Binary Journal of Data Mining & Networking 5 (2015) 01-04
http://www.arjournals.org/index.php/bjdmn/index

Review Article

ISSN: 2229 -7170

Time-Series Data Mining:A Review

Suman H. Pal[1], Jignasa N. Patet[1]

*Corresponding author:

Suman H. Pal

[1]Department of Information Technology Shri S'ad Vidya Mandal Institute of Technology Bharuch 392-001, Gujarat, India.

Abstract

Data mining refers to the extraction of knowledge by analyzing the data from different perspectives and accumulates them to form useful information which could help the decision makers to take appropriate decisions. Classification and clustering has been the two broad areas in data mining. As the classification is a supervised learning approach, the clustering is an unsupervised learning approach and hence can be performed without the supervision of the domain experts. The basic concept is to group the objects in such a way so that the similar objects are closer to each. Time series data is observation of the data over a period of time. The estimation of the parameter, outlier detection and transformation of the data are some of the basic issues in handling the time series data. An approach is given for clustering the data based on the membership values assigned to each data point compressing the effect of outlier or noise present in the data. The Possibilistic Fuzzy C-Means (PFCM) with Error Prediction (EP) are done for the clustering and noise identification in the time-series data.

Keywords: Data Mining, Clustering Algorithm,FCM ,PFCM .

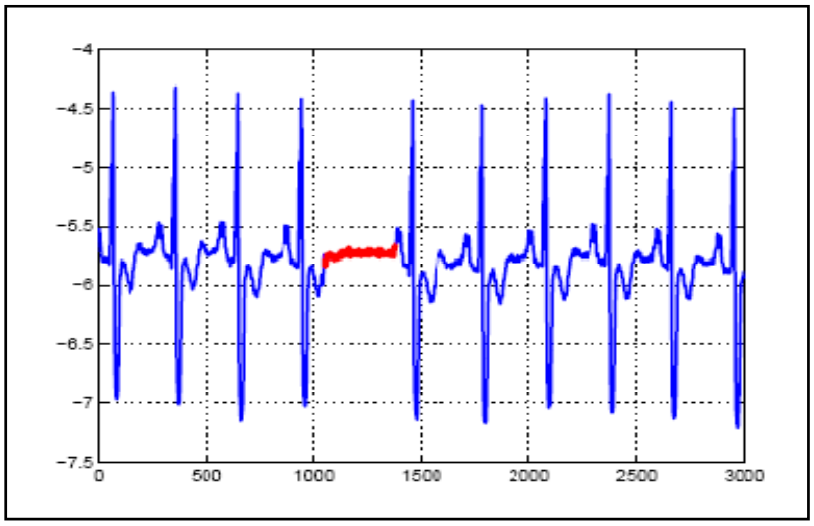
Introduction

The time series data is a set of values observed over a period of time. The data has similar characteristics and can be identified by finding the correlation among these points. The explosion of time series data mining has become an important area of research due to the increasing dimensionality and availability of data[1]. Another problem with time series data is the similarity measure which is based on human perception. From various domains such as intrusion detection, fraud detection, flight safety, health care, etc., the data is collected in form of sequence values over time period. The analysis of such data extracts the information that helps in forecasting of temperature, whether or seasons and the stock exchange prediction can also be done. A fault in such values may result in error of the instrument or an undesired scenario. The identification of such abnormal conditions is termed as Anomaly Detection or Error Prediction (EP).

Mathematical definition of time series data is given as follows [2]:

Time Series: A time series T of length n is a sequence of pairs

$$T=[(p_1,t_1),(p_2,t_2)..(p_i,t_i)..(p_n,t_n)]||(t_1<t_2<...t_1<.....t_1)$$



Figur.1Time Series Data

Some challenge that the times series data observes are:

Forecasting

The central logical problem in forecasting is making predictions never form a random sample from the same population as the time periods about which the predictions is done. A sudden drop is observed in the stock due to news development which cannot be predicted. It is now widely believed that a major cause of the drop was the newly-introduced widespread use of personal computers programmed to sell stock options whenever stock prices dropped slightly.

