ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Hybrid Clustering Algorithm for Time Series Data Stream: Current State of the Art

T.Rajesh#, Dr. K.V.G Rao*

Research Scholar, Department of CSE, JNTUH, Hyderabad, India#

Professor, Department of CSE, G.Narayanamma Institute of Technology &Science, Hyderabad, India*

ABSTRACT: Clustering time series data is a trouble that has applications in an extensive variety of areas and has recently evoked a large amount of research. Time series data may contain large and outliers. In addition, time series data is a one kind of special data set where attributes have a temporal ordering. Therefore clustering of time series data is a good issue in the data mining process. Different techniques and various clustering algorithms have been proposed to assist clustering of time series data sets also different kinds of non-developmentary optimization techniques and clustering algorithms have been applied for clustering multivariate time series data in some applications, usually they produces poor efficient results due to the dependency on the initial set of values and their poor performance in manipulating multiple objectives. Sometimes Time series data doesn't contain same length and they usually have missing values, the basic measure Euclidean distance and dynamic time warping cannot be applied for such datasets to measure the similarity of objects. The clustering algorithms and its effectiveness on various applications are compared to develop a new method to solve the existing problem. Henceforth, this presents a survey on dissimilar clustering algorithms available for time series datasets. Moreover, the uniqueness and ceiling of previous researches are also discussed and further it contains short descriptive taxonomy of Data Mining Techniques and Tools.

KEYWORDS: Clustering techniques, Time Series Data, Pitfalls, Secular Trend, ML Algorithms,Tools

I. INTRODUCTION

Clustering is considered the most important unsupervised learning problem. The clustering of time series data is particularly advantageous in exploratory data analysis and summary generation. Time series clustering is also a preprocessing step in either another time series mining task or as part of a complex system. Researchers have shown that using well-known conventional algorithms in the clustering of static data, such as partitional and hierarchical clustering, generates clusters with an acceptable structural quality and consistency and is partially efficient in terms of execution time and accuracy [1] [17]. However, classic machine learning and data mining algorithms are ineffective with regard to time series data because of the unique structure of time series, that is, its high dimensionality, very high feature correlation, and (typically) large amount of noise [1] [18–20]. Accordingly, numerous research efforts have been conducted to present an efficient approach to time series clustering. However, the focus on the efficiency and scalability of these methods in handling time series data has come at the expense of losing the usability and effectiveness of clustering [1] [21]. Now a Days managing of Time Series data has become a significant research in the area of data mining. Especially, performing of clustering technique on time series data has most attracted area of various researchers. Usually Data mining suffers with three limited number of resources. They are Time, Memory and Sample size. Recently two resources time and memory seems to be bottleneck for different kinds of Data Mining methods. Clustering technique is an unsupervised learning process for combining a dataset into subgroups. A time series data is a series of numbers, every number in the series indicates a some value at any given point of time. Regularly Data travels from a data stream at quick speed, Real world applications produces more examples over the time. A classical clustering algorithm doesn't support to the high speed of data arrival in the case of time series datasets. This is a reason; Hybrid clustering algorithms have been developed for processing the real time data. Every

Application domain produces Time series data at unique speed e.g., stock market fluctuations , Health care , scientific experiments, Electrical industries, Geographical location based services for moving objects, reading data in sensor networks, Biological and Medical experiments, etc. A classical clustering technique uses various kinds of batch procedures. There exist two main categories for clustering techniques based on different parameters. First one is Partition clustering and the second one is Hierarchical clustering algorithms. Accuracy and effectiveness for time series data sets are achieved by applying clustering techniques. Working with the time series datasets is very costly why because it uses multiple arrays.

Clustering techniques of time series data sets provides an efficient mechanism to retrieve hidden patterns, similarity measures and used to predict the forecast the values in future for temporal data . However, because of excess dimensionality, it is normally requires a more memory for performing clustering technique, moreover excess dimensionality is the big challenge when working with the multi variate time series data. Most of the classical clustering applications work with the multi variate data. Classical clustering algorithms like k-means, hierarchical and expectation maximization are used to apply the clustering process on the time series data sets to discover the similar features between the given input data. However, those algorithms and other types of non-developmentary techniques usually produce the poor results due to their dependency on the initial set of solutions, less performance in handling multiple objectives.

