International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Pre-processed Hierarchical Clustering for Time Series Data Streams

V. Kavitha, A. V. Senthil Kumar, N. Revathy, C. Daniel Nesa Kumar

P. Hemashree

Abstract*: The behaviour of the human body is based on the signals of chemical, electrical origin. These signals afford information that may not be directly perceptible but some information is hidden in the structure of the signal. These hidden signal information has to be translated in some way before the signals can be given useful analysis. The transformation of human body signals has been discovered useful in explaining and identifying various pathological conditions. The process of transformation is comfortable to perform since involves a limited manual effort like visual investigation of the signal generated as a result. In spite of these signals with their complexity is often considered and consequently biomedical signal processing has become an essential task for extracting significant clinical information hidden from the original signal[1]. Time series data streams constitute numerous dimensions and noisy features. Therefore, detecting the original clusters in high dimensional noisy features time series data stream is a dispute task. The challenging task involved in time series data stream are noisy and high dimensional. The existing technique is incapable of handling noisy high dimensional data stream. The most important key objective of this research part is to develop a novel pre-processing feature selection technique for discarding the noisy data is a vital successful process. Therefore this technique achieves minimum time complexity. Pre-processing feature selection is an established technique to deal with the time series data stream with noisy and high dimensional [3]. Furthermore, this innovative feature selection approach is boost up the cluster process without noisy and also it accomplishes the quality clusters with minimal time interval.

***Keywords* : Feature Selection, High Dimensionality, Time Series Data Stream, Preprocessing, Fuzzy Logic**

Manuscript published on 30 September 2019

* Correspondence Author

Dr.V.Kavitha*, PG & Research Department of Computer Applications, Hindusthan College of Arts and Science, Coimbatore, India, Email kavithahicas@gmail.com

Dr.A.V.Senthil Kumar, PG & Research Department of Computer Applications, Hindusthan College of Arts and Science, Coimbatore, India, Email avsenthilkumar@yahoo.com

Dr.N.Revathy, PG & Research Department of Computer Applications, Hindusthan College of Arts and Science, Coimbatore, India, Email drnrevathy@gmail.com

Mr.C.Daniel Nesa Kumar, PG & Research Department of Computer Applications, Hindusthan College of Arts and Science, Coimbatore, India, Email danielnesakumar@gmail.com

Mrs.P.Hemashree, PG & Research Department of Computer Applications, Hindusthan College of Arts and Science, Coimbatore, India, Email mantar.253@gmail.com

I. INTRODUCTION

The behavior of the human body is based on the signals of chemical, electrical origin. These signals afford information that may not be directly perceptible but some information is hidden in the structure of the signal. These hidden signal information has to be translated in some way before the signals can be given useful analysis. The transformation of human body signals has been discovered useful in explaining and identifying various pathological conditions. The process of transformation is comfortable to perform since involves a limited manual effort like visual investigation of the signal generated as a result. In spite of these signals with their complexity is often considered and consequently biomedical signal processing has become an essential task for extracting significant clinical information hidden from the original signal.

Traditionally, biomedical signals are evaluated visually and based on manual procedure manner; those measurements are relatively inadequate for diagnosing the real factor. Extracting real features from the biomedical signal processing is a clinical challenge. The innovative feature selection methods can be designed to imitate the manual measurement and supporting the human expert for diagnosing the disease. Moreover, the minute variations in heart beat rate cannot be perceived by the human eye but they have been discovered to contain very effective and beneficial clinical information.

II. FORMULATION OF THE PREPROCESSED HIERARCHICAL CLUSTERING

The research of this part is utilizing the new feature selection technique of pre-processing feature selection approach. This approach is applied with the time series data stream for the purpose of discarding the unwanted noisy data. This pre-processing leads to afford the high quality cluster with the least possible time complexity.

III. PERSPECTIVE OF THE PREPROCESSED HIERARCHICAL CLUSTERING

This part of research work is predictable to find out the noisy feature and reduce the dimensionality of the time series data stream for experimental clustering. Hence, the work is superior with the prominent technique of pre-processing feature selection for high dimensional noisy data stream.

