Recent Techniques of Clustering of Time Series Data: A Survey

Sangeeta Rani
Department of Computer Science and Engineering,
Dr. B R Ambedkar National Institute of Technology,
Jalandhar
Punjab, India

Geeta Sikka
Department of Computer Science and Engineering,
Dr. B R Ambedkar National Institute of Technology,
Jalandhar
Punjab, India

ABSTRACT
Time-Series clustering is one of the important concepts of data mining that is used to gain insight into the mechanism that generate the time-series and predicting the future values of the given time-series. Time-series data are frequently very large and elements of these kinds of data have temporal ordering. The clustering of time series is organized into three groups depending upon whether they work directly on raw data either in frequency or time domain, indirectly with the features extracted from the raw data or with model built from raw data. In this paper, we have shown the survey and summarization of previous work that investigated the clustering of time series in various application domains ranging from science, engineering, business, finance, economic, health care, to government.

Keywords
Clustering, Time series data, Data mining, Dimensionality reduction, Distance measure.

1. INTRODUCTION
We Clustering of Time-Series data is the unsupervised classification of a set of unlabeled time series into groups or clusters where all the sequences grouped in the same cluster should be coherent or homogeneous. It can be shape-level if it is performed on the many individual time-series or structure-level if it works on single long-length time-series. The major issues related to time-series clustering are high dimensionality, temporal order and noise. Time-series clustering is Temporal-Proximity-Based Clustering if it works directly on raw data either in frequency or time domain; Representation-Based Clustering if it works indirectly with the features extracted from the raw data and Model-Based if it works with model built from raw data.

Han and Kamber [1] classified clustering methods developed for handing various static data into five major categories: partitioning methods, hierarchical methods, density based methods, grid-based methods, and model-based methods. Three of the five major categories of clustering methods for static data, specifically partitioning methods, hierarchical methods, and model based methods, have been utilized directly or modified for time series clustering.

Two key aspects for achieving effectiveness and efficiency when managing time series data are representation methods and similarity measures. Time series are essentially high dimensional data and directly dealing with such data in its raw format is very expensive in terms of processing and storage cost. It is thus highly desirable to develop representation techniques that can reduce the dimensionality of time series, while still preserving the fundamental characteristics of a particular data set. Many techniques have been proposed in the literature for representing time series with reduced dimensionality [2], such as Discrete Fourier Transformation, Single Value Decomposition, Discrete Cosine Transformation, Discrete Wavelet Transformation, Piecewise Aggregate Approximation, Adaptive Piecewise Constant Approximation, Chebyshev polynomials, Symbolic Aggregate approximation, Indexable Piecewise Linear Approximation etc.

In conjunction with this, there are over a dozen distance measures [2] for similarity of time series data in the literature, e.g., Euclidean distance (ED), Dynamic Time Warping (DTW), distance based on Longest Common Subsequence (LCSS) ,Edit Distance with Real Penalty (ERP) , Edit Distance on Real sequence (EDR) , DISSIM , Sequence Weighted Alignment model (Swale) , Spatial Assembling Distance (SpADe) and similarity search based on Threshold Queries (TQuEST) . Moreover, there must be some criteria on the basis of which the performance of time series clustering method must be evaluated. Two different categories of evaluation are given based on known ground truth and unknown ground truth [3].

The rest of this paper is organized as follows: Section 2 explains Major time series approaches in three subsections. Subsection 2.1 explains Taxonomy-based-approaches. In subsections 2.2 and 2.3 Representations-based and Model-based approaches are explained respectively. Section 3 concludes this paper.

2. MAJOR TIME SERIES CLUSTERING APPROACHES
Time series clustering methods have been divided into three major categories depending upon whether they work directly with raw data, indirectly with features extracted from the raw data, or indirectly with models built from the raw data.

2.1 Literature Survey of Temporal-Proximity-Based Clustering Approach
This approach usually 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 with an appropriate one for time series.

M. Kumar [25] proposed a distance function based on the assumed independent Gaussian models of data errors and used a hierarchical clustering method to group seasonality sequences into a desirable number of clusters. The experimental results based on simulated data and retail data showed that the new method outperformed both k-means and Ward's method that do not consider data errors in terms of (arithmetic) average estimation error. They assumed that data used have been preprocessed to remove the effects of non-

seasonal factors and normalized to enable comparison of sales of different items on the same scale.

T.W. Liao [26] developed a two-step procedure for clustering multivariate time series of equal or unequal length. The first step applies the k-means or fuzzy c-means clustering algorithm to time stripped data in order to convert multivariate real-valued time series into univariate discrete-valued time series. The converted variable is interpreted as state variable process. The second step employs the k-means or FCM algorithm again to group the converted univariate time series, expressed as transition probability matrices, into a number of clusters. The traditional Euclidean distance is used in the first step, whereas various distance measures including the symmetric version of Kullback–Liebler distance are employed in the second step.

