ISSN No. 0976-5697

Volume 4, No. 8, May-June 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Comparative Analysis of Various Data Stream Mining Procedures and Various Dimension Reduction Techniques

Diksha Upadhyay, Susheel Jain, Anurag Jain
Department of Computer Science RITS Bhopal, India
diksha.du31@gmail.com, jain_susheel65@yahoo.co.in, anurag.akjain@gmail.com

Abstract: In recent years data mining is contributing to be the great research area, as we know data mining is the process of extracting needful information from the given set of data which will be further used for various purposes, it could be for commercial use or for scientific use .while fetching the information (mined data) proper methodologies with good approximations have to be used .In our survey we have provided the study about various data stream clustering techniques and various dimension reduction techniques with their characteristics to improve the quality of clustering, we have also provided our approach(our proposal) for clustering the streamed data using suitable procedures ,In our approach for stream data mining a dimension reduction technique have been used then after the Fuzzy C-means algorithm have been applied on it to improve the quality of clustering.

Keywords: Data Stream, Dimension Reduction, Clustering

I. INTRODUCTION

In the growing economy ,as there is the tremendous growth in Information technology so as in data repositories which will be in any form for e.g. From the prospective of commercial sector each firm will store the information(data) regarding their turnover, about their customer statistics ,about challenging factors for an enterprise, profit areas etc of several previous years in the data repositories .data mining is the research area in which the important information will be dig out from the given repository of data store .The generated (produced) information ,then be further used by the business leaders (business analysts) of an enterprise to take the strategic decisions regarding the future policies ,but this is only the commercial application of data mining it can be used mostly in scientific research also. In this study our intension is on presenting the comparison between several data stream mining techniques .Data stream mining is the hot research area now a days in which the main task is to form the cluster (Fuzzy nature).

There is continuous generating data stream [1] from many sources such as from sensor networks, from web user click events, from the flow of internet traffic etc. The streamed data generated from various sources will exhibits the similar characteristics such as infinite arriving nature, uncertain arriving speed, and it have to be scanned in the single pass, the mining phenomena have to be performed on data stream with some conventions such as limited memory space, and with in limited time constraint so the more efficient mining algorithms have to be drawn.

In mining the data clusters have been found using clustering, which is the method of generating several chunks of provided input data , with each chunk (cluster) composed up of elements(data objects) having similar characteristics. Here we are providing the snap shot of various data stream clustering algorithms. In [2] STREAM procedure have been presented to cluster data stream ,To cluster binary stream data a procedure [8] have been presented to cluster data stream where the modification on K-means have been proposed for better result .A time- series clustering mechanism has been presented in [4] to generate the hierarchy of clusters and many more.

A. Dimension reduction:

As we know the generated streamed data (upcoming) from various sources can be of many dimensions from the prospective of geometrical overview, or it may be possible that it will contain dozen of dimensions, so to cluster such data stream is very critical using conventional clustering procedures such as K-means algorithm because we know clustering have to be done with in limited memory capacity and time constraint.

From the survey it have been identified that the clustering algorithms are less efficient with the high dimensional datasets, so to improve their efficiency data sets will passed through some respective dimension reduction procedure [16] such as SVD, PCA, SVM etc, and then after on the reduced dataset the clustering algorithm will be applied for better clustering, we have also provided the comparative analysis for various dimension reduction techniques.

B. Objective of our work:

In this study various data stream mining techniques and dimension reduction techniques have been evaluated on the basis of their usage statistics, application parameters, working mechanism etc.

C. Structure of This paper:

After providing the overview of various concepts in section1,the comparison between various data stream mining algorithms and dimension reduction techniques have been presented in section 2,and then after our proposal have been presented with the working model in the section3,finally section 4 will concludes the paper with possible future enhancements.

