CHAPTER -2
2.1 INTRODUCTION

Time series arises in many important areas. Some examples include the evolution of stock price in CES, share prices or commodity prices, the sales figures for a particular good or service, micro economic and demographic indicator, image sequencing, acceleration measures of a sensor on a bridge, ECG or EEG recordings, gene expression measurements, meteorological studies, pollution control and monitoring etc., (Jeroen Boets et al. (2005)). Time Series data is perhaps the most commonly encountered kind of data explored by data miners. Supervised Learning and Unsupervised Learning algorithms are the most frequently used data mining techniques being useful in exploratory data analysis methods. Hence, time series data mining has attracted an extraordinary amount of attention.

Jiawei Han and Micheline Kamber (2001) classified clustering methods for handling various static data into five major categories namely 1) Partitioning methods 2) Hierarchical methods 3) Density based methods 4) Grid based methods and 5) Model based methods. Time series clustering requires a clustering algorithm or procedure to form clusters on a given set of unlabeled data objects and the choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. As far as time series data are concerned, distinctions can be made as to whether the data are discrete-valued or real-valued, uniformly or non-uniformly sampled, univariate or multivariate and whether the data series are of equal or unequal length. Non-uniformly sampled data must be converted into uniformed data before clustering operations can be performed. This can be achieved by a wide
range of methods, from simple down sampling based on the roughest sampling interval to a sophisticated modeling and estimation approach.

Various algorithms have been developed to clustering and classification of different type of time series data. Several authors developed various algorithms for time series data mining. Eamonn Keogh and Shruti Kasetty (2002) and T. Warren Liao (2005) have reviewed the literature available on time series data mining. Putting their difference aside, it is far to say that in spirit they all try to modify the existing algorithms for clustering static data in such a way that time series data can be handled or to convert it into the form of static data so that the existing algorithms for clustering static data can be directly used (T. Warren Liao (2005)). The former approach usually works directly with raw time series data is called raw-data based approach and the major modification lies in replacing the distance/similarity measure for static data with appropriate one for time series. The later approach first converts a raw time series data either into a feature vector of low dimension or a number of model parameters and then applies a conventional clustering algorithm to the extracted feature vectors or model parameters, called feature- and model-based approach.

Among these three methods, model based approaches for time series data mining are more efficient compared to other methods. Hence in this chapter, some contributions in time series data mining based on model based approaches are reviewed in order to highlight the present work in its right perspective. The available literature on model based data mining in time series can be classified in to 3 categories. They are 1) Time series data mining based on ARIMA models 2) Time
series data mining based on neural networks and 3) Time series data mining based on other models.

2.2 REVIEW ON TIME SERIES DATA MINING BASED ON ARIMA MODELS

In this section, the review of some contributions on time series data mining based on autoregressive integrated moving average model is presented. The autoregressive integrated moving average model is a generic one which include stationary and non stationary time series as particular cases. The stationary models like ARMA, AR and MA are particular cases of ARIMA (p, d, q) when the parameter d=0.

Elizabeth Ann Maharaj (1996) stated that ARMA models can be used to classify stationary time series into groups when the Euclidean distance between the parameter estimates of autoregressive expansions are given. A test of hypothesis is given to determine whether two stationary series in a particular group have significantly different generating processes. They developed a clustering algorithm.

H. Brian Hwarng and H.T. Ang (2001) designed a study to investigate a neural network solute a class of time series corresponding to ARIMA (p, d, q) structures. They also studied the significance of matching the input window size with the nature of time series. The study adopted a simulation approach in conjunction with an experimental design. Konstantinos Kalpakis, Dhiral Gada and Vasundhara Puttagunta (2001) considered the linear predictive coding (LPC) cepstrum of time series for clustering ARIMA time series, by using Euclidean distance between the LPC cepstra of two time series as their dissimilarity measure. The clustering of time
series was done using Partition Around Medoids method (PAM) with various similarity measures.

Lon-Mu Liu, Sidhartha Bhattacharyya, Stanley L. Sclove, Rong Chen and William J. Lattyak (2001) employed Box-Jenkins seasonal ARIMA models to analyze and forecast the time series. They employed a univariate ARIMA models to forecast the demand. They aggregated the data into daily intervals presented in time in order to forecast daily demand of ingredients so as to facilitate better inventory management.

Dingfei Ge, Narayanan Srinivasan and Shankar M Krishnan (2002) developed an AR model and a GLM for classification and shown that they are effective for the classification of cardiac arrhythmias in critically ill patients and aid in the diagnosis of heart disease. They suggested AR model and GLM models are suitable for real-time implementations and can be used for compression as well as diagnosis. Fang-Mei Tseng and Gwo-Hshiung Tzeng (2002) developed a fuzzy seasonal ARIMA forecasting model which combines the advantage of seasonal time series ARIMA model and fuzzy regression model. It is used to forecast two seasonal time series data of total production values of the Taiwan machinery industry and soft drink time series. The FSARIMA is useful for decision makers to forecast the best and worst estimates based on fewer observations than the SARIMA model.

