INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATION & MANAGEMENT

IJRCM

A Monthly Double-Blind Peer Reviewed (Refereed/Juried) Open Access International e-Journal - Included in the International Serial Directories

Indexed & Listed at:

Ulrich's Periodicals Directory ©, ProQuest, U.S.A., EBSCO Publishing, U.S.A., Cabell's Directories of Publishing Opportunities, U.S.A.

as well as in Open J-Gage, India [link of the same is duly available at Inflibnet of University Grants Commission (U.G.C.)]

Registered & Listed at: Index Copernicus Publishers Panel, Poland

Circulated all over the world & Google has verified that scholars of more than 1500 Cities in 141 countries/territories are visiting our journal on regular basis.

Ground Floor, Building No. 1041-C-1, Devi Bhawan Bazar, JAGADHRI – 135 003, Yamunanagar, Haryana, INDIA

www.ijrcm.org.in

CONTENTS

CALL FOR MANUSCRIPTS

We invite unpublished novel, original, empirical and high quality research work pertaining to recent developments & practices in the area of Computer, Business, Finance, Marketing, Human Resource Management, General Management, Banking, Insurance, Corporate Governance and emerging paradigms in allied subjects like Accounting Education; Accounting Information Systems; Accounting Theory & Practice; Auditing; Behavioral Accounting; Behavioral Economics; Corporate Finance; Cost Accounting; Econometrics; Economic Development; Economic History; Financial Institutions & Markets; Financial Services; Fiscal Policy; Government & Non Profit Accounting; Industrial Organization; International Economics & Trade; International Finance; Macro Economics; Micro Economics; Monetary Policy; Portfolio & Security Analysis; Public Policy Economics; Real Estate; Regional Economics; Tax Accounting; Advertising & Promotion Management; Business Education; Management Information Systems (MIS); Business Law, Public Responsibility & Ethics; Communication; Direct Marketing; E-Commerce; Global Business; Health Care Administration; Labor Relations & Human Resource Management; Marketing Research; Marketing Theory & Applications; Non-Profit Organizations; Office Administration/Management; Operations Research/Statistics; Organizational Behavior & Theory; Organizational Development; Production/Operations; Public Administration; Purchasing/Materials Management; Retailing; Sales/Selling; Services; Small Business Entrepreneurship; Strategic Management Policy; Technology/Innovation; Tourism, Hospitality & Leisure; Transportation/Physical Distribution; Algorithms; Artificial Intelligence; Compilers & Translation; Computer Aided Design (CAD); Computer Aided Manufacturing; Computer Graphics; Computer Organization & Architecture; Database Structures & Systems; Digital Logic; Discrete Structures; Internet; Management Information Systems; Modeling & Simulation; Multimedia; Neural Systems/Neural Networks; Numerical Analysis/Scientific Computing; Object Oriented Programming; Operating Systems; Programming Languages; Robotics; Symbolic & Formal Logic and Web Design. The above mentioned tracks are only indicative, and not exhaustive.

Anybody can submit the soft copy of his/her manuscript **anytime** in M.S. Word format after preparing the same as per our submission guidelines duly available on our website under the heading guidelines for submission, at the email address: infoijrcm@gmail.com.

GUIDELINES FOR SUBMISSION OF MANUSCRIPT

1. **COVERING LETTER FOR SUBMISSION:**

DATED: _____

THE EDITOR
IJRCM

Subject: **SUBMISSION OF MANUSCRIPT IN THE AREA OF .**

(e.g. Finance/Marketing/HRM/General Management/Economics/Psychology/Law/Computer/IT/Engineering/Mathematics/other, please specify)

DEAR SIR/MADAM

Please find my submission of manuscript entitled '_____' for possible publication in your journals.

I hereby affirm that the contents of this manuscript are original. Furthermore, it has neither been published elsewhere in any language fully or partly, nor is it under review for publication elsewhere.

I affirm that all the author (s) have seen and agreed to the submitted version of the manuscript and their inclusion of name (s) as co-author (s).

Also, if my/our manuscript is accepted, I/We agree to comply with the formalities as given on the website of the journal & you are free to publish our contribution in any of your journals.

