2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)

Shape Based Time Series Reduction using PCA

Gajanan Gawde
Computer Engineering Department
Goa College of Engineering
Ponda, India
gg@gec.ac.in

Jyoti Pawar
Department of Computer Science and Technology
Goa University
Panaji, India
jyotidpawar@gmail.com

***Abstract*—Moving objects generate huge number of time series trajectories and length of such time series trajectories are large. To process such lengthy time series trajectories is very time consuming. There is need to reduce the dimension of the time series trajectory so that overall execution time would get reduced. There are few techniques proposed to tackle dimension reduction issue and none of them are working on a shape feature of the time series trajectory. PCA and SVM techniques are used to reduce the time series trajectories. This methods are not considering shape of the trajectories while reduction. In this paper, we have proposed Shape_PCA dimension reduction technique using Principle Component Analysis with shape as a feature vector. Our proposed technique identifies shapes which are repeated more number of times in the trajectory and accordingly reduce the time series trajectory. Thorough experimental study was carried out with different datasets. Shape_PCA method was compared with PCA and SVM without shape feature. The experimental results show that, Shape_PCA technique reduction ratio is higher side compared to PCA and SVM method. Thus, Shape_PCA is efficient method compared to PCA and SVM technique.**

***Index Terms*— Shape Based Dimension Reduction, Time Series Trajectories , Principal Components Analysis.**

I. INTRODUCTION

Reducing the dimension of time series trajectory is essential and this would reduce the overall length of trajectory. The reduction of time series is directly linked to the reduction of execution time of time series algorithm. Moving objects are generating very large trajectory and such trajectories contain duplicate or redundant data which can be eliminated. To eliminate such kind of redundant data is very challenging and we need to use proper algorithm. We need to keep in mind that there should not loss of original data while reducing the dimension of time series trajectories. That mean, we should be able generate the original data back using reverse process. This property is very important and essential for dimension reduction.

There are many applications of the searching of similar time series trajectories such as Hand Written Character Recognition, ECG classification, Route Identification etc. In hand written character recognition, characters are compared with standard characters and converted into digital form. ECG signal of the patient is recorded and are compared with the normal ECG signal. If the signal is matched with the normal signal, then patient is normal else patient is having problem. Now a day’s people are making movement from one place to other and constantly visiting places of their interest. It is important that, people should know which routes are popular or preferred by the most of people. The routes traveled by peoples are compared with each other to see if they are same. If there is match between the two trajectories, then the route are same. The route is called as popular if the route is preferred by many people. In this manner, the popular route is identified.

Reducing the dimension of time series trajectory is very challenging issue. There are few technique proposed in the literature survey but none of them can be applied directly to time series trajectory directly based on the shape feature. Rather these techniques are following regression line principle to reduce the length of the trajectory. In our proposed technique, initially we preprocess the input trajectories and then extract the shape of time series trajectory and make it as feature vector. The feature vector is representing the unique feature of the time series trajectories. Extracted feature vector is passed to the PCA for dimension reduction. The Eigen vectors are computed by PCA and the most relevant Eigen vectors are retain and remaining unimportant vectors are skipped.

The dimension reduction process is explained in figure 1. The input time series trajectories are given as input and passed to the feature extraction block. The feature extraction block extract feature vector from input time series trajectories. The feature vector extracted from this stage is passed to the next stage where PCA is modeled with feature vector to reduce dimension of datasets. The reduced dimensions of datasets are output from PCA modeling block as a final output.

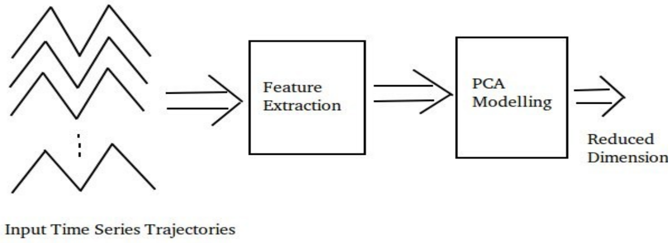

Figure 1 : Dimension Reduction Model

II. Related Work

Dimensionality reduction of high dimensional dataset is required to improve the performance. There are few researchers who have contributed towards this field. In paper [1], Authors have proposed dimensionality reduction technique using local dimension reduction technique and later indexing is done to further improve performance. In paper [2], Authors have proposed dimension reduction technique for non linear datasets. Piecewise Constant Approximation (PCA) technique is proposed to reduce dimension of time series datasets in [3]. In paper [4] , input data is reduced for prediction purpose using PCA technique. Global characteristic method is used to proposed dimension reduction in [5] of time series datasets. In paper [6][12][17], Authors have carried out survey of existing dimension reduction techniques of time series sequence. In paper [8], Authors have proposed APAC technique to reduce the dimension reduction of time series datasets.

