Time Series Clustering- Introduction to Healthcare System

T. Rajesh, K.V.G.Rao

Abstract— A clustering technique is an appropriate solvable approach for classifying information while no existence of premature information pertaining to class labels, using promising techniques like cloud based computing and big data over latest years. Investigating awareness was gradually piled up with unsupervised methods such as clustering approaches to pull out useful information from the data set available. Time series based clustering data was used in most of the technical domains to extract information enriched patterns to power the data analysis which extracts useful essence from complicated as well as large data sets. It is mostly not possible for large datasets using classification approach whereas clustering approach will resolve the problem with aid of unsupervised techniques. In the proposed methodology, main spotlight on time series health care datasets, one of the kind of admired data in clustering approaches. This summary will expose 4 major components of Time series approaches.

Keywords: clustering, Time-series, Health care evaluation measure, Representation.

I. INTRODUCTION

An unsupervised mining approach is termed as Clustering where related information is grouped without prior information. Groups are created by combining objects which has been more similar with other kind of objects. It is a helpful technique for analyzing the data as it can identify the patterns by combining data which is unlabeled in to other groups. Clustering technique is utilized for analyzing the data with aid to reports generation and pre processing process for extra mining techniques or as an integral core of an expert systems.

Due to the advancement of data storage along with processors, many applications found the opportunity to store as well as to keep data available for more time interval. Data in most of the applications are stored as time series representation format. The examples for such are finance, stock data, weather, health care (eg:- ECG , BP measurements etc). Accordingly, Many Researchers are working in a variety of fields such as finance, Health care and meteorological. This huge volume of time series data supported the chance for performing analysis of time series for numerous mining researchers.

Consequently, researchers have worked in different domains for detection of anamoloy, motif discovery of motif, segmentation and analysis of trend.

Time series clustering (TSC):-

Time series is a temporal progression, which is a continuous or real valued data, it is large dimensional and high volumes of data in size. Mainly Time series clustering objectives are as below:

Such data consists of useful essence which can be collected through pattern detection. Cluster analysis arises as best suitable answer to reveal those patterns in time series. The data set consists of high volumes of data and cannot be handled effectively by users hence most of the users prefer to use structured result.

This cluster analysis is widely used approach and as a common subroutine in many complicated algorithms for mining such as classification and rule discovery indexing. Its visualization can help to recognize about data, clustering and datasets anomalies more quickly.

Definition: TSC N-time data series, D={R1,R2,R3,----Rn}. The process of clustering of D onto C={C1,C2,C3, C4,….C_k}, such that similar time-series are combined using a resemblance distance measure is called TSC.

C_j cluster, while $D=U_{j=1}^{X}C_j$ & $C_j \cap C_K = \Phi$ such that $j \neq k$

Based on Research work Time series clustering techniques are categorized in to three categories those are while TSC, subsequence time series clustering (SC), Time point clustering (TPC) are represented in fig 1.

Fig 1. Hierarchy of TSC

WTSC:-

WTSC is represented like grouping of time series sets pertaining to the resemblance measure.

Sub-sequence TSC:-

It is grouping of a sub sequence of time series which are collected through sliding window.

TPC:-

It is other category of clustering technique. It is used to group time points by combining the temporal proximity time point values..

Whole TSC uses one of the methodologies as given below to group the time series (TS) data.

a) Customize the standard clustering approaches inorder that they are well-matched with time series (TS) data. In this technique, resemblance measure is updated to compatible with TS data.

b) Translating TS data into objects as input of standard clustering algorithms.

c) Multi-step clustering approach accepts input in the structure of multi resolutions of TS data.