T.W. Liao [27] applied several clustering algorithms including K-means, fuzzy c-means, and genetic clustering to multivariate battle simulation time series data of unequal length with the objective to form a discrete number of battle states. The original time series data were not evenly sampled and made uniform by using the simple linear interpolation method.

C. S. Möller-Level [28] , in their study of DNA Microarray data, proposed short time series (ST S) distance to measure the similarity in shape formed by the relative change of amplitude and the corresponding temporal information of uneven sampling intervals. All series are considered sampled at the same time points. By incorporating the STS distance into the standard fuzzy c-means algorithm, they revised the equations for computing the membership matrix and the prototypes (or cluster centers), thus developed a fuzzy time series clustering algorithm.

Shumway [29] proposed the clustering of nonstationary time series by applying locally stationary versions of Kullback–Leibler discrimination information measures that give optimal time–frequency statistics for measuring the discrepancy between two non-stationary time series. To distinguish earthquakes from explosions, an agglomerative hierarchical cluster analysis was performed until a final set of two clusters was obtained.

Vit Niennattrakul and Chotirat Ann Ratanamahatana for the clustering of multimedia time series [5] applied K-means and K- medoids algorithms with dynamic time warping and demonstrated that K-means is much more generic clustering method when Euclidean distance is used, but it failed to give correct results when dynamic time warping is used as distance measure in averaging the shape of the time series. As the results of their experiments, they have confirmed that dynamic time warping should not be used as subroutine with K-means algorithm and K-medoids with dynamic time warping gives satisfactory results.

For clustering time series gene expression data Pooya Sobhe Bidari [6] presented two phase functional clustering as a new approach in gene clustering. The proposed approach is based on finding functional patterns of time series using Fuzzy C-Means and K-means algorithms.

Pearson correlation similarity measure is used to extract the expression patterns of genes. In this approach, genes are clustered by K-means and FCM methods according to theirs time series expression, then patterns of gene behavior are extracted. Then, new features are defined for the genes and by calculating Pearson correlation between new feature vectors, genes with similar expression behavior are obtained which can lead to find interconnections between genes.

For detecting climate change in multivariate data Hardy Kremer [7] proposes novel clustering and clustering tracing techniques. In this novel clustering approach, time series is split into disjoint, equal length intervals and then density based subsequence clustering approach is applied, and dynamic time warping is used as a distance measure.

Jian Yin [9] proposed a clustering algorithm for time series data. He proposed a encoded-bitmap-approach-based swap is used to improve the classical hierarchical method. Traffic flow data is used as time series and grey relation is used a similarity measure. After getting K clusters, encoded-bitmap-approach based swap is used to refine the K clusters and get the new K clusters. Experiments show that the proposed method has a better performance on the change trend of time series than classic algorithm.

Ville Hautamaki and Pekka Nykanen [10] defined an optimal prototype as an optimization problem and proposed a local search solution to it. They applied two Euclidean space clustering methods to time-series clustering: random swap and hierarchical clustering followed by k-means fine-tuning and it provided 10-22% improvements to k-medoids.

S. Chandrakala and C. Chandra Sekhar [11] proposed a density based method for clustering of multivariate time series of variable length in kernel feature space. Kernal DBSCAN algorithm with Euclidean distance measure is used. They presented heuristic methods to find the initial values of the parameters used in our proposed algorithm. The performance of the proposed method is compared with the spectral clustering and kernel k-means clustering methods. Besides handling outliers, the proposed method performs as well as the spectral clustering method and outperforms the kernel k-means clustering method.

Dacheng Nie [13] analyzed time series by using NLCS (Normalized Longest Common Subsequence). NLCS is a similarity measurement widely used in comparing character sequences. In this paper, he developed the NLCS and presented a novel algorithm to precisely calculate the similarity of time series. The algorithm used the sum of all common subsequence instead of longest common subsequence which can't represent the similarity of sequences accurately. The experiments based on synthetic and real-life datasets showed that the proposed algorithm performed better in comparing the similarity of time series. Comparing with Euclidean distance on four cluster validity indices, the results lead to a better performance by k-means or self-organize map.