Noisy data are ubiquitous in time series data stream within real classes which leads to increase the time complexity. Hence, the irrelevant feature will find out inside the time series data stream and it must be discarded using the technique of pre-processing feature selection. This leads to attain the nominal time complexity with high quality clusters.

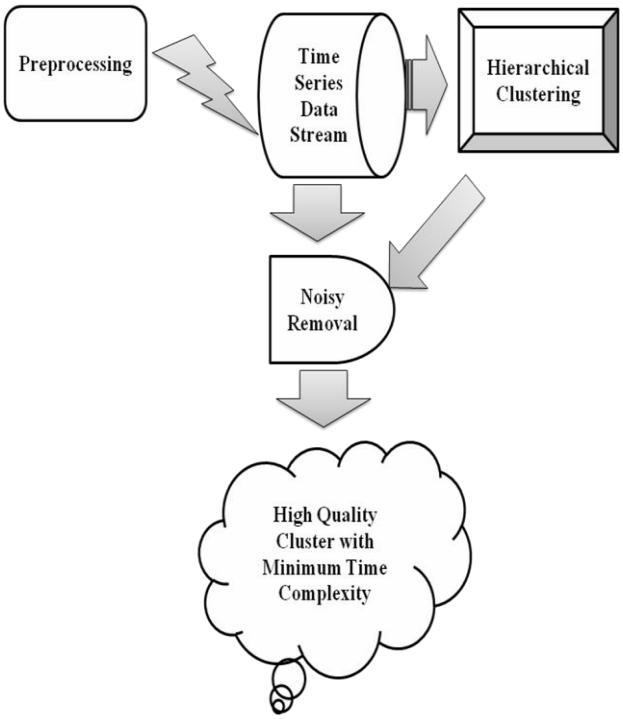

Fig. 1 Approach of Preprocessed Time Series Data Stream

Definitions and Notations

Notation 1: A Feature Space is a set of feature subsets capable of with a vector space interpretation and a distance metric described on it. The valuable information may not loss in the feature space in the case of unsupervised feature selection.

Notation 2: In the dispersion region of spatial Π_{SL} and Ω are given

Where T_i - the term of union of all partitions.

Ω - is a feature space which is bounded

Π_{SL} - is an partition of SL – level

$\Omega = va_1 \cup va_2 \ldots..\cup va_\rho$

$SL = \{sl_1, sl_2, \ldots\ldots.sl_n | sl_i \in Z\}$

Z = Feature space

$\rho = \prod_{i=1}^{n} sl_i$

$\Pi_{SL} = \{ va_1, va_{2,\ldots\ldots} va_n\}$ of Ω

Notation 3: SE – Spatial Entropy is SL level of term T_i

$O_{SL}^{(x)}(T_i)$ | x = 1,2,….. ρ and $O(T_i)$

and entropy of the term Ti described as

$$H_{SL}(T_i) \equiv -\sum_{x=1}^{\rho} P_i(va_x) \log P_i(va_x)$$

Where P is termed as the probability of the term T_i and

$P(va_k) \equiv O_{SL}^{(x)}(T_i) \; / \; O(T_i)$ which is occurred in the partition va_k.

Notation 4: SV – Spatial Variance is SL level of the term T_i

Given that $O_{SL}^{(x)}(T_i)$ | x = 1,2,….. ρ

Hence, the expression SV denoted as

$Var_{SL}(T_i) \equiv Var([\; O_{SL}^{(1)}(T_i) \;, O_{SL}^{(2)}(T_i), \ldots\ldots.. O_{SL}^{(\rho)}(T_i)\;])$,

Where the right side of the equation denotes for the variance of local number occurrence $O_{SL}^{(1)}(T_i)$, $O_{SL}^{(2)}(T_i), \ldots\ldots.. O_{SL}^{(\rho)}(T_i)$.