Authors of [9] have proposed SVM technique to reduce dimension of time series sequence. Functional regression technique is used to reduce dimension of time series sequence in [14]. PAA technique is proposed in [15] to reduce dimension of time series sequence. In paper [18], PCA is indexed further to reduced dimension of time series sequence. Symbolic representation is used to reduce the dimension of time series sequence in [19]. Spectral method is used to reduce the dimension of time series sequence in [20].

III. Dimension Reduction using Principal Component Analysis

Problem Definition

Let 'N' be number of time series trajectories of moving objects. Principal Components are identified using PCA technique in order to reduce dimensions of time series trajectory. The Principal Components are identified by computing Eigen Values and Eigen Vector for time series trajectory datasets.

Shape Based Feature Vector from Time Series Trajectories

The feature vector extraction from time series trajectories is an important task of dimension reduction process. This feature vector decides the amount of dimension reduction of the given datasets. Feature Vector is represented in two dimensional space and it has rows and columns. Polygons of the time series trajectory are represented row wise and column wise three parameters are defined. The numbers of segments of polygons, turning function of polygons and area of polygons. Figure 1 shows the input time series trajectories and feature vector extracted from time series trajectory. Feature vector contains three columns such as numbers of segments, turning function and area of polygons and each polygon is represented row wise.

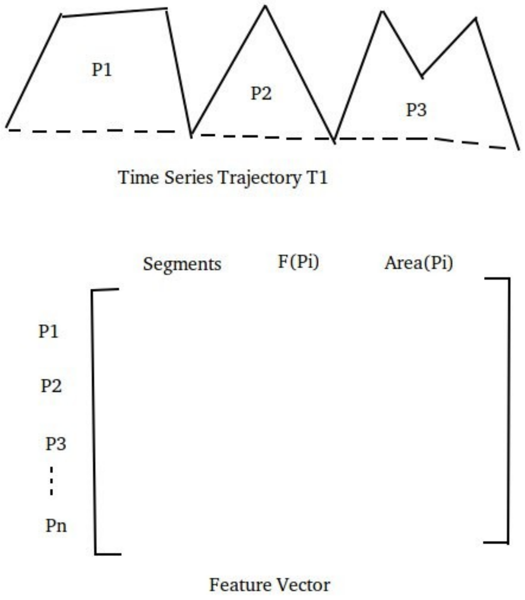

Figure 2: Feature Vector of Time Series Trajectories

Dimension Reduction using PCA

The main idea of principal component analysis is to reduce the dimension of a time series datasets consisting of large number of interrelated variable. The variables with maximum variation are retained and remaining variables are reduced. This is achieved by converting to new set of variables , the principal components (PCS), which are uncorrelated and which are ordered so that first few contain maximum variation which are retained and remaining are skipped.

Initially the time series trajectories are processed to filter the unwanted information from the input trajectories. Preprocessing of the input trajectory helps in reducing the length of trajectory and also eliminates the unwanted noise or information present in time series trajectory. The most popular preprocessing technique is Minimum Description Length technique (MDL). MDL is replacing the segments of time series trajectory if MDL distance of partition is less than MDL distance of non partition. The replacement of segment is done in such that data loss due to replacement is minimized.

The PCA technique is using statistical methods to reduce the dimension of datasets. Initially, Covariance of the matrix is computed and then using this matrix, Eigen vectors are computed. The Eigen Values which are having maximum value are retained and remaining values are ignored. The maximum variation is present in the Eigen values with high values.

Principal Component Analysis: Mathematical Formulation

Input time series trajectories are mapped into the S matrix. The covariance matrix is computed using S matrix as follows:

$$C = S^T S \quad (1)$$

The covariance matrix C is contains the variation present in the actual data and this matrix is used to compute Eigen values and Eigen Vectors. Following equation is used to compute Eigen values and Eigen Vectors.

$$Cu = \lambda u \quad (2)$$

Where C is covariance matrix and λ is Eigen Value. The Eigen Vectors are used to define the principal components. Let P1 and P2 be two principal components and are defined as follow:

$$P1 = u1\ x + u2\ y \quad (3)$$

$$P2 = v1x + v2\ y \quad (4)$$

where u and v are Eigen vectors computed in previous step.

