A Measure Of Similarity Of Time Series Containing Missing Data Using the Mahalanobis Distance

Dinkar Sitaram, Aditya Dalwani, Anish Narang, Madhura Das, Prafullata Auradkar
Dept. of Computer Science and Engineering
PES Institute of Technology
Bangalore, India.
dinkars@pes.edu, prafullatak@pes.edu, das.madhura94@gmail.com

***Abstract*— The analysis of time series data is of interest to many application domains. But this analysis is challenging due to many reasons such as missing data in the series, unstructured nature of the data and errors in the data collection procedure, measuring equipment, etc. The problem of missing data while matching two time series is dealt with either by predicting a value for the missing data using the already collected data, or by completely ignoring the missing values. In this paper, we present an approach where we make use of the characteristics of the Mahalanobis Distance to inherently accommodate the missing values while finding the best match between two time series. Using this approach, we have designed two algorithms which can find the best match for a given query series in a candidate series, without imputing the missing values in the candidate. The initial algorithm finds the best nonwarped match between the candidate and the query time series, while the second algorithm is an extension of the initial algorithm to find the best match in the case of warped data using a Dynamic Time Warping (DTW) like algorithm. Thus, with experimental results we go on to conclude that the proposed warping algorithm is a good method for matching between two time series with warping and missing data.**

Keywords*— *Time Series; Missing Values; Mahalanobis Distance; Similarity Measure

I. INTRODUCTION

A regular time series is one in which values are collected over regular intervals of time. This can often degenerate into an irregular time series, due to issues in the collection mechanism of data itself or due to human errors in data entry and change in collection frequencies of the data. Such data is quite common in applications involving financial time series data such as stocks and future contracts [1], climate monitoring and prediction [2], remote sensing systems [3] and in the data pertaining to several other fields. The usual methods that deal with missing values in time series data either fill up the missing values or ignore them. In the former case, the missing values are predicted by techniques such as Mean/Mode Imputation [4][5][6] or k-Nearest Neighbours [4][7] and so on. Techniques like smoothing that make use of moving averages [8][9] are also commonly used to handle such irregularities in time series data. Other available methods of missing data prediction [10] have considerable overhead in the data preparation. This paper presents an algorithm that handles the data with missing values per se to measure the similarity component between two given time series. This paper is the result of an effort to reduce the data preparation time without compromising the accuracy. The proposed algorithms are designed based on the Mahalanobis Distance measure. These algorithms extend flexibility to the user to control the quality of the matches and also improve overall efficiency by pruning the candidate.

II. RELATED WORK

Analyses dealing with time series data that may contain missing data currently employ certain techniques to handle the missing values. These techniques can be broadly categorized as: i) Case Deletion: The missing point is skipped over completely while using the dataset with missing data. ii) Interpolation: In a geometric sense, a line is drawn between the ends of sequences on either side of the missing data. iii) Imputations: These involve the fixing of a value at the point where it is missing by using certain algorithms such as k-means, imputations of most common values in the series, Mean/Median Imputations and so on. iv) Smoothing: A moving average calculation technique is applied on the data to smooth it into one continuous curve over which known techniques for processing of regular smooth curves may be applied.

The methods mentioned above are those that are most commonly used. Even though many such methods exist to fix missing data, each of them do have certain drawbacks involving one or more of the following: (i) Information from the original data set is lost in the following phases of processing of the time series data. (ii) A new value is assumed to have been in the spot of the missing one which involves a certain amount of error that is difficult to predict as the missing value itself may be difficult to predict. (iii) The original data representation is completely changed into a new one which might not correctly express the same information as the original data.(iv) There is a large computational pre-processing overhead involved in applying the method to handle the anomalies in the data.

The drawbacks of these existing methods strongly suggest a need for better methods to process missing data in time series. The algorithms presented in this paper neither have the overhead of changing the data representation, nor the missing values are completely ignored in the matching process. Previously, a new distance metric was formulated using the Bhattacharyya Distance measure in combination with the k-

978-1-4799-1734-1/15 $31.00 © 2015 IEEE
DOI 10.1109/ICACCE.2015.14

Nearest Neighbours imputation technique [7] to be used on time series data that had missing values. However, since it does perform an imputation for the missing value, we do not compare the technique with the one we propose. There is mention, however, of the Bhattacharya Distance degenerating to the Mahalanobis distance in certain cases which is an affirmation of sorts of our techniques which use this measure. An idea of a threshold is also included which manages the sensitivity of the analysis of the time series [11].