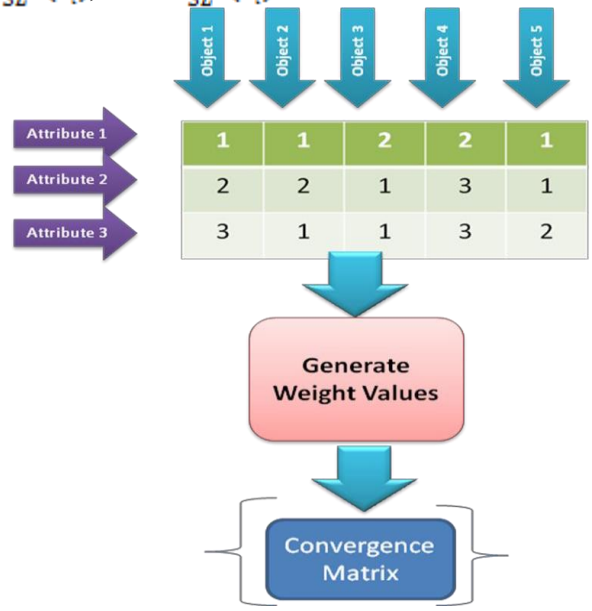

Fig. 2 Work flow of Pre-Processing

Pre-Processing Feature Selection Algorithm

Input: x_0, a, A Set of real feature points

Output: Ā - The set of results with feature selection variables.

STEP 1: If SE and SV are involved, then build a SL-level partition of subset Π_{SL} of A with x_0 parts through a spatial index method, and then design a subset list FP_{SL} (Π_{SL} $\Pi_{SL,}$ $\Pi_{SL,}$ . . . Π_{SL}) in which the i^{th} component denotes the subset with respect to the i^{th} feature; if 1-d-SE and 1-d-SV are involved, then build a 1-d subset $\Pi_{SL(i)}$ of A with x_0 parts for every feature T_i, and then design a subset list FP_{SL-1} = ($\Pi_{SL(1,)}$ $\Pi_{SL(2)1,}$ $\Pi_{SL(3),}$ . . . , $\Pi_{SL(n)}$) where the i^{th} component refers the subset with respect to the i^{th} feature.

STEP 2: For each feature T_i, compute the local frequency number $O_{SL}^{(x)}(T_i)$ for SE and SV or $O_{SL}^{(x)}(T_i)$ for 1-d-SE and 1-d-SV for x = 1, ... ρ based on the subset partition list $FP_{SL\ or}$ FP_{SL} - 1created by step 1, and then determine the equivalent invariant metric statistic.

STEP 3: Calculate the coverage matrix CM with respect to A.

STEP 4: Solution to solve the weight wt based on the model with consequence factor variables namely af and the parameter bf.

STEP 5: Create a hard selection (0/1) or soft (weighted) selection on A in accordance with weight value wt and built a new set of feature vectors Ā. A feature ratio refers the ratio of the particular features should be formerly defined.

In the above diagram describes about the work flow of pre-processing feature selection technique. This technique is applied to the raw data of time series data stream.

Retrieval Number: C3961098319/19©BEIESP
DOI:10.35940/ijrte.C3961.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The weight value will be generated for each and every attribute of the stream data then the mean value will be calculate for each object. According the weight values the convergence matrix will be formed. Based on the convergence matrix the noisy feature is discovered.

IV. EVALUATION OF PERFORMANCE FACTORS

The performance factors of time complexity, cluster quality of inter and intra clusters are assessed for the empirical evaluation of the performance analysis. Likely more number of performance factors are utilised to evaluate the preprocessed system.

The following figure charts explains the experimental analysis of the performance factor of time complexity. Which is compared among the two clustering techniques of traditional fuzzy hierarchical clustering and Pre- processed fuzzy hierarchical clustering[2]. Table 1 denotes the comparison between the two hierarchical based clustering technique of Traditional Fuzzy Hierarchical Clustering(FC) and Pre- processed Fuzzy Hierarchical clustering(FS).

