Increasing Efficiency of Time Series Clustering by Dimension Reduction Techniques

Saeid Bahadori [1] and Nasrollah Moghadam Charkari [2]

Tarbiat Modares University, Industrial Engineering Department, Tehran

Summary
Finding similar time series has attracted a lot of interest and much research has been done recently as a result [1]. For the reason of high dimension of the time series data, finding a good answer to this problem is difficult. Encounter with these high dimensional data requires us to use dimension reduction techniques, and then performing data mining tasks on reduced dataset. Several time series dimension reduction techniques have been proposed before, such as DFT [2], DWT [3], SVD [4], PAA [5] and [31], APCA [6], PLA [7], SAX [8] and many others, but we cannot simply choose an arbitrary compression algorithm [9]. Each one of these algorithms has different answers to a unique problem. The main contribution of this paper reviewing the time series data mining and data mining literature. In this research we have been compared result of clustering after reducing dimension of data set with two different well-known algorithms, DFT and DWT techniques. Finally, we proposed energy ratio algorithm to find out the most efficient number of dimensions in a new space.

Key words:
Time series, Data mining, Dimension reduction, Cluster validity measures.

1. Introduction

In the last decade, we have seen an increase in the importance placed on time series data mining [10]. If you look around, you will be able to find a lot of things that change over time dimension.If we are able to see patterns of change in the past, then we can better forecast the patterns of change in the future. For example, changes of weather temperature, value of the stock, usage of a website, bank account and other variables by time can help us to discover some useful patterns. So the list of time series databases that need to be mined is expanded. Hundreds of papers have been published covering all aspects of time series data mining, namely, dimension reduction or representation techniques, indexing, clustering, classification, novelty detection, motif discovery, etc. According to the research in this field, the main tasks of time series data mining methods are: forecasting, indexing, clustering, classification, novelty detection, motif discovery and rule discovery [11] and [6].
First, in most of the above tasks it is necessary to define a similarity measure between two time-series. A second issue that arises in time series data mining and interrelated with the choice of a similarity measure is the representation of a time series. The aim at this paper is to serve as an overview of advances in time series data mining and dimension reduction of time series and have a comparison from techniques through implementing these techniques on data sets to find similar time series and group them into a specific cluster.

1.1. Introduction to data mining

Data mining refers to extracting knowledge from large amounts of data.[33] The process of data mining consists of 3 main steps: Data pre-processing, Data analysis and Result Interpretation. [34] Fig.1 showed a detailed view of the some different methods that is using for each step. [32] As Han described in his book, pre-processing is critical and important part which playing important role in success of data mining project.Depending on type of data being mined, Pre-processing consists of different sub-tasks. One of the major sub-tasks is dimension reduction which we will discuss about it later. As shown in Fig.1, in analysis step, based on the problem and resolution that we are going to do, we faced with different methods and algorithms such as classification techniques – k nearest neighborhood, Decision Trees, Bayesian Networks, SVM – Association Rules and clustering techniques such as k-Means, Density-based, Message-passing and Hierarchical. As described before, since we intend to group similar time series in one group of objects we focused on clustering analysis with k-Means algorithm for the reason for its simplicity and popularity of application.

Manuscript received May 5, 2018
Manuscript revised May 20, 2018

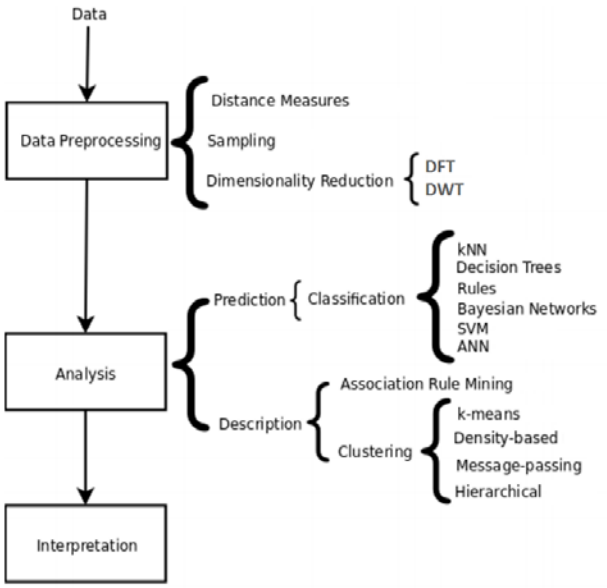


Fig.1 Main steps and methods in Data Mining problem, with their correspondence methods

1.2. Time series data mining concepts and tasks

A time series is a collection of observations made sequentially through time. As mentioned in Section 1, the most common tasks of time series data mining methods are: indexing [12] and [13], clustering [14] and [15], classification, novelty detection [16], motif discovery [17], [18] and [19] and rule discovery [20]. A brief description of each task is given below:

Indexing: Find the most similar time series in a database to a given query time series.

