time Series Data

S.Sridevi[#1], Dr.S.Rajaram[*2], C.Parthiban[&3], S.SibiArasan[&4], C.Swadhikar[&5]

[#]*Assistant Professor, Department of CSE*
[*]*Associate Professor, Department of ECE*
[&]*Student, Department of CSE*
Thiagarajar College of Engineering, Madurai
Tamilnadu, India
[1]sridevi@tce.edu
[2]rajaram_siva@tce.edu
[3]swads117@gmail.com

***Abstract*— Data preprocessing plays an important and critical role in the data mining process. Data preprocessing is required in order to improve the efficiency of an algorithm. This paper focuses on missing value estimation and prediction of time series data based on the historical values. A number of algorithms have been developed to solve this problem, but they have several limitations. Most existing algorithms like KNNimpute (K-Nearest Neighbours imputation), BPCA (Bayesian Principal Component Analysis) and SVDimpute (Singular Value Decomposition imputation) are not able to deal with the situation where a particular time point (column) of the data is missing entirely. This paper focuses on autoregressive-model-based missing value estimation method (ARLSimpute) which is effective for the situation where a particular time point contains many missing values or where the entire time point is missing. Data preprocessing output is given to the input of the prediction techniques namely linear prediction and quadratic prediction. These techniques are used to predict the future values based on the historical values. The performance of the algorithm is measured by performance metrics like precision and recall. Experimental results on real-life datasets demonstrate that the proposed algorithm is effective and efficient to reveal future time series data.**

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

I. INTRODUCTION

Data mining is the process of extracting or “mining” knowledge from large amounts of data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Data mining is concerned with analyzing large volumes of unstructured data to discover interesting regularities or relationships, which in turn lead to better understanding of the underlying processes. Collection of dataset is a difficult task in data mining process. Data preprocessing plays an important role in data mining task. Data preprocessing which includes data selection, attribute selection, data cleaning, data integration, data summarization, data transformation and construct a final dataset from a raw set. Data cleaning which involves missing value estimation [1] and noise removal. Data mining can be used in variety of fields like weather prediction, stock market prediction [6], banking, fraud detection, targeted marketing and scientific data analysis. Temporal data mining is the subdivision of data mining which is defined as extraction of knowledge or information from the data base with respect to the time information.

For the case of temporal data mining, these tasks may be grouped as follows: (i) prediction, (ii) classification, (iii) clustering, (iv) search & retrieval and (v) pattern discovery. Of the five categories listed above, the first four have been investigated extensively in traditional time series analysis and pattern recognition. Unlike in search and retrieval applications, in pattern discovery there is no specific query in hand with which to search the database. The objective is simply to unearth all patterns of interest. Temporal data mining [8] utilizes temporal databases or time series databases. Temporal databases and time series databases both store time related data. A temporal database usually stores relational data that include time related attributes. These attribute may involve several timestamps, each having different semantics. A time series database stores sequence of values or events obtained from repeated measurements of time. The task of time-series prediction has to do with forecasting [10] future values of the time series based on its past samples. In order to do this, one needs to build a predictive model for the data. Primary goals of data mining are prediction and description. Prediction makes use of existing variables in the database to predict unknown or future values of interest, and description focuses on finding patterns[11] describing the data and the subsequent presentation for user interpretation. The relative emphasis of both prediction and description differ with respect to the application and the technique.

Although the term prediction refers to both numeric and class label prediction, in this paper we use it to refer primarily to numeric prediction. Numeric prediction is the task of predicting continuous or ordered values for given input. Straight line Regression Analysis is a statistical methodology that is most often used for numeric prediction. Regression Analysis[12] can be used to model the relationship between one or more independent or predictor variables and a

978-1-4577-0590-8/11/$26.00 ©2011 IEEE

dependent or response variable. In general the values of the predictor variables are known. The response variable is what we want to predict.

The rest of this paper is organized as follows. Section II describes related work, section III defines proposed method. Experimental result and performances of the proposed method are reported in section IV and section V covers conclusion and future work.