Figure 3 demonstrates the time complexity between the fuzzy hierarchical clustering and pre-processing feature selection. In this investigational outcome analysis illustrates that non pre processing hierarchical technique attain excess of execution time for as of high dimensionality and noisy feature of the time series stream data set. In proposed technique of Pre-processing feature selection technique elucidate about its enrichment of time complexity performance due to reduction of noisy features from the time series data stream.

TIME COMPLEXITY					
Technique / Number of Clusters	2	4	6	8	10
Traditional Fuzzy Hierarchical Clustering	1.6789	1.6878	1.7001	1.7222	1.7681
Pre-Processed Fuzzy Hierarchical Clustering	1.4836	1.5299	1.5541	1.5763	1.5997

Table1 : Time Complexity between Traditional Fuzzy Hierarchical Clustering and Pre-Processed Fuzzy Hierarchical Clustering

Existing system of coupled fuzzy hierarchical clustering designate the blue line and the new proposed work of pre processing indicates the red line in the resulting chart. In the x axis specifies the number of clusters with the values of 2,4,6,8 and 10 as an input parameter. And the y axis denotes the time in seconds. The proposed system takes least time performance than the existing related work.

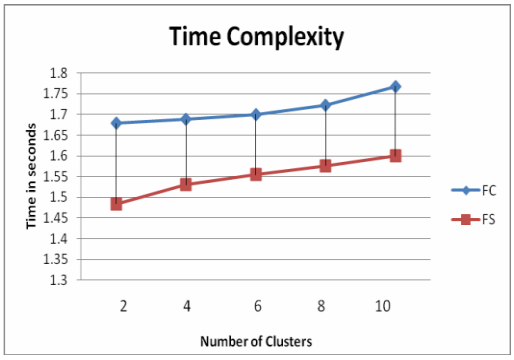

Fig. 3 Time complexity of Pre-Processed Hierarchical Clustering

The following tables describe the detail experimental analysis about inter cluster quality and intra cluster quality. The intra cluster quality denotes that the object distance must be condensed with a particularized cluster. For this distance reduction is suggested for the cluster quality improvement. In the graph figure explain about the difference between the performance of existing technique of Traditional fuzzy hierarchical clustering technique which is indicated in blue line and the proposed technique of pre-processing feature selection is indicated in red line.

V. CONCLUSION

Large number of noisy features occurred in the time series data stream is performed poor clusters due to dominating reason of unsupervisied learning. Suppose that the features are gathered through different measurements such as gram or kilogram, Fahrenheit of Celsius, liter or milli liter and so on. These various kinds of measurements are in different form of feature scales. Therefore, a efficient pre-processing technique is to stabilize all kinds of features. While the stabilization step can eradicate the possibility of noisy features from the stream data with the dominating dimension, it may lead to the loss of valuable scale information appropriate to the clustering process[13]. Most of the existing unsupervised approaches of feature selection methods are dependent on the similarity measure metrics which is defined on the space of the feature.

The performance of the pre-processing technique is determined through the percentage of noisy level. Weather the noisy level is very low and the feature space is suggested as reasonable feature space. For this process the traditional feature selection approach is adequate to differentiate the discriminitive meaningful features from noisy features. Otherwise the noisy level is decided as very high which is refered as unreasonable feature space it needs the effective feature selection approach for the process of discriminative meaningful features from noisy.

Naturally time series data streams constitute countless elements with noisy features. Therefore, diagnosing the genuine clusters in high dimensional noisy features time series data stream is a dispute task. The challenging task involved in time series data stream are noisy and high dimensional.

The existing technique of traditional clustering is incapable of handling noisy high dimensional data stream. Pre-processing feature selection is an established technique to deal with the time series data stream with noisy and high dimensional. Therefore this technique achieves minimum time complexity. Furthermore, this innovative feature selection approach will boost up the cluster process without noise and also it accomplishes the quality clusters with minimal time interval.