NAME OF CORRESPONDING AUTHOR:
Designation:
Affiliation with full address, contact numbers & Pin Code:
Residential address with Pin Code:
Mobile Number (s):
Landline Number (s):
E-mail Address:
Alternate E-mail Address:

NOTES:

a) The whole manuscript is required to be in ***ONE MS WORD FILE*** only (pdf. version is liable to be rejected without any consideration), which will start from the covering letter, inside the manuscript.
b) The sender is required to mention the following in the **SUBJECT COLUMN** of the mail:
New Manuscript for Review in the area of (Finance/Marketing/HRM/General Management/Economics/Psychology/Law/Computer/IT/ Engineering/Mathematics/other, please specify)
c) There is no need to give any text in the body of mail, except the cases where the author wishes to give any specific message w.r.t. to the manuscript.
d) The total size of the file containing the manuscript is required to be below **500 KB**.
e) Abstract alone will not be considered for review, and the author is required to submit the complete manuscript in the first instance.
f) The journal gives acknowledgement w.r.t. the receipt of every email and in case of non-receipt of acknowledgment from the journal, w.r.t. the submission of manuscript, within two days of submission, the corresponding author is required to demand for the same by sending separate mail to the journal.

2. **MANUSCRIPT TITLE:** The title of the paper should be in a 12 point Calibri Font. It should be bold typed, centered and fully capitalised.

3. **AUTHOR NAME (S) & AFFILIATIONS:** The author (s) **full name**, **designation**, **affiliation** (s), **address**, **mobile/landline numbers**, and **email/alternate email address** should be in italic & 11-point Calibri Font. It must be centered underneath the title.

4. **ABSTRACT:** Abstract should be in fully italicized text, not exceeding 250 words. The abstract must be informative and explain the background, aims, methods, results & conclusion in a single para. Abbreviations must be mentioned in full.

5. **KEYWORDS:** Abstract must be followed by a list of keywords, subject to the maximum of five. These should be arranged in alphabetic order separated by commas and full stops at the end.

6. **MANUSCRIPT:** Manuscript must be in ***BRITISH ENGLISH*** prepared on a standard A4 size ***PORTRAIT SETTING PAPER***. It must be prepared on a single space and single column with 1" margin set for top, bottom, left and right. It should be typed in 8 point Calibri Font with page numbers at the bottom and centre of every page. It should be free from grammatical, spelling and punctuation errors and must be thoroughly edited.

7. **HEADINGS:** All the headings should be in a 10 point Calibri Font. These must be bold-faced, aligned left and fully capitalised. Leave a blank line before each heading.

8. **SUB-HEADINGS:** All the sub-headings should be in a 8 point Calibri Font. These must be bold-faced, aligned left and fully capitalised.

9. **MAIN TEXT:** The main text should follow the following sequence:

INTRODUCTION

REVIEW OF LITERATURE

NEED/IMPORTANCE OF THE STUDY

STATEMENT OF THE PROBLEM

OBJECTIVES

HYPOTHESES

RESEARCH METHODOLOGY

RESULTS & DISCUSSION

FINDINGS

RECOMMENDATIONS/SUGGESTIONS

CONCLUSIONS

SCOPE FOR FURTHER RESEARCH

ACKNOWLEDGMENTS

REFERENCES

APPENDIX/ANNEXURE

It should be in a 8 point Calibri Font, single spaced and justified. The manuscript should preferably not exceed ***5000 WORDS***.

10. **FIGURES & TABLES:** These should be simple, crystal clear, centered, separately numbered &self explained, and **titles must be above the table/figure**. **Sources of data should be mentioned below the table/figure**. It should be ensured that the tables/figures are referred to from the main text.

11. **EQUATIONS:**These should be consecutively numbered in parentheses, horizontally centered with equation number placed at the right.

12. **REFERENCES:** The list of all references should be alphabetically arranged. The author (s) should mention only the actually utilised references in the preparation of manuscript and they are supposed to follow **Harvard Style of Referencing**. The author (s) are supposed to follow the references as per the following:

- All works cited in the text (including sources for tables and figures) should be listed alphabetically.
- Use (**ed.**) for one editor, and (**ed.s)** for multiple editors.
- When listing two or more works by one author, use --- (20xx), such as after Kohl (1997), use --- (2001), etc, in chronologically ascending order.
- Indicate (opening and closing) page numbers for articles in journals and for chapters in books.
- The title of books and journals should be in italics. Double quotation marks are used for titles of journal articles, book chapters, dissertations, reports, working papers, unpublished material, etc.
- For titles in a language other than English, provide an English translation in parentheses.
- The location of endnotes within the text should be indicated by superscript numbers.