II. RELATED WORK

Troyanskaya *et al.* [1] summarize two imputation methods, namely *k*-Nearest Neighbours imputation (KNNimpute) and Singular Value Decomposition imputation (SVDimpute), where the former is shown to be outperformed by the latter from the biological viewpoint. The advantages of KNN imputation are: (i) k-nearest neighbor can predict both qualitative attributes (the most frequent value among the k nearest neighbors) and quantitative attributes (the mean among the k nearest neighbors). (ii) It does not require to create a predictive model for each attribute with missing data. The drawbacks of KNN imputation are the choice of the distance function and it searches through all the dataset looking for the most similar instances. In order to overcome this problem ARLSimpute was introduced.

Little and Rubin [2] introduces mean imputation method to find out missing values. The drawbacks of mean imputation are (i) Sample size is overestimated, (ii) variance is underestimated, (iii) correlation is negatively biased, and (iv) the distribution of new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean. Replacing all missing records with a single value will deflate the variance and artificially inflate the significance of any statistical tests based on it. The Fixed-Rank Approximation Algorithm (FRAA) proposed by Friedland *et al.* [3] carries out the estimation of all missing entries in the dataset. The results of FRAA will similar to the mean imputation method. This method will work in all situations, but their imputation results are very poor. ARLSimpute is used to solve above problems.

The LLSimpute(Local Least Squares Imputation) algorithm [4] uses the KNN process to select the most correlated genes and then predicts the missing value using the least squares formulation for the neighborhood gene and the non-missing entries. It works well but the time complexity is higher. Due to above disadvantages in [5] they discussed an autoregressive-model-based missing value estimation method that takes into account the dynamic property of microarray temporal data and the local similarity structures in the data. This method is especially effective for the situation where a particular time point contains many missing values or where the entire time point is missing.

Zhang et al. [6], introduce a new algorithm namely DIAL (Dynamic Interdimension Association rules for Local -scale weather prediction) to discover potential relations between the special change tendency and the severe weather. Updating weather dataset is a difficult task. Previous work on predicting future data mainly use fuzzy methods [7] or data mining techniques [8] to extract features from the training data set and perform the prediction tasks on real-time series. However, in order to achieve accurate results that are insensitive to the evolving data, these methods usually require training the predictors on the actual data at high cost.

To predict the future time series vales using clustering or training the neural network [10] , however incur a very high update cost for either mining fuzzy rules or training parameters in different models. Therefore, they are not applicable to efficient online processing in the stream environment, which requires low prediction and training costs. To overcome this drawbacks Xiang et al [12] proposed three approaches namely polynomial, Discrete Fourier Transform (DFT) and probabilistic, to predict the unknown values that have not arrived at the system and answer similarity queries based on the predicted data. They also applied efficient indexes, that is, a multidimensional hash index and a B+-tree, to facilitate the prediction and similarity search on future time series, respectively.

III. PROPOSED METHOD

This section explores missing value estimation and the method to predict time series data. It focuses on autoregressive-model-based missing value estimation method (ARLSimpute) which is effective for the situation where a particular time point contains many missing values or where the entire time point is missing. Data pre-processing output is given to the input of the prediction techniques namely linear prediction and quadratic prediction. These techniques are used to predict the future values based on the historical values. The flowchart of the proposed method is shown in figure 1.

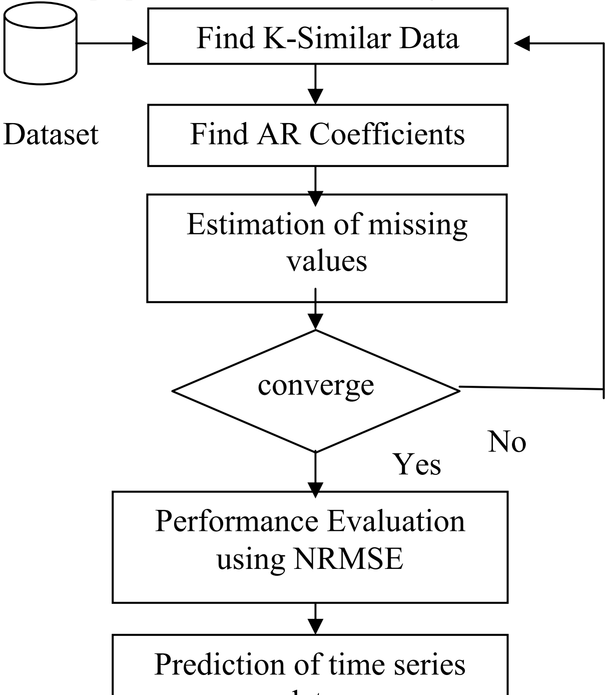

Dataset
Find K-Similar Data
Find AR Coefficients
Estimation of missing values
converge
No
Yes
Performance Evaluation using NRMSE
Prediction of time series data
Precision – Recall Analysis