The rest of this paper is organised as follows. In Section III, we define the Mahalanobis Distance. Section IV presents the implementation methodology employed in achieving our goal. Section V presents the Nonwarping algorithm and Section VI presents the Warping algorithm that extends the Nonwarping one. Each of these sections also contains the corresponding visualizations and results for the sample dataset chosen. We go on to provide the conclusion about the proposed algorithm in Section VII and the plans on developing these concepts further in Section VIII. Finally, we list out the literature that was referred to while working on this implementation.

III. The Mahalanobis Distance

Mahalanobis Distance was first used to find similarities between skulls based on measurements [12]. A version of it known as asymmetric Mahalanobis Distance was used to propose a method for handwritten character recognition [13]. It is also widely used in various data mining techniques such as clustering and classification techniques.

The Mahalanobis Distance is a distance measure between a point P and a distribution of points D. By definition, it is the number of standard deviations the point P is away from the mean of the distribution D.

$$\lambda = \left(\sum_{i=1}^{N_D} D_i\right)/N_D \qquad (1)$$

$$\sigma = \sqrt{\sum_{j=1}^{N_D}(D_j - \lambda)^2/N_D} \qquad (2)$$

From Equation (1) and Equation (2), we get the bare bones representation of the Mahalanobis Distance as

$$\gamma = \frac{|P-\lambda|}{\sigma} \qquad (3)$$

The use of standard deviation in this definition is what helps to inherently account for the missing data in the time series under question by capturing the dispersion trends on either side of the missing values when comparing a point in the query to the distribution of a localized section of points from the candidate. This crucial property of the Mahalanobis Distance measure allows us to do away with any sort of processing to explicitly handle missing values in the data. Another advantage of the Mahalanobis Distance is that it is scale invariant and hence no rescaling of data is required.

IV. Implementation Methodology

We develop the algorithm in two phases. In the first phase, we assume an exact match between the candidate with missing points and the query time series. The results are presented in the form of best match and exact match for each algorithm. It is seen that Mahalanobis Distance gives the same results as exact match with a good number of missing points implying that it is a reasonable way to extend matching algorithms for missing values. In the second phase, we extend the algorithm to also cater to warped time series and propose an algorithm that handles warped data using a method similar to the one used in DTW [14].

V. Nonwarping Algorithm

The algorithm takes in three input parameters i.e. the measured data as the candidate time series which may have missing data, a regular time series as the query to be matched against the candidate and a quality parameter that the resultant match provided by the algorithm has to comply with. This quality parameter is a threshold value which is set by the user as a maximum percentage of points in the matching region of candidate that may be missing. If the chosen run of data has more percentage of missing points compared to the query, the whole selection is skipped over and the next set is chosen.

The candidate is chosen with a length equal to the number of query points. It is chosen in the form a sliding window of size equal to the query length through the whole candidate series. The query time series is iterated over and a range of points in the candidate equal to a window size of 10 is chosen which includes any missing points. The Mahalanobis Distance is then calculated between the point and this distribution window using (3). This is then repeated for every point in the query and added. The sum is then compared with a globally maintained distance value and the location of match updated to the current one if the obtained sum is lesser than that of a global value.

The procedure is then continued for various alignment possibilities of the candidate and query and finally the best match location is given to the user. Therefore, in a single pass over the candidate data with missing values, the best match between the candidate and query is obtained.

ALGORITHM I.

```
mahalDist ← ∞
q ← length(query)
for i in 1 to length(candidate)-q+1 do
  if i in range(5) then
      block ← candidate[i:i+q]
  else do
      block ← candidate[i-5:i+q]
  dist_local ← 0
  count_missingpoints ← count(block)
```

```
if((count_missingpoints/q)>threshold)
    go to next block

for j in range(5) then
 temp←mahalanobis(block[0:10],query[j])
 if temp not equal to ∞ then
    dist_local←dist_local + temp
start ← 5
end ← 15

for j in 5 to q do
    temp←mahalanobis(block[start:end],qu
ery[j])
     start++
     end++
     if temp not equal to ∞ then
        dist_local ← dist_local + temp

for j in (q-4) to q do
    temp←mahalanobis(block[start:end],qu
ery[j])
    if temp not equal to ∞ then
          dist_local ←dist_local + temp

if dist_local < mahalDist then
    mahalDist ← dist_local
     location ← i
```

A. *Best match results of Nonwarping Algorithm*

The best match results of both algorithms are presented using a standard EEG trace corpus recorded in the ECT Lab at Duke, on a patient undergoing ECT therapy for clinical depression [15]. The candidate dataset consists of 3600 points of electric potential recordings. Another segment of EEG data of 400 recordings was used as the query dataset to find the best match. Missing points in the data are obtained by removing the recorded values at random locations in the time series. After removing 30% of the data points missing from the candidate time series with a threshold value of 50%, the two proposed algorithms were run and the results are displayed graphically in Fig1. The first 200 points of the match are displayed in the image.