II. COMPARATIVE ANALYSIS

A. *Overview of various Fuzzy and other clustering methods (algorithms):*

Table: 1

S. No	Name of technique	Type of data on which it operates	Working mechanism	Implemented on (tool used)
1	Istrap [6]	Stream data	Static analysis is performed, and the methods have been proposed to remove outliers.	Matlab
2	WFCM [5]	Stream data	Data stream is divided in to chunks as per the arrival .the main idea of WFCM is renewing the weighted clustering centers by iterations till the cost function get a satisfying result	C++
3	Improved FCM [7]	Synthetic and real images	A trade-off fuzzy factor and kernel method is used	Matlab
4	DSCLU [9]	Stream data (real &synthetic data)	It separates the process in phases in online phase and offline phase and then operates	JAVA 1.6
5	IFCM [10]	Interval valued symbolic data (with real & synthetic symbolic interval)	This method is objective function based calculation technique	Fuzzy corrected rank index
6	DCWFP-miner [11]	Data stream (continuous , unbounded and high speed data streams)	It uses divide and conquer strategy and from bottom to top ,depth first recursive method to mine the closed weighted frequent pattern of CWFP tree in sliding window	C language
7	VLFCM [12]	Very large data	Methods have been used to extend the fuzzy C-means algorithm	Matlab
8	HSW stream [13]	High dimensional stream data	Projected clustering technique is used	Microsoft Visual c++
9	Modified fuzzy C-means [14]	Stream data	Mahalanobis distance algorithm is used	Matlab
10	Density WFCM [15]	Various featured and relational datasets are used	The weighted and relational algorithm is proposed in it (WNERFCM) .	Matlab

B. *Comparative overview of various Dimension reduction techniques:*

Table: 2

S.No	Name of technique	Nature	Usage statistics	Working mechanism
1	Singular value decomposition(SVD)	unsupervised	Most of these algorithms has the primary objectives of applying it are identification and extraction of structural constitution within the data	A gene selection procedure have been used (it is factorization method)
2	Principle component analysis (PCA)	unsupervised	Most of these algorithms has Feature extraction objectives	The main focus is to convert the values into set of linear combinations
3	Linear discriminant analysis(LDA)	unsupervised	Most of these algorithms are used in small sample size problems (for feature selection and feature transformation)	Classification based approach
4	Support vector machines (SVM)	supervised	Most of these algorithms will be specifically appropriate for certain specific datasets where the sample data set is much smaller in comparison to large number of features(genes)	Its main purpose is the analysis of gene expression data
5	Independent Component Analysis(ICA)	unsupervised	Most of the ICA algorithms will require whitened data by means of an identity covariance matrix	It works on the principle of additive subcomponents and separation of singular units from a large multivariate source
6	Canonical Correlation Analysis (CCA)	unsupervised	Most of CCA algorithms are used DNA micro array assess expressions of numerous thousand of genes	Its main focus is on discovering linear combinations from two sets of variables and then by approximates correlations amongst these variables

III. OUR PROPOSAL

A. *Problem statement:*

To increase the efficiency of data stream clustering algorithms (reduced weighted fuzzy C-means algorithm in our case) [5] and for visualization purpose, the dimension reduction technique [3] have to be used which will resolve the given input data in to two dimensions for better clustering quality. We will apply our proposed technique on the data having streaming behavior and higher dimensions.

B. *Proposed approach:*

In this study we propose a dimension reduced weighted fuzzy c-means algorithm. The algorithm will be applicable for those high dimensional data sets that have streaming behavior. An example of such data sets is live high-definition videos in internet. These data have two special properties which separate them from other data sets: a) They have streaming behavior and b) They have higher dimensions. Firstly, we will apply the dimension reduction technique [3] to convert the higher dimensional data stream in to less dimensions (Two dimensions) and then after the weighted fuzzy C-means algorithm [5] will be applied for better data stream clustering.

C. *Algorithm:*

STEP 1: [Input] Input the data set components (o_1, o_2… $o_{k)}$ having both high dimensions and Streaming behavior.

STEP 2: [Reduction] Apply the proposed dimension reduction technique [3] on the given input Data.

STEP 3: [Clustering] Apply the Fuzzy C-means algorithm on the data received in the step2.

STEP 4: [Output] we get data set components (o_1, o_2… $o_{k)}$ with lesser dimensions.

D. *Proposed Technology Used:*

We will implement our proposed algorithm in Java and will test its efficiency with some real high definitional data sets which have the streaming behavior.