Fig .1 Flowchart of the Proposed System

A. Estimation of AR Coefficients

In the input dataset all the missing values are initialized by setting them to zero. The AR model in matrix form can be described as

$$y_j = Y_j\, a_j + \varepsilon_j \qquad (3.1)$$

Let $S = \{y_1, . . . , y_t, . . . , y_n\}$ be a stationary time series that follows an AR model of order p, a_j is the AR Coefficients and ε_j is the is a noise sequence that we assume to be normally distributed, with zero mean. The formula can be rewritten as

$$\begin{pmatrix} y[p+1] \\ y[p+2] \\ \cdots \\ y[n] \\ y[1] \\ y[2] \\ \cdots \\ y[n-p] \end{pmatrix} = \begin{pmatrix} y[p] & y[p-1] & \cdots & y[1] \\ y[p+1] & y[p] & \cdots & y[2] \\ \cdots & & & \\ \cdots & & & \\ y[n-1] & y[n-2] & \cdots & y[n-p] \\ y[2] & y[3] & \cdots & y[p+1] \\ \cdots & \cdots & & \\ y[n-p+1] & y[n-p+2] & \cdots & y[n] \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ .. \\ .. \\ .. \\ a_p \end{pmatrix} + \varepsilon_j \qquad (3.2)$$

The forward–backward linear prediction method is used instead of forward or backward prediction only because this algorithm increases the number of equations to determine the coefficients. We assume that the strongly correlated datas have the same AR coefficients. Correlation between data can be found by using the measure lift. If lift value is equal to one there is no correlation between data. If the value is greater than one the data are strongly correlated else it is negatively correlated. This method has been proven to be effective in improving the accuracy of the estimated frequency. With the combination of *k* datas, we try to find the jointly modeled AR coefficients using a least-square solution based on SVD [4]. In this way one can improve the stability and accuracy of the estimated AR coefficients [5] by zeroing the small singular values. The co-expressed data's are identified based on Euclidean distance, which has been proven to outperform other similarity measures.

B. Estimation of Missing Data

Let us assume that $(y_1,y_2,….y_s)$ are the observed data and $\{x_1, . . . , x_m\}$ are the missing data. Estimation of missing data in matrix form is given by

$$e=Az \qquad (3.3)$$

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

$$A = \begin{pmatrix} -a_p & \cdots & -a_1 & 1 & 0 & \cdots & 0 \\ 0 & -a_p & \cdots & -a_1 & 1 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & -a_p & \cdots & -a_1 & 1 \end{pmatrix} \qquad (3.4)$$

If we separate the observed data from missing data and split A in the block matrix , the equation can be written as

$$e=Bx+Cy \qquad (3.5)$$

whereB = $[B_1 , B_2 \cdots]$ and C = $[C_1, C_2 \cdots]$ 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

$$x= -B^{\#}\, Cy \qquad (3.6)$$

C. 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

$$\text{NRMS} = \sqrt{\frac{\sum_{i=1}^{m} \sum_{j=1}^{n} [\, E(i,j) - Y(i,j)]^2}{\sum_{i=1}^{m} \sum_{j=1}^{n} [\, Y(i,j)]^2}} \qquad (3.7)$$

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

D. Prediction Techniques

Prediction technique uses H historical values$(x_1, . . . , x_H)$to predict Δt consecutive values in the future. Without loss of generality, let x_1 be the value at time stamp 1, x_2 at time stamp 2, and so on. In linear prediction, which assumes that all the H+Δt values can be approximated by a single line in the form

$$x=a.t + b \qquad (3.8)$$

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

$$a = 12 \,.\, \sum_{i=1}^{H} (i - (H+1)/2).x_i \,/\, H(H+1)(H-1) \qquad (3.9)$$

$$b = 6 \,.\, \sum_{i=1}^{H} (i - (2H+1)/3).x_i \,/\, H(1-H) \qquad (3.10)$$

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

$$x= a.t^2+b.t+c \qquad (3.11)$$

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 samples. In order to do this, one needs to build a predictive model .when the approximating curve (either linear or quadratic) intersects with the lower/upper bound min/max, the predicted values after this time stamp would become meaningless. Furthermore, if the number of predicted values that are meaningful is not greater than Δt, curves of higher orders may have to be used.