Clustering: Find groups of time series in a database where time series of the same group are similar to each other and time series from different groups are dissimilar to each other.

Classification: Assign a given time series to a predefined group in a way that is more like other time series of the same group than it is to time series from other groups.

Novelty detection: Find all sections of a time series that contain a different behavior than the expected with respect to some base model.

Motif discovery: Detect previously unknown repeated patterns in a time series database.

Rule discovery: Infer rules from one or more time series describing the most possible behavior that they might present at a specific time point (or interval).

Clustering and classification of time series rely heavily on finding similar time series. Since finding similar time series is strongly dependent on the representation scheme selected, thus, representation of time series data have critical rule in time series data mining tasks efficiency. As with most problems of computer science, the suitable choice of representation greatly affects the ease and efficiency of time series data mining. [6]

2. Problem Definition

Finding similar groups with high dimension of the time series data, is very difficult. Facing with these high dimensional data requires us to use dimension reduction techniques, and then performing data mining tasks such as clusting on reduced dataset to find similar time series. Several time series dimension reduction techniques have been proposed, but choosing proper dimension reduction technique is the problem that we are going to address in this research. On the other hand number of dimension of new space is second problem that we encounter after we select the dimension reduction techniqe.

2.1. Clustering techniques

Clustering techniques considers data tuples as objects. They partition objects into some groups, so that object of a cluster is similar to one another. Similarity is defined in terms of how objects are close to the space to anther based on a distance function. One of the well-known clustering methods is k-Means. k-Means takes k as an input parameter and partitions the data set of objects of k clusters and the goal of this calculation is having clusters of high inter cluster similarity and low intra-cluster similarity. This algorithm is described in most of the papers and we do not explain it in detail. But one of the main things that is very affecting the result of the clustering result measures the quality of the clustering result. Many validity measures have been proposed to evaluating clustering results. Some of these validity measures are [29]:

1. Dunn's measure (DI)
2. Davies-Bouldin's measure (DB)
3. Partition coefficient (PC)

Clustering is not a total solution to resolving the problem and depending on the problem statement and the application domain we may consider different aspects. For instance, for a specific application and problem statement, it may be important to have well separated clusters while for another issue, we consider more to have compactness of the clusters. Therefore, selecting the proper cluster analysis measure is the willingness of how we expect from the cluster and their shapes. [29] In this research because we want to have compactness clusters, we select Davies-Bouldin's measure (DB) for measuring the clustering result which in some papers that analysis and compares different clustering validity measure, DB measure has the best result.

[29] This measure is a function of the ratio of the sum of within cluster scatter to between-cluster separation, and it uses both the clusters and their sample means.

$$d_{ij} = d(v_i, v_j) \tag{1}$$

$$S_i = \frac{1}{\|c_i\|} \sum_{x \in c_i} d(x, v_i) \tag{2}$$

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} \tag{3}$$

$$R_i = \max_{\substack{j=1,..,k \\ i \neq j}} (R_{ij}) \tag{4}$$

$$DB = \frac{1}{k} \sum_{i=1}^{k} R_i \tag{5}$$

In this indicator, R_{ij} is cluster similarity and S_i is about distance between each cluster and d_{ij} will be dissimilarity of clusters. Based on this formula DB measure will be a criterion which shows that clustering is the optimum.

2.2. Time series clustering

Clustering on time series data is one of the interested subjects in data mining. Various approaches have been developed to cluster different types of time series data. The former approach works directly with raw time series data, thus called raw-data-based approach, and the major modification lies in replacing the distance/similarity measure for static data onto an appropriate one for time series. The latter approach first converts a raw time series data either into a feature vector or a number of model parameters, and then applies a conventional clustering algorithm to the extracted feature vectors or model parameters, thus called feature- and model-based approach, respectively. The three different approaches: raw-data-based, feature-based and model-based [15]. Another approach that Liao didn't mention in his research is using dimensional reduction in preparation phase. So our aim is clustering time series data onto dimensional reduction.

