Research Journal of Applied Sciences 11 (10): 1089-1094, 2016
ISSN: 1815-932X

A New Polynomial Curve Fiting Based on Segmentation of Variable Point and Variable Modes for Reconstructing Missing Values

Nabeel H. Al-A'araji, Eman S. Al-Shamery and Alyaa Abdul Hussein
Faculty of Information Technology, University of Babylon, Babylon, Hillah, Iraq

Abstract: The lost values are a common problem in many pivotal real application like data-mining machine learning and pattern detection algorithms. In now days ,there are huge streams of information which contain a missing values for many reasons for instance a malfunction in a piece of equipment, a tissue section on a slide was not stained properly,a technician forgot to enter a value in a spreadsheet, etc. If they are neglected, they will be loosed as well as, another information related. This research is suggested a new mathematical approach for solving the missing value problem based on a polynomial fitting model. The proposed model (polynomial curve fitting based on segmentation of variable point and variable modes SVPVM algorithm)is applied on PAMAP2 that contain 2872532 records each of them have 54 values. Also system is implemented on samples of dataset after some points are removed. Finally, the proposed approach is executed on known function such as sine wave. The results are compared with another methods for retrieval missing points such as linear and mean methods. In general the proposed model has been produced good results according toevaluation methods such as mean square error. So, the proposed method is expected to give promising results in the field of information loss.

Key words: Detection algorithms, PAMAP2, SVPVM, proposed method, promising results

INTRODUCTION

Missing values are problem in analyzing data for a large group of research areas. when an attribute value is hidden the value is defined as Missing value (Goeij *et al*., 2013). This mean there is no value, but does not known. investigator eliminate missing points of data or fill it with 'reasonable estimate' when analyze datasets with missing values . removal missing values is applied simply but just sensible when the dataset is big enough so that the lost values are not cause clear effect on the analysis. Streams are Considered a time-series or multidimensional processes. In data flow processing, data are processed in real-time at local interim windows, it has been proposed that data value are used at the moment to take the systems' Characteristics. (Enders, 2010; Rahman *et al*., 2014). Streaming data is not had a fixed length Compared to static data. Study (Lemnaru, 2012) is said that he majority of the classifiers recruit easy and ineffective approaches when processing missing data. Such that , all missing values are removed by classifiers, or is replaced with NULL or is considered as especially value. Taking into account the facts before, pretreatment phase is needed to deal with the lost in data in order to reduce the danger of bad classification. Millsap and Olivares, 2009). Filter-based is used by researcher to remove missing data.

Filtering is preferred to improve the work when the amount of missing data is few. The imputation methods are tried to substitute the lost values of features with values which may more appropriate in a certain position, therefor learning process is reinforced. Missing values are replaced in Imputation strategies by emanating available data, for bring out the full data set study (Magnani, 2004) is showed this method.

Hua and Pei (2007) is proposed two methods to preprocess missing value these are list-wise deletion, all cases are removed which is had leastwise one missing point and pairwise deletion, a case points are removed if the missing values are founded in variables that is used currently. In (Bishop, 2006; Allison, 2001) many methods are used to replace the value of each of the missing with the calculated value of one. Mean or mode imputation (MEI), is used as Simpler and more effective way which is filled the missing point by as for feature mean or mode. Nearest Neighbor methods are applied by recognition the more like cases to the one at a lost value based on the observed values at the rest of the variables at this case. They are applied a pre-determined basis (e.g. average weight in (Troyanskaya *et al*., 2001) or seed function (e.g. exponential in (Yu *et al*., 2011) to replace a value based on these instances. In (Zhang *et al*., 2011) methods are considered more precise than MEI but full sufficient cases are required to identify the neighbors and number of neighbors are needed to determine.

Polynomial fitting is used as one of three different data preprocessing techniques proposed by Imad Rahal which performs appropriate more efficiently when in use

Corresponding Author: Nabeel H. Al-A'araji, Faculty of Information Technology, University of Babylon, Babylon, Hillah, Iraq

K-Means and evaluating outcome clusters, this technique is fitted data values in each record attribute with the single polynomial that is passed using a specific values .