G. Peter Zhang (2003) considered a hybrid methodology that combines both ARIMA and ANN models to take the advantage of unique strength of ARIMA and ANN models in linear and non linear modeling. The experimental results proved that the combined model performs better than either of the models used separately. Yi...
Xiong and Dit-Yan Yeung (2004) studied the clustering of data patterns that are represented as sequences or time series possibly of different lengths. They developed a model-based approach using mixtures of ARMA models. EM algorithm for univariate data has been derived for learning the mixing coefficients and parameters of the component models. The model selection problem was addressed by using Bayesian information criterion (BIC) and used to determine the number of clusters in the data.

Chorng-Shyong Ong, Jih-Jeng Huang and Gwo-Hshiung Tzeng (2005) identified three main stages to build an ARIMA model. They are model identification, model estimation and model checking. They provided a genetic algorithm (GA) based model identification to overcome the problem of local optima. Ping-Feng Pai and Chih-Sheng Lin (2005) studied the support vector machines with neural network techniques in solving non linear regression estimation problems. Using this method ARIMA model in forecasting stock prices is considered. Data sets of stock prices were used to examine the forecasting accuracy of the developed model.

A.K.Mishra and V.R. Desai (2006) compared the linear stochastic models (ARIMA/SARIMA), recursive multi step neural network (RMSNN) and direct multi-step neural network (DMSNN) for drought forecasting. The models were applied to forecast droughts using standardized precipitation index (SPI) series as drought index in the Kansabati river basin which lies in West Bengal.

Theodoros Koutroumanidis, Lazaros Iliadis and Georgios K. Sylaios (2006) used ARIMA models for optimal forecasting techniques and decision support systems based on fuzzy mathematics in order to predict the general trend of a given
fish landings time-series. They applied these modeling methods to forecast anchovy
fish catches landed in a given port during 1979-2000 and hake and bonito total fish
catches during 1982-2000. They assessed the accuracy of the model by comparing
the model results to the actual monthly fish catches of the year 2000.

Guoqi Qian and Xincgong Zhao (2007) studied the model selection for time
series data where millions of candidate ARMA models may be eligible for selection..
They developed a computing method based on the Gibbs sampler. By this method
model selection is performed through a random sample generation algorithm, and
given a model of fixed dimension the parameter estimation is done through the
maximum likelihood method. Marcella Corduas and Domenico Piccolo (2007)
investigated the statistical properties of the $AR$ distance between ARIMA processes,
the asymptotic distribution of the squared $AR$ distance and an approximation which is
computationally efficient are derived. They suggested that clustering not to be
confined within the bounds of exploratory methods.

Mohd Zamri Ibrahim, Roziah Zailan, Marzuki Ismail and Muhd Safiih
Lola (2009) used a Box-Jenkins ARIMA approach in order to predict the status of
future air quality in Malaysia using the time series of monthly maximum 1h carbon
monoxide and nitrogen dioxide concentrations in the east coast states of Peninsular
Malaysia to a comparison with the representative west coast state represent of Hulu
Kelang. Mehmet Tektas (2010) performed a comparative study of statistical and
neuro-fuzzy network models for forecasting the weather. Adaptive Network Based
Fuzzy Inference System (ANFIS) and Auto Regressive Moving Average (ARIMA)
processes have been applied for developing the models. The developed models are
validated with nine year data 2000-2008 comprising daily average temperature (dry-wet), air pressure, and wind-speed.

2.3 REVIEW ON TIME SERIES DATA MINING BASED ON NEURAL NETWORKS

In this section, a brief review on time series data mining based on neural networks is given. Artificial neural networks are used as forecasting models. After learning the data presented to them they can correctly predict the unseen data even in the presence of noise. The neural networks are non linear models and suited to capture the true nature of many natural processes with applications in social, environment, economical engineering, foreign exchange, medical and so forth.

Fredric M. Ham and Soowhan Han (1996) have analyzed two different conditions normal and abnormal premature ventricular contraction (PVC) and studied the electrocardiogram (ECG) data using fuzzy adaptive resonance theory mapping (ARTMAP) neural network for classification of normal and abnormal PVC conditions of cardiac arrhythmias. Based on MITBIH database annotations, cardiac hearts for normal and abnormal QRS complexes were extracted from this database, scaled, and Hamming windowed after bandpass filtering to yield a sequence of 100 samples for each QRS segment. From these sequences, two linear predictive coding (LPC) coefficients were generated using Burg’s maximum entropy method. The two LPC coefficients with the mean-square value of the QRS complex segments were utilized as features for each condition to train and test a fuzzy ARTMAP.

Howard R. Kirby, Susan M. Watson and Mark S. Dougherty (1997) discussed the relative merits of neural networks and time series methods for traffic
forecasting and summarize the findings from a comparative study of their performance for motorway traffic in France with that of the earlier method. Guoqiang Zhang, B. Eddy Fatuwo and Michael Y. Hu (1998) presented a survey of ANN applications in forecasting. Their findings include the characteristics of ANNs adaptability, ability, nonlinearity, arbitrary function mapping ability which make ANN suitable and useful for forecasting task.

J.M. Corchado and C. Fyfe (1999) presented the results of using a Negative Feedback Neural Network for extraction of models of the thermal structure of oceanographic water masses and to forecast time series in real time. The results are compared with Linear Regression and an ARIMA model. Sandy D. Balkin and J. Keith Ord (2000) developed a procedure for selecting the architecture of an artificial neural network for forecasting. The ANNs perform better provided there is sufficiently long time series to detect the nonlinearity and to provide reliable estimates of the parameters using statistical procedures.