An efficient suitable approach to overcome the drawbacks of the non-developmentary clustering techniques is to use the concept of collective intelligence. One challenge in applying Clustering Techniques for clustering MVTS data is to choose a suitable metric to measure similarity between series of the data. The basic common similarity measurement is used for the time series data is based on Euclidian distance and Dynamic Time Warping (DTW). Moreover using Euclidian distance in this area requires the similar length of time series data and contains exactly the same dimensions; this situation is not true in the case of MVTS data. While DTW measure is used to make the length of time series is similar, this technique does not provide accurate results because of missing data items. Moreover, the association between time series is one kind of issue which is mostly ignored when using the basic Euclidean distance. There are very less number of existing studies focused on producing a similarity measure and also producing a new algorithm for clustering MVTS data.

A time series is eventually a collection of periodic recording of data. The data series based on time can be a collection from various sources with symmetric or dissymmetric nature depending on the nature of the data [1]. For example in case of random data sources, the weightage average of the data cannot produce any significant point to make any prediction, thus making it more difficult to use it for any statistical analysis [2] [16]. In order to get rid of the problem, the use of moving average and considering the moving average as new time series data for predictive analysis is the most popular way [3]. Nevertheless, the moving average or MA method is prone to error for multiple reasons. In this work, the problems of the MA methods are been highlighted and suitable solutions are also been recommended.

With the focus of time series data [4], this work defines the basic nomenclature of the time series. The use of time series for making long term decision and also the use in making just in time decisions makes it highly popular. The long term analysis of the time series can range from 5 years to 20 years for any given organization giving the opportunity for the other department of the organization to make plans for financials and human capital related decisions [5]. In the other hand, Just in time decisions are ranging from 1 week to 1 month time to enable less mission critical decisions for the organization [6]. The time series can show high rate of unpunctuality for a longer duration. Hence it is to be understood that, the errors or the noises will affect the long term analysis rather than the Just in time analysis. More and more organizations are tend to plan for long term strategies and depend on the time series, thus making the time series noise free is the major demand of any research.

Conversely, the facet understanding of the time series is must for any type of operations on time series and a very less number of literatures address that issue. Consequently, in order to proceed further, in this domain of research, this work produces a major understanding of the time series and its basic properties.

The rest of the paper organized as follows, Section II explains about the relevant approaches of the time series data, Section III Spotlight on Time Series Data Mining and its techniques, Section IV Discusses the conclusion of the paper with fewer discussions.

II. CURRENT STATE OF THE ART

We have less number of existing algorithms for performing clustering techniques for time series data sets. This part of the article gives an overview of the previous clustering techniques for time series datasets. Ville Haulamati at el[7], presents about the problems related to clustering of time series data in Euclidean space using Random Swap (RS) and Agglomerative Hierarchical clustering followed by k-mean fine-tuning algorithm to compute locally optimal prototype.
Outcome: It provides good clustering accuracy. And also provide more extensions to the k-medoids. The drawback of this algorithm will gives less cluster quality.

Pedro Pereiva Rodrigous at el[6], proposed an incremental system for clustering streaming of time series data sets; a top-down strategy using Online Divisive Agglomerative Clustering system continuously maintains a tree-like structure of clusters. Cluster's diameter measure calculated by Using ODAC cluster quality. The most relevant dissimilarity between different objects of the same cluster is defined as diameter. The ODAC strength is no need to specify a predefined number of target clusters.
Outcome: It provides an excellent performance on finding the exact number of clusters obtained by a number of runs of k-Means. The disadvantage of this system is when the tree structure enlarges, the variables should move from root to leaf, while splitting the variables there is no mathematical confidence on the decision of assignment.

Xiang Lian at el[2], analyzes that all types of time series data applications requires an efficient and accurate similarity search over stream of data is essential. Three approaches namely Polynomial, Discrete Fourier Transform (DFT) and Probabilistic are used to predict the unknown values that have not arrived at the system and answer similarity queries based on the predicted data. These techniques can provide good offline prediction accuracy and it is not fit for online stream environment.
Outcome: Because online needs low prediction and training costs. These techniques are straight forward for asking most general solutions, also it gives relevant confidence for predicting data.
Outcome: They can explicitly provide confidence to predict values, The polynomial approach that predicts future values on the basis of the approximated curve of recent values. The Discrete Fourier Transform (DFT) estimates the future values using approximations in the frequency domain.
Outcome: The probabilistic approach can produces to predict the values and it can be adjustable to the change of data. The group probabilistic approach is makes use of the correlations among stream of time series data. The limitation of this probabilistic approach, it needs more time to forecast the future values.