Dimension Reduction using PCA: Algorithm

```
Step 1:- Collect the data from time series trajectory in the form of matrix.
        Let R is a trajectory and A is a feature matrix extracted from R.
Step 2:- Subtract the matrix with mean.
        $A = A – mean$
Step 3:- Compute the Covariance Matrix.
        Covariance (A)
Step 4:- Compute the Eigen Vector using Covariance Matrix computed in step 3.
        $A v = \lambda v$
Step 5:- Choosing components and forming feature vectors.
        Feature Vector = $[\lambda1 , \lambda2, \lambda3,\ldots, \lambda n]$
Step 6:- Reduced datasets using feature vector computed in previous step.
```

IV. Experimental Study

Configuration

We have used python and C++ languages to write programs and Ubuntu 12.04 operating system. The hardware configuration used for experimental study was Pentium V processor, 4 GB ram and 500 GB harddisk space. We have written C++ and python programs and were tested for correctness with different datasets and found to be correct.

Datasets

GPS Trajectories: - The dataset has been feed by Android app called Go Track. The dataset is composed by two tables. The first table go_track_tracks presents general attributes and each instance has one trajectory that is represented by the table go_track_trackspoints using latitude, longitude and altitude parameters.

ECG datasets: - Concerning the study of H. Altay Guvenir: "The aim is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in either one of the group. Class 01 refers to 'normal' ECG class and Class 02 refers to “abnormal ECG” class.

Istanbul Stock Exchange datasets: - Data sets includes returns of Istanbul Stock Exchange with seven other international index; SP, DAX, FTSE, NIKKEI, BOVESPA, MSCE_EU, MSCI_EM from Jun 5, 2009 to Feb 22, 2011.

Synthetic Time Series Dataset:- This is a synthetic dataset generated using C program. The temperature values were taken on y axis and time on the x axis. Initially, values were initialize to some constant values and then temperature and time values were updated using random function.

Results

Our proposed dimension reduction technique using Shape_PCA compared with SVM and PCA methods. There were three datasets used for experimental study such as ECG datasets, Istanbul Stock Exchange datasets and GPS trajectories. Each datasets was divided into three different categories such as Datasets with size 1000, 2000 and 3000. This different size datasets was used to see the performance of dimension reduction with increase in size of datasets.

Three datasets ECG, GPS and Stock Exchange datasets were selected without noise and the reduction ration was recorded. Reduction ratio of PCA and SVM without shape feature is less compared to Shape_PCA. The average reduction ratio using SVM is 4.37. The average reduction ratio of PCA is 4.48. The average reduction of Shape_PCA is 8.45. Thus, Shape_PCA is efficient compared to PCA and SVM.

Datasets	Dataset Size	Reduction Ratio (%)		
		SVM	PCA	Shape_PCA
ECG	1000	3.91	4.41	8.00
	2000	4.20	4.80	8.34
	3000	4.60	5.10	9.10
Stock	1000	4.10	3.98	7.89
	2000	4.45	4.20	8.12
	3000	4.60	4.89	8.99
GPS	1000	4.56	4.23	8.10
	2000	4.71	4.60	8.32
	3000	4.20	4.14	9.20

Table 1: Performance of Dimension Reduction without Noise

Three datasets ECG, GPS and Stock Exchange datasets were selected with noise and the reduction ration was recorded. Reduction ratio of PCA and SVM without shape feature is less compared to Shape_PCA. The average reduction ratio using SVM is 3.99. The average reduction ratio of PCA is 3.84. The average reduction of Shape_PCA is 7.05. Thus, Shape_PCA is efficient compared to PCA and SVM. The reduction ratio with noise is slightly at lower side compared to without noise. Figure 3 shows the reduction ratio of SVM, PCA and Shape_PCA with four different datasets.

Datasets	Dataset Size	Reduction Ratio (%)		
		SVM	PCA	Shape_PCA
ECG	1000	3.21	3.67	6.90
	2000	3.50	3.89	7.20
	3000	3.45	4.10	7.34
Stock	1000	4.12	3.67	6.88
	2000	4.40	3.87	6.98
	3000	4.55	3.90	7.21
GPS	1000	4.31	3.78	7.01
	2000	4.12	3.68	6.89
	3000	4.25	4.01	7.10