Vincent Cho (2003) investigated the application of three time-series forecasting techniques namely, exponential smoothing, univariate ARIMA and Elman’s Model of Artificial Neural Networks (ANN) to predict travel demand from different countries to Hong Kong. Exponential smoothing and ARIMA are statistical time series forecasting techniques where as Neural Networks is an artificial intelligence technique. Mevlut Ture and Imran Kurt (2006) compared time series prediction capabilities of three artificial neural networks (ANN) algorithms Multi-Layer Perceptron (MLP), Radial Basis Function (RBF), Time Delay Neural Networks (TDNN)) and ARIMA model to Hepatitis A virus (HAV) forecasting.
Henry C. Co and Rujirek Boosarawongse (2007) compared the performance of Artificial Neural Networks (ANNs) with exponential smoothing and ARIMA models in forecasting rice exports from Thailand. They evaluated aggregate measures of forecast error MAE, MSE, MAPE and RMSE during validation process of the models. ANN was able to track the dynamic non-linear trend and seasonality and interactions between them. R.N. Yadav, P.K. Kalra and J. John (2007) developed a neuron model based on a polynomial architecture. A single aggregation function is used instead of considering the higher order terms. The aggregate function is considered as a product of linear function in different dimension space. The functional mapping capability was demonstrated through time series prediction problems and compared with multi layer neural network.

P. Ram Kumar, M.V. Ramana Murthy, D. Eashwar and M. Venkatdas (2008) developed Back Propagation Algorithm (BPA) that learns by computing an error signal and then propagating the error backward throw the network. The BPA method is applied to statistical model ARIMA to test the efficiency and then applied to actual geophysical data. The results of the analysis are tested simulating/modeling on large amount of data. Iffat A. Gheyas and Leslie S. Smith (2009) developed an ensemble learning technique that combines the advice from several Generalized Regression Neural Networks. The algorithm is compared with the other algorithms on real and synthetic datasets.

Florin Leon and Mihai Horia Zaharia (2010) developed a hybrid model for time series forecasting. It is a stacked neural network, containing one normal multilayer perceptron with bipolar sigmoid activation functions and the other with an exponential activation function in the output layer. The case studies are shown on
developed stacked hybrid neural model. They also studied combination of weights of the two stack components that leads to optimal performance.

2.4 REVIEW ON TIME SERIES DATA MINING BASED ON OTHER MODELS

In this section, a brief review on time series data mining based on other than the Markovian and Neural networks models is given.

H.Tong and P. Dabas (1990) presented 12 time series models reported in literature for the annual Canadian lynx data of 1820-1934. They studied the respective goodness of fit of each model. Jeffrey D. Banfield and Adrian E. Raftery (1993) identified the reparameterization of a covariance matrix to specify some features be the same for all clusters. They outlined the non-Gaussian clustering and described a mean of incorporating noise in the form of a Poison process. They also gave a Bayesian method for choosing the number of clusters.

H A Guvenir, B Acar, G Demiroz and A Cekin (1997) presented a machine learning algorithm VFI5 for Voting Feature Intervals for the diagnosis of cardiac arrhythmia from standard 12 lead ECG recordings. VFI5 is a supervised and inductive learning algorithm for inducing classification knowledge from examples. The input to VFI5 is a training set of records containing clinical measurements from ECG signals and information such as sex, age and weight along with the decision of an expert cardiologist. The knowledge representation is based on a technique called Feature Intervals, where a concept is represented by the projections of the training cases on each feature separately. Classification in VFI5 is based on a majority voting among the class predictions made by each feature separately.
Claudio Bettini, X. Sean Wang, Sushil Jajodia and Jia-Ling Lin (1998) studied algorithms to solve an event structure that have temporal constraints with multiple granularities. The basic components of the algorithms include timed automata with granularities (TAGs) and a number of heuristics for testing whether a specific temporal pattern called a candidate complex event type appears frequently in a time sequence. Heuristics are presented aiming at reducing the number of candidate event types and reducing the time spent by the TAGs testing. These heuristics exploit the information provided by explicit and implicit temporal constraints with granularity in the given event structure.

A.K. Jain, M.N. Murty and P.J. Flynn (1999) presented pattern clustering methods in image segmentation, object recognition and information retrieval from a statistical pattern recognition perspective. The taxonomy of clustering techniques and identification of cross-cutting themes are also described. Ruy L. Milidiu, Ricardo J. Machado and Raul P. Renteria (1999) described a system formed by a mixture expert model (MEM) for time-series forecasting. The advantage of MEM method was the use of the Haar wavelets transform to perform a base change of the input vector space giving an overall shape description of each pattern to the clustering algorithm.

Byoung-Kee Yi and Christos Faloutsos (2000) presented an indexing scheme for time sequences, when the distance function is any of arbitrary $L_p$ norms ($p = 1, 2, \ldots, \infty$). The one index structure method is compared to the other competitor of $L_2$ and $L_\infty$ norms and faster for $L_1$ norm. Shyi-Ming Chen and Jeng-Ren Hwang (2000) developed a two-factor time-variant fuzzy time series model for handling forecasting problems. They compared the time complexity of the developed
algorithms as $O(cwm)$, where $c$ is the number of partitioned groups in the historical data, $w$ is the window basis, and $m$ is the number of elements in the universe of discourse.