Sudipto Guha at el, explained about the streaming algorithm that successfully clusters huge data streams. To perform analysis of such data, the capability to process the data in a single pass, or a small number of passes, while using less memory, is decisive. STREAM algorithm based on Divide and Conquer that achieves a constant factor estimated in less space. This STREAM algorithm is location based on a place algorithm that might produce excess number of k-centers.
Outcome: The main use of STREAM algorithm is tradeoff between cluster accuracy and running time. This algorithm is contrast with BIRCH Algorithm and demonstrated that BIRCH appears to do a reasonable fast and dull job.

Ashish Singhal, Dale E. Seborg at el[8], combinely documented about the technique of batch fermentation that calculates the similarity levels among the different multi variate time series data sets it uses the two similarity measures. First one is principal component analysis. Second one is similarity measure is belongs to Mahalanobis distance between the datasets. This algorithm is uses similarity measure along with PCA will produces an effective result.while performing clustering on multivariate time series datasets batch fermentation is better than previous

techniques.
Outcome: It produces better clustering performance and also produces accurate results and also the clustering performance is sensitive.
Hui Zhang at el, documented an unsupervised clustering algorithm which is used to perform feature extraction which uses the orthogonal wavelets to perform the selection of dimensionality features automatically.
Observation: The drawback of calculating the feature dimensionality is circumvented by choosing the appropriate scale of the wavelet transform. When the dimensionality is minimized the information may be misplaced.
Outcome: This feature extraction algorithm controls the lower dimensionality and less error by choosing the scale within which the nearest less scale.
Outcome: The main use of this feature extraction is chosen automatically. And the standards of clustering with extracted features are superior than that with features corresponding to the scale prior and posterior scale moderately for the pre-owned data sets.

Bagnall at el[9], demonstrated a technique of clipping ,It is the process of representing binary sequences of data on the basis of median value to assess the dimensionality effect, clustering on clipping data contains many advantages those are it generates good clusters irrespective of outliers.
Outcome: It requires less memory and oerations to perform the task, distance measures are also calculated quickly.
Outcome: It produces accurate clusters and the problems of using clipped data is it contains huge collection of data in dataset, it reduces the speed of the clustering algorithm.

Ernst at el, reported an algorithm for clustering short time series gene expression data [7]. Many numbers of clustering algorithms are not achievable to make a distinction between real and random patterns. They documented an algorithm specifically designed for performing clustering of short time series expression data. Their algorithm acts by assigning genes to a predefined set of model profiles that catch the potential definite patterns that can be look for from the experiment, also explained how to achieve a set of profiles and how to decide the significance of each of these profiles.
Outcome: Notable profiles are retained for further analysis and can be grouped to form clusters. Also they try out their technique on both simulated and real biological data.
Outcome: Using immune response data they exhibited that their algorithm can accurately detect the temporal profile of similar functional categories.
Outcome: Using Gene Ontology analysis the results exhibited that their algorithm is better than both general clustering algorithms and algorithms specifically developed for clustering time series gene expression data.

Li at el, discussed a new clustering method for time series data sets. Clustering Technique for streaming of time series data is a critical task [9]. The huge part of the classical clustering algorithms contains outliers and huge amount of not ordered data. In their article, they demonstrated a new clustering method, which performs clustering of streamed Bi-clipped (CBC) data. Mainly It have three phrases, those are, reducing dimensionality by applying piecewise aggregate approximation (PAA), Bi-clipped technique is used to perform clip the real valued series by bisecting the value column, and clustering.
Outcome: Through the similar kind of experiments, they established that CBC gains huge quality solutions in minimum time when compared with the Mclipped method that is used to clip the real value series through the mean, and unclipped methods. This position is well defined when streaming of time series data that contain outliers.