Curve fitting functions: Curve fitting, known as regression analysis, has been used "best fit" line or curve for a set of data value. Curve fit is generate an equation that is used to find points during the more than the curve Curve fitting contain the following types:

Linear : This function is fitted a consecutive set of data points , the form y = μ+x. is used . where (μ) is (the value of y when x = 0) , (?) is slop , (X) is Interpretative variable and (Y) is the dependent variable. Data limitation are not associated with this curve fit.

Polynomial: Dataisfitted by a curve throughdata, of the form

$$y = \alpha_0 + \sum_{k=1}^{j} \alpha_k x^k$$

Where (α_0, α_k) polynomial coefficients are selected with minimum error between data and fitted function and (j) is order of polynomial. Greater polynomial order is required for more complicated the curvature of the data, to fit it.

Best polynomial model is chosen frequently based on trial and error. If the curve is followed the direction ofdata, pick a higher order equation. If it oscillate too much, pick a lower order equation. Other types of curve fitting are (exponential, logarithmic and power), (Chaitanya Datta).

MATERIALS AND METHODS

The proposed method: In this study curve fitting technique has been employed for estimating missing value . Polynomial fitting has been exploited in away differ from other research studys by using variable polynomial models rather than unique model for whole data points. Variable segmentation in new approach has been utilized based on threshold with overlapping for more precision. Also the new direction in the proposed system has been defined by variable models for each segment of data. Inverse operation take into account the variable models of the data in reconstruction of missing values.

Missing values has been estimated by fitting all the values in each feature with a polynomial variable by the time and then extended the resulting polynomial to fit the major sized data piece in a process known as polynomial fitting.

The degree of polynomial has been specified based on number of known points that have been determined by threshold. The threshold has been chosen in a way that reduce error.

The general formula of polynomial function has been used with order of (threshold -1). For each segment of data polynomial fitting model has been constructed and missing values have been founded depending on that model , then next segment have been taken and the same thing have been done until all missing value have been founded.

The proposed Algorithm (SVPVM)

Algorithm: Polynomial curve fiting based on segmentation of variable point and variable modes (SVPVM)

Input: The input for proposed algorithm include matrix (Vij) of time series data with missing points , where (i) represent time and (j) represent attribute.
Output: $\bar{V}_{ij}$: Matrix of data without missing points,
C_{pq}: Matrix store coefficient of polynomial model ,where
p :Counter determine the model number
q :Polynomial coefficients p=1……(T-1) where T is threshold.
Step1: Extract two vectors from V_{ij} data matrix

- Vector x_t : where t Time value.
- Vector y_a : where a is attribute value at time t.

Step2: Form vector R_t depend on y_a , if y_a = NaN then R_t=1 otherwise R_t=0.

Step3: Locate the start and end for each missing value by finding known segment(B_{kn})where k represent number of known point (T) and (n=1,2,3,….. 6)represent values that related to known point k
Do
If current point is not NaN (R_t = 1)then store the following in matrix B_{kn} otherwise increment the counter for NaN points N.

$B_{k1}=x_t$ (time value)
$B_{k2}=y_a$ (attribute value)
$B_{k3}=t$ (location of not NaN value)
$B_{k4}=t - N$ (start of NaN)
$B_{k5}=t-1$ (end of NaN)
$B_{k6}=k$ (counter of known point)
k=k+1
N=0;

while k <= Length(R_t)
step 4:
-From B_{kn} matrix select number of points according to threshold (T) with shift and overlapping and do the following:

Do

- Form two vectors from selected points

p1 from (x1 , x2 ,x3, …… xT)
p2 from (y1 , y2 ,y3, …… yT)
-Apply equation of polynomial fitting model on above segment points in p1 and p2 to find the polynomial coefficients of a M of order (T-1) using the following equation:

$$y = \alpha_0 + \sum_{m=1}^{(T-1)} \alpha_m x^m$$

Where (α_0, α_m) are coefficient of polynomial and stored in matrix C_{pq}x represent vector p1, y represent vector p2

-For each missing point between known point (T) find $\bar{y}i$ apply the founded model for the current segment by substitution the time of missed value and replace NaN by the result of applied model.