PLEASE USE THE FOLLOWING FOR STYLE AND PUNCTUATION IN REFERENCES:

BOOKS

- Bowersox, Donald J., Closs, David J., (1996), "Logistical Management." Tata McGraw, Hill, New Delhi.
- Hunker, H.L. and A.J. Wright (1963), "Factors of Industrial Location in Ohio" Ohio State University, Nigeria.

CONTRIBUTIONS TO BOOKS

- Sharma T., Kwatra, G. (2008) Effectiveness of Social Advertising: A Study of Selected Campaigns, Corporate Social Responsibility, Edited by David Crowther & Nicholas Capaldi, Ashgate Research Companion to Corporate Social Responsibility, Chapter 15, pp 287-303.

JOURNAL AND OTHER ARTICLES

- Schemenner, R.W., Huber, J.C. and Cook, R.L. (1987), "Geographic Differences and the Location of New Manufacturing Facilities," Journal of Urban Economics, Vol. 21, No. 1, pp. 83-104.

CONFERENCE PAPERS

- Garg, Sambhav (2011): "Business Ethics" Paper presented at the Annual International Conference for the All India Management Association, New Delhi, India, 19–22 June.

UNPUBLISHED DISSERTATIONS AND THESES

- Kumar S. (2011): "Customer Value: A Comparative Study of Rural and Urban Customers," Thesis, KurukshetraUniversity, Kurukshetra.

ONLINE RESOURCES

- Always indicate the date that the source was accessed, as online resources are frequently updated or removed.

WEBSITES

- Garg, Bhavet (2011): Towards a New Natural Gas Policy, Political Weekly, Viewed on January 01, 2012 http://epw.in/user/viewabstract.jsp

AN APPROACH ON PREPROCESSING OF DATA STREAMS

AVINASH L. GOLANDE
STUDENT
DR. DY PATIL INSTITUTE OF ENGINEERING & TECHNOLOGY
PIMPRI

RAJESH D. BHARATI
ASST. PROFESSOR
DR. DY PATIL INSTITUTE OF ENGINEERING & TECHNOLOGY
PIMPRI

PRASHANT G AHIRE
ASST. PROFESSOR
PIMPRI CHINCHWAD COLLEGE OF ENGINEERING
PUNE

RAHUL A. PATIL
ASST. PROFESSOR
PIMPRI CHINCHWAD COLLEGE OF ENGINEERING
PUNE

ABSTRACT

The recent advances in hardware and software have enabled the capture of different measurements of data in a wide range of fields. These measurements are generated continuously and in a very high fluctuating data rates. Examples include sensor networks, web logs, and computer network traffic. The storage, querying and mining of such data sets are highly computationally challenging tasks. Mining data streams is concerned with extracting knowledge structures represented in models and patterns in non stopping streams of information. The research in data stream mining has gained a high attraction due to the importance of its applications and the increasing generation of streaming information. Applications of data stream analysis can vary from critical scientific and astronomical applications to important business and financial ones. Algorithms, systems and frameworks that address streaming challenges have been developed over the past three years. In this review paper, we present the state of- the-art in this growing vital field.

KEYWORDS
component, data stream, VFDT.

INTRODUCTION

The intelligent data analysis has passed through a number of stages. Each stage addresses novel research issues that have arisen. Statistical exploratory data analysis represents the first stage. The goal was to explore the available data in order to test a specific hypothesis. With the advances in computing power, machine learning field has arisen. The objective is to find computationally efficient solutions to data analysis problems. Along with the progress in machine learning research, new data analysis problems have been addressed. Due to the increase in database sizes, new algorithms have been proposed to deal with the scalability issue. Moreover machine learning and statistical analysis techniques have been adopted and modified in order to address the problem of very large databases. Data mining is that interdisciplinary field of study that can extract models and patterns from large amounts of information stored in data repositories.