3. Experimental Results and Discussion

3.1. Time series dimensionality reduction techniques

Dimensionality reduction is an effective approach to downsizing data.[35] It is a method that attempts to convert high dimensional vectors to a lower dimension space while retaining their behavior. There has been several time series dimension reduction techniques proposed in the literature such as DFT [2] , DWT [3] , SVD [4] , PAA [5] and [31] , APCA [6] , PLA [7] , SAX [8] and many others. A general framework, GEMINI, for indexing time series using dimension reduction techniques is presented in Faloutsos [21] and [22]. It uses the following steps:

1. Map the set of time series into a reduced dimensionality space.
2. Use an indexing technique to index the new space.
3. For a given query time series X (a) map the query X to the new space; (b) find nearest neighbors to X in the new space using the index; and (c) compute the real distances and keep the closest.
4. All of the above dimension reduction techniques support lower bounding lemma [21]. Discrete Fourier Transform (DFT) was one of the first representation schemes proposed to data mining context [2]. DFT expressed based on the idea that every signal can be represented as a superposition of sine and cosine waves. DFT decomposes a time-series into summation of sine wave and keep first N coefficients. DFT transforms a time series from the time domain into the frequency domain. The formula for transforming time domain to frequency domain via the DFT technique can be seen below:

$$DFT: \underset{k=0,1,...N-1}{X(f_{k/N})} = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi kn}{N}} \tag{6}$$

$$IDFT: \underset{k=0,1,...N-1}{X(n)} = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(f_{k/N}) e^{\frac{-j2\pi kn}{N}} \tag{7}$$

Parseval's Theorem states that energy in the frequency domain equals the energy in the time domain. It follows that computing the distances using k coefficients provides a lower bound on the Euclidean distance of the original sequences [11] and [27]. Time complexity of DFT for a given time series with n dimension is $O(n^2)$, but some other new algorithms based on DFT proposed which tried to decreased the complexity. This algorithm named FFT with the complexity of $O(n \log n)$. [27]

Another technique which is like DFT and transforms it into the time/frequency or space/frequency domain is Discrete Wavelet Transform (DWT) [3]. The DWT can be calculated efficiently, and an entire dataset can be indexed in $O(mn)$. DWT has some limitations. One of the limitations is it is only defined for sequences whose length are an integral power of two.

3.2 Comparison of dimensionality reduction techniques and experiment results

To our knowledge, there is no paper that compares the above techniques with one another. However, there exists on some excellent tutorials by Keogh [23], [26] and [27] and Faloutsos [24]. Therefore, in this section we will compare two of more popular techniques DWT and DFT with e another based on calculating time, results and ease of understanding. For this purpose, we will implement the above-mentioned techniques on a time series data. In this phase, we will need to answer two questions that arise. First, is which techniques have superior advantages or better results from the others? Second, is what number of dimensions is the most efficient in a new space? We implement these techniques on the time series data onto 19106 selected subscribers of a major ISP company in Iran during a 2-week period, and collected their upload and download internet traffics every one hour. At the end, each customer has 672 data, and the total number of data collected are 12.83 million. Then we use k-Means to clustering the data set.

We use k-Means for clustering because this algorithm is easy to understand and implementation does not consume a lot of time, and thus, it allows us to conduct many such experiments. We check the results of clustering by Davis-Bouldin cluster validation technique [28] and [29]. This technique finds clusters of high compactness. Members of similar characteristics are sequence together in clusters. As a result, there is a large separation between the many clusters which allows us to check the characteristics of the many clusters.

Fig. 2 and Fig.3 present download usage of a subscriber and a new representation of lower dimension format through DFT and DWT techniques.