While k<= number of points in B_{kn}
Step 5: Return $\bar{V}_{ij}$ with new attribute $\bar{y}i$

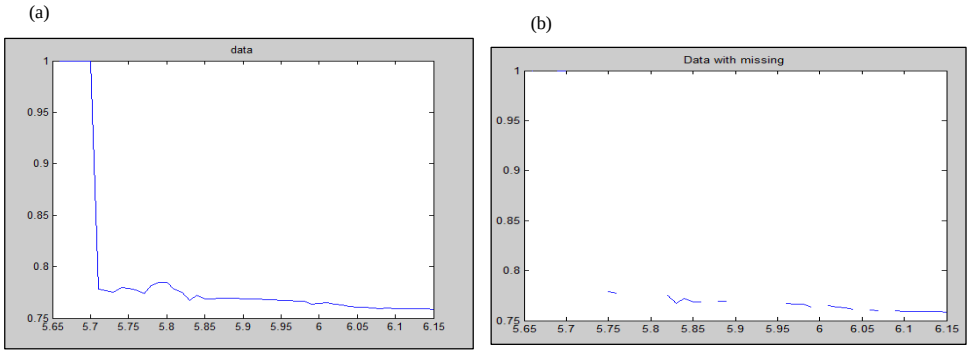

Fig. 1: a) Testing set and b) Training set

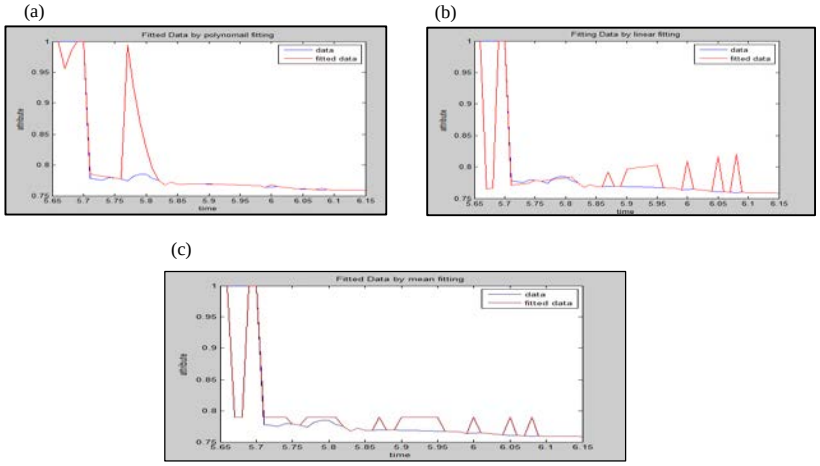

Fig. 2: a) Proposed method with MSE = 0.0016; b) Linear fitting with MSE = 0.0025 and c) Mean with MSE = 0.0019

Data set: Real dataset that has been taken from have been named (PAMAP2) for physical activity have been used in this study .The PAMAP2 dataset contains data of nine healthful persons, three Inertial Measurement Units (IMUs) are worn by each person's and a heart rate is monitored. Each IMU measures accelerometer, temperature . The data have been sampled at 100Hz and are passed to computer via a 2.4GHz wireless network. Person wore one IMU one on the dominant ankle, on the dominant wrist and one on the chest. All nine persons have been labeled in the dataset as subject101 through subject 109. Data contain 2872532 measurements each have 54 attribute values. Values that are missed, stores as a NaN (Not a Number) value .The performed activities are, cycling ,standing, vacuum cleaning, ironing, walking, running, , Nordic walking, lying, sitting ascending stairs, descending stairs and rope jumping. The proposed system results have been checked on this dataset but it's also possible to apply the system on any other numerical dataset for generality.

RESULTS AND DISCUSSION

Experiments and results: The proposed method has been applied on three different samples of dataset and to evaluate the accuracy of the suggested method Mean Square Error (MSE) method have been used.