Jian at el[11], established a clustering algorithm for time series data. Intelligent Traffic System, the research demonstrates about the reasoning of time series of traffic flow is important and informative. Clustering methods are used to examine time series data not only can find some complicate patterns of traffic flow, but also can combine the divisions of highway by their distinct flow of characteristics.
Outcome: They are given an Encoded-Bitmap-approach-based swap technique to increase the classic hierarchical method. Furthermore, their test results demonstrated that their proposed method has an excellent performance on the change trend of time series than classic algorithm.

Beringer at el, place forward a clustering algorithm for parallel data streams. Now a days, the management and

preparing of so-called data streams has become a subject of aggressive research in many fields of computer science such as, e.g., distributed systems, database management systems, and data mining [11]. A data stream can nearly be thought of as a volatile, progressively increasing series of time stamped data.
Outcome: They examined the question of clustering parallel streams of real-valued data, that is to say, progressively evolving time series. In other words, they are interested in combining data streams the development over time of which is worthy of comparison in a specific sense.
Outcome: In order to sustain an up-to-date clustering structure, it is necessary to examine the incoming data in a wired manner, tolerating not more than a fixed time delay.
Outcome: For this purpose, they refined a resourceful wired version of the classical K-means clustering algorithm. Their method's effectiveness is mainly due to a adaptable wired transformation of the authentic data which allows for a quick computation of exact distances between streams.

Wang at el[4], explained characteristics based clustering of time series data sets. They proposed a technique for clustering of time series based on their constructional characteristics. Unlike other options, their suggested technique does not cluster point values using a distance measure, preferably it clusters based on universal features take out from the time series data sets. The feature measures are extracted from each single series and can be fed into random clustering algorithms, along with an unsupervised neural network algorithm, self-organizing map, or hierarchal clustering algorithm. Universal measures are used for outlining the time series are acquired by assigning statistical operations that most excellent capture the latent differentness: trend, seasonality, periodicity, serial correlation, skewness, kurtosis, chaos, nonlinearity, and self-similarity. Since the method clusters using obtained Universal measures, it decreases the dimensionality of the time series data and is much smaller impressible to missing or noisy data. They additionally gives a search technique to identify the better selection from the feature set that must be used as the clustering inputs. Their method has been examined using standard time series datasets already announced for time series clustering and a collection of time series datasets with popular uniqueness.
Outcome: The experimental results proved that their method is able to achieve considerable clusters. The developed clusters are worthy of comparison to those produced by other techniques, but with some hopeful and entertaining variations that can be naturally made clear with knowledge of the universal characteristics of the time series datasets.

Hirano at el, explained an algorithm for clustering the medical time series data. Their work narrates a cluster analysis technique for multidimensional time-series data on clinical testing room's examinations. Their technique symbolizes the time series of experimental results as trajectories in multidimensional space, and contrasts their structural correspondence by using the multiscale contrasting method.
Outcome: It enables us to identify the part-to-part correlation between two trajectories, considering an account the relationships between different experiments.
Outcome: The development distinction can be used in future for performing clustering algorithms to identify the clusters of most similar cases. The technique was used to perform cluster analysis of Albumin-Platelet information in the chronic hepatitis time series dataset.
Outcome: The test results showed that it could generate an interesting cluster of cases that have high equivalence to the fibrotic stages.

Bagnall at el[9], explained clustering of time series clipped data. They presented that the simple technique of clipping the time series data decreases the memory requirements and very much speeds up clustering without removing the accuracy of clusters. They also explained that clipping improves accuracy of clustering when there are outliers in the data, hence serving as a means detection of outlier and a technique of analyzing model misspecification. They treated simulated information from polynomial, autoregressive moving average and hidden Markov models and given that the expected measures of the clipped data used in grouping tend, asymptotically, to those of the unclipped data. Furthermore they showed analytically that, if the sequences are large enough, the efficiency on clipped data is not significantly smaller than the efficiency on unclipped data, and if the sequence consists outliers then clipping outputs in significantly best clusters.
Outcome: Finally, they decorated how to use clipped sequence can be of practical advantage in identifying model misspecification and outliers on two real world time series data sets: an electricity generation bid data set and an ECG

DOI: 10.15680/IJIRCCE.2017. 0503256

data set.