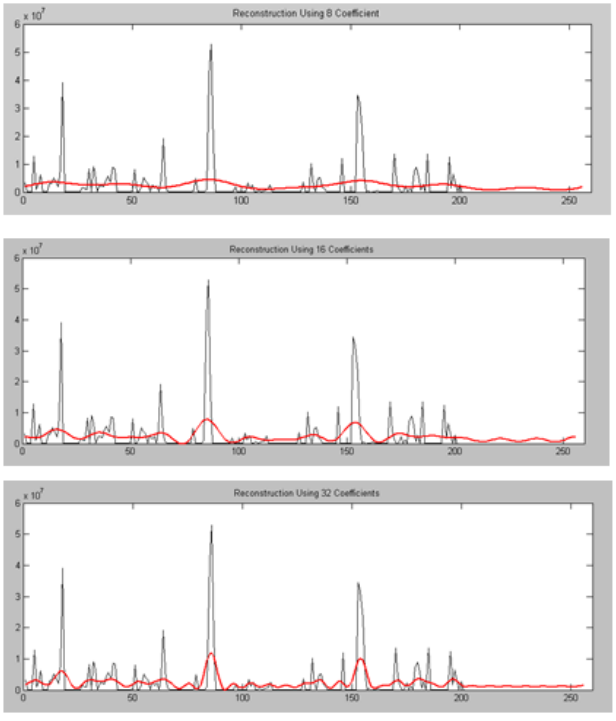


Fig. 2 The DFT representation of a time series sequence of length 256 which is reduced to 8,16 and 32 dimensions.

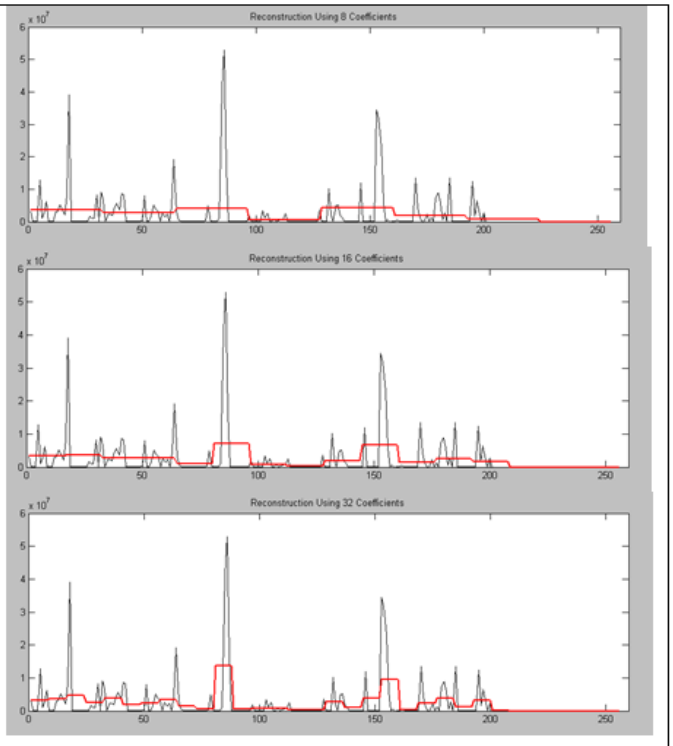
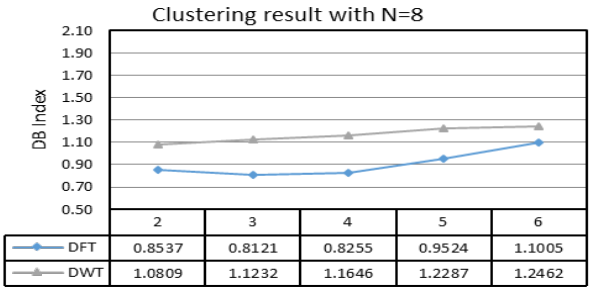


Fig. 3 The DWT representation of a time series sequence of length 256 which is reduced to 8,16 and 32 dimensions.

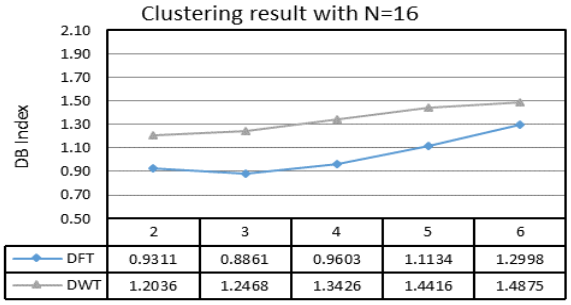
Fig.4 shows the results of implementing various dimensional reduction techniques using the same number of dimensions (N=8). The results prove that the DFT technique has the lowest DB index in different number of

clusters. By using the DFT technique we realized that a DB index of three clusters was the most efficient. Fig. 5 shows the same result using a different number of dimensions (N=16). Below, you will be able to find Fig.6 and 7 showing cluster results with different dimensions (N=32 and N=64). We reached an important conclusion, when observing Fig.7 which demonstrates cluster results of the dimension of N=64. In Fig.7, we observed that all techniques are most efficient using three clusters when dealing with 64 dimensions. Collectively, the results point out that DFT is the most efficient technique because it has the same number of clusters of a variable number of dimensions.