Two methods have been used in addition to the proposed method to estimate missing valuesthese methods are linear fitting and mean method .

Applied system on sample data without missing points: The data have been divided into two sets first ,for training of the proposed method and the second for test or evaluate the result of the estimation process. The two sets have been shown in Fig. 1. In the training some of data are removed then proposed method has been applied the result has been shown in Fig. 2a ,the result of application of other method such as linear fitting and mean method have been shown in Fig. 2b and c respectively. The MSE also mentioned with each figure.

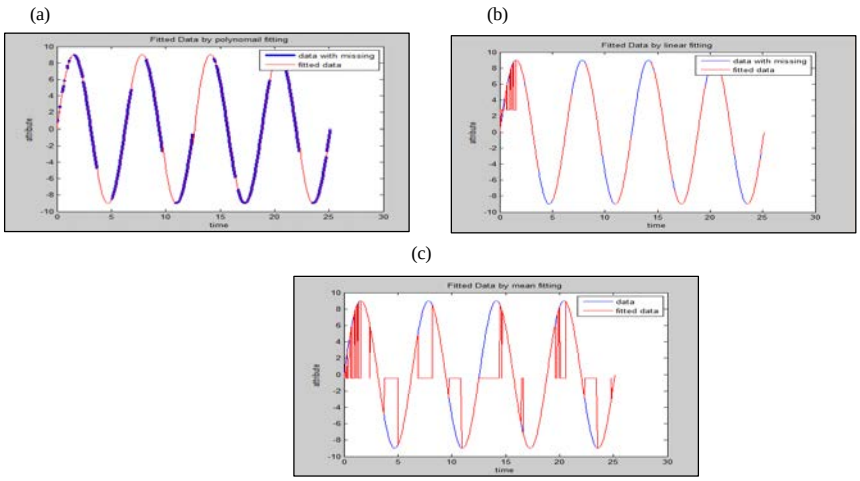

Fig. 3: a) Proposed method with MSE = 0.0197; b) Linear method with MSE = 0.7883 and c) Mean method with MSE = 18.2377

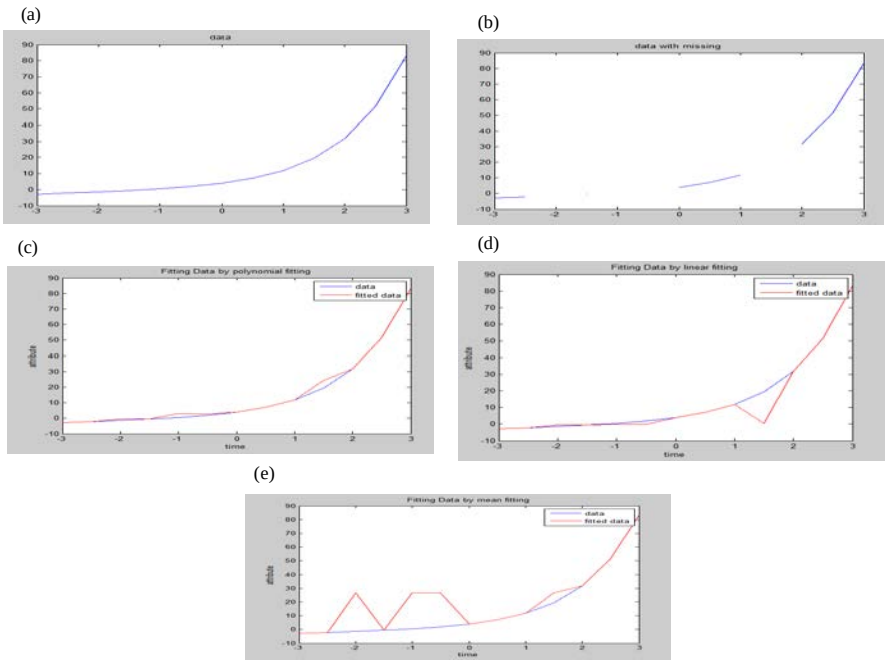

Fig. 4: a) Data points; b) Data with missing; c) Proposed fitting method; d) Linear fitting method and e) Mean fitting method

The application of proposed method , linear and mean method for estimating missing values have been shown in Fig. 2.