Table 2.1 Summary of Temporal-Proximity-Based Clustering Approach

Paper	Distance Measure	Algorithm	Application
M. Kumar	Based on the assumed independent Gaussian models of data errors	Agglomerative Hierarchical	Seasonality pattern in retails
T.-W. Liao	Euclidean and symmetric version of Kullback–Liebler distance	K-Means and Fuzzy C-Means	Battle simulations
T.-W. Liao	Dynamic Time Warping	K-Medoids Based Genetic Clustering	Battle simulations
C.S. Möller-Levet	Short time series (STS) distance	Modified Fuzzy C-Means	DNA microarray
Shumway	Kullback–Leibler discrimination information measure	Agglomerative Hierarchical	Earthquakes and mining explosions
Vit Niennattrakul	Dynamic Time Warping	K-Means, K-Medoids	Multimedia time series
Pooya Sobhe Bidari	Pearson Correlation	K-Means, Fuzzy C-Means	Pattern extraction in genes
Hardy Kremer	Dynamic Time Warping	Density Based Subsequence Clustering	Detecting climate change
Jian Yin	Grey Relation	Hierarchical Clustering	Change trend of traffic flow data
S. Chandrakala	Euclidean	Kernal DBScan	Multivariate time series clustering
Aurangzeb Khan	Euclidean	K-Mean+ MFP(Most Frequent Pattern)	Stock and inventory data
Mengfan Zhang	CVT(Computational Verb Theory)	K-Means	Stock market data
S.R.Nanda	Euclidean	K-Means	Portfolio management
Jianfei Wu	N/A	K-Means	Stock data

Aurangzeb Khan [16] used hybrid clustering algorithm to mine the frequent pattern in the stock or inventory data. He proposed an algorithm for mining patterns of huge stock data to predict factors affecting the sale of products. In the first phase of his method, he applied k-means algorithm to divide the stock data into three different clusters i.e. Dead Stock (DS), Slow Moving (SM) and Fast Moving (FM) on the basis of product categories. In the second phase, he applied Most Frequent Pattern (MFP) algorithm to find frequencies of property values of the corresponding items. MFP provides frequent patterns of item attributes in each category of products and also gives sales trend in a compact form. The experimental result showed that the proposed hybrid k-mean plus MFP algorithm can generate more useful pattern from large stock.

Songpol Ongwattanakul [15] and Dararat Srisai introduced a new variation of Dynamic time warping distance measure for time series shape averaging classification. According to them resampled DTW and Hybrid DTW give better accuracy and high performance than original DTW but to improve the accuracy further they introduced Contrast Enhanced Dynamic Time Warping (CEDTW) that reduces the effect from data points that have non-trivial contribution to the measured distance and improves the accuracy in similarity measure.

Xueyan WU [17] proposed a method of data stream clustering for stock data analysis. The method aimed to retain shape and tend features during the clustering process. He divided the process of the proposed method into two parts i.e. online clustering and offline macro clustering. Online clustering extracted data flow characteristics and maintains the clustering feature vectors and offline macro clustering is the process which responded to user requests and achieved clustering. Experiments showed that the method had good results.

Aurangzeb Khan [16] used hybrid clustering algorithm to mine the frequent pattern in the stock or inventory data. He proposed an algorithm for mining patterns of huge stock data to predict factors affecting the sale of products. In the first phase of his method, he applied k-means algorithm to divide the stock data into three different clusters i.e. Dead Stock (DS), Slow Moving (SM) and Fast Moving (FM) on the basis of product categories. In the second phase, he applied Most

Frequent Pattern (MFP) algorithm to find frequencies of property values of the corresponding items. MFP provides frequent patterns of item attributes in each category of products and also gives sales trend in a compact form. The experimental result showed that the proposed hybrid k-mean plus MFP algorithm can generate more useful pattern from large stock.

Xueyan WU [17] proposed a method of data stream clustering for stock data analysis. The method aimed to retain shape and tend features during the clustering process. He divided the process of the proposed method into two parts i.e. online clustering and offline macro clustering. Online clustering extracted data flow characteristics and maintains the clustering feature vectors and offline macro clustering is the process which responded to to user requests and achieved clustering. Experiments showed that the method had good results.

Mengfan Zhang [18] and Tao Yang applied the computational verb theory (CVT) to analyze the stock market data. His goal was to find the most representative trends in the intra-day stock market data. First round of computational verb clustering algorithm was used to categorize the stock market data and in the second round of computational verb k-means clustering algorithm is used to refine the representative trends in the stock market data. Experiments showed that the applied method yielded the most representative curves in the stock market data.