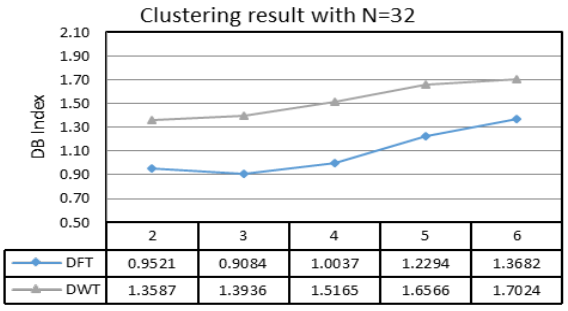
	2	3	4	5	6
DFT	0.8537	0.8121	0.8255	0.9524	1.1005
DWT	1.0809	1.1232	1.1646	1.2287	1.2462

Fig. 4 The results of implementing various dimensional reduction techniques using the same number of dimensions (N=8)



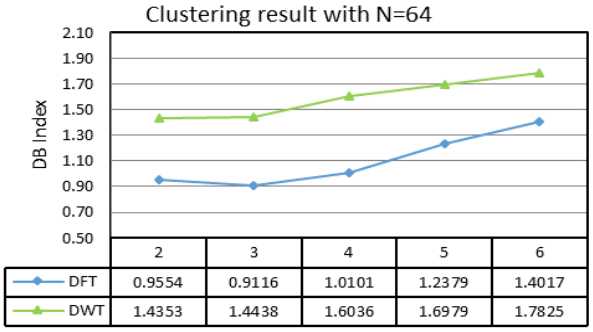
	2	3	4	5	6
DFT	0.9311	0.8861	0.9603	1.1134	1.2998
DWT	1.2036	1.2468	1.3426	1.4416	1.4875

Fig. 5 The results of implementing various dimensional reduction techniques using the same number of dimensions (N=16)



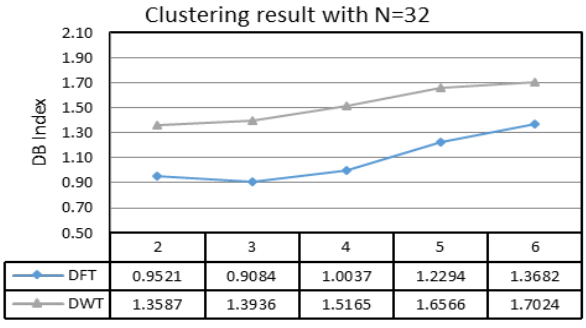
	2	3	4	5	6
DFT	0.9521	0.9084	1.0037	1.2294	1.3682
DWT	1.3587	1.3936	1.5165	1.6566	1.7024

Fig. 6 The results of implementing various dimensional reduction techniques using the same number of dimensions (N=32)



	2	3	4	5	6
DFT	0.9554	0.9116	1.0101	1.2379	1.4017
DWT	1.4353	1.4438	1.6036	1.6979	1.7825

Fig. 7 The results of implementing various dimensional reduction techniques using the same number of dimensions (N=64)



	2	3	4	5	6
DFT	0.9521	0.9084	1.0037	1.2294	1.3682
DWT	1.3587	1.3936	1.5165	1.6566	1.7024

Fig. 8 The results of implementing various dimensional reduction techniques using the same number of dimensions (N=64).

The other criteria in finding proper technique is calculation time, in our 20 times experiment, result shows that DFT technique has the lowest computing time with different Dimensions (N). So based on the result of clustering with DB measure and time of calculation DFT still is better

approached in dimension reduction of time series. For attaining the number of dimensions that have proper results we proposed energy percentage which is calculated through the formula seen below:

In this equation $\bar{A}$ is the signal in new space with deducted dimensions and A is the original signal. By this equation we can see that in which coefficient or dimension the ratio is higher than a threshold. In our experiment we assume that if the energy ratio is equal to 90 percent then the related dimension is quite enough dimension which has most similarity with the original data.

$$Energy = \bar{A}^2 \quad (1)$$

$$Ratio = Energy / A^2 \quad (2)$$

Fig. 9 is the curve of energy percentage and the number of new spaces. This figure illustrates that there is higher accuracy when using higher dimensions. At a certain point, the accuracy is 100 percent, and it is stable at perfection regardless of how much higher the dimensions are. Therefore, we would like to argue that 40 dimensions are the most efficient dimension and, therefore, there is no need to go any higher than that because you do not meet higher accuracy at higher dimensions.