A second problem that arises in time-series forecasting is that the distribution of data is not always known in prior. Such a simplification may produce more serious errors in time-series work than in other areas. Even if forecast is which is updated after each new time period, the forecasts are made one at a time, and if becomes wrong forecast may result in bankrupt of company.

Impact of single events

When you try to assess the impact of a single event, the major problem is that there are always many events occurring at any one time. While study of the time series data the effects of one variable on another, they usually have at least two time series--one for the independent variable and one for the dependent-variable.

One problem with such study is the observations within each series are not independent of each other; the probability of finding a high correlation between the two series may be higher than is suggested by standard formulas.

A second problem is that it is rarely reasonable to assume that the time sequence of the causal patterns matches the time periods in the study. Thus if increased unemployment typically produced an increase in crime exactly six months later but not five months later, then it would be fairly easy to discover that relationship by correlating monthly changes in unemployment with monthly changes in crime six months later.

Apart from the various challenges that exist for mining of the time series data a huge number of application areas are detailed below:

Detecting abnormal pattern from ECG data

The ECG data is a periodic time series data. The anomaly detected in such case is the non-conforming patterns depending upon amplitude, indicating health problem.

Detection of anomalous flight sequence from aircrafts

The flight data obtained from different sensors are recorded at every interval of time. For a sudden change that may occur in course of flight may cause an unhappening. Hence, such deviation in behaviour can be a anomaly.

Detecting outliers in periodic stars

The periodic variability in the stars by observing the standard deviation from the rest is the anomalies. These outliers are interesting patterns changing over period or amplitude, introduces noise in the light curve

Eco-system

The earth science data such as vegetation and temperature are also studied for the variability in production of vegetation and its temperature effects.

Anomaly Detection

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behaviour [3]. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains. Anomaly detection finds extensive use in a wide variety of applications such as fraud detection for credit cards, insurance or health care, intrusion detection for cyber-security, fault detection in safety critical systems, and military surveillance for enemy activities.

The importance of this approach is that anomalies in data translate to significant actionable information in various application areas. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination [4]. An anomalous MRI image may indicate presence of malignant tumors. Anomalies in credit card transaction data could indicate credit card or identity theft or anomalous readings from a space craft sensor could signify a fault in some component of the space craft.

Anomalies can be defined as patterns in data that do not conform to a well defined notion of normal behaviour. Fig 1 illustrates anomalies in a two-dimensional data set. The data has two regions, N1 and N2. Points that are adequately far away from the regions, e.g., points o1 and o2, and points in region O3, are anomalies.

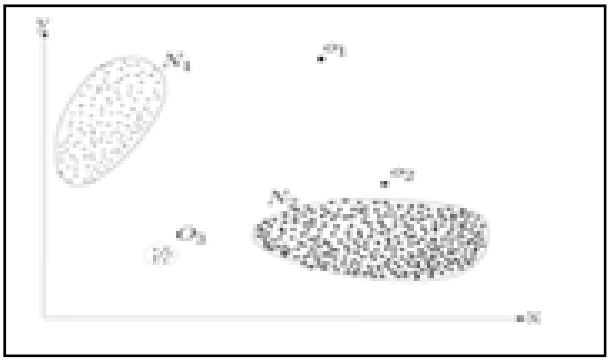


Figure .1 An example of Anomaly Detection

The rest of the paper continues with Section III discussing state-of-art survey on the existing systems. Section IV,V discusses future work Conclusion.

Litrature Review

Saeed Aghabozorgi, Mahmoud Reza Saybani and The Ying Wah [5], in this paper the clustering is done in an incremental manner. The basic step is to perform the dimensionality reduction with Discrete Wavelet Transform (DWT) and the hierarchical clustering is done for determination of the number of clusters. The LCSS is taken as a distance measure for the hierarchical clustering. The clustering is done with the Fuzzy C-Means (FCM) Clustering.