Jianfei Wu [19] introduced an algorithm that used stock sector information directly in conjunction with time series subsequences for mining core patterns within the sectors of stock market data. He used the stream sliding window concepts. Two adjacent sliding windows were used, namely training window and evaluation window. The algorithm detected significant sectors in the training window, and built core patterns for the significant ones. The algorithm identified whether a stock sector currently shows coherent behavior. When coherent behavior of a stock sector was detected, core patterns were extracted. The core patterns were more stable than clusters found by some clustering algorithm DBScan. Through comparing with DBScan, we show the effectiveness of the proposed algorithm.

Huawang Shi [20] proposed a novel unascertained C-means clustering algorithm. He used the theory and method of unascertained measure and established clustering weights and a novel unascertained C-means clustering algorithm. Experimental results showed that the proposed method performed more robust to noise than the fuzzy C-means (FCM) clustering algorithm do.

S.R.Nanda [23] applied clustering to stock market data for portfolio management. She used stock returns at different times along with their valuation ratios and results analysis showed that k-means cluster analysis builds the most compact cluster as compared to SOM and Fuzzy C-means for stock classification data. She then selected stocks from clusters to build portfolio, minimizing portfolio risk and compared the returns with that of benchmark index.

2.2 Literature Survey of Representation-Based Clustering Approach

It is not easy to work directly with the raw data that are highly noisy. Feature based approach first converts a raw time series data into a feature vector of lower dimension and then clustering algorithms are applied.

T.-C. Fu [30] described the use of self-organizing maps for grouping data sequences segmented from the numerical time series using a continuous sliding window with the aim to discover similar temporal patterns dispersed along the time series. They introduced the perceptually important point (PIP) identification algorithm to reduce the dimension of the input data sequence in accordance with the query sequence. The distance measure between the PIPs found was defined as the sum of the mean squared distance along the vertical scale (the magnitude) and that along the horizontal scale (time dimension). To process multi resolution patterns, training patterns from different resolutions are grouped into a set of training samples to which the SOM clustering process is applied only once. Two enhancements were made to the SOM: filtering out those nodes (patterns) in the output layer that did not participate in the recall process and consolidating the discovered patterns with a relatively more general pattern by a redundancy removal step.

Dong Jixue [8] for mining the financial time series uses the wave cluster, which is a kind of grid cluster and the density cluster unify. In this, basic details and methods of phase space reconstruction are analyzed in details. All of these provided the theoretical basis and technical feasibility to time series data mining based on phase space reconstruction. After contrasting the different means of time series pattern mining, the problem of Time Series Data Mining framework TSDM is pointed out, and the temporal patterns mining method based Wave cluster is systematically presented. By the multiresolution property of wavelet transformations and the grid-based partition method, it could detect arbitrary-shape clusters at different scales and levels of detail.

Huiting Liu [12] proposed a new similar pattern matching method. Firstly, trends of time series are extracted by empirical mode decomposition, and the trends are translated into vectors to realize dimension reduction. Secondly, the vectors are clustered by a forward propagation learning algorithm. Finally, all the series that is similar with the query are found by calculating Euclidean distance between the query and the series that belong to the same category with it. Experimental results showed that EMD outperforms the FFT (Fast Fourier Transform) when they are used to reduce dimension.

Table 2.2 Summary of Representation-Based Clustering Approach

Paper	Features	Distance Measure	Clustering Algorithm	Application
T.-C. Fu	Perceptually important points	Sum of the mean squared distance along the vertical and horizontal scales	Modified SOM	Hong Kong stock market
M. Vlachos	Haar wavelet transform	Euclidean	Modified k-means	Non-specific
Huiting Liu	Empirical mode decomposition	Euclidean	Forward propagation learning algorithm	Non-specific
Chonghui GUO	Independent component analysis	Euclidean	Modified k-means	Real world stock time-series
Jian Xin Wu	Independent component analysis	N/A	support vector regression	Financial time-series
Geert Verdoolaege	Wavelet transform	Kullback- Liebler divergence	k-means	Detection of activated voxels in FMRI data
Liu Suyi	Hough transform	N/A	Mean shift algorithm	Feature recognition of underwater images
Dong Jixue	Wavelet transform	N/A	Grid-based partitioning method	Financial time-series

M. Vlachos [31] presented an approach to perform incremental clustering of time series at various resolutions using the Haar wavelet transform. First, the Haar wavelet decomposition is computed for all-time series. Then, the k-means clustering algorithm is applied, starting at the coarse level and gradually progressing to finer levels. The final centers at the end of each resolution are reused as the initial centers for the next level of resolution. Since the length of the data reconstructed from the Haar decomposition doubles as we progress to the next level, each coordinate of the centers at the end of level i is doubled to match the dimensionality of the points on level i+1. The clustering error is computed at the end of each level as the sum of number of incorrectly clustered objects for each cluster divided by the cardinality of the dataset.