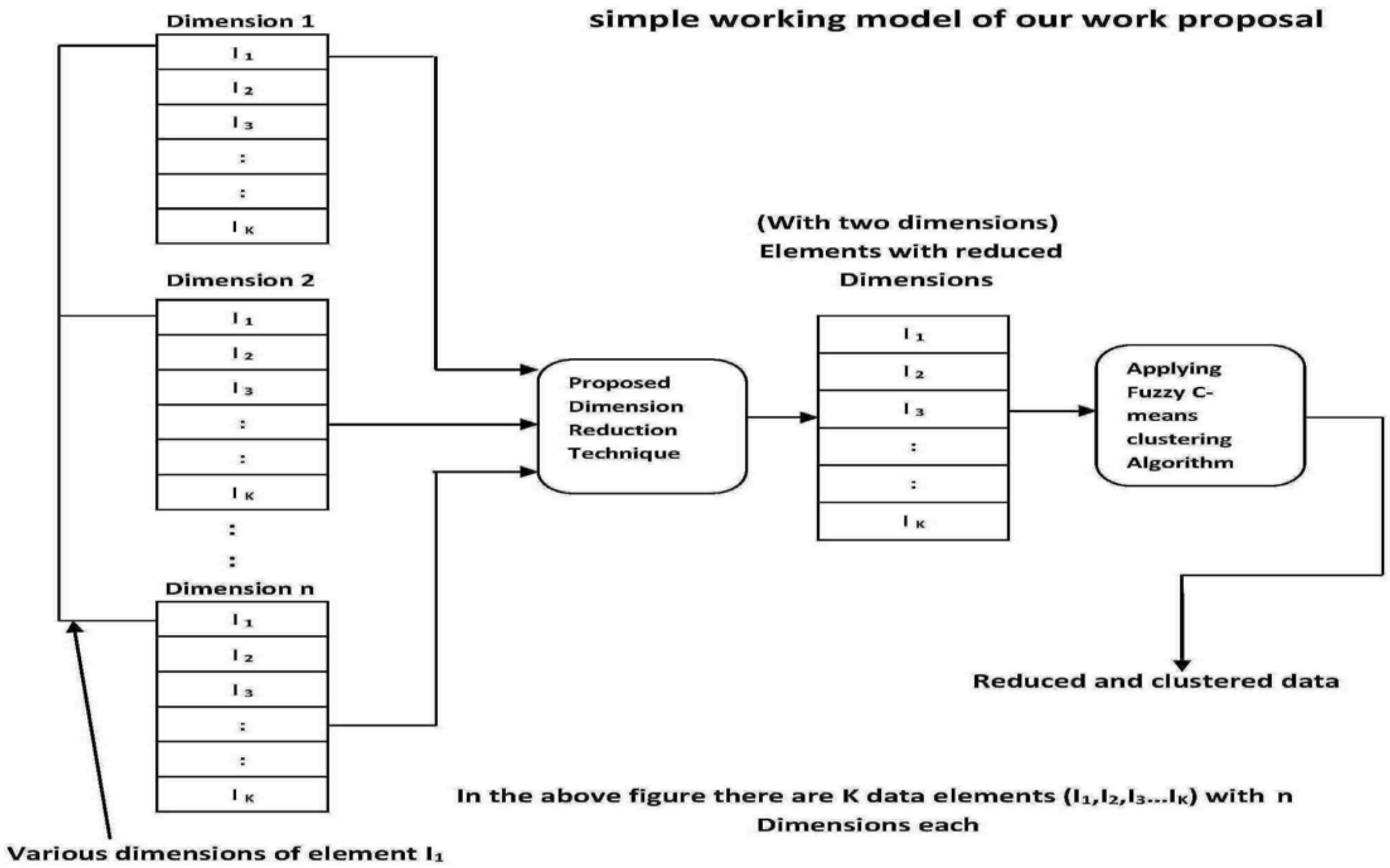
simple working model of our work proposal
Dimension 1
I₁
I₂
I₃
:
:
Iκ
Dimension 2
I₁
I₂
I₃
:
:
Iκ
:
:
Dimension n
I₁
I₂
I₃
:
:
Iκ
Proposed Dimension Reduction Technique
(With two dimensions) Elements with reduced Dimensions
I₁
I₂
I₃
:
:
Iκ
Applying Fuzzy C-means clustering Algorithm
Reduced and clustered data
In the above figure there are K data elements (I₁,I₂,I₃...Iκ) with n Dimensions each
Various dimensions of element I₁