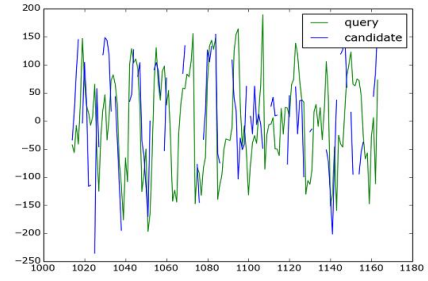

Fig. 1. Best Match Result of Nonwarping Algorithm

B. *Exact match results for Nonwarping Algorithm*

For an exact match, we have taken 400 points from the EEG candidate series to be the query and used it to match it against the EEG candidate dataset to establish a relationship between the threshold and the missing values present in the data set.

The exact match results were tested for three cases.

Case A: This denotes a low number of missing points and high threshold or tolerance value. In this case, we can expect a match at the region of the candidate from which the query was chosen. Our results showed the Location of match and actually observed results are the same.

Case B: This denotes a case where a larger number of missing values is present in the dataset and the threshold value is reasonably low. For this case, the result is highly dependent on the spread of missing values in the data. It depicts the quality control of the match. It may or may not find the exact match depending on the number of missing values present in the match area. This is because the region is populous with a number of missing values due to which the region was not considered for matching as in the result presented in case B and whereas it will match if there is a different spread of missing values and hence continue to match in a similar manner as in Case A. This makes the best match highly dependent on the spread of missing values in the candidate time series.

Case C: In this case, we give a high number of missing points and a high tolerance value which again must behave as in Case A as the skipping criteria applied in Case B has been relaxed by a higher threshold value. However, for this combination we can see that the Nonwarping algorithm works better than the Warping approach presented later. This is mostly due to the fact that a distribution of points, including missing points in the candidate, are mapped to a point in the query whereas in warping we are choosing a distribution by ignoring the missing points and considering a warped mapping. The expected match was to be found at location 800.

VI. WARPING ALGORITHM

This method proposes an incremental warping approach to search for the closest match in the candidate time series with respect to the query. Similar to the Nonwarping method, the query time series is iterated over and data of length equal to

TABLE I. EXACT MATCH RESULTS FOR NONWARPING ALGORITHM

Case	Observations		
	Missing Points (%)	***Threshold (%)***	***Location***
A	30	70	800
B	50	50	995
C	70	70	800

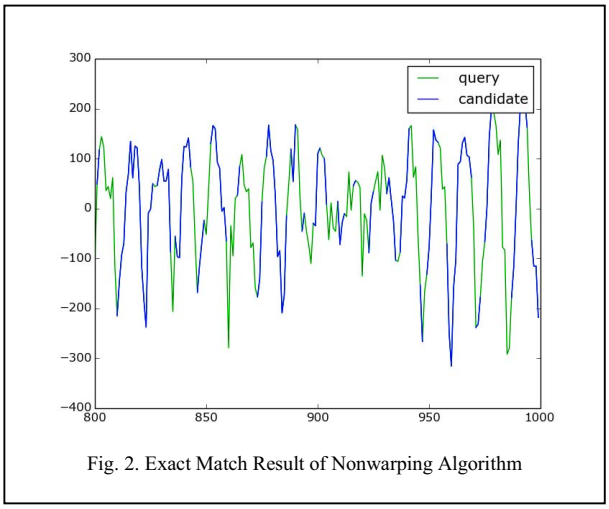

Fig. 2. Exact Match Result of Nonwarping Algorithm

the query length is considered from the candidate. A threshold value is also taken as in the previous algorithm to control quality of the resultant matching regions. In this algorithm, however, the window size is determined with respect to the threshold value provided by the user. The window is then filled with considered points from the candidate without the missing points. For the first and last pairs of points present in the window, Euclidean distance is calculated as warping is not possible for them. The remaining points of the query are then iterated over starting from the second one.