Recently, the data generation rates in some data sources become faster than ever before. This rapid generation of continuous streams of information has challenged our storage, computation and communication capabilities in computing systems. Systems, models and techniques have been proposed and developed over the past few years to address these challenges.

In this paper, we review the theoretical foundations of data stream analysis, mining data stream systems, techniques are critically reviewed. Finally, we outline and discuss research problems in streaming mining field of study. These research issues should be addressed in order to realize robust systems that are capable of fulfilling the needs of data stream mining applications. The paper is organized as follows. Section 2 presents the theoretical background of data stream analysis. In sections 3 and 4 mining data stream techniques and systems are reviewed respectively. Open and addressed research issues in this growing field are discussed in section 5. Finally section 6 summarizes this review paper. section 7 enlist the references.

THEORETICAL FOUNDATIONS

Research problems and challenges that have been arisen in mining data streams have its solutions using well established statistical and computational approaches. We can categorize these solutions to data-based and task-based ones. In data-based solutions, the idea is to examine only a subset of the whole dataset or to transform the data vertically or horizontally to an approximate smaller size data representation. At the other hand, in task-based solutions, techniques from computational theory have been adopted to achieve time and space efficient solutions. In this section we review these theoretical foundations

DATA-BASED TECHNIQUES

Data-based techniques refer to summarizing the whole dataset or choosing a subset of the incoming stream to be analyzed. Sampling, load shedding and sketching techniques represent the former one. Synopsis data structures and aggregation represent the later one. Here is an outline of the basics of these techniques with pointers to its applications in the context of data stream analysis.

Sampling : Sampling refers to the process of probabilistic choice of a data item to be processed or not.

Load Shedding: ItLoad shedding refers to the process of dropping a sequence of data streams.

Sketching: Sketching is the process of randomly project a subset of the features.

Synopsis Data Structures: Creating synopsis of data refers to the process of applying summarization techniques that are capable of summarizing the incoming stream for further analysis.

Aggregation: Aggregation is the process of computing statistical measures such as means and variance that summarize the incoming stream.

TASK-BASED TECHNIQUES

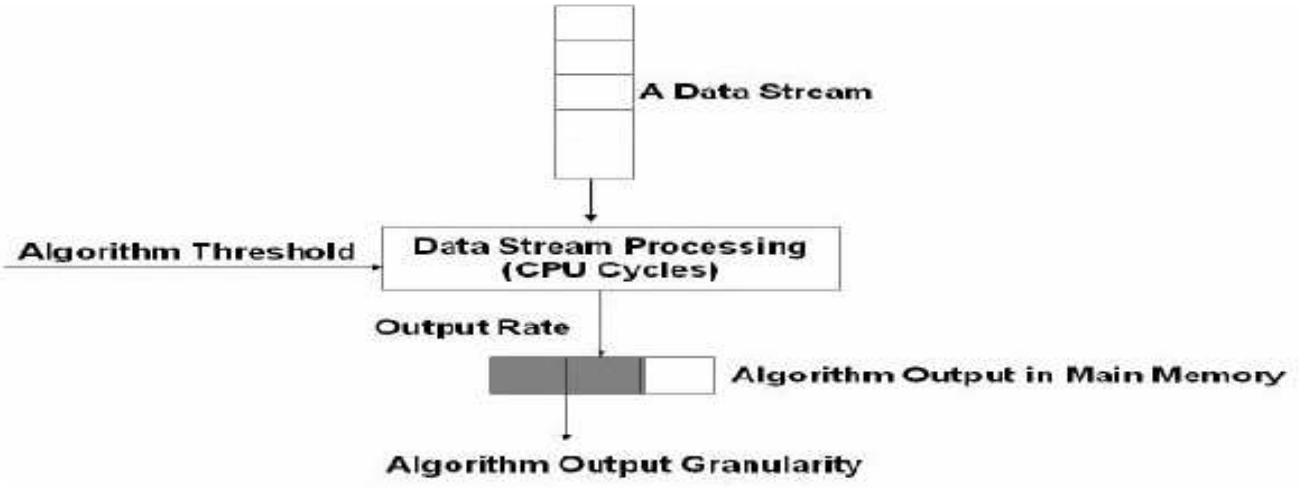
Task-based techniques are those methods that modify existing techniques or invent new ones in order to address the computational challenges of data stream processing. Approximation algorithms, sliding window and algorithm output granularity represent this category. In the following subsections, we examine each of these techniques and its application in the context of data stream analysis.