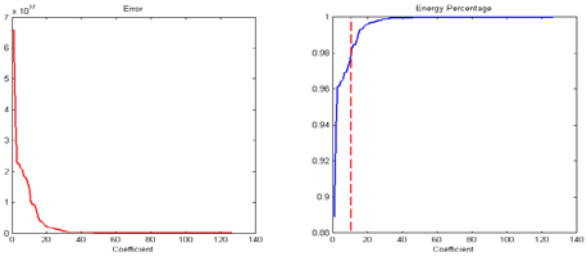


Fig. 9 The curve of energy percentage and the number of new space, Trade off between accuracy and number of dimensions in new space.

4. Conclusion

In this work we reviewed the first time series dimension reduction techniques, DFT and DWT, in the literature. Then we answered two questions that had been aroused. First, was which techniques have superior advantages or better results from the other? We compared with techniques with one another based on calculating time, results and ease of understanding. For this purpose, we implemented the abovementioned techniques on a time series data and using k-Means for clustering. We evaluated the results of clustering by Davis-Bouldin cluster validation technique. Collectively, the results point out that DFT is the most efficient technique because it has better results and lower time consumption. Second, was what number of dimensions is themost efficient in a new space? For attaining the number of dimensions that have proper results we propose Energy Percentage method.

References

[1] Keogh, E., Chakrabarti, K., Pazzani, M. & Mehrotra, S.: Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. Knowledge and Information Systems 3(3): pp. 263-286, (2000)

[2] Agrawal, R., Faloutsos, C., & Swami, A.: Efficient similarity search in sequence databases. Proc. of the 4th Conference on Foundations of Data Organization and Algorithms. (1993)

[3] Chan, K. & Fu, W.: Efficient Time Series Matching by Wavelets. Proceedings of the 15th International Conference on Data Engineering. (1999)

[4] Wu, D., Agrawal, D., El Abbadi, A. Singh, A. & Smith, T. R.: Efficient retrieval for browsing large image databases. Proc of the 5th International Conference on Knowledge Information. pp 11-18, (1996)

[5] Keogh, E. & Pazzani, M.: A simple dimensionality reduction technique for fast similarity search in large time series databases. Proceedings of Pacific- Asia Conf. on Knowledge Discovery and Data Mining, pp 122-133, (2000)

[6] Keogh, E., Chakrabarti, K., Pazzani, M. & Mehrotra, S.: Locally adaptive dimensionality reduction for indexing large time series databases. Proceedings of ACM SIGMOD Conference on Management of Data, May. pp 151-162, (2001)

[7] Keogh, E., Chu, S., Hart, D. & Pazzani, M.: An Online Algorithm for Segmenting Time Series. Proceedings of IEEE International Conference on Data Mining. pp 289-296. (2001)

[8] Lin J., Keogh E., Lonardi S., Chiu B.: A Symbolic Representation of Time Series, with Implications for Streaming Algorithms, Proc. DMKD 2003, pp. 2-11, San Diego California (USA), June 2003.

[9] Ratanamahatana, C., Keogh, E., Bagnall, T. and Lonardi, S.: A Novel Bit Level Time Series Representation with Implications for Similarity Search and Clustering. PAKDD 05. (2005)

[10] E. Keogh, J. Lin, and W. Truppel.: Clustering of Time Series Subsequences is Meaningless: Implications for Past and Future Research. Proceedings of the 3rd IEEE International Conference on Data Mining. Melbourne, FL. Nov 19-22. pp 115-122. (2003)

[11] Nong Ye : The Handbook Of Data Mining, Lawrence Erlbaum Associates Ltd,2003

[12] Agrawal R., Faloutsos C., and Swami A.: Efficient Similarity Search In Sequence Databases, In: Proc. FODO 1993, pp.69-84, Chicago, Illinois (USA), October 1993

[13] Oates T.: Identifying Distinctive Subsequences in Multivariate Time Series by Clustering, In: Proc.ACM SIGKDD 1999, pp. 322-326, San Diego, California (USA), August 1999.