Time series data stream encompass with high dimensional and noisy stream data which is a problematical process for congregation of dynamic fruitful cluster. Hence, the pre-processing is enhanced with the hierarchical clustering then the innovative technique is fashioned with the name of Pre-processed Hierarchical Clustering technique. The obstacle of dimensionality and noisy feature detection is dealing with the new techniques which preside over to attain the absolute cluster.

REFERENCES

1. Pantelis n.Karamolegkos, Charalampos Z.Patrikakis Nikolaos D.Doulamis Panagiotis, "An Evaluation Study of Clustering Algorithms in the Scope of user Communities Assessment" Computers $ Mathematics with Applications, Elsevier, Vol No 58, issue no 8, October 2009, Pages 1498 - 1519.
2. Man Abdel - Maksoud, Mohammed Elmogy, Rashid Al-Awadi, "Brain Tumor Segmentation Based on a Hybrid Clustering Technique", Egyptian Informatics Journal, Vol No 16, Issue no 1, March 2005, Pages 1 - 81.
3. Madjid Khalilian, Norwati Mustapha, "Data Stream Clustering: Challenges and Issue", Proceedings of the International Multi conference of Engineers and Computer Scientists 2010 Vol No1, IMECS 2010,March 17-19 2010.
4. Maryam Mousavi1 , Azuraliza Abu Bakar, and Mohammadmahdi Vakilian, "Data Stream Clustering Algorithms: A Review", International Journal of Advance Soft Computer Applications Vol 7, Issue No 3, November 2015, ISSN 2074-8523.
5. Jose R. Fernandez," A Framework and Algorithm for Data Stream Cluster Analysis", International Journal of Advanced Computer Science and Applications, Vol No 2, Issue No11, Pages 87, 2011.
6. Twinkle B Ankleshwaria, " Mining Data Streams: A Survey", International Journal of Advance Research in Computer Science and Management Studies, Vol No 2, Issue No 2, Feb 2014, ISSN: 2321-778.
7. Amineh Amini, Teh Ying Wah, "Density Micro-Clustering Algorithms on Data Streams: A Review", Proceedings of the International MultiConference of Engineers and Computer Scientists 2011 Vol No 1, IMCES 2011, March 16-18, 2011.
8. Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., de Carvalho, A. C. P. L. F., and Gama, J, "Data stream clustering: A survey", ACM Computing Surveys, Vol No 46, Issue No1, Article 13, October 2013, Pages 31.
9. DoniaAugustine, "A Survey on Density based Micro-clustering Algorithms for Data Stream Clustering", International Journal of Advanced Research in Computer Science and Software Engineering Research, Vol No 7, Issue No 1, January 2017.
10. Dure Supriya Suresh, Prof. Wadne Vinod, "Survey Paper on Clustering Data Streams Based on Shared Density between Micro-Clusters", International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395 -0056, Vol No 04 ,Issue No 01, January 2017.
11. Amini A, Wah TY, Saboohi H, "On density-based data streams clustering algorithms: A survey",Journal of Computer Science and Technology, Pages 116–141, January 2014, DOI 10.1007/s11390-013-1416-3.
12. Safal V Bhosale, "A Survey: Outlier Detection in Streaming Data Using Clustering Approache", International Journal of Computer Science and Information Technologies, Vol No 5,2014, 6050-6053 ISSN 0975 - 9646.
13. Prashant V. Desai, Vilas S. Gaikawad, "Novel approach for data stream clustering through micro-clusters shared Density",International Journal of Computer Sciences and Engineering Volume-5, Issue-1 E-ISSN: 2347-2693.
14. M.S.B.PhridviRaj, C.V.GuruRao, "Data Mining - Past, Present and Future - A Typical Survey on Data Streams", Elsevier Procedia Technology", Vol No 12, 2014, Pages 255 - 263.
15. Yisroel Mirsky, Bracha Shapira, Lior Rokach, and Yuval Elovici, "pcStream: A Stream Clustering Algorithm for Dynamically Detecting and Managing Temporal Contexts", Springer International Publishing Switzerland 2015, PAKDD 2015, Part II, LNAI 9078, pp. 119–133, 2015. DOI: 10.1007/978-3-319-18032-8_10.
16. Shufeng Gong, Yanfeng Zhang, Ge Yu1, "Clustering Stream Data by Exploring the Evolution of Density Mountain", PVLDB, 11(4) 2017. DOI: 10.1145/3164135.3164136.