Approximation algorithms – they have their roots in algorithm design. It is concerned with design algorithms for computationally hard problems.

Sliding Window - The inspiration behind sliding window is that the user is more concerned with the analysis of most recent data streams.

Algorithm Output Granularity -The algorithm output granularity (AOG) introduces the first resource-aware data analysis approach that can cope with fluctuating very high data rates according to the available memory and the processing speed represented in time constraints.

FIG. 1: THE AOG ALGORITHM



MINING TECHNIQUES

Mining data streams has attracted the attention of data mining community for the last three years. Number of algorithms has been proposed for extracting knowledge from streaming information. In this section, we review clustering, classification, frequency counting and time series analysis techniques.

Clustering*:* Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

Classification: GEMM algorithm accepts a class of models and an incremental model maintenance algorithm for the unrestricted window option, and outputs a model maintenance algorithm for both window-independent and window dependent block selection sequence. Domingos et al. have developed VFDT. It is a decision tree learning systems based on Hoeffding trees.

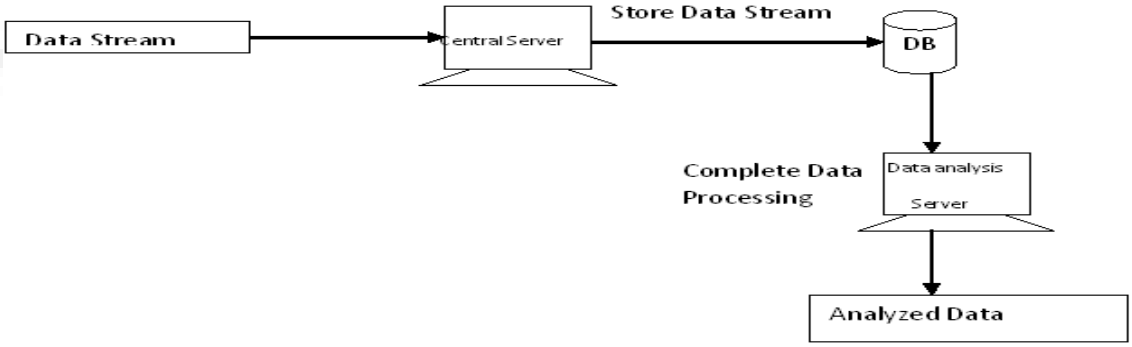
1. **FrequencyCounting**: Lightweight frequency counting LWF. It has the ability to find an approximate solution to the most frequent items in the incoming stream using adaptation and releasing the least frequent items regularly in order to count the more frequent ones.
2. **Time Series Analysis**: This representation allows dimensionality/numerosity reduction. They have demonstrated the applicability of the proposed representation by applying it to clustering, classification, indexing and anomaly detection. The approach has two main stages. The first one is the transformation of time series data to Piecewise Aggregate Approximation followed by transforming the output to discrete string symbols in the second stage

EXISTING SYSTEMS

In many scientific domains, especially those using sensors, the data is generated as a data stream, that is, the data arrives in sequential order in an infinitely long stream. As this data indicates the state of the system, we are interested in on-line, any-time algorithms which can analyze the data as it is being collected to determine if there are any anomalous events in the data.

Data stream mining is the process of extracting knowledge structure form continuous, rapid data records. Recent years, the researchers in data stream mining have received attention due to the importance of its applications and the increasing generation of streaming information. A variety of systems, models and technique have been proposed and developed to address the challenges in data stream mining.

FIG. 2: TRADITIONAL DATA STREAM ANALYSIS SYSTEM



Above fig. shows the existing data analysis processing system. In above system, data is analyzed statically. First, data stream from various sources are temporarily stored in central database server and then complete data is send to data analysis server, which perform all data-based technique on that data. After completing analyzing data is stored in another server.