Table 2: Performance of Dimension Reduction with Noise

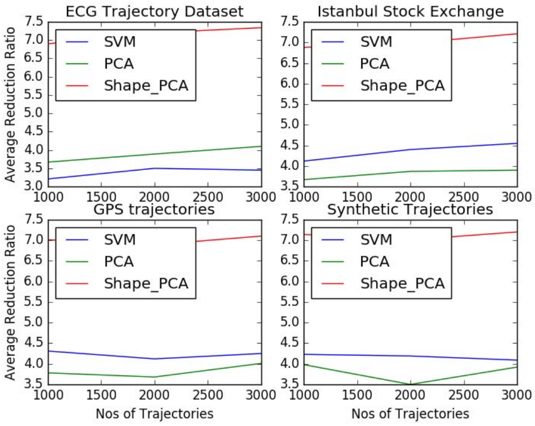

Figure 3: Reduction Ratio of SVM, PCA and Shape_PCA

V. Conclusion

We have proposed Shape_PCA dimension reduction technique for time series trajectories using PCA. Feature vector was extracted successfully from input time series trajectories datasets and passed to the PCA method. Shape_PCA technique was compared with SVM and PCA techniques. Experimental study was carried out on three different datasets such as ECG, Istanbu Stock Exchange and GPS trajectories datasets. Experimental results revealed that Shape_PCA average reduction ratio without noise was 8.45 and with noise was 7.05. SVM average reduction without noise was 4.37 and with noise was 3.99. PCA average reduction ratio without noise was 4.48 and with noise was 3.84. Thus, Shape_PCA was efficient compared to SVM and PCA techniques.

References

[1] Kaushik Chakrabarti , SharadMehrotra, 2000, Local Dimensionality Reduction: A New Approach to Indexing High Dimensional Spaces. VLDB .

[2] David DeMers , Garrison Cottrell , 2004, Non–Linear Dimensionality Reduction. Conference on Machine Learning.

[3] Vasileios Megalooikonomou Guo Li Qiang Wang , 2004, A Dimensionality Reduction Technique for Efficient Similarity Analysis of Time Series Databases. CIKM.

[4] A. Lendasse1, J. Lee2, E. de Bodt3, V. Wertz1, M. Verleysen2, 2001, Input Data Reduction for the Prediction of Financial Time Series. ESANN.

[5] Xiaozhe Wang, Kate A. Smith, and Rob J. Hyndman, 2005,Dimension Reduction for Clustering Time Series Using Global Characteristics. ICCS.

[6] I. K. Fodor , Survey on Dimension reduction techniques.

[7] Rainer Hegger, Holger Kantz , Practical implementation of nonlinear time series methods: The TISEAN package.

[8] Wing Kam Fung, Xuming He, Li Liu1 and Peide Shi, 2002, DIMENSION REDUCTION BASED ON CANONICAL CORRELATION. Statistica Sinica.

[9] Eamonn Keogh Kaushik Chakrabarti Sharad Mehrotra Michael Pazzani , 2002, Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. Transaction on Database System.

[10] Jinbo Bi, Kristin P. Bennett ,Mark Embrechts ,Curt M. Breneman , Minghu Song , Dimensionality Reduction via Sparse Support Vector Machines.

[11] K. V. Ravi Kanth Divyakant Agrawal Amr El Abbadi Ambuj Singh, 1999, Dimensionality Reduction for Similarity Searching in Dynamic Databases. SIGMOD.

[12] Miguel A , Carreira-Perpinan , 2002,A Review of Dimension Reduction Techniques. SIGMOD.

[13] U. Amatoa, A. Antoniadisb,∗, I. De Feisa, 2003, Dimension reduction in functional regression with applications.

[14] Eamonn Keogh, Chotirat Ann Ratanamahatana , 2002, Exact indexing of dynamic time warping. VLDB.

[15] Eamonn Keogh Kaushik Chakrabarti Michael Pazzani Sharad Mehrotra, 2000, Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. VLDB.

[16] Ming Yuan , Ali Ekici , Zhaosong Lu , Renato Monteiro , 2006, Dimension reduction and coefficient estimation in multivariate linear regression.

[17] GIUSEPPE REGA , HANS TROGER , 2000,Dimension Reduction of Dynamical Systems: Methods, Models, Applications. Modeling and Simulation of Micro system.

[18] Eamonn J. Keogh and Michael J. Pazzani , 2000, A Simple Dimensionality Reduction Technique for Fast Similarity Search in Large Time Series Databases. KDD.

[19] Jessica Lin Eamonn Keogh Stefano Lonardi Bill Chiu, 2003, A Symbolic Representation of Time Series, with Implications for Streaming Algorithms.SIGMOD.

[20] Lawrence K. Saul , Kilian Q. Weinberger , Fei Sha , Jihun Ham , Daniel D. Lee , Spectral Methods for Dimensionality Reduction. Annual Conference on Computational Learning.