E. Performance Measure of Prediction Techniques

Prediction techniques can be measured by two parameters namely precision and recall. To calculate Precision and Recall TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative), P (No of Positives), N (No of Negatives) values are taken into account. Precision, Recall, TP rate, FP rate can be calculated by

Precision=TP/ (TP+FP) (3.12)

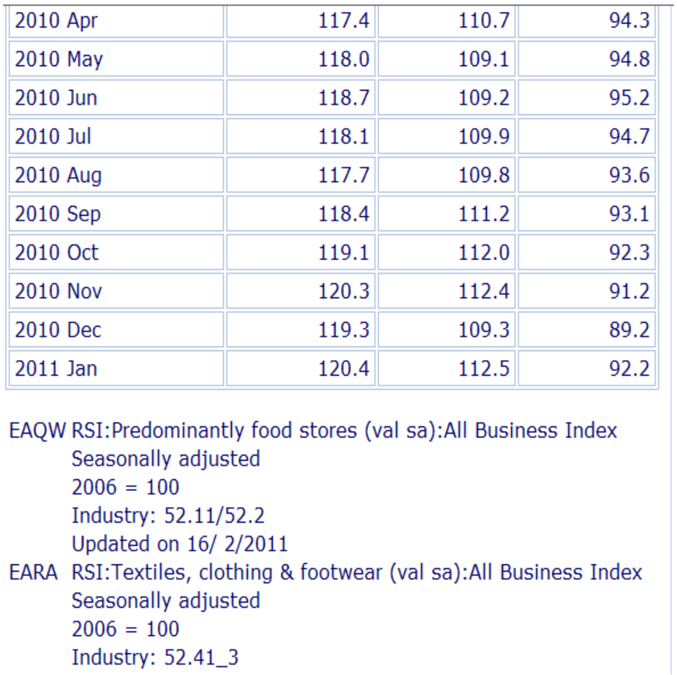
Recall= TP/ (TP+FN) (3.13)

TP rate=TP/P (3.14)

FP rate =FP/ N (3.15)

IV. EXPERIMENTAL RESULTS

Missing values are estimated for stock dataset, UK statistics dataset, sales dataset and weather dataset using Auto Regressive (AR) Model. The sample dataset and the output of AR model when order = 3 and order =4 are shown in figure 2-4. This algorithm is used where the situation where a particular column contains many missing values, and even when values in an entire column are missing. This imputation method takes into account the dynamic behaviour of the microarray time series data where each observation may depend on prior ones.

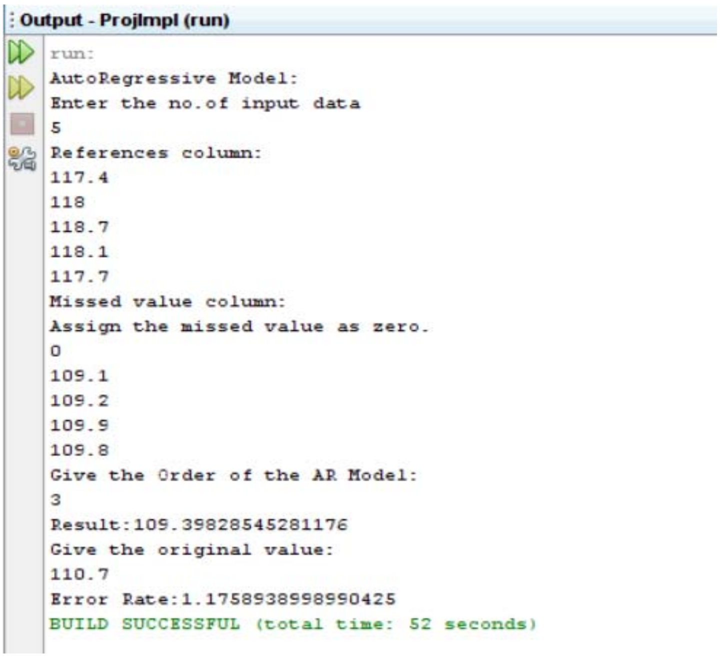


2010 Apr	117.4	110.7	94.3
2010 May	118.0	109.1	94.8
2010 Jun	118.7	109.2	95.2
2010 Jul	118.1	109.9	94.7
2010 Aug	117.7	109.8	93.6
2010 Sep	118.4	111.2	93.1
2010 Oct	119.1	112.0	92.3
2010 Nov	120.3	112.4	91.2
2010 Dec	119.3	109.3	89.2
2011 Jan	120.4	112.5	92.2