[14] Liao T. W.: Clustering of time series data – a survey, In: Pattern Recognition, 38 pp1857-1874, (2005)

[15] Keogh E., Lonardi S., Chiu B. Y.: Finding surprising patterns in a time series database in linear time and space, In: Proc. ACM SIGKDD 2002, pp. 550-556, Edmonton, Alberta (Canada), July 2002.
[16] Staden R.: Methods For Discovering Novel Motifs, In: Nucleic Acid Sequences, Computer Applications in Biosciences, vol. 5, no. 4, pp. 293-298, October 1989.
[17] Lin J., Keogh E., Lonardi S., Patel P.: Finding Motifs in Time Series, In: Proc. 2nd workshop on Temporal Data Mining ACK SIGKDD 2002, pp. 53-68, Edmonton, Alberta (Canada), July 2002.
[18] Tanaka Y., Iwamoto K., Uehara K.: Discovery of Time Series Motif from Multi-Dimensional Data Based on MDL Principle, In: Machine Learning, vol. 58, no. 2-3, pp. 269-300, February 2005.
[19] Das G., Lin K.I., Mannila H., Ranganathan G., Smyth P.: Rule Discovery from Time series, In Proc. ACM SIGKDD 1998, pp. 16-22, New York, New York (USA), August 1998.
[20] Chakrabarti, K & Mehrotra, S.: Local dimensionality reduction: A new approach to indexing high dimensional spaces. In: Proceedings of the 26th Conference on Very Large Databases, Cairo, Egypt. (2000)
[21] Faloutsos, C., Ranganathan, M., & Manolopoulos, Y.: Fast subsequence matching in time-series databases. In: Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data. Minneapolis. (1994)
[22] Keogh E.: Data Mining and Machine Learning in Time Series Databases, In: Tutorial ECML/PKDD 2003, Cavtat-Dubrovnik (Croatia), September 2003.
[23] Faloutsos C.: Mining Time Series Data, In: Tutorial ICML 2003, Washington DC (USA), August 2003.
[24] E. Keogh, S. Chu, D. Hart and M. Pazzani: Segmenting Time Series: A Survey and Novel Approach, In: World Scientific Publishing Co. Pte. Ltd. pp 1- 21, (2004)
[25] E. Keogh: A Tutorial on Indexing and Mining Time Series Data, IEEE International Conference on Data Mining, (2001)
[26] M. Vlachos: A practical Time-Series Tutorial with MATLAB, In:16th European conference on machine learning and 9th European conference on principles and practice of knowledge discovery in databases, (2005)
[27] Ferenc Kovács, Csaba Legány and Attila Babos: Cluster Validity Measurement Techniques, In: 6th International Symposium of Hungrim Researchers on Computational Intelligence, (2005)
[28] Mu-Chun Su, Chien-Hsing Chou and Eugene Lai: A New Cluster Validity Measure for Clusters and its application to image compression, In: Pattern Analysis and Application , (2004)
[29] M. L. Hetland: A Survey of Recent Methods for Efficient Retrieval of Similar Time Sequences, pp 23-41, (2004)
[30] Mark Last, Yaron Klein, and Abraham Kandel: Knowledge Discovery in Time Series Databases, In: Systems, Man, and Cybernetics, Part B, IEEE Transactions,Volume 31, Issue 1, pp 160 – 169, Feb 2001
[31] Xavier Amatriain, Alejandro Jaimes, Nuria Oliver, and Josep M. Pujol , Data Mining Methods for Recommender Systems (2010)
[32] Jiawei Han, Micheline Kamber, Jian Pei , Data mining concepts and techniques 3rd, 2012
[33] Pyle, D., Data Preparation for Data Mining. Morgan Kaufmann, second edition (1999)
[34] G.N.Rramadevi, K.Usharani, Study on dimensionality reduction techniques and applications, Problems & Application in Engineering Research, (2013)
[35] Xavier Amatriain, Alejandro Jaimes, Nuria Oliver, and Josep M. Pujol , Data Mining Methods for Recommender Systems (2010)
[36] Imola K. Fodor, Chandrika Kamath ,Dimension reduction techniques and the classification of bent double galaxies (2002)

Saeid Bahadori is post graduate student in Information Technology Engineering at Tarbiat Moddares University, Iran. His research is about time serices datamining.

Nasrollah Moghadam Charkari is Associate Professor of computer science at Tarbiat Moddares University, Iran. His research is about image and video processing data mining sensor networks distributed and parallel computing.