Nicole Powell [14] compared unsupervised classification techniques such as k-means clustering with supervised classification techniques such as support vector machines for stock prices forecasting .He used Principal component analysis to reduce the dimension of the data set to select the component which have the biggest effect and concluded that for this application both method give comparable results but unsupervised classification techniques are better for stock trend forecasting because unsupervised methods fine pattern in data that is usually seen as uncorrelated.

Anthony J. T .Lee [24] presented an effective approach (Hierarchical agglomerative and recursive k-means clustering) to stock market prediction. The proposed method converted each financial report to feature vector and used hierarchical agglomerative clustering to divide this feature vector into clusters and then, for each sub-cluster so that most feature vectors in each sub-clusters belonged to the same class. Then, for each sub-cluster, a centroid was chosen as the representative feature vector and finally this feature vector was employed to predict the stock price movements.

Chonghui GUO [22] presented a novel feature based approach to time series clustering which first converted the raw time series into feature vector of lower dimension by using ICA (Independent Component Analysis) algorithm and then applied a modified k-means clustering algorithm to the extracted feature vector . Finally to validate the feasibility and effectiveness of the proposed approach he used it to analyze the real world stock time series and achieved the reasonable results.

Jian Xin Wu [32] proposed a combination of ICA (independent component analysis) with SVR (support vector regression) for predicting the financial time series. The proposed method used SRM (structure risk minimization) principle. It removed the problem of learning algorithms based on ERM (empirical risk minimization) principle, which always have good fit for the training samples, but bad prediction for future samples. He used non-linear SVR (Support vector regression) for the prediction, and before applying SVR, he used ICA for the feature extraction. ICA considers independence which is a more strict condition than PCA which takes into account uncorrelated between features.

Geert Verdoolaege [33] proposed a new method for the detection of activated voxels in event related BOLD FMRI data. Firstly, he derived wavelet histograms from each voxel

time series and modeled the derived statistics through GGD (generalized Gaussian distribution). Finally, he performed the K-Means clustering of the GGD's characterizing the voxel data in a synthetic data set using the KLD (Kullback- Liebler divergence) as a similarity measure.

The main issue of similarity search in time series database is to improve the search performance since time series data is usually is of high dimension and for this is important to reduce the search space for the efficient processing of similarity search. In this paper, D.Muruga Radha Devi [34], proposed a combination of using Vari-DWT and Polar wavelet. Vari-DWT is fast to compute and requires little storage for each sequence; it preserves Euclidean distance and allows good approximation with a subset of coefficients. But its drawbacks are, it shows poor performance for locally distributed time series data since it uses averages to reduce the dimensionality of the data and another limitation is it works best when the length of the time series is 2^n, and hence becomes the reason of using Polar wavelet, it uses polar co-ordinates which are not affected from averages and, it works with the time sequences of any length without distorting the original signal. She evaluated the effectiveness of this approach by using real weather data and synthetic datasets.

Liu Suyi [35] did feature recognition for underwater images. Firstly, the preprocessing of underwater image was done to improve image quality so that seam feature could be recognized easily. Weld featured image was effectively segmented by Mean Shift Algorithm, and finally Hough Transform was used to recognize the feature of underwater weld images.

2.3 Literature Survey of Model-Based Clustering Approach

This class of approaches considers that each time series is generated by some kind of model or probability distributions. Time series are considered similar when the models characterizing individual series or the remaining residuals after fitting the model are similar.

Baragona [36] evaluated three meta-heuristic methods for partitioning a set of time series into clusters in such a way that (i) the cross-correlation maximum absolute value between each pair of time series that belong to the same cluster is greater than some given threshold, and (ii) the k-min cluster criterion is minimized with a specified number of clusters. The cross-correlations are computed from the residuals of the models of the original time series. Among all methods evaluated, Tabu search was found to perform better than single linkage, pure random search, simulation annealing and genetic algorithms based on a simulation experiment on ten sets of artificial time series generated from low-order univariate and vector ARMA models.

K. Kalpakis [37] studied the clustering of ARIMA time series, by using the Euclidean distance between the Linear Predictive Coding cepstra of two time-series as their dissimilarity measure. The cepstral coefficients for an AR*(p)* time series are derived from the auto-regression coefficients. The partition around medoids method that is a *k*-medoids algorithm was chosen, with the clustering results evaluated with the cluster similarity measure and Silhouette width. Based on a test of four data sets, they showed that the LPC cestrum provides higher discriminatory power to tell one time series from another and superior clustering than other widely used methods such as the Euclidean distance between (the first 10 coefficients of) the DFT, DWT, PCA, and DFT of the auto-correlation function of two time series.