Applied system on known signal: the second sample of data that have been used formed by the following equation y = 9sin(x) where x has been taken in interval (0-8p) then some points have been removed and are replaced by (NaN) value the results to found the missing points is shown in the Fig. 3. The comparison for data in (7.1) and (7.2) based on error has been shown in Table (1)

Example with clear values to apply the system: The data have been displayed in Table 2 with simple values ,the suggested method , linear fitting and mean fitting methods have been applied where some of points have been removed and have been replaced with NaN to make missing points that have been estimated by these methods the result have been shown in Fig.4.

Applied system on life data set: The application of proposed method on PAMAP2 has been shown in Fig. 5.

Table 1: Comparison between different method with proposed method

	Proposed method			Linear method			Mean method		
Variables	Error average	Error max	Mean square error	Error Average	Error max	Mean square error	Error Average	Error max	Mean square error
Data 1 in (7.1)	0.000032639	0.0009778	0.0016	0.000050093	0.0011	0.0025	0.000037339	0.00086856	0.0019
Data 2 in (7.2)	0.000024615	0.0024	0.0197	NaN	0.0482	0.7883	-0.4321	9	18.2377

Table 2: Data points, data with missing values and estimated values with different methods and associated error with each method

Points	x	y	y with missing	$\bar{y}$ fit by system	$\bar{y}$ fit by linear	$\bar{y}$ fit by mean
1	-3.0	-2.80	-2.80	-2.80	-2.80	-2.80
2	-2.5	-2.17	-2.17	-2.17	-2.17	-2.17
3	-2.0	-1.46	NaN	-0.59	-0.41	26.69
4	-1.5	-0.61	-0.61	-0.61	-0.61	-0.61
5	-1.0	0.47	NaN	2.89	-0.20	26.69
6	-0.5	1.93	NaN	2.60	-0.10	26.69
7	0.0	4.00	4.00	4.00	4.00	4.00
8	0.5	7.09	7.09	7.09	7.09	7.09
9	1.0	11.87	11.87	11.87	11.87	11.87
10	1.5	19.43	NaN	24.33	0.30	26.69
11	2.0	31.56	31.56	31.56	31.56	31.56
12	2.5	51.23	51.23	51.23	51.23	51.23
13	3.0	83.34	83.34	83.34	83.34	83.34
				MSE= 2.39	MSE= 28.57	MSE= 164.99