Eamonn Keogh, Kaushik Chakrabarti, Sharad Mehrotra and Michael Pazzani (2001) introduced Adaptive Piecewise Constant Approximation (APCA) technique to approximate each time series by a set of constant value segments of varying lengths such that their individual reconstruction errors are minimized. APCA representation was indexed using a multidimensional index structure. This approach also allows faster approximate querying on the same index structure. Marco Ramoni, Paola Sebastiani and Paul Cohen (2001) introduced a Bayesian method for clustering dynamic processes. The method models dynamics as Markov chains and then applies an agglomerative clustering procedure to discover the probable set of clusters capturing different dynamics.

Mark Last, Yaron Klein and Abraham Kandel (2001) introduced a methodology for knowledge discovery in time series database that includes cleaning and filtering of time series data, identifying the most important predicting attributes and extracting a set of association rules used to predict the time series behavior. The method is based on signal processing techniques and the information-theoretic fuzzy approach to knowledge discovery. Monica Adya, Fred Collopy, J. Scott Armstrong and Miles Kennedy (2001) developed an expert system rule based forecasting that uses features of time series to select and weight extrapolation techniques. They developed an automated heuristics to detect six features identified in RBF as outliers, left shifts, change in basic trend, unstable recent trend, unstable recent trend, unusual last observation and functional form. These heuristics rely on simple statistics, first
differences and regression estimates. The use of automated feature detection heuristics reduced the costs of using RBF without negatively affecting forecast accuracy.

Quinton J. Nottingham and Deborah F. Cook (2001) used Local Linear Regression (LLR) as a tool for predicting future values of process parameters based on historical values. LLR procedure fits seasonal time series data as well as traditional ARIMA models. Robert Baragona (2001) considered the problem of finding the best partition of a set of time series according to the cross-correlation estimated from the pre-whitened series. The dissimilarity matrix entries were computed to include the cross-relationships up to a pre-specified lag. The k-min cluster criterion was a natural choice and the clusters were constrained to include only time series with pair wise cross correlation absolute values exceed a given threshold.

Sang Jun Lee and Keng Siau (2001) discussed challenges of data mining and described major data mining techniques as statistics, artificial intelligence, decision tree approach, genetic algorithm and visualization. They narrated that right information at right time is crucial in making the right decision. Hence organizations will be competing in generating information from data and not in collecting data. Eamonn Keogh and Shruti Kasetty (2002) have reviewed the literature on time series data mining. They have re-implemented the contribution of various papers and tested them on real world data sets. The results suggested the need for a set of time series bench marks and more careful empirical evaluation in data community.

James V. Hansen and Ray D. Nelson (2002) analyzed historical time series using stacked generalization, a methodology devised to aid in developing models that
generalize well to future time periods. Stacked generalization is compared to ARIMA and to stand alone neural networks. John F. Roddick and Myra Spiliopoulou (2002) investigated the confluence of data mining and temporal data bases. They addressed techniques which are applicable to temporal data sets, series of static data sets and temporally oriented data warehouses.

Liadan O'Callaghan, Nina Mishra, Adam Meyerson, Sudipto Guha and Rajeev Motwani (2002) described streaming algorithm that clusters large data streams. They provided the algorithm's performance on synthetic and real data streams. Lorenzo Giada and Matteo Marsili (2002) studied the problem of data clustering by introducing an unsupervised, parameter-free approach based on maximum likelihood principle. The likelihood provides a parameter-free measure for cluster structures which can be taken as the basis of clustering algorithms for large data sets. The cluster structure is determined by internal correlations of the data set. They analyzed two different data sets namely time series of financial market returns and gene expression data.

S.L.Ho, M.Xie and T.N.Gosh (2002) performed a comparative study of Box-Jenkins ARIMA models and artificial neural networks in predicting the failures of repairable system. The evaluated neural networks architecture are the multi layer feed forward network and recurrent network. Venkatesh Ganti, Johannes Gehrke and Raghu Ramakrishnan (2002) described a generic algorithm for model maintenance algorithm that any traditional incremental data mining that allows restriction on a temporal subject of the data base and another described a generic framework for change detection that quantifies the difference between two data sets in items of the data mining models they induce.
Yixin Chen, Guozhu Dong, Jiawei Han, Benjamin W.Wah and Jianyong Wang (2002) investigated methods for online multi-dimensional regression analysis of time series stream data where only a small number of numerical values of data are need to be registered for multi-dimensional analysis. The online analysis framework uses tilt time frame, explores minimal interesting and observation layers and adopts an exception-based computation method. Anna C. Gilbert, Yannis Kotidis, S. Muthukrishnan and Martin J. Strauss (2003) presented methods for capturing various linear projections and used them to provide point wise and range sum estimation of data streams. These methods use small amounts of space and per-item time while streaming through the data and provide accurate representation.

Aurelio Tobias, Marc Saez, Inaki Galan, Michael J. Campbell (2003) analysed the relationship between photochemical air pollutants namely nitrogen dioxide and ozone and emergency room admissions for asthma in Madrid for the period 1995–1998 using the statistical models commonly used to studying the short-term effects of air pollution on health. Jonathan Alon, Stan Sclaroff, George Kollios and Vladimir Pavlovic (2003) fitted a finite mixture of hidden Markov models (HMMs) to the motion data using the expectation-maximization (EM) framework. The algorithm allowed each sequence to belong to more than a single HMM with some probability.