Xiong and Yeung [38] proposed a model-based method for clustering univariate ARIMA series. They assumed that the time series are generated by *k* different ARMA models, with each model corresponds to one cluster of interest. An expectation-maximization (EM) algorithm was used to learn the mixing coefficients as well as the parameters of the component models that maximize the expectation of the complete-data log-likelihood. In addition, the EM algorithm was improved so that the number of clusters could be determined automatically.

L. Wang [39] presented a framework for tool wear monitoring in a machining process using discrete hidden Markov models. The feature vectors are extracted from the vibration signals measured during turning operations by wavelet analysis. The extracted feature vectors are then converted into a symbol sequence by vector quantization, which in turn is used as input for training the hidden Markov model by the expectation maximization approach. Yun Yang and Ke Chen [4] presented an unsupervised ensemble learning approach to time series clustering by combining RPCL (rival-penalized competitive learning) with different representations. This approach first exploits its capability of RPCL rule in clustering analysis of automated model selection on individual representations and then applies ensemble learning for the synergy of reconciling diverse partitions resulted from the use of different representations and augmenting RPCL network in automatic model selection and overcoming its inherent limitation. They evaluated their approach on 16 benchmark time series data mining task. Simulation results demonstrated that their approach yielded favorite results in clustering analysis of automatic model selection.

Yu-Chia Hsu [4] proposed a approach using self organizing map (SOM) for time series data clustering and similar pattern recognition to improve the optimal hedge ratio (OHR) estimation. Taiwan stock market hedging is used. Five SOM based models and two traditional models were compared in this approach. Experiments demonstrated the SOM approach provides a useful alternative to the OHR estimation.

Xin Huang proposed [40] a research on predicting agriculture drought based on fuzzy set and R/S analysis model. He used fuzzy clustering iteration method to cluster the data of many years rainfall and then considered sensitiveness coefficient as the foundation of calculating weight, which affected the crop output by valid rainfall in each growth stage. The results of application showed that the model is convenient and feasible in the application of forecasting the years of occurrence in agriculture drought.

Yupei Lin [41] tried to improve the prediction accuracy with correcting two deficiencies, sub intervals failing to well represent the data distribution structures and a single antecedent factor in the fuzzy relationship in current fuzzy time series model. First, he partitioned the universe of discourse in subintervals with the midpoints of two adjacent clusters centers, and the subintervals are employed to fuzzify the time series into fuzzy time series. Then, the fuzzy time series model with multi factors high order fuzzy relationship is built-up to forecast the stock market. The results showed that the model improved the prediction accuracy when compared with the benchmark model.

Shan Gao [42] analyzed ARCH (Autoregressive Conditional Heteroscedasticity) effects of wind data series with Eviews

software. Firstly, he built an ARMA (Autoregressive Moving Average) model of wind speed time series and, tested the ARCH effect of residual ARMA Model by Lagrange Multiplier. Lastly, he compared the forecasting performances of ARMA-ARCH model with ARMA model and proved that ARMA-ARCH model possesses higher accuracy.

Table 2.3 Summary of Model-Based Clustering Approach

Paper	Model	Distance measure	Clustering algorithm	Application
Baragona	ARMA	Cross-correlation based	Tabu search, GA and	Non-specific
K. Kalpakis	AR	Euclidean	Partitioning around medoids	Public data
Xiong and Yeung	ARMA mixture	Log-liklihood	EM learning	Public data
L. Wang	Discrete HMM	Log-liklihood	EM learning	Tool condition monitoring
Xin Huang	Fuzzy set and R/S analysis model	N/A	Fuzzy clustering iteration method	Predicting agriculture drought
Shan Gao	ARMA-ARCH	N/A	N/A	To analyze the effects of wind data series

3. CONCLUSIONS

In recent years, there has been variety of interests in mining time series. Particularly, the clustering of time series has attracted the interest of researchers. As Time series data are frequently large and may contain outliers, therefore careful examination of the proposed algorithms is necessary. In this paper we have studied most recent techniques on the subject of time series clustering. The uniqueness and limitation of previous research are discussed and several possible topics for future research are identified. It is hoped that this review will serve as the steppingstone for those interested in advancing this area of research.

4. REFERENCES

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.

[2] H.Ding, " Querying and Mining of Time Series Data: experimental comparison of representations and distance measures". Proceedings of the VLDB Endowment VLDB Endowment Hompage archive Volume 1 Issue 2, August 2008, pp 1542-1551.

[3] T.W. Liao, Clustering of time series data—survey, Pattern Recognition 38 (2005), pp. 1857–1874.