Nakamoto at el, documented a speed clustering algorithm for time series data sets. Their work suggested a speed clustering method for time-series data sets on the basis of a data structure: TWS (Time Warping Squashing) tree. A clustering method is based on a comprehensive search technique is time-taking although its outputs produced high quality. BIRCH, which decreases the number of experiments by data suppressing based on a data structure: CF (Clustering Feature) tree, shows an effective results for such a technique when the input data set contains only numerical attributes. For time-series data sets, moreover, an honest technique of BIRCH is based on a classical Euclidean distance for a pair of series, desperately fails since such a distance measure typically distinct from human's vision. A distance measure is based on DTW (Dynamic Time Warping) is having advantage, but no techniques have been identified for time series data sets in the situation of data-suppressing clustering.
Outcome: In order to get around this problem, they given the TWS tree, which make use of a distance measure based on DTW, and compresses series to an average series. An average series is achieved by a novel method which expected correct shrinkage of a output of DTW.
Outcome: Tests based on the Australian sign language data showed the superiority of the identified technique in terms of accuracy of clusters, while its time taking and the efficiency is insignificant.

Clustering of time series data sets uses gene expression and the data using smoothing spline derivatives was implemented by Dejean at el. Microarray data achieved during time consuming experiments allows the temporal variations in gene expression to be watched carefully. An imaginative postprandial fasting test was handled in the mouse and the expression of 200 genes was monitored with a dedicated macroarray at 11 time points between 0 and 72 hours of fasting. The goal of their study was to hand over an applicable clustering of gene expression temporal profiles. This was acquired by concentrating on the shapes of the curves moderately on the supreme level of expression.
Outcome: In point of fact, they grouped spline smoothing and initial derivative calculation with hierarchical and partitioning clustering techniques.
Outcome: A heuristic approach was identified to tune the spline smoothing measure using both statistical and biological discussions. Clusters are illuminated a posteriori through principal component analysis and heatmap techniques.
Observation: Most outputs were organized to be in correspondence with the literature on the belongings of fasting on the mouse liver and hand over promising instructions for coming up biological observations.
Observation: Currently numerous algorithms are available for performing clustering on pre-processed time series datasets.
Observation: In existing literature states that K-means clustering algorithm for Time series datasets attained the less accuracy and there is an extensive scope of refinement.
Observation: Also there may be a chance to implement the hybrid clustering algorithm for time series datasets, which may fruitfully bring about the additional amount of accuracy.

III. TIME SERIES AND DATA MINING TECHNIQUES

Time series clustering is an important solution to various problems in numerous fields of research, including business, medical science, and finance. However, conventional clustering algorithms are not practical for time series data because they are essentially designed for static data. This impracticality results in poor clustering accuracy in several systems.
To measure distances between time series data in the second level, Dynamic Time Warping (DTW) was used on data with varying lengths, and Euclidean distance (ED) was used on data of equal length. However, CAST algorithm is used twice in this approach, once to generate initial clusters and the other to split each cluster into sub clusters, which is rather complex. The authors in also propose a new multilevel approach for shape-based time series clustering. First, time series data are selected from a generated one-nearest-neighbor network. To generate the time series network, the authors propose a triangle distance measurement to calculate the similarity between time series data. Hierarchical clustering is then performed on the selected time series data. Second, the data size is reduced by approximately 10% using this approach. This algorithm requires a nearest-neighbor network in the first level. The complexity in generating

DOI: 10.15680/IJIRCCE.2017. 0503256

a nearest-neighbor network is , which is rather high. As a result, the authors attempt to reduce the search area by data preclustering (using k-Means) and limit the search to each cluster only to reduce the creation network. However, generating the network itself remains costly, rendering it inapplicable in large datasets. Additionally, the solution to the challenge of generating the prototypes via k-Means when the triangle is used as a distance measure is unclear. In this study, the low quality problem in existing works is addressed by the proposal of a new Two-step Time series Clustering (TTC) algorithm, which has a reasonable complexity. In the first step of the model, all the time series data are segmented into sub clusters. Each sub cluster is represented by a prototype generated based on the time series affinity factor. In the second step, the prototypes are combined to construct the ultimate clusters. To evaluate the accuracy of the proposed model, TTC is tested extensively using published time series datasets from diverse domains. This model is shown to be more accurate than any of the existing works and overcomes the limitations of conventional clustering algorithms in determining the clusters of time series data that are similar in shape. With TTC, the clustering of time series data based on similarity in shape does not require calculation of the exact distances among all the time series data in a dataset; instead, accurate clusters can be obtained using prototypes of similar time series data.