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

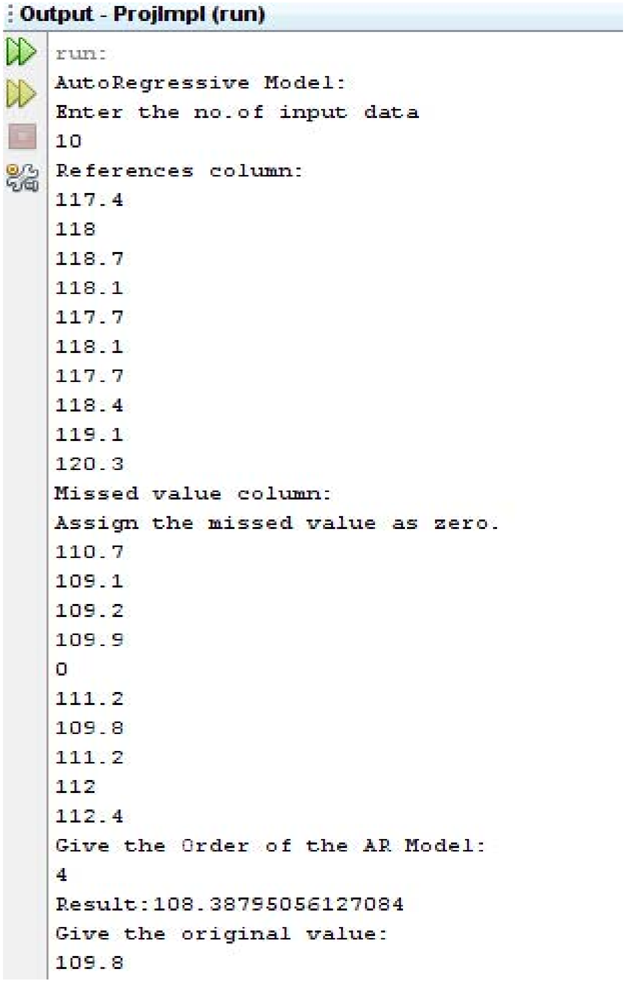
Fig. 2 Sample Dataset



Output - ProjImpl (run)

```
run:
AutoRegressive Model:
Enter the no.of input data
5
References column:
117.4
118
118.7
118.1
117.7
Missed value column:
Assign the missed value as zero.
0
109.1
109.2
109.9
109.8
Give the Order of the AR Model:
3
Result:109.39828545281176
Give the original value:
110.7
Error Rate:1.1758938998990425
BUILD SUCCESSFUL (total time: 52 seconds)
```