[4] Y. Yang and K. Chen," Time-Series Clustering via RPCL Network ensemble with different representations", IEEE Transactions On Systems, Man, And Cyzbernetics—Part C: Applications And Reviews, Vol. 41, No. 2, March 2011, pp. 190-199.

[5] V. Niennattrakul; C.A. Ratanamahatana , "On Clustering Multimedia Time Series Data Using K-Means and Dynamic Time Warping," Multimedia and Ubiquitous Engineering, 2007. MUE '07. International Conference on , vol., no., pp.733-738, 26-28 April 2007.

[6] P. Sobhe Bidari ; R. Manshaei ; T. Lohrasebi; A. Feizi; M.A. Malboobi,; J. Alirezaie; , "Time series gene expression data clustering and pattern extraction in Arabidopsis thaliana phosphatase-encoding genes,"BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on , vol., no., pp.1-6, 8-10 Oct. 2008.

[7] H. Kremer; S. Gunnemann; T. Seidl,; , "Detecting Climate Change in Multivariate Time Series Data by Novel Clustering and Cluster Tracing Techniques," Data Mining Workshops (ICDMW), 2010 IEEE International Conference on , vol., no., pp.96-97, 13-13 Dec. 2010

[8] D. Jixue; , "Data Mining of Time Series Based on Wave Cluster," Information Technology and Applications, 2009. IFITA '09. International Forum on , vol.1, no., pp.697-699, 15-17 May 2009.

[9] J. Yin; D. Zhou; Q.-Q. Xie; , "A Clustering Algorithm for Time Series Data," Parallel and Distributed Computing, Applications and Technologies, 2006.

PDCAT '06. Seventh International Conference on , vol., no., pp.119-122, Dec. 2006

[10] V. Hautamaki; P.Nykanen; P. Franti; , "Time-series clustering by approximate prototypes," Pattern Recognition, 2008. ICPR 2008. 19th International Conference on , vol., no., pp.1-4, 8-11 Dec. 2008

[11] S. Chandrakala; C. Chandra Sekhar; , "A density based method for multivariate time series clustering in kernel feature space," Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on , vol., no., pp.1885-1890, 1-8 June 2008.

[12] H. Liu; Z. Ni; J. Li; , "Time Series Similar Pattern Matching Based on Empirical Mode Decomposition," Intelligent Systems Design and Applications, 2006. ISDA '06. Sixth International Conference on , vol.1, no., pp.644-648, 16-18 Oct. 2006.

[13] N. Dacheng; F. Yan; Z. Junlin; F. Yuke; X. Hu; , "Time series analysis based on enhanced NLCS," Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference on, vol., no., pp.292-295, 23-25 June 2010.

[14] N. Powell; S.Y. Foo; M. Weatherspoon; , "Supervised and Unsupervised Methods for Stock Trend Forecasting," System Theory, 2008. SSST 2008. 40th Southeastern Symposium on , vol., no., pp.203-205, 16-18 March 2008.

[15] S. Ongwattanakul; D. Srisai;," Contrast Enhanced Dynamic Time Warping Distance for Time Series Shape Averaging Classification" ICIS 2009, November 24-26, 2009 Seoul, Korea Copyright © 2009 ACM 978-1-60558-710-3/09/11.

[16] A. Khan; K. Khan; B.B. Baharudin; , "Frequent Patterns Minning of Stock Data Using Hybrid Clustering Association Algorithm," Information Management and Engineering, 2009. ICIME '09. International Conference on , vol., no., pp.667-671, 3-5 April 2009.

[17] X. Wu; D. Huang;, "Data stream clustering for stock data analysis," Industrial and Information Systems (IIS), 2010 2nd International Conference on , vol.2, no., pp.168-171, 10-11 July 2010.

[18] M. Zhang; T. Yang; , "Application of computational verb theory to analysis of stock market data," Anti-Counterfeiting Security and Identification in Communication (ASID), 2010 International Conference on , vol., no., pp.261-264, 18-20 July 2010.

[19] W. Jianfei; A. Denton; O. Elariss; X. Dianxiang; , "Mining for Core Patterns in Stock Market Data," Data Mining Workshops, 2009. ICDMW '09. IEEE International Conference on , vol., no., pp.558-563, 6-6 Dec. 2009.

[20] H. Shi; "A Novel Unascertained C-Means Clustering with Application," Intelligent Computation Technology and Automation, 2009. ICICTA '09. Second International Conference on , vol.1, no., pp.134-137, 10-11 Oct. 2009.