Data Mining is mainly used to extract the meaningful information or knowledge from huge amount of databases. Data mining techniques and tools are mainly used to discover unknown patterns and movements from the input data set. Its main aim is to automatically discover the hidden patterns in the dataset with fewer amounts of user effort and inputs. Main contribution is used to make the decision and also used in estimating future tendency of market. Many business organizations use data mining as a tool these days for knowledge discovery and performing analysis as it easily assess patterns and trends of related to market and generate efficient outputs. The following are familiar techniques in Data mining:

Association: Association is the well-known and well established technique for performing data mining. It is also called as relation method because which is used to discover patterns and paradigms based on the relationship of items from input datasets.
Classification: Classification is a classical data mining method which is used to classify each tuple in a data set into one of already defined group of classes. It is a classical data mining approach and it is mainly based on machine learning.
Clustering: Clustering is a basic data mining approach it is used to create cluster of elements that must consists of more similar behavior using preprogrammed method. Clustering and classification is moderately non identical techniques.
Prediction: Prediction is one of the useful data mining method which is used to discover relationships among different kinds of independent variables also discovers the association among different dependent , independent variables.
Sequential Patterns: Sequential Patterns is one of the predefined data mining methods that is used to finding homogeneous patterns or systematic events in transactional data over a business point of time.The following are familiar tools in Data mining:

WEKA: Weka is a most popular data mining and machine learning tool which was developed by university of Waikato of New Zealand that consists of several predefined preprocessing and algorithms of Data Mining and Machine Learning, Weka tool is developed by using a programming language of JAVA. WEKA is also provides various kinds of machine learning algorithms to the different kinds of data mining problems. These predefined algorithms may directly apply to the input dataset. WEKA will supports various kinds of input file formats like ARFF, CSV, C4.5, etc.
TANAGRA: TANAGRA is one of the most popular open source software for researchers, they can access to the pseudo code of the algorithms and also they can add their developed methods and techniques also performs the comparison between different algorithms, if we consists of standard software distribution license.
MATLAB: Matlab is also one of the data mining tools which were developed by using high level language. It offers an interactive executable environment for performing visualization, mathematical computation and programming. This consists of predefined mathematical functions; Matlab is used to explore various techniques and helps to achieve a best solution quicker than with the spreadsheet of most popular programming languages like C, C++ and JAVA. Also used to perform analysis of data, writing algorithms, and finding new models and applications.
.NET FRAMEWORK: It is one of the useful software framework introduced by the Microsoft company which is

basically works with windows environment. It establishes reliable communication and homogeneous applications. It is also supports interoperability feature among various programming languages.

IV. CONCLUSION

This paper appraised contemporary art in Clustering and Time series data stream assessment of a Software System. Assessment illustrates that there is huge progress in utilization of Time series for clustering. Along with that it is also clear that the new dimensions in usage of clustering and invention of new metrics [16] gives a greater scope for Research in Time series data stream directions and strategies. By seeing the growing fame of Time series data stream, possibility of developing new models are very high, which would be useful in all aspects to build a system in an efficient manner. Considering the present research attempts in this domain, this work identifies the pitfalls and proposes the solutions,

Firstly, the time series data is most captured from a random source. Hence formatting of the data to reduce the noise from the captured data and make it more analysis ready is the demand of the work. Secondly, an algorithm will be proposed to reduce the visible data points in order to maintain a piecewise aggregate approximation sets for the time series data. Thirdly, a novel predictive analytical hybrid clustering model is to be designed and implemented.Lastly, the existing approaches such as K-Means, Bisecting K-Means, DBScan, OPTICS to be compared with the proposed method for establishing the novelty and better performance.

ACKNOWLEDGEMENTS

We are very thankful to all the authors in the reference list for preparing this CSA paper in a realistic manner.

REFERENCES

[1] Saeed Aghabozorgi, Teh Ying Wah, Tutut Herawan, Hamid A. Jalab, Mohammad Amin Shaygan, and Alireza Jalali, “A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique”, Scientific World Journal. 2014.