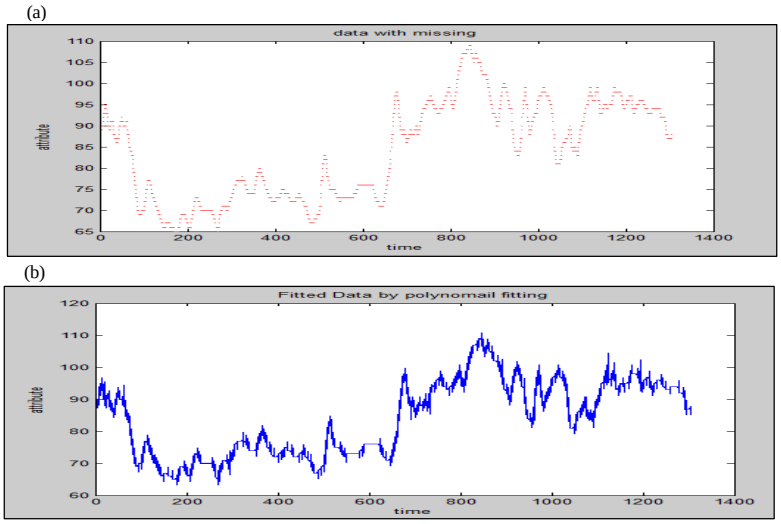

Fig. 5: Proposed method of PAMAP2

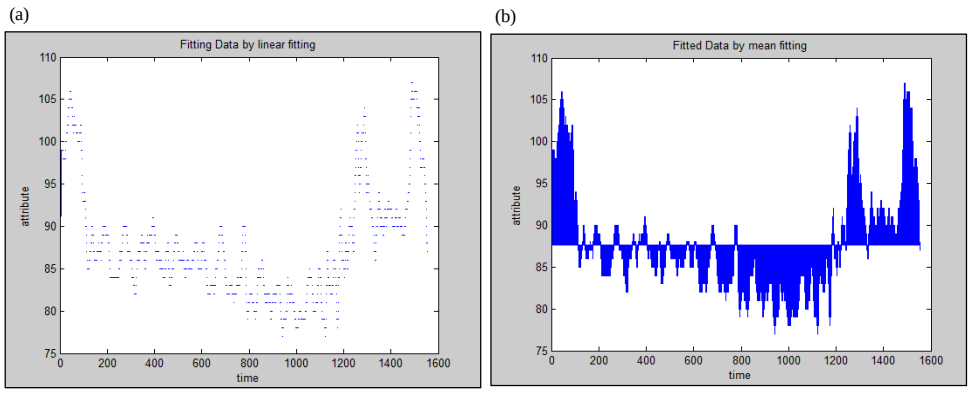

Fig. 6: a) Estimated missing points by linear fitting and b) Estimated missing points by mean method

The result of estimate missing values of data set PAMAP2 by linear fitting and mean method has been shown in Fig. 6.

CONCLUSION

Overall, study of thethe proposed technique of polynomial fitting and additional methods of linear and mean methods. The proposed method is much better from the other methods.

There is no constraints on the type of data (numerical data). The system increase the accuracy in addition overlapping operation . The variable models according to the coefficients is proposed as new approach also support to stream process.

REFERENCES

Allison, P.D., 2001. Missing Data. Sage Publication, Thousand Oaks, California,.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. 1st Edn., Springer, Heidelberg, ISBN: 0-387-31073-8.

Enders, C.K., 2010. Applied Missing Data Analysis. Guilford Press, New York, USA.,.

Goeij, M.C.D., M.V. Diepen, K.J. Jager, G. Tripepi and C. Zoccali *et al*., 2013. Multiple imputation: Dealing with missing data. Nephrology Dialysis Transplantation, 28: 2415-2420.

Hua, M. and J. Pei, 2007. Cleaning disguised missing data: A heuristic approach. Proceedings of the 13th International Conference On ACM SIGKDD Knowledge Discovery and Data Mining, August 12-15, 2007, ACM, New York, USA., ISBN: 978-1-59593-609-7, pp: 950-958.

Lemnaru, C., 2012. Strategies for Dealing with Real World Classification Problems. Ph.D Thesis, Technical University of Cluj-Napoca, Cluj-Napoca, Romania.

Magnani, M., 2004. Techniques for dealing with missing data in knowledge discovery tasks. Department of Computer Science, University of Bologna, Italy, pp: 1-10.

Millsap, R.E. and A.M. Olivares, 2009. The Sage Handbook of Quantitative Methods in Psychology. Sage Publications, Thousand Oaks, California, pp: 72-89.

Rahman, S.A., Y. Huang, J. Claassen and S. Kleinberg, 2014. Imputation of missing values in time series with lagged correlations. Proceedings of the 2014 IEEE International Conference on Data Mining Workshop, December 14-14, 2014, IEEE, New York, USA., ISBN:978-1-4799-4274-9, pp: 753-762.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown and T. Hastie *et al*., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics, 17: 520-525.

Yu, T., H. Peng and W. Sun, 2011. Incorporating nonlinear relationships in microarray missing value imputation. IEEE. ACM. Trans. Comput. Biol. Bioinf., 8: 723-731.

Zhang, S., Z. Jin and X. Zhu, 2011. Missing data imputation by utilizing information within incomplete instances. J. Syst. Software, 84: 452-459.