Syed A. Pasha and Philip H.W. Leong[6], the paper proposed a procedure for performing cluster analysis using joint information of temporal and continuous-valued data for sampling. The proposed procedure uses the Lomb-Scargle method to perform cluster analysis based on the joint information of the temporal distribution and the process observed at the irregular sampling times. Cluster analyses of multivariate tick data of major stock market indices and currencies are presented.

Min Ji, Fuding Xi, and Yu Ping [7], highlighted the drawbacks of the Fuzzy C-means approach and proposed an algorithm named dynamic fuzzy clustering applied for time series data in order to form groups in which the class labels of data are indistinct and require further separation. The advantage of this approach over the existing clustering methodology is that the property of time series data belonging to different clusters can be discovered partially over time which makes proper groups dynamically. This approach incorporates the information on the key point to the time series data with the associated fuzzy values representing the uncertainty of the data to different clusters.

Saeed Aghabozorgi and Teh Ying Wah[8], proposed clustering approach for the time series data. The method addressed the problem of assigning the cluster class label that was vague. The paper does a two-level clustering method. In the first step the Symbolic Aggregation Approximation is done to convert the time series data to symbolic forms. Further the similarity between the two symbolized time series is done using Longest Common sub Sequences approach. The Fuzzy C-Means clustering is done for clustering the data. In

FCM approach has different advantages FCM is done largely due to the advantage of the degree of membership of a time series to the clusters in clustering process. It is used to facilitate the detection of changes in prototypes. Moreover, fuzzy sets have a more realistic approach to address the concept of similarity than classical sets. A fuzzy set is a set with fuzzy boundaries where each element is given a degree of membership to each set.The FCM works by partitioning a collection of n vectors into c fuzzy groups and finds a cluster center in each group such that the cost function of dissimilarity measure is minimized.

Saeed Aghabozorgi and Teh Ying Wah [9], performed a similar research with a new proposed methodology with Dynamic Time Wrapping (DTW) as similarity measure and a new Multi-step Time series Clustering model (MTC) to make clusters based on similarity in shape. This method tries to overcome the clustering problems in finding cluster of similar time series in shape. The MTC is initially pre-processed to group low resolution time series data. The data undergoes through z-score Normalization which makes time series invariant to scale and offset. The Symbolic Approximation is adopted to reduce the dimensionality of time series data after normalization. In the second step, split the pre-clusters and making prototypes. Moreover, the number of time series in the dataset (cardinality) is reduced by defining a prototype for each group of very similar time series. It decreases the complexity of MTC to a large extent. In the final step merging the clusters is performed from the sub-clusters from the above two steps. The mapping activity is performed to assign the cluster with each prototype. Different advantages are The algorithm is accurate for clustering of large time series based on similarity in shape, Providing a clear understanding of the domains by its ability to generate hierarchical clusters of large time series data.

GoktugT.Cinar and Jose C. Principe[10], the proposed method consists of a measurement equation and multiple state transition equations with Hierarchical Linear Dynamical System (HLDS). The new HLDS model consists of one observation layer and multiple hidden state layers. Each hidden state layer acts as the driving input/cause to the layer below it, with only the lowest layer relating to the observations via a linear model. However the system suffers high dimensionality. In which advantages are The method is adaptive and self-organizing, previous exposure to the acoustic input is the only requirement for learning and recognition. Moreover there is no need of selecting the number of clusters.

SaeedAghabozorgi,TehYingWah,TututHerawan,HamidA.Jalab,Mohammad Amin Shaygan,andAlirezaJalali [11], illustrated the drawbacks of traditional clustering algorithms for time-series data. First, the parameters should be set in the algorithm and it is based on the user's assumptions which may be false, results in inaccurate clusters. Second, the processing time of the model-based clustering is very slow with respect to large datasets. In the proposed method the sub-cluster grouping is done on the basis of time similarity. These sub-clusters are merged with the Medoid algorithm on the basis of their shape. The algorithm is evaluated against various syntactic and real-world time series datasets.