[2] Xiang Lian, “Efficient Similarity Search over Future Stream Time Series,” IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 1, pp. 40-54, Jan 2008.

[3] L. M. Talbot, B. G. Talbot, R. E. Peterson, H. D. Tolley, and H. D. Mecham, “Application of fuzzy grade-of membership clustering to analysis of remote sensing data,” Journal on Climate, vol. 12, pp. 200– 219, 1999.

[4] X. Z. Wang and R. F. Li, “Combining conceptual clustering and principal component analysis for state space based process monitoring,” Ind. Eng. Chem. Res. Vol. 38, pp. 4345–4358, 1999.

[5] L. Kaufman, and P. R. Rousseeuw, “Finding Groups in Data: an Introduction to Cluster Analysis,” John Wiley: NY, 1990.

[6] Pedro Pereira Rodriguess and Joao Pedro Pedroso, “Hierarchical Clustering of Time Series Data Streams,” IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 5, pp. 615-627, May 2008.

[7] Villie Hautamaki, Pekka Nykanen and Pasi Franti, “Time Series Clustering by Approximate Prototypes,” IEEE 2008.

[8] Ashish Singhal, and Dale E Seborg, “Clustering Multivariate Time Series Data,” Journal of Chemometrics, vol. 19, pp. 427-438, Jan 2006.

[9] A. J. Bagnall, and G. J. Janacek. “Clustering time series from ARMA models with Clipped data,” ACM Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 49-58, 2004.

[10] A. Doucet S. Godsill C. Andrieu "On sequential Monte Carlo sampling methods for Bayesian filtering" vol. 10 no. 3 pp. 197-208, 2000.

[11] J. B. Tenenbaum V. de Silva J. C. Langford "A global geometric framework for nonlinear dimensionality reduction" vol. 260 pp. 2319-2323, 2000.

[12] M. Belkin P. Niyogi "Laplacianeigenmaps for dimensionality reduction and data representation", vol. 15 pp. 1373-1396 2003.

[13] R. Talmon R. Coifman "Empirical intrinsic geometry for nonlinear modeling and time series filtering" National Academy of Science, vol. 10 no. 31, 2013.

[14] T. Berry J. Harlim "Semiparametric forecasting and filtering: correcting low-dimensional model error in parametric models" Journal of Computer. vol. 308 pp. 305-321, 2016.

[15] T. Berry J. Harlim "Forecasting turbulent modes with nonparametric diffusion models: Learning from noisy data" Phys. D: Nonlinear Phenom, vol. 320, pp. 57-76, 2016.

[16] Amjan Shaik et al, “Object Oriented Software Metrics and Quality Assessment: Current State of the Art”, 2012, IJCA

[17] Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Computing Surveys. 1999;31(3):316–323.

[18] Keogh E, Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Mining and Knowledge Discovery.2003;7(4):349–371.

[19] Lin J, Vlachos M, Keogh E, Gunopulos D. Advances in Database Technology—EDBT 2004. 2004. Iterative incremental clustering of time series; pp. 106–122.

[20] Rani S, Sikka G. Recent techniques of clustering of time series data: a survey. International Journal of Computational and Applied. 2012;52(15):1–9.

[21] Ratanamahatana C. Multimedia retrieval using time series representation and relevance feedback. Proceedings of the 8th International Conference on Asian Digital Libraries (ICADL '05); 2005; pp. 400–405.

BIOGRAPHY

Mr.T.Rajesh is graduated with B.Tech in 2006 from JNT University, India and completed M.Tech from CBIT, India during 2009. He is presently working as Assistant Professor of the Department of Computer Science and Engineering, G. Narayanamma Institute of Technology & Science for women College, India. He has about 7 years of teaching experience. His areas of interest include Data Mining, Machine Learning, Software Engineering.

Dr. K. Venu Gopala Rao is presently working as Professor of the Department of Computer Science and Engineering, G. Narayanamma Institute of Technology & Science for women for women College, India .He has published more than twenty papers in national/International journals. His areas of interest include E-Learning, Software Engineering, Data Mining, Networking and etc. He has about 25 years of teaching experience. He is guiding many research scholars and has published many papers in national and international conference and in many international journals.