[21] Y.-C. Hsu; A.-P. Chen; , "Clustering Time Series Data by SOM for the Optimal Hedge Ratio Estimation," Convergence and Hybrid Information Technology, 2008. ICCIT '08. Third International Conference on , vol.2, no., pp.1164-1169, 11-13 Nov. 2008

[22] C. Guo; H. Jia; N. Zhang; , "Time Series Clustering Based on ICA for Stock Data Analysis," Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on , vol., no., pp.1-4, 12-14 Oct. 2008.

[23] S.R Nanda;B Mahanty;M.K tiwari; Clustering Indian stock market data for portfolio management "published in journal Experts Systems with Application: An international journal archive Volume 37 Issue 12,December, 2010.

[24] Anthony J.T. Lee; M.-C. Lin; R.-T. Kao; and K.-T. Chen, , "An Effective Clustering Approach to Stock Market Prediction" (2010). PACIS 2010 Proceedings. Paper 54.

[25] M. Kumar, N.R. Patel, J. Woo, Clustering seasonality patterns in the presence of errors, Proceedings of KDD '02, Edmonton, Alberta, Canada.

[26] T.W. Liao, Mining of vector time series by clustering, Working paper, 2005.

[27] T.W. Liao, B. Bolt, J. Forester, E. Hailman, C. Hansen, R.C. Kaste, J. O'May, Understanding and projecting the battle state, 23rd Army Science Conference, Orlando, FL, December 25,2002.

[28] C.S. Möller-Levet, F. Klawonn, K.-H. Cho, O. Wolkenhauer, Fuzzy clustering of short time series and unevenly distributed sampling points, Proceedings of the 5th International Symposium on Intelligent Data Analysis, Berlin, Germany, August 28–30, 2003.

[29] R.H. Shumway, Time–frequency clustering and discriminant analysis, Stat. Probab. Lett. 63 (2003) 307–314.

[30] T.-C. Fu, F.-L. Chung, V. Ng, R. Luk, Pattern discovery from stock time series using self organizing maps, KDD 2001 Workshop on Temporal Data Mining, August 26–29, San Francisco, 2001, pp. 27–37.

[31] M. Vlachos, J. Lin, E. Keogh, D. Gunopulos, A waveletbased anytime algorithm for *k*-means clustering of time series, Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, May 1–3, 2003.

[32] J. X. Wu, J. L. Wei, "Combining ICA with SVR for prediction of finance time series", Proceedings of the IEEE International Conference on Automation and Logistics August 18 - 21, 2007, Jinan, China, pp 95-100.

[33] G. Verdoolaege and Y. Rosseel," Activation Detection In Event-Related Fmri Through Clustering Of wavelet Distributions", Proceedings of 2010 IEEE 17th International Conference on Image Processing, September 26-29, 2010, Hong Kong, pp 4393-4395.

[34] D.M.R. Devi; V. Maheswari; P. Thambidurai; , "Similarity search in Recent Biased time series databases using Vari-DWT and Polar wavelets," Emerging Trends in Robotics and Communication Technologies (INTERACT), 2010 International Conference on , vol., no., pp.398-404, 3-5 Dec. 2010

[35] L. Suyi, Z. Hua," Feature Recognition for Underwater Weld Images", Proceedings of the 29th Chinese Control

Conference, July 29-31, 2010, Beijing, China, pp 2729-2734.

[36] R. Baragona, A simulation study on clustering time series with meta-heuristic methods, Quad. Stat. 3 (2001) 1–26.

[37] K. Kalpakis, D. Gada, V. Puttagunta, Distance measures for effective clustering of ARIMA time-series, Proceedings of the 2001 IEEE International Conference on Data Mining, San Jose, CA, November 29–December 2, 2001, pp. 273–280.

[38] Y. Xiong, D.-Y. Yeung, Mixtures of ARMA models for model-based time series clustering, Proceedings of the IEEE International Conference on Data Mining, Maebaghi City, Japan, 9–12 December, 2002.

[39] L. Wang, M.G. Mehrabi, E. Kannatey-Asibu Jr ., Hidden Markov model-based wear monitoring in turning, J. Manufacturing Sci. Eng. 124 (2002) 651–658.

[40] X. Huang, H.-l. LI," Research on Predicting Agricultural Drought Based on Fuzzy Set and RlS Analysis Model", 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE),pp 186-189.

[41] Y. Lin; Y. Yang;, "Stock markets forecasting based on fuzzy time series model," Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on, vol.1, no., pp.782-786, 20-22 Nov. 2009.

[42] S. Gao; Y. He; H. Chen;, "Wind speed forecast for wind farms based on ARMA-ARCH model," Sustainable Power Generation and Supply, 2009. SUPERGEN '09. International Conference on, vol., no., pp.1-4, 6-7 April 2009.