Generally Time series clustering contain 3 different approaches namely Model-based, feature based, shape based techniques. Shape based techniques, Time series data shapes

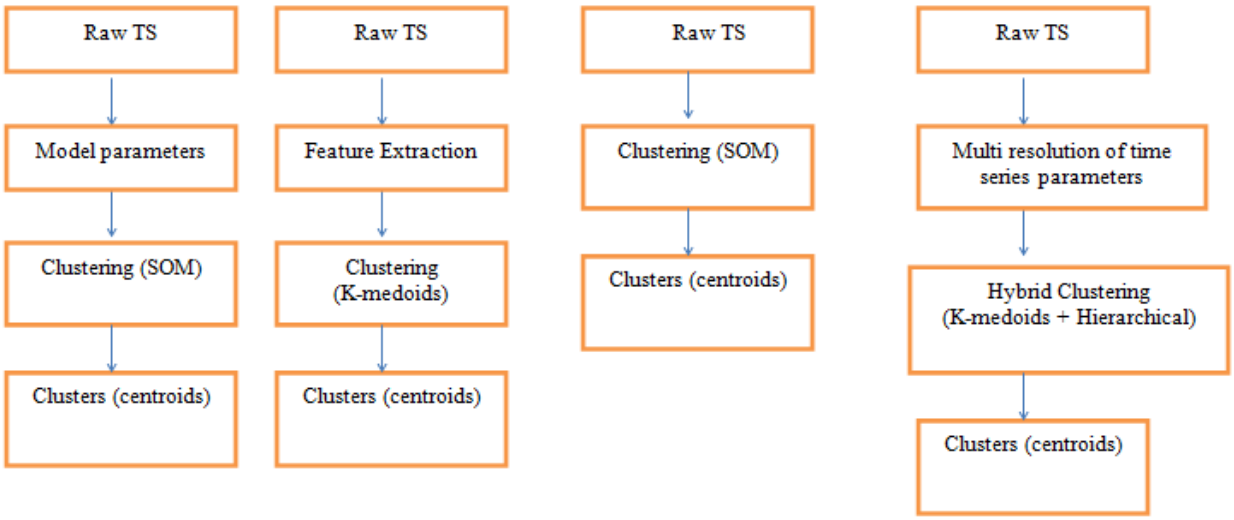
Revised Manuscript Received on October 15, 2019.

Mr.T.Rajesh, Asst Professor, GNITS, Shaikpet , Hyderabad, Telangana, India.

Dr. K.V.G.Rao, Professor, GNITS, Shaikpet , Hyderabad, Telangana, India.

are matched as well as possible. Feature based techniques, Time series raw data are translated in to feature vector of low level dimension. Model based techniques, Time series raw data are converted in to the model parameters.

Whole Time-series clustering



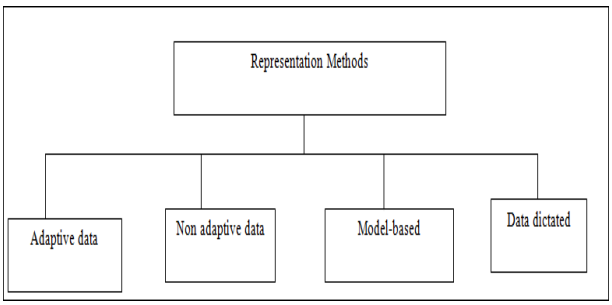
II. PAPER ORGANIZATION

Remaining of this paper, we are providing a major component for TS clustering approaches in addition to the evaluation techniques and similarity or resemblance procedures are existing to check clustering validity. Section III provides the approaches for representing TS data, sections III, IV, V are used to provide information about cluster analysis prototypes and algorithms respectively, section VI are explained about evaluation similarity measures and finally the paper contains the conclusion in section VII.

Time series clustering mainly contains four major components: Representation approach or (dimension reduction), resemblance measures, TSC along with cluster evaluation.

III. TS REPRESENTATION TECHNIQUES

The first and foremost part of clustering of TS data is dimension decline. TS representation is also called as TS Dimensionality reduction.



Adaptive data

It is applied on all most all TS data sets and it is used to limit the global construction error using non-equal segments.