Figure: 1

IV. CONCLUSION

In this paper we have presented various data stream mining techniques, as it is the popular research area now a days ,we have also discussed various characteristics of stream data and its possible sources .As the streamed data can be high dimensional various dimension reduction techniques were applied on it prior to it clustering .In our survey we have provided the comparative analysis of various stream mining procedures and dimension reduction techniques ,also the working model for mining data stream have been discussed in our proposal section of this paper ,after reading this work one can get the quick overview about stream data mining and various dimension reduction techniques available.

V. REFERENCES

[1]. Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, Jennifer Widom (2002). ”Models and issues in

data stream systems". Proceedings of the 21th ACM SIGACT-SIGMOD SIGART Symposium on Principles of Database Systems. pp. 1–16.

[2]. Liadan O'Chalaghan, Nina Mishra, Adam Meyerson, Sudipto Guha, Rajeev Motwani (2002). "Streaming data algorithms for high quality clustering". Proceedings of the 18th international conference on data engineering. pp.685– 694.

[3]. P.S. Bishnu , V. Bhattacherjee. A Dimension Reduction Technique for K-Means Clustering Algorithm; 1st International Conf. on Recent Advances in Information Technology | RAIT- 2012 |.

[4]. Pedro Rodrigues, Joao Gama, Joao Pedro Pedroso (2004). "Hierarchical time-series clustering for data streams". Proceedings of the first international workshop on knowledge discovery in data streams. pp. 22–31.

[5]. Renxia Wan, Xiaoya Yan, Xiaoke Su; A Weighted Fuzzy Clustering Algorithm for Data Stream. 2008ISECS International Colloquium on Computing, Communication, Control, and Management.

[6]. Lingjuan Li, Xiong Li College of Computer Nanjing University of Posts and Telecommunications Nanjing,Improved Online Stream Data Clustering Algorithm 2012 Second International Conference on Business Computing and Global Informatization.

[7]. Maoguo Gong, Member, IEEE, Yan Liang, Jiao Shi, Wenping Ma, and Jingjing MaFuzzy C Means Clustering With Local Information and Kernel Metric for Image Segmentation IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 22, NO. 2, FEBRUARY 2013

[8]. Carlos Ordonez (2003). "Clustering binary data streams with k-means". Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery.pp.12–19.

[9]. Amin Namadchian Gholamreza Esfandani DSCLU: a new Data Stream CLUstring algorithm for multi density environments 2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing.

[10]. Arthur F. M. Alvim, Renata C.M.R de Souza A Fuzzy Weighted Clustering Method for Symbolic Interval Data IEEE.

[11]. Wang Jie, Zeng Yu DCWFP-Miner: Mining Closed Weighted Frequent Patterns over Data Streams 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012).

[12]. Timothy C. Havens, Senior Member, IEEE, James C. Bezdek, Life Fellow, IEEE, Christopher Leckie, Lawrence O. Hall, Fellow, IEEE, and Marimuthu Palaniswami, Fellow, IEEE Fuzzy c-Means Algorithms for Very Large DataIEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 20, NO. 6, DECEMBER 2012.

[13]. Weiguo Liu, Jia OuYang Clustering Algorithm for High Dimensional Data Stream over Sliding Windows 2011 International Joint Conference of IEEE TrustCom-11/IEEE ICESS- 11/FCST-11.

[14]. Shao-Hong Yin, Min Li Study on a modified Fuzzy C-Means Clustering Algorithm 2010 International Conference On Computer Design And Appliations (ICCDA 2010).

[15]. Richard J. Hathaway and Yingkang Hu Density-Weighted Fuzzy c-Means Clustering IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 17, NO. 1, FEBRUARY 2009.

[16]. Nebu Varghese1, Vinay Verghese2, Prof. Gayathri. P3 and Dr. N. Jaisankar4 A SURVEY OF DIMENSIONALITY REDUCTION AND CLASSIFICATION METHODS International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.3, June 2012.