Lijuan Cao (2003) developed the support vector machines (SVMs) experts for time series forecasting. The generalized SVMs experts have two-stage neural network architecture. In the first stage, self-organizing feature map (SOM) is used as a clustering algorithm to partition the whole input space into several disjointed regions. A tree-structured architecture is adopted in the partition to avoid the problem
of predetermining the number of partitioned regions. In the second stage, multiple SVMs, also called SVM experts, that best fit partitioned regions are constructed by finding appropriate kernel function and the optimal free parameters of SVMs.

Michael L. Raymer, Travis E. Doom, Leslie A. Kuhn and William F. Punch (2003) used a classifier based on Bayes discriminant function. The discriminant function classifier gain in computational efficiency by estimating the class-conditional feature value distributions based on the training data rather than storing every training sample. Sangjun Lee, Dongseop Kwon and Sukho Lee (2003) considered the problem of similarity search based on the minimum distance in large time series databases. They developed a dimensionality reduction technique for indexing time series based on the minimum distance known as Segmented Sum of Variation (SSV) indexing. The performance of SSV indexing is verified through real time stock moment data.

Spiros Papadimitriou, Anthony Brockwell and Christos Faloutsos (2003) developed arbitrary window stream modeling method (AWSOM) which allows sensors in remote or hostile environments to discover interesting patterns and trends. AWSOM deal duality with problem of unsupervised stream mining and pattern. A.Sfetsos and C.Siriopoulos (2004) discussed the application of clustering algorithms in combinatorial forecasting for time series data. The developed cluster-based combinatorial forecasting schemes were examined in a single-step ahead prediction of the pound-dollar daily exchange rate.

Charu C. Aggarwal, Jiawei Han, Jianyong Wang and Philip S. Yu (2004) developed a classification system in which the training model can adapt to the
changes of the underlying data stream. They developed an on-demand classification process which can dynamically select the appropriate window of past training data to build the classifier. Charu C. Aggarwal, Jiawei Han, Jianyong Wang and Philip S. Yu (2004a) developed a high dimensional data stream clustering method called HP Stream. The method incorporates a fading cluster structure and the projection based clustering methodology. It is incrementally updatable and is scalable on both the number of dimensions and the size of data streams. A numerical illustration with time series data is given.

Francesco Pattarin, Sandra Paterlini and Tommaso Minerva (2004) identified three steps in classification procedure as dimensionality reduction step based on principle component analysis, a clustering step that exploits a robust evolutionary clustering method and style identification step via constrained regression model. A classification algorithm for mutual funds style analysis is developed. Funds are partitioned into different classes through clustering algorithm that runs on pre-selected principal components and then constrained regression model is used in characterizing styles for each class. Richard J. Povinelli (2004) demonstrated the methods on the TSDM framework. He stated that these methods successfully characterize and predict complex, non-periodic, irregular and chaotic time series from both the engineering and financial domains. The TSDM framework was able to characterize and predict metal droplet releases from a given multi-dimensional time series generated by sensors on a welding station. Also, in financial domain, the framework was able to generate a trading-edge by characterizing and predicting stock price events.
Saif Ahmad T. Taskaya-Temizel and Khurshid Ahmad (2004) presented a time series summarization and prediction framework to analyze non-stationary volatile and high frequency time series data. Multiscale wavelet analysis is used to separate the trend, cyclical fluctuations and autocorrelation effects. The framework can generate verbal signals to describe each effect. The output is used to reason the future behavior of the time series and to give a prediction. Experiments on the intraday European currency spot exchange rates are described. The results are compared with a neural network prediction framework. Anthony Bagnall and Gareth Janacek (2005) showed a procedure of clipping the time series reduce the memory requirements and speeds up clustering without decreasing clustering accuracy. They considered simulated data from polynomial, ARMA and hidden Markov models and showed that the estimated parameters of the clipped data used in clustering tend asymptotically to those of the unclipped data. They illustrated using clipped series can have practical benefit in detecting model misspecification and outliers.

Ashish Singhal and Dale E. Seborg (2005) developed a methodology of clustering multivariate time-series data on calculating the degree of similarity between datasets using two similarity factors. One similarity factor is based on principal component analysis and the angles between the principal component subspaces while the other is based on the Mahalanobis distance between the datasets. The K-means clustering algorithm clusters multivariate time-series datasets using similarity factors. C.S. Moller-Levet, F. Klawonn, K.-H. Cho, H. Yin and O. Wolkenhauer (2005) developed the fuzzy short time-series (FSTS) clustering algorithm to cluster profiles based on the similarity of their relative change of expression level and the
corresponding temporal information. They developed the fuzzy short time-series (FSTS) clustering algorithm that clusters profiles based on the similarity of relative change of expression level and the corresponding temporal information. The advantages of fuzzy clustering are that genes can belong to more than one group, revealing distinctive features of each gene's function and regulation. The fuzzy $c$-means, $k$-means, average linkage hierarchical algorithm and random clustering are compared to the proposed FSTS algorithm.