Non data adaptive

It is applied on fixed length TS data (length is equal)

Model-based:-

Used to represent TS data in a statistical model.

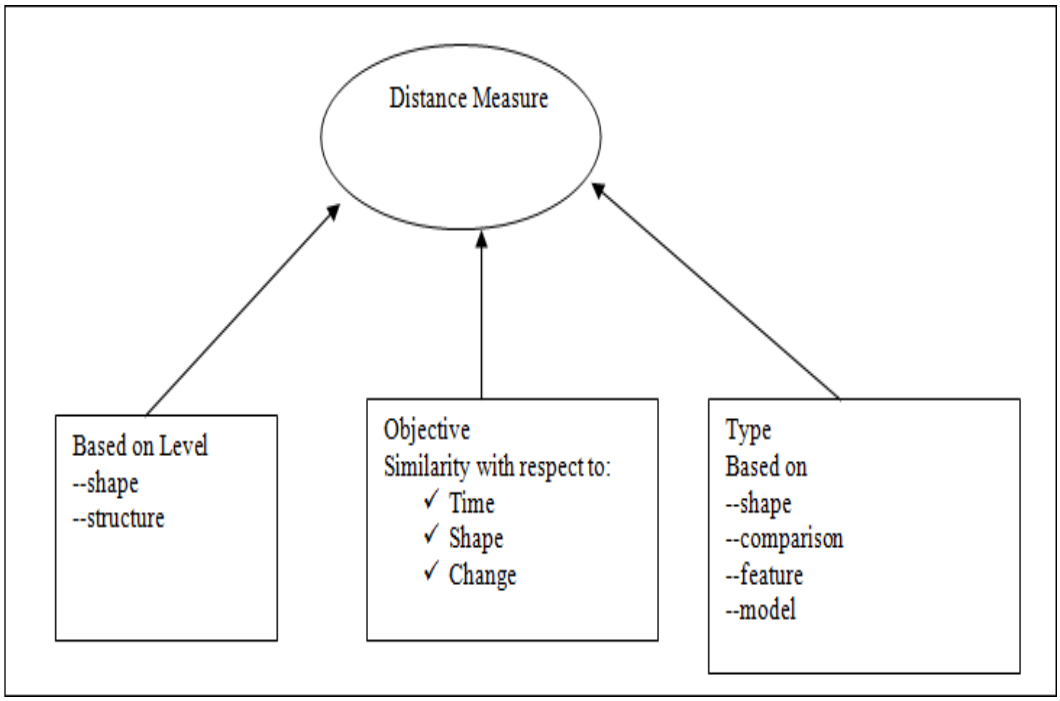
Data dictated

In this method, compression ratio is represented involuntarily based on clipped TS data.

Dissimilarity (distance) Measures:-

Time series clustering highly depends on distance measures. There exist several dissimilarity measures to calculate the distance among TS. The Euclidean distance (ED), Dynamic Time warping (DTW), HMM-based distance, Hausdorf distance, Modified Hausdorf (MODH), longest common subsequence (LCSS) is the commonly available dissimilarity measures used to estimate the distance between TS data.

Selection of an appropriate dissimilarity measure method relies on the feature of TS data, TS length, TS depiction approach and objective of TSC to a maximum extent. It is represented in figure shown below.



TSC prototypes:-

The representation object of cluster is required in clustering of TS data. The methods to solve the problem of cluster prototype inaccuracy, specifically in partition clustering techniques i.e. k-means, k-medoids or hierarchical approach.

Generally we have 3 different techniques to define prototypes:-

Using Time series medoid.

Time series average.

prototype based on local search.