RESEARCH ISSUES

Data stream mining is a stimulating field of study that has raised challenges and research issues to be addressed by the database and data mining communities. The following is a discussion of both addressed and open research issues. The following is a brief discussion of previously addressed issues is as follows:

1. Unbounded memory requirements due to the continuous flow of data streams.
2. Required result accuracy.
3. Transferring data mining results over a wireless network with a limited bandwidth.
4. Modeling changes of mining results over time.
5. Developing algorithms for mining results' changes
6. Visualization of data mining results on small screens of mobile devices.
7. Interactive mining environment to satisfy user requirements.
8. The integration between data stream management systems and the ubiquitous data stream.
9. Mining approaches.
10. The needs of real world applications.
11. Data stream pre-processing.
12. Model over fitting.
13. Data stream mining technology
14. The formalization of real-time accuracy evaluation.

RESULTS AND DISCUSSION

These techniques are used for reduction of large data set into small datasets. Sampling and sketching for the different data sets gives different output.
For Ex: Given dataset consist of D records. If we implemented Sampling and Sketching on this dataset then it gives following output:

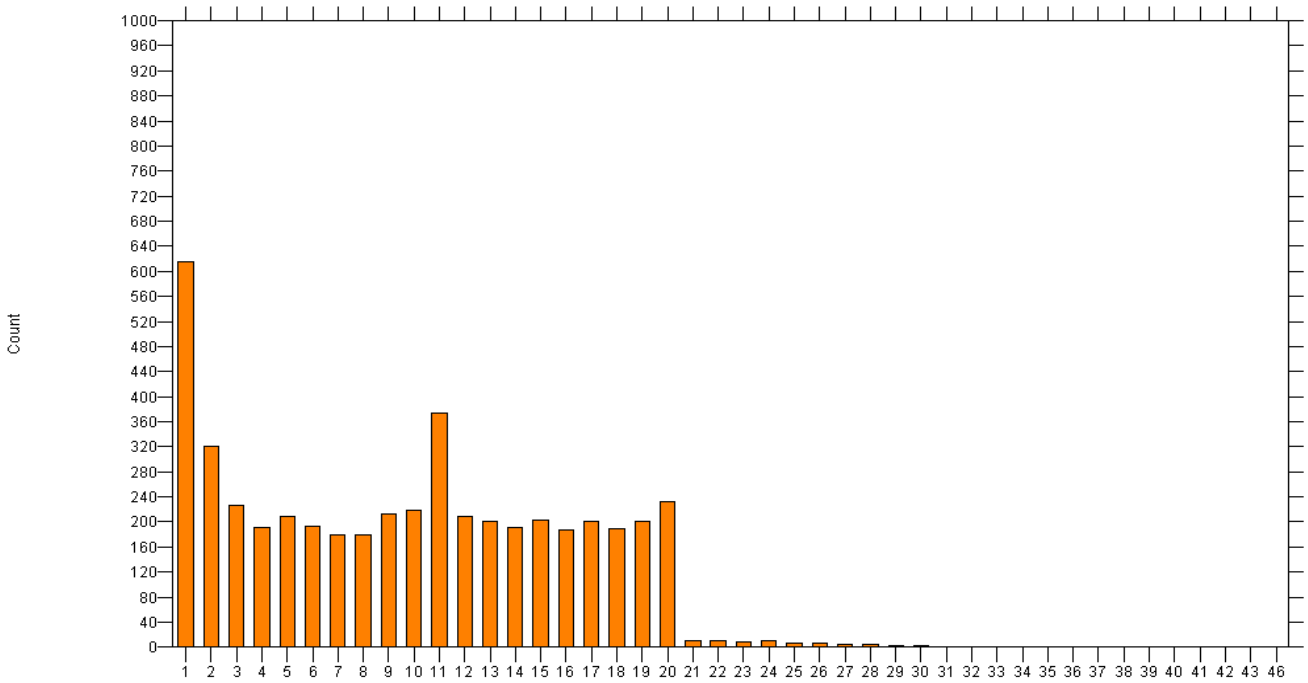
SAMPLING

1) **Nth Sampling**: In this method, we skipped number of record as per the value of N. If N is 'n' then output in record for above dataset is D/n. It means it will reduce the data (100-(D/n))%.
2) **Random Sampling**: In this method, we skipped number of record randomly. Random number is selected by this method and it will reduce the data It will skip the data on that random number. It will reduce the data approximately 50%. Reduction is depend upon the random number.
3) **Strata Sampling**: In this method, we skipped number of record depending upon the even and odd numbers. It generate random cluster and select even or odd cluster as per user decide. It will reduce the data approximately 60%.
4) **Biased-L2 Sampling**: In this method, we use sampling rate and penalty function. This penalty function is generated by random cluster. Penalty function reduces the constant number of record and useful for error reduction. It will reduce the data approximately 60%.