Hua-Fu Li, Suh-Yin Lee and Man-Kwan Shan (2005) developed a single pass algorithm called MFC-append (Mining Frequency Changes of Append only data streams) for discovering the frequent frequency-changed items, vibrated frequency changed items and stable frequency changed items over continuous append only data streams. The data structure change-sketch is developed to compute frequency changes between two continuous data streams. MFC-append based algorithm called MFC-dynamic (Mining Frequency Changes of dynamic data streams) is developed to find the frequency changes over dynamic data streams. Jeroen Boets, K.DE Cock, M.Espinoza and B.De.Moor (2005) developed methodology to cluster time series based on measurement data that uses subspace angles within a model and between two models. This distance is used to obtain clustering the set of time series and shown how it is related to the mutual information of the past and future output process and to a previously defined cepstral distance. The methodology is applied on clustering time series of power consumption with in Belgian electricity grid.

Jiawei Han, Yixin Chen, Guozhu Dong, Jian Pei, Benjamin W. Wah, Jianyong Wang and Y.Dora Cai (2005) developed an on-line analytical processing of stream data for on-line computation of multi-dimensional, multi-level stream cube.
The method uses a tilted time frame, explores minimal interesting and observation layers storage of stream cube to facilitate OLAP analysis of stream data. The stream cubing methodology has been implemented in the MAIDS (Mining Alarming Incidents in Data Streams) project at NCSA (National Centre for Supercomputing Applications) at University of Illinois and tested using online stream data sets.

Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy (2005) stated data stream mining has gained a high attraction due to importance of its application and increasing generation of streaming information. Applications of data stream analysis vary from critical scientific and astronomical applications to impart business and financial ones. They also reviewed various techniques in data stream mining. Mohamed Medhat Gaber, Shonali Krishnaswamy and Arkady Zaslavsky (2005) developed a output granularity (AOG) algorithm in mining data streams. AOG approach pays attention to the data stream rate with respect to the available resources. They used NASA data to demonstrate the efficiency of the developed algorithm.

T. Warren Liao (2005) summarized previous works on clustering of time series data in various application domains. He presented general-purpose clustering algorithms commonly used in time series clustering studies, the criteria for evaluating the performance of the clustering results, and the measures to determine the similarity/dissimilarity between two time series in the forms of raw data, extracted features or some model parameters. Tugba Taskaya-Temizel, and Matthew C. Casey (2005) studied the relation between hybrid models and better than single models. They discussed the drawbacks associated with data mining algorithms, ARIMA models and neural networks. They studied the relationship between the
models, linear and non linear components. They demonstrated the use of autoregressive linear and time delay neural network model using 9 data sets.

Xiaolei Li, Jiawei Han, Xiaoxin Yin and Dong Xin (2005) developed a Gaussian transformation-based regression model to capture time-variant relationships between customers and products. Such relationships in a multi-dimensional space compute multi-dimensional aggregates in a data cube environment. Yutao Shou, Nikos Mamoulis and David W. Cheung (2005) considered a technique to decompose the time series sequences into a number of segments. They presented progressively lighter bounds relying on the existence or not of warping constraints. They developed an index and multi step technique that uses the proposed bounds and performs two levels of filtering to process similarity queries.

A.M. Alonso, J.R. Berrendero, A. Hernandez and A. Justel (2006) developed a time series clustering algorithm based on probabilistic density of the forecasts. A resampling method combined with a non parametric Kernel estimator provides estimates of the forecast densities. A measure of discrepancy is defined between these estimates and the resulting dissimilarity matrix used to carry out the required cluster analysis. Anthony Bagnall, Chotirat “ANN” Ratnamahatana, Eamonn Keogh, Stefano Lonardi and Gareth Janacek (2006) stated clipping as a useful and flexible transformation of exploratory analysis of large time dependent data sets. They demonstrated how the time series stored as bits can be very efficiently compressed and manipulated. Their results demonstrated clipping is a useful transformation for time series data mining based on similarity.
C. Harpham and C.W. Dawson (2006) demonstrated the variation of test set error between six recognized basis functions. The tests were carried out and each RBF network was learned using a two-stage approach utilizing k-means clustering algorithm for the first stage and singular value decomposition for the second stage. They studied the effect of different basic functions on a radial basis function network for time series prediction. A comparative study of various models is also presented. Dina Goldin, Ricardo Mardales and George Nagy (2006) presented a distance measure for time series subsequence clusters based on cluster shapes. The cluster shape is defined as the sorted list of the pairwise Euclidean distances between their centroids.

Fabian Morchen (2006) described a method for the understandable description of local temporal relationships in multivariate data called Time Series Knowledge representation. The patterns have a hierarchical structure each level corresponds to a single temporal concept. Lowest level intervals are used to represent duration. Overlapping parts of intervals represent coincidence on the next level. Several such blocks of intervals are connected with a partial order relation on the highest level. They proposed algorithms for the discovery of the patterns. Fatih Altiparmak, Hakan Ferhatosmanoglu, Selnur Erdal and Donald C. Trost (2006) developed an approach for information mining. They first applied data mining algorithm over homogeneous subsets of data and identify common or distinct patterns over the information gathered. The approach is implemented for heterogeneous and high dimensional time series clinical trails data. They considered a copy of utilizing frequent item set mining, clustering, declustering techniques with novel distance measure for measuring similarity between time series data. By clustering, the groups which are strongly correlated are identified.
Geoffroy Simon, John A. Lee, Michel Verleysen (2006) studied the limitations of clustering in time series. This problem was illustrated with various raw time series. The usefulness of unfolding preprocessing is illustrated for various time series. They introduced distance to diagonal criterion to measure how well a time series regressor distribution is unfolded. Hui Zhang, Tu Bao Ho, Yang Zhang and Mao-Song Lin (2006) developed an unsupervised feature extraction algorithm using orthogonal wavelet transform for choosing the dimensionality of features. The feature extraction algorithm selects the feature dimensionality by leveraging two conflicting requirements, lower dimensionality and lower sum of squared errors between the features and original time series.