IV. TSC ALGORITHMS

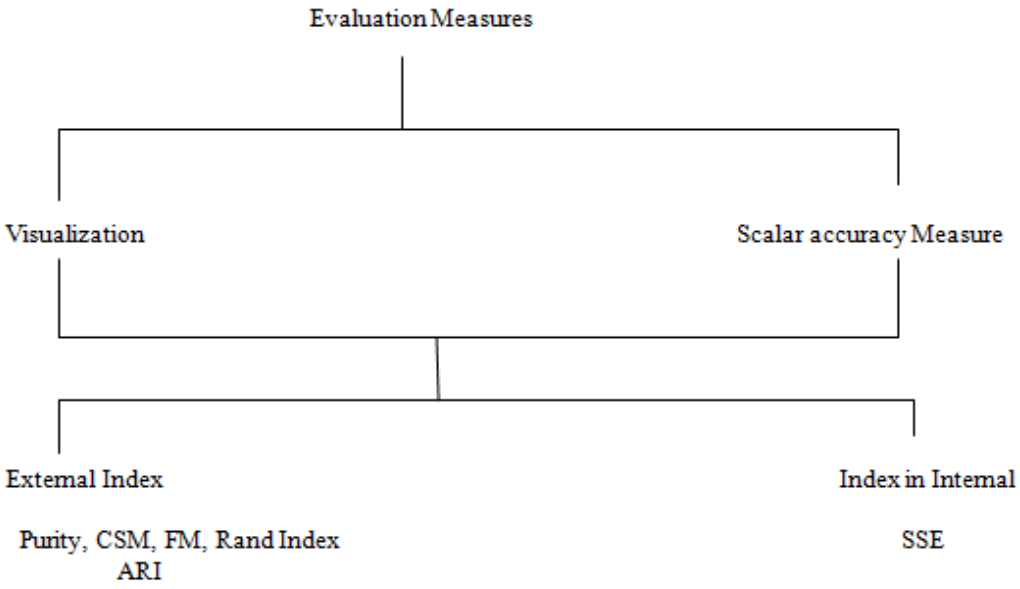
In Mining clustering techniques are categorized in to six types: partition techniques, hierarchical, model-based, and grid-based, multi-step and density based techniques.

V. CLUSTERING ALGORITHMS

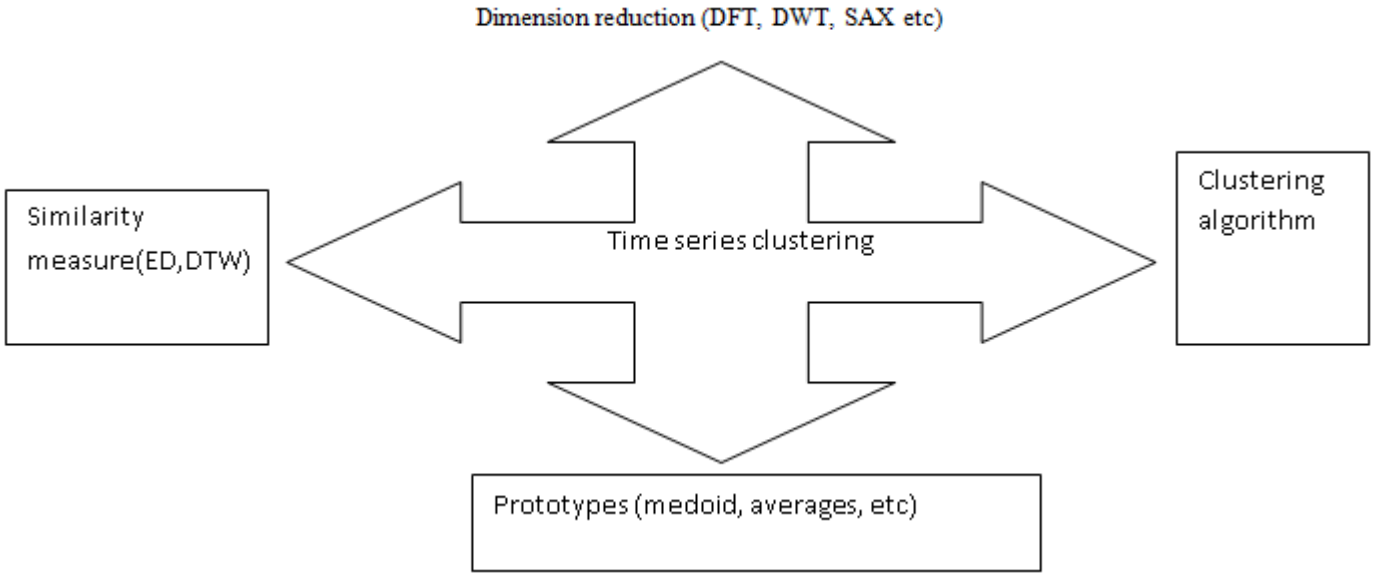
Partition techniques: k-means, k-medoid, clara, clarans.
Hierarchical technique: Top-down, Bottom-up
Model based: COBWEB, SOM, and ARIMA
Density based: DBSCAN, OPTICS
Grid based: STING,Wave cluster
Multi-step: To enhance the cluster representation quality, dissimilarity measure, and cluster prototype by a new approach (Usually hybrid technique) for grouping of TS data.