Tiantian Yang and Jun Wang [12], proposed a method hybridizing the phase shift and the clustering approach and developed phase shift weighted spherical K-means Algorithm. The algorithm finds optimal phase shift between two subsequences. A fuzzy feature weight is added to the data thereby minimizing the Lagrange's multipliers. The results are performed on the ECG data signal and results satisfactory.

Some of the basic problems existing in the literature are discussed below:

The most prominent problems arise from the high dimensionality of time-series data and the difficulty in defining a form of similarity measure based on human perception.

One of the major problems affecting time-series systems is the large numbers of parameters induced by the method. The user is

usually forced to tune the settings in order to obtain best performances[4].
FCM membership for data clustering does not represent the degree of belongingness properly. As the sum of all the clusters across the data set should be 1, results in abnormality.
The FCM produces inaccurate results when data contains noise.
Noisy points are forced to be included in clusters without detection and removal.

Conclusion

A time series is a sequence of real data, representing the measurements of a real variable at time intervals. The main motivation for representing a time series in the form of clusters is to better represent the main characteristics of the data. In this work the state of art review of the existing algorithms. The proposed work is outline with the basic idea of possibilistic fuzzy c-means algorithm incorporating the error prediction for analysis of the behavior and fault in the series data. The proposed idea needs to analyzed experimentally.

Future Work

In future we look forward to implement the approach and perform the comparative analysis With clustering approach for the time series data, the proposed idea is to be implemented in coming days.

References

[1]. Robert Darkins, Emma J. Cooke , Zoubin Ghahramani , Paul D. W. Kirk1,David L. Wild, Richard S. Savage, "Accelerating Bayesian Hierarchical Clustering of Time Series Data with a Randomised Algorithm", PLOS ONE, 2013; 8,4; 1-8.

[2]. Sampasetty Saravanan and Gulam Mohideen Kadhar Nawaz, "Ensemble-based Time Series Data lustering for High Dimensional Data", International Journal of Innovative Computing, Information and Control, 2014; 10,4 ; 1457-1470.

[3]. Hesam Izakian and Witold Pedrycz, "Anomaly Detection in Time Series Data using a Fuzzy C-Means Clustering", IFSA World Congress IEEE, 2013; 1513-1518,

[4]. Kavitha V, Punithavalli M. "Improved Hierarchical Clustering Using Time Series Data", International Journal of Emerging Technology and Advanced Engineering, 2013; 3,1; 569-573.

[5]. Saeed Aghabozorgi, Mahmoud Reza Saybani and The Ying Wah, "Incremental Clustering of Time-Series by Fuzzy Clustering", Journal of Information Science and Engineering,2012;28; 671- 688.

[6]. Syed A. Pasha and Philip HW. Leong, "Cluster Analysis of High-Dimensional High Frequency Financial Time Series", IEEE Conference on Computational Intelligence for Financial Engineering &Economics(CIFEr), 2013; 68-76.

[7]. Syed A. Pasha and Philip H.W. Leong, "Cluster Analysis of High-Dimensional High Frequency Financial Time Series", IEEE Conference on Computational Intelligence for Financial Engineering &Economics(CIFEr), 2013; pp. 68-76

[8]. Saeed Aghabozorgi and Teh Ying Wah, "Effective Clustering of Time-Series Data Using FCM", International Journal of Machine Learning and Computing, 2014; 4 ,2.

[9]. Saeed Aghabozorgi and The Ying Wah, "Clustering of large time series datasets", Intelligent Data Analysis, 2014; pp.793-817.

[10]. GoktugT.CinarandJoseC.Principe, "Clustering of Time Series Using Hierarchical Linear Dynamical System",IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP),2014; 6791-6796 .

[11]. Saeed Aghabozorgi, Teh Ying Wah, Tutut Herawan, Hamid A. Jalab, Mohammad Amin Shaygan, and Alireza Jalali, "A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique", The Scientific World Journal, Vol. 2014; pp. 1-12.

Tiantian Yang and Jun Wang, "Clustering Unsynchronized Time Series Subsequences with Phase Shift Weighted Spherical k-means Algorithm", Journal of Computers, 2014; 9,5 ; pp. 1103-1108,.