Jorge Caiado, Nuno Crato and Daniel Pena (2006) studied a measure of distance between time series based on the normalized periodogram. They discussed the classification of time series as stationary or as non-stationary. They also considered the use of both hierarchical and non-hierarchical clustering algorithms. Jurgen Beringer and Eyke Hullermeier (2006) studied the problem of clustering parallel streams of real-valued data, continuously evolving time series. The incoming data is analyzed to maintain an up-to-date clustering structure in an online manner. An online version of the classical K-means clustering algorithm was developed.

Madalina Olteanu (2006) introduced a descriptive method for time series analysis to determine the number of regimes in a switching autoregressive model. This problem was translated as a classification problem and defines a criterion for hierarchical clustering different model fittings. Self-organizing maps are introduced to mix the regimes if the regression hyper planes are too close. Markus Nilsson, Peter Funk, Erik M.G. Olsson, Bo Von Scheele and Ning Xiong (2006) developed
a decision support system for clinicians in analyzing respiratory sinus arrhythmia (RSA). The decision support system contains a classification system HR3 model that classifies respiratory sinus arrhythmia. The system performs an analysis of the concerned physiological time series.

Paul Hungman Kim and Yun-hwan Lim (2006) introduced an approach to transition from unsupervised clustering of ubiquitous data streams to interpreting and using the results in an online and continuous manner in a mobile device environment. They extended online and resource aware ubiquitous data stream clustering algorithm Light-Weight Clustering (LWC) with support for labeling of clusters in real-time using fuzzy rules based on domain/expert knowledge.

Srinivasan Laxman and P.S. Sastry (2006) presented techniques of temporal data mining. They discussed an algorithm for pattern discovery in sequential data streams to rule discovery. They also discussed sequential patterns and episodes for frequent pattern discovery. Tak-chung Fu, Chi-wai Law, Kin-kee Chan, Fu-lai Chung and Chak-man Ng (2006) developed SB-tree data structure to represent time series according to the importance of data points. Their investigation on the size of SB-tree is reported. Two SB-tree optimization approaches are proposed to reduce the size of SB-tree while the overall shape of time series can be preserved. They demonstrated suitability of the approach for different categorization and clustering applications.

Xiaozhe Wang, Kate Smith and Rob Hyndman (2006) developed a method for clustering of time series based on their structural characteristics on global features extracted from the time series. The feature measures are obtained from each
individual series. The developed technique was tested using benchmark time series datasets. **Yanchang Zhao and Shichao Zhang** (2006) have studied clustering of time series and incrementally pattern maintaining. They developed a generalized dimension reduction framework for biased approximations in time series analysis. The framework is designed in equi segmented scheme and vari-segmented scheme. In both schemes, time series data are partitioned into segments and dimension reduction technique was applied.

**Allou Same, Christophe Ambroise and Gerard Govaert** (2007) presented an online clustering algorithm based on mixture models in the context of a real-time flaw-diagnosis application for pressurized containers which uses data from acoustic emission signals. The developed algorithm is a stochastic gradient algorithm derived from the classification version of the EM algorithm. **Defeng Wang, Lin Shi, Daniel S.Yeung, Eric C.C. Tsang and Pheng Ann Heng** (2007) proposed an extension of SVC (Support Vector Clustering) to step further toward a data-sensitive detector. This Ellipsoidal Support Vector Clustering (ESVC) is performed on features extracted from the fMRI time series via Fourier transform.

**Mahesh Kumar and Nitin R. Patel** (2007) developed a statistical model and algorithms for clustering data in the presence of errors. The errors associated with data follow a multivariate Gaussian distribution and are independent between data points. The model used univariate data and provided a new metric for clustering. This metric is used to develop two algorithms for error-based clustering, hError and kError, that are generalizations of Ward’s hierarchical and k-means clustering algorithms. They demonstrated the applications of the developed methodology using four statistical models namely sample analysis, multiple linear regression, ARIMA models...
for time series and Markov Chains and showed error based clustering performed better than traditional clustering methods.

Paolo Capitani and Paolo Ciaccia (2007) developed Stream-DTW (SDTW), which unlike DTW can be updated at each time step. SDTW speed up the monitoring process by a factor that grows linearly with the size of the window sliding over the stream was demonstrated experimentally. P.-A.Cornillon, W.Imam and E.Matzner-Lobelr (2007) developed two forecasting methods of time series. The first method is an application of spline principal component analysis with respect to instrumental variables and second one is an adaptation of spline principal component analysis with respect to instrumental variables. In modified version, the used criteria according to the unknown value that need to be predicted are differentiated.