VI. TS EVALUATION

TS visualization and scalar measures are the main methods for evaluating cluster quality. Methods are used to evaluate time series cluster quality is shown in figure below.



Four aspects of focusing Time series clustering are



Time series clustering in Health care Domain:

Recently, drastically change of internet and computer technology leads to high volumes of time series information is available in various areas such as weather change, outlier's analysis, speech, and Health care (ECG) text mining and so on. Among from all these areas Healthcare field is essential.

In our Research plan, an ECG data is chosen as TS data set and TS clustering algorithms will be applied to form clusters. Many Researchers addresses TS clustering is the clustering of separate TS data such as heartbeats.

Our Research work is to focus on ECG data to identify the abnormal heart beat problems which leads to detection of cardiovascular diseases early, Hence it is useful to avoid premature deaths.

VII. RESULTS

The clustering algorithms was applied on the section of MIT BIH values into 2 data sets as utilized in Lannoy [10] is explained. The AAMI standards are analyzed. The experiment results are compared to the works that uses AAMI based inter patient classification. There are two major challenges in the classification of Heart beat.

Inter patient heart beat variations, 2) Intra patient heart beat variations. Each class of dataset notifies inter patient heart beat regulations i.e. across various patients, Heart beats of similar class may have changes because of the

patient related data. Intra-patient data, every class of dataset may contain changes. Existing methods such as random projection (PR) and support vector machine (SVM) classification techniques are used to ensemble to identify the V class then ratio of the RR regular interval to mean RR interval values was compared to an already computed threshold values to identify SVEB on the data values of MIT BIH Arrhythmia dataset, but it uses three classes with aid of experimental analysis.

	Sensitives				+ve predictive values (PPV)				TCA %
Approach	N	S	V	F	N	S	V	F	%
Haung [6]	99.2	91.1	93.9	-	95.2	42.2	90.9	-	93.8
SVMensemble [12]	88.9	79.1	85.5	93.8	98.9	35.9	92.8	13.74	86.6
Weighted CRF+L1 Lannoy [10]	79.8	92.6	85.2	84.5	-	-	-	-	85.4
Weighted LDA [3]	95	77	81	-	98	38	87	-	93
MLP [5]	89.6	83.2	86.8	61.1	99.3	33.5	75.9	16.6	89
Hierar, SVM [4]	86.3	82.6	80.9	54.9	-	-	-	-	85.6

VIII. CONCLUSION

Many researchers have been explored the clustering of TS data, traditional clustering techniques are fail to work efficiently to cluster TS data. Due to large dimensionality reduction, high volumes of feature correlation, high amount of noise. We explained the cluster representation techniques in this paper can be concludes that important goal of clustering is to minimize the construction error.