Consider *i* as the current chosen point of the query. The distance is calculated between this point in and a run of the candidate time series. The run of the candidate time series is constructed by taking points from 0 to i-2 and then selecting either *the (i-1)th, ith* or *(i+1)th* value for the final position based on which one gives a lower value for Mahalanobis Distance in that position. The algorithm does not consider the points in previous windows as the nonlocalised trends captured in cases where they are considered tend to hamper the performance of the algorithm. All the minimum distances chosen for each point in the query for the window are added. The same procedure is applied for the corresponding windows until we have reached the end of the chosen portion of the candidate series. All these distances calculated are summed up. If this sum is less than a globally maintained minimum value, the corresponding index value in the candidate data is saved as the location identifying the best match so far, which also accounts for possible distortions between the two sequences under consideration.

ALGORITHM II.

```
Winsize←(((1threshold)/2)x(length(query))
mahalDist ← ∞
q ← length(query)

for i in 0 to (length(candidate)-q+1) do
   block ← candidate[i to (i+q)]
   dist_local ← 0
   count_missing_points ← count(block)
   if((count_missing_points/q)>threshold)
        go to next block
   while end of block not reached do
        dist_local←0
   candidate_window← data block of winsize
   query_window ← query data of winsize
   candidate_temp←[]
   d←ED(candiate_window[firstpoint],query_
   window[firstpoint])+ED(candiate_window[
   lastpoint],query_window[lastpoint])
   dist_local ← dist_local + ed
  candidate_temp ← data_slice

  for j in 1 to length(candidate_temp)-1
do
       back ← ∞
       straight ← ∞
       forward ← ∞
       templist←candidate_temp[0 toj-1]
       back←mahalanobis(templist,query_wind
    ow[j])
       templist←candidate_temp[0 to j
without j-1]
  straight←mahalanobis(templist,query_win
dow[j])
            templist ← candidate_temp[0 to
j+1 without j-1 and j]
            forward ←
mahalanobis(templist,query_window[j])
            minval ←
minimum(back,straight,forward)
            if minval equals ∞ then
               go to next calculation
            else do
               dist_local ← dist_local +
minval
    if dist_local < mahalDist then
       mahalDist ← dist_local
       location ← i
```

A. *Best match results for Warping Algorithm*

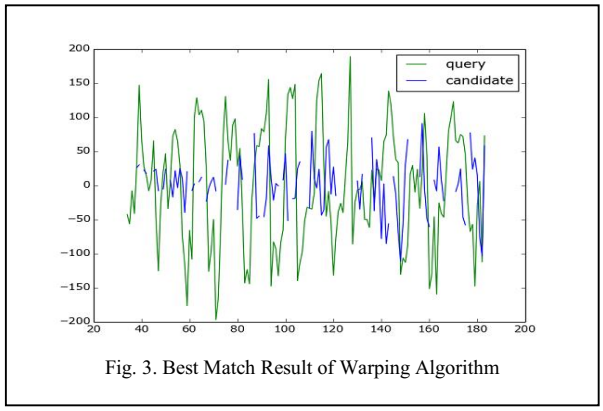

Fig. 3. Best Match Result of Warping Algorithm

The best match results for the Warping algorithm is presented in Fig. 3, with the same dataset and tested with the

same parameters in the best match results for the Nonwarping algorithm.

B. *Exact match results for Warping Algorithm*

The exact match results for the Warping algorithm is presented with the same dataset and tested with the same parameters as in the best match results for the Nonwarping algorithm. It is visually represented in Fig. 4.

The Warping algorithm takes longer to compute the best match compared to the Nonwarping method as the Mahalanobis Distance needs to be computed thrice for each point in the query in order to achieve the warping effect, as opposed to a single computation in the case of the Nonwarping algorithm.

From the above experiments, one may infer that the two proposed algorithms produce good results even when the time series has missing points. However, the quality of the matches does reduce as the number of missing values in the matching regions increase as there is lesser data to capture trends from. Comparing the performance of both our algorithms, the Nonwarping algorithm is more stable to the changes in number of missing data values. However, the Warped algorithm caters to time series where the sampling rate at certain locations may not be identical. Both algorithms give the corresponding best match between the candidate and query time series, one assuming the best case of uniform sampling rates, and one assuming the worst case of localised nonuniform sampling rates.