P.K. Dash, Maya Nayak, M.R.Senapati and I.W.C.Lee (2007) presented a comparison between different wavelet feature vectors for data mining of non stationary time series that occurs in an electricity supply network. An approach S-transform-based time frequency analysis in processing power quality disturbance data is narrated. Neural networks are used to compute the classification accuracy. T.Warren Liao (2007) developed a two step procedure for the exploratory mining of real-valued vector time series using partition-based clustering methods. They tested the proposed procedure with model-generated data, multiple sensor-based process data, as well as simulation data. The developed procedure produced better clustering results than a hidden Markov model (HMM) based clustering method if there is a priori knowledge about the number of clusters in the data.
B. Chandra, Manish Gupta and M. P. Gupta (2008) developed an approach based on dynamic time wrapping and parametric Minkowski model, to find similar crime trends among various crime sequences of different crime locations and used this information for future crime trends prediction. The Indian crime data is considered for illustration. Parvathi Chundi and Daniel J. Rosenkrantz (2008) considered an item-set time series to facilitate the temporal analysis of software version histories, email logs, stock market data etc. Segmentation of a time series partitions the time series into a sequence of segments where each segment is constructed by combining consecutive time points of the time series. Each segment is associated with an item set that is computed from the item sets of the time points in that segment, using a function called as a measure function. The segment difference measures the difference between the item set of a segment and the item sets of the time points in that segment. The segment difference values are required to construct an optimal segmentation of the time series. An algorithm is described to compute segment difference values for each of the measure functions. The item-set time series segmentation techniques are used to analyze the temporal content of three different data sets—Enron email, stock market data, and synthetic data sets.

Pei-Gee Peter Ho and C.H. Chen (2008) used Time Series statistical models namely two-dimensional (2-D) Autoregressive (AR) mode in describing the texture and contextual information of a remote sensing image. They used 2-D univariate time series based imaging model was derived mathematically to extract the features for further terrain segmentations. The effectiveness of the model was demonstrated in region segmentation of a multispectral image of the Lake Mulargias region in Italy.
Sylvia Fruhwirth-Schnatter and Sylvia Kaufmann (2008) pooled multiple time series into several groups using finite-mixture models. They estimated the groups of time series simultaneously with the group-specific model parameters using Bayesian Markov chain Monte Carlo simulation methods. Venkataramana Kini B and C. Chandra Sekhar (2008) developed a method for time series pattern classification based on the generative modeling using AR model and optimizing the boundaries between these models using the large margin concepts. The developed model captures the correlations in the time series data. Multi-class classification can be performed directly without performing binary classification. The optimization is performed using genetic algorithm for obtaining global optimal parameters. The developed method is applied on simulated and ECG data.

Aydin, M. Karakose and E. Akin (2009) stated that time series data mining combines the fields of time series analysis and data mining techniques. They developed a prediction algorithm using time series data mining based on fuzzy logic. A Gaussian shaped fuzzy membership function was used in order to determine the temporal pattern clusters. The truth of prediction algorithm was studied through earthquake and Lorentz series. Ignacio Rojas, Osama Salamesh and Mai Handon (2009) studied the problem of time series prediction from a given set of input/output data. The problem consists of prediction of feature values based on past and present data using radial basis function. This approach is based on clustering of centres of the radial basis function neural network. The method uses error committed in every cluster. Using real output of radial basis function neural network trying to concentrate more clusters in those input regions where error is bigger and move the cluster instead of input values of the input/output data.
Sunlit Goswami, Sudeshna Sarkar and Mayur Rustagi (2009) reported the difference in blogging for gender and age group variation. The results are based on two independent features namely use of slang words, variation in average length of sentences across various age groups and genders. These two features are augmented with previous study results for stylometric analysis of gender and age. Both these features when augmented with other features increases with prediction accuracy.

Vincent S. Tseng, Chun-Hao Chen, Pai-Chieh Huang and Tzung-Pei Hong ((2009) developed a time series segmentation approach by combining the clustering technique, the discrete wavelet transformation and the genetic algorithm to automatically find segments and patterns from a time series. The genetic algorithm is used to find the segmentation points for deriving appropriate patterns. In fitness evaluation, the developed approach first divides the segments in a chromosome into $k$ groups according to their slopes by using clustering techniques. The Euclidean distance is then used to calculate the distance of each subsequence and evaluate a chromosome. The discrete wavelet transformation is also used to adjust the length of the subsequences for calculating. The evaluation results are utilized to choose appropriate chromosomes for mating. Akash Rajak and Kanak Saxena (2010) considered a study on diabetes mellitus type 1 patients on the bases of time series based representation of clinical temporal data. They simulated a case on an interactive PC-based freeware computer program AIDA, which contains a simple model of glucose - insulin interaction in the body. They also suggested the possible treatment for the patient having insulin dependent diabetes. Similarly, the simulator can be used in planning the therapy and monitoring the blood-insulin level of diabetes patient.
Christoph Pamminger and Sylvia Fruhwirth-Schnutter (2010) discussed two approaches for model based clustering of categorical data based on time homogeneous first order markov chains. For markov chain clustering the individual transition probabilities are fixed to a group specific transition matrix. The proposed approach Dirichlet multinomial clustering the rows of the individual transition matrices deviates from group mean and follow a dirichlet distribution with unknown group-specific hyper parameters.