AUTHORS PROFILE

Dr.V.Kavitha had pursued B.Sc., Computer Science from Bharathiar University in 1998, Coimbatore and Master of Computer Applications (MCA) from Bharathidasan University, Trichy in 2009, Master of Philosophy in Computer Science from Alagappa University, Coimbatore in 2005 and Ph.D in Computer Science from Karpagam University, Coimbatore in the year 2014. Area of research is Data Mining. Serving as a Reviewer and Editor in various International and National Journals. At present working as a professor in the Department of PG and Research Department of Computer Applications (MCA) at Hindusthan College of Arts and Science, Coimbatore-641 028. She published 37 papers in International Journals, presented 40 papers in International Conferences and National Conferences. She has 16 years of teaching experience and 10 yrs of Research experience.

A.V.Senthil Kumar obtained his Ph.D in Computer Science. He has to his credit 9 Book Chapters, 174 papers in International Journals, 4 papers in National Journals, 25 papers in International Conferences, 5 papers in National Conferences, and **edited five books in Data Mining, Mobile Computing, Fuzzy Expert Systems, Biometric Authentication and Web Mining (IGI Global, USA).** He is an Editor-in-Chief for 4 International Journals. and **Key Member** for India, Machine Intelligence Research Lab (MIR Labs). He is an Editorial Board Member and Reviewer for various International Journals. He is also a Committee member for various International Conferences.

Dr Revathy Nanjappan had completed B.Sc., Computer Science (2000) and Master of Computer Applications (MCA) 2003 under Bharathiar University. Completed M.Phil. in Computer Science from Alagappa University in the year 2005 and Ph.D in Computer Science from Mother Teresa Women's University, Kodaikanal in the year 2013. Area of research is Neural Networks, Data Mining and Artificial Intelligence. Published a book titled "System Software" at 2018. Received "Teacher in Computer Science and Engineering Award" from Global Out Reach Education Awards 2018-2019 and "Outstanding Educator Award 2018" from International Institute of Organized Research I2OR Awards. Delivered Guest Lecture at various Engineering and Arts Colleges. Completed 7 Online Certification courses in the areas like Java and Python Programming & Data Mining conducted by E & ICT Academy, IIT, Kanpur. Also serving as Reviewer and Editor in various International Journals. At present working as a professor in the Department of PG and Research Department of Computer Applications (MCA) at Hindusthan College of Arts and Science at Coimbatore-641 028 and published 40 papers in International Journals, presented 8 papers in International Conferences and 56 papers in National Conferences.

Mr. Daniel Nesa Kumar C pursed Bachelor of Science from Bharathiar University, Coimbatore in 2006 and Master of Computer Applications from Bharathidasan University, Trichy in 2009 and Master of Philosophy in Computer Science from Bharathiar University, Coimbatore in 2013 and currently working as an Assistant Professor in Department of Computer Applications, Hindusthan College of Arts and Science, Coimbatore, Since 2009.

He has published more than 15 research papers in reputed journals International journals and Conferences. His main research work focuses on Networking, Data Mining, Image Processing. He has 10 years of teaching experience and 5 yrs of Research experience.

Mrs. P. Hemashree has obtained her B Sc (Computer Science, Mathematics, Statistics) in April 2012 from Mount Carmel College, Bangalore and obtained the MCA degree from Coimbatore Institute of Technology, Coimbatore in 2015. She has qualified for the post of Assistant Professor through the State Level Eligibility Test (SET) and the National Eligibility Test (NET) in 2018. She has to her credits 4 International Publications in reputed Journals. She is associated with Hindusthan College of Arts and Science as an Assistant Professor in the PG and Research Department of Computer Applications since 2016. She published 4 papers in International Journals. She has 4 years of teaching experience.