Fig .3 Output of AR Model when order =3

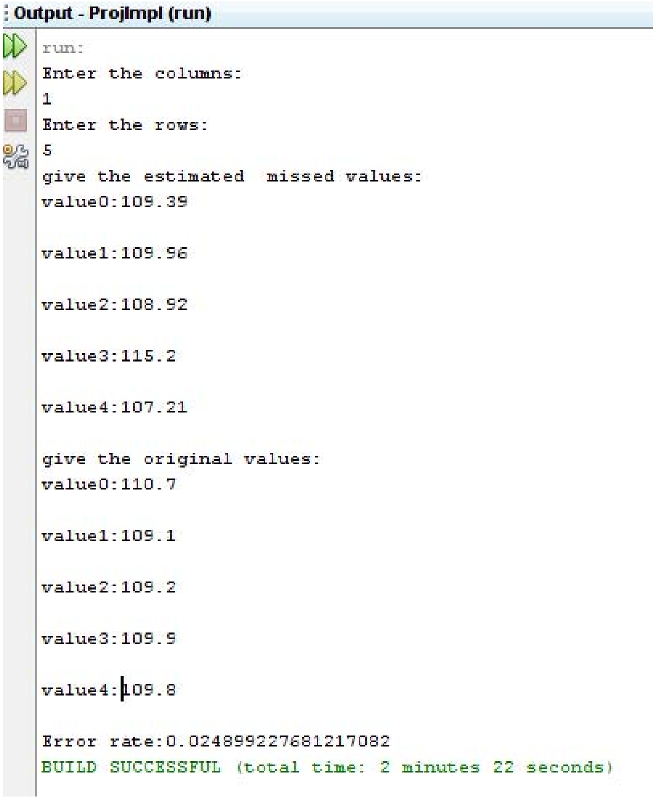


Output - ProjImpl (run)

```
run:
AutoRegressive Model:
Enter the no.of input data
10
References column:
117.4
118
118.7
118.1
117.7
118.1
117.7
118.4
119.1
120.3
Missed value column:
Assign the missed value as zero.
110.7
109.1
109.2
109.9
0
111.2
109.8
111.2
112
112.4
Give the Order of the AR Model:
4
Result:108.38795056127084
Give the original value:
109.8
```

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.



```
run:
Enter the columns:
1
Enter the rows:
5
give the estimated  missed values:
value0:109.39

value1:109.96

value2:108.92

value3:115.2

value4:107.21

give the original values:
value0:110.7

value1:109.1

value2:109.2

value3:109.9

value4:109.8

Error rate:0.024899227681217082
BUILD SUCCESSFUL (total time: 2 minutes 22 seconds)
```

Fig. 5 NRMSE output

TABLE I
NRMSE RESULT 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.

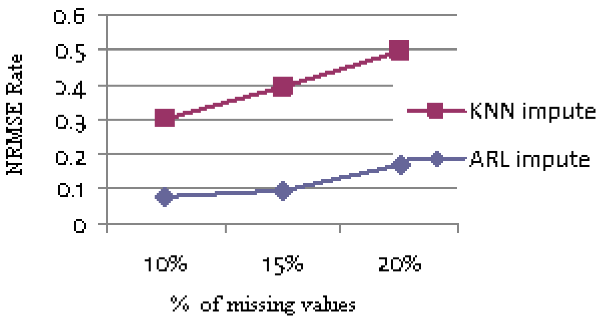


Fig .6 NRMSE Comparison of KNN impute and ARLSimpute when order=3

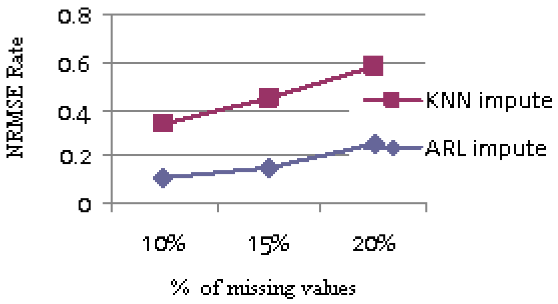
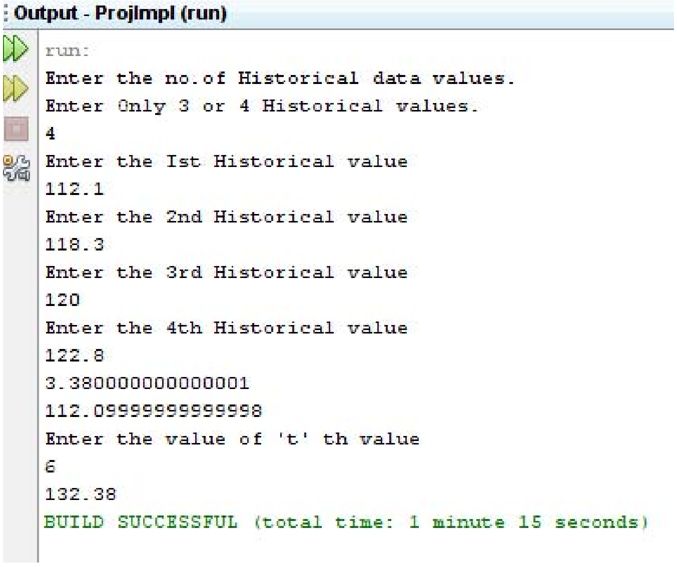