SKETCHING

Sketching is the process of randomly project a subset of the features. It is the process of vertically sample the incoming stream. Sketching has been applied in comparing different data streams and in aggregate queries. It gives the frequency of key attribute in whole data stream.

FIG. 3: HISTOGRAM FOR SKETCHING



We use KDD data set with 10000 records. Key attribute is srv_count. Graph shows the frequency of different values of server count (srv_count) for the given dataset.

CONCLUSIONS

The objective behind designing this tool was to make easy and familiar the concept of data analysis in data mining. We have successfully able to implement data based technique for data stream analysis. We have able to implement sampling, sketching and load shedding technique on data stream. In sampling, we have implemented four algorithms named Nth sampling, random sampling, Strata sampling and Deterministic Biased-L2 sampling. The main aim of our software is to analysis of data stream dynamically. When the data is dynamically analyzed then we don't need to store it in any intermediate server. We also increase the

efficiency of algorithms so that it will take less time to analyze the data. We use the queue which holds the 1000 records at a time. User can change the size of the queue.

SCOPE FOR FUTURE RESEARCH

We have successfully implemented two techniques of statistical model that is sampling and sketching. In future, user can implement remaining three techniques. We have implemented these techniques dynamically with higher efficiency.

REFERENCES

1. An analytical framework for data stream mining techniques based on challenges and requirements Mahnoosh Kholghi (Department of Electronic, Computer and IT, Islamic Azad University, Qazvin Branch, Qazvin, Iran and member of Young Researchers Club), Mohammadreza Keyvanpour (Department of Computer Engineering Alzahra University Tehran, Iran)
2. C. Aggarwal, J. Han, J. Wang, P. S. Yu, A Framework for Clustering Evolving Data Streams, Proc. 2003 Int. Conf. on Very Large Data Bases, Berlin, Germany, Sept. 2003.
3. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, On Demand Classification of Data Streams, Proc. 2004 Int. 24 SIGMOD Record, Vol. 34, No. 2, June 2005 Conf. on Knowledge Discovery and Data Mining, Seattle, WA, Aug. 2004.
4. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, A Framework for Projected Clustering of High Dimensional Data Streams, Proc. 2004 Int. Conf. on Very Large Data Bases, Toronto, Canada, 2004.
5. DATA STREAM MINING - A Practical Approach Albert Bifet and Richard Kirkby, August 2009 Hebah H. O. NasereddinDepartment of computer Information system Faculty of IT Amman Arab University for Graduate Studies,Amman – Jordan
6. Gaber, M.M., Krishnaswamy, S., and Zaslavsky, A. (2006). On-board Mining of Data Streams in Sensor Networks, In Advanced Methods of Knowledge Discovery from Complex Data, S. Badhyopadhyay, et al., Editors., Springer. pp. 307-335.
7. Mining Data Streams: A Review Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy Centre for Distributed Systems and Software Engineering, Monash Universitys
8. R. Bhargava, H. Kargupta, and M. Powers, Energy Consumption in Data Analysis for On-board and Distributed Applications, Proceedings of the ICML'03 workshop on Machine Learning Technologies for Autonomous Space Applications, 2003.

WEBSITES

9. http://researcher.ibm.com
10. http://domino.watson.ibm.com

REQUEST FOR FEEDBACK

Dear Readers

At the very outset, International Journal of Research in Computer Application and Management (IJRCM) acknowledges & appreciates your efforts in showing interest in our present issue under your kind perusal.

I would like to request you to supply your critical comments and suggestions about the material published in this issue as well as on the journal as a whole, on our E-mail **infoijrcm@gmail.com** for further improvements in the interest of research.

If you have any queries please feel free to contact us on our E-mail **infoijrcm@gmail.com**.

I am sure that your feedback and deliberations would make future issues better – a result of our joint effort.

Looking forward an appropriate consideration.