In Future work, we want to present Time series clustering techniques on ECG data also need to identify the unsupervised methods for long-term ECG data monitoring to detect cardiovascular diseases early to avoid premature deaths.

REFERENCES

1. AAMI, “Testing and reporting performance results of cardiac rhythm and st segment measurement algorithms.ANSI/AAMI EC38:1998” ANSI/AAMI EC13:2002, Arlington, VA Association for the Advancement of Medical Instrumentation, 2002.
2. Database MITBIH Arrhythmia (available online) http://www.physionet.org/physio-bank/database/mitdb/
3. Llamedo Soria M, Martinez JP, “Heartbeat classification using feature selection driven by database generalization criteria". IEEE Trans Biomed Eng, 58:616625, 2011
4. Park KS, Cho BH, Lee DH, Song SH, Lee JS, Chee YJ, Kim IY, Kim SI, “Hierarchical Support Vector Machine Based Heartbeat Classification Using Higher Order Statistics and Hermite Basis Function." Proceedings of 35 Annual Computers in Cardiology Conference, IEEE Bologna,2008.
5. Mar Tanis, Zaunseder S, Martinez JP, “Optimization of ECG classi_cation by means of feature selection." IEEE Transactions Biomedical Engineering,58, pp.2168-2177,2011.
6. Huang et al. “A new hierarchical method for inter-patient heartbeat classification using random projections and RR intervals." Biomedical Engineering On Line 2014 13:90.doi:10.1186/1475-925X-13-90 (2014)
7. G. Huang, S. Song, J. N. D. Gupta, and C. Wu, “Semi-supervised and Unsupervised Extreme Learning Machines,” IEEE Transactions on Cybernetics, 2014.
8. Yeh Y.C., Wang, W.J.,Chiou, C.W. “A novel Fuzzy c-means method for classifying Heartbeat cases from ECG signals" .Measurement, 43,pp 1542-1545, 2010.
9. K. Tan, K. Chan, and K. Choi, “Detection of the QRS complex, P wave and T wave in electrocardiogram." Advances in medical signal & information processing, pages 41-47, 2000.
10. G de Lannoy JD D Francois, Verleysen M, “Feature Relevance Assessment in Automatic Inter- patient Heart Beat Classification." In Proceedings of the International Conference on Bio-inspired Systems and Signal Processing, Valencia, Spain, 13-20, 2010.
11. de Lannoy G, Francois D, Delbeke J, Verleysen M, “Weighted Conditional random fields for supervised inter-patient heartbeat Classification." IEEE Transactions in Biomedical Engineering, 59:241-247,2012.
12. P. de Chazal and R. B. Reilly, ”A patient-adapting heartbeat classifier using ECG morphology and heartbeat interval features" IEEE Transactions on Biomedical Engineering, vol. 53, no. 12, pp. 2535--2543, (2006)

13. Zhang ZC, Dong J, Luo XQ, Choi KS, Wu XJ: "Heartbeat classification using disease-specific feature selection." Comput Biol Med, 46:79-89. (2014)
14. T. Rajesh, Y.S. Devi, K.V. Rao, Hybrid clustering algorithm for time series data-a literature survey, in: 2017 International Conference on Big Data Analytics and Computational Intelligence, 2017, pp. 343–347.
15. Mohamed Elgendi, Mirjam Jonkman, Friso De Boer, "Premature Atrial Complexes Detection using The Fisher Linear Discriminant." 7th IEEE International Conference on Cognitive Informatics (ICCI), pp. 83-88, Aug.2008.

AUTHOR PROFILE

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 CSE, G. Narayanamma Institute of Technology & Science for women College, India. He has about 7 years of teaching experience. His areas of interest includes Data Mining, Machine Learning, Software Engineering, and etc.

Dr. K. Venu Gopala Rao is presently working as Professor of the Dept. of Computer Science; G. Narayanamma Institute of Technology & Science 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.