Fig .7 NRMSE Comparison of KNN impute and ARLSimpute when order=3

Data pre-processing output is given to the input of the prediction techniques namely linear prediction and quadratic prediction. These techniques are used to predict the future values based on the historical values. The output of linear prediction was shown in figure 8.



```
run:
Enter the no.of Historical data values.
Enter Only 3 or 4 Historical values.
4
Enter the Ist Historical value
112.1
Enter the 2nd Historical value
118.3
Enter the 3rd Historical value
120
Enter the 4th Historical value
122.8
3.380000000000001
112.09999999999998
Enter the value of 't' th value
6
132.38
BUILD SUCCESSFUL (total time: 1 minute 15 seconds)
```

Fig .8 Prediction Output

The performance of the prediction was measured by Precision-Recall curve and ROC (Receiver Operating Characteristics) curves which are shown in Fig 9 and 10. Based on these curves, accuracy of the proposed algorithm can be measured. Precision- Recall curve can be obtained by plotting recall in x axis and precision in y axis. Fig 9 shows that recall value is decreased with respect to increasing value of precision. ROC curve can be obtained by plotting FP rate in x axis and TP rate in y axis. Fig 10 shows that FP rate is increased with respect to increasing value of TP rate.

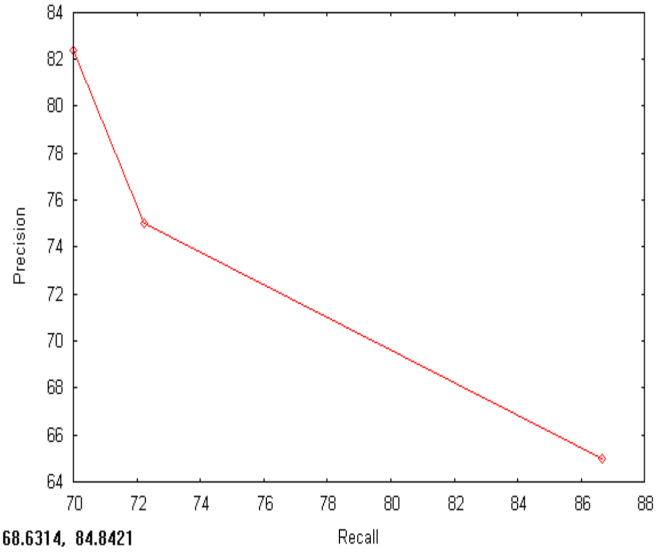


Fig . 9 Precision-Recall Curve

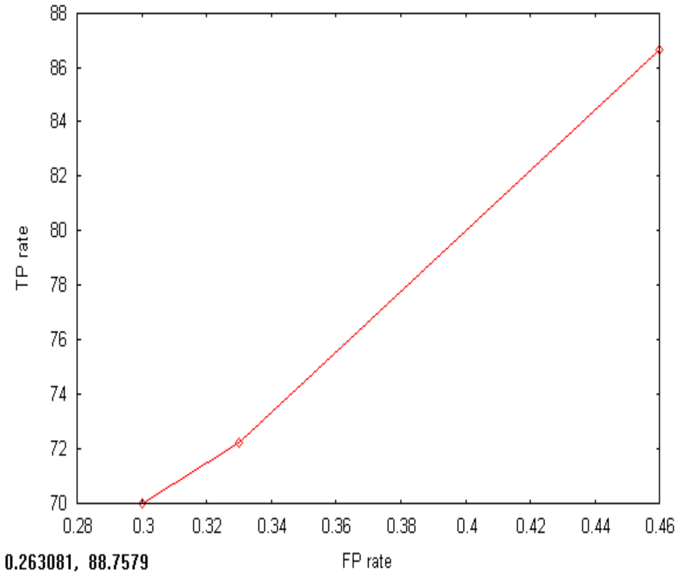


Fig . 10 ROC Curve

V. CONCLUSION AND FUTURE WORK

We have developed Autoregressive method to find out missing values and prediction techniques to estimate the future values based on historical values. Autoregressive method is used where the situation where a particular column contains many missing values, and even when values in an entire column are missing and it is found to shows competitive results when compared to other techniques such as KNN impute method, Row average method, and mean imputation method. Experimental results on real-life datasets demonstrate that the proposed algorithm is effective and efficient to reveal future time series data. Our future work aims at finding the missing values for more than two columns of missing data and predicting the future values using the algorithms like Probabilistic , Group Probabilistic and Prediction using Fuzzy logic and genetic algorithms.

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.