TABLE II. EXACT MATCH RESULTS FOR WARPED ALGORITHM

Case	Observations		
	Missing Points (%)	***Threshold (%)***	***Location***
A	30	70	800
B	50	50	468
C	70	70	1257

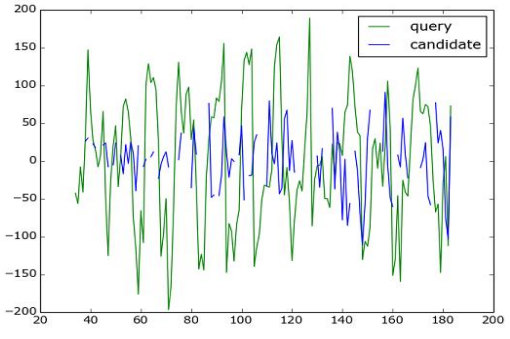

Fig. 4. Best Match Result of Warping Algorithm

VII. CONCLUSION

From the above findings, we conclude that the Warping algorithm that we have presented is a good method to find the match between two time series without performing any imputations in the candidate to handle missing values. The Nonwarping algorithm, which accounts for missing data is extended to a warping algorithm by using ideas from DTW. The Warping algorithm can also be considered an extension of DTW to fit missing data. In a single pass over the data, the algorithm finds the best match between the two time series as opposed to two or more passes that are necessary in the existing implementations. Hence, the computationally expensive preprocessing methods that are involved in handling missing data are not necessary to find the match.

VIII. FUTURE WORK

The DTW algorithm has a set of optimization techniques to improve its performance. The UCR Suite [14] further added to this set of algorithms and highly optimized the performance of the DTW algorithm such that it could handle much larger datasets in a shorter time period without compromising on accuracy. As part of our future work, we plan to extend our algorithm to cater to the needs of Big Data by devising techniques along similar lines. We also plan to make a parallelized implementation of the proposed algorithm to further improve its performance with the aim of it becoming the norm in dealing with matching of time series that may contain missing values.

REFERENCES

[1] Boetticher, Gary D. "Teaching financial data mining using stocks and futures contracts." Journal of Systemics, Cybernetics and Informatics 3, no. 3 (2006): 26-32.

[2] Ferrari, G. Tatiana, and V. Ozaki. "Missing data imputation of climate datasets: implications to modeling extreme drought events." Revista Brasileira de Meteorologia 29, no. 1 (2014): 21-2

[3] Kandasamy, Sivasathivel, F. Baret, A. Verger, P. Neveux, and M. Weiss. "A comparison of methods for smoothing and gap filling time series of remote sensing observations–application to MODIS LAI products."Biogeosciences 10, no. 6 (2013): 4055-4071

[4] Myrtveit, Ingunn, E. Stensrud, and U. H. Olsson. "Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods." Software Engineering, IEEE Transactions on 27, no. 11 (2001): 999-1013.

[5] AbdAllah, Loai. "Unsupervised Distances over Complete and Incomplete Datasets and Their Applications." PhD diss., University of Haifa, 2014

[6] Little, Roderick JA, and D. B. Rubin. Statistical analysis with missing data. John Wiley & Sons, 2014.

[7] AbdAllah, Loai, and I. Shimshoni. "A Distance Function for Data with Missing Values and its Application." In Proceedings of the 2013th International Conference on Data Mining and Knowledge Engineering. 2013.

[8] A. Eckner. "A framework for the analysis of unevenly-spaced time series data." Preprint. Available at: http://www. eckner. com/papers/unevenly_spaced_time_series_analysis (2012).

[9] A. Eckner. Algorithms for unevenly-spaced time series: Moving averages and other rolling operators. Working Paper, 2012.

[10] J. Scheffer. "Dealing with missing data." (2002).

[11] Rita Faria, Manuel Gomes, David Epstein, Ian. R. White (2014); A guide to handling missing data in Cost-Effective Analysis Conducted

within Randomised Controlled Trials, PharmacoEconomics, Vol 32 Issue 12, pp. 1157-1170. Springer 2014

[12] Mahalanobis, P. Chandra (1927); Analysis of race mixture in Bengal, Journal and Proceedings of the Asiatic Society of Bengal, 23:301–333

[13] Kato, Nei, M. Suzuki, S. Omachi, H. Aso, and Y. Nemoto. "A handwritten character recognition system using directional element feature and asymmetric Mahalanobis distance." Pattern Analysis and Machine Intelligence, IEEE Transactions on 21, no. 3 (1999): 258-262

[14] Rakthanmanon, Thanawin, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. "Searching and mining trillions of time series subsequences under dynamic time warping." In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 262-270. ACM, 2012.

[15] EEG (ElectroEncephaloGram) recordings. https://stat.duke.edu/~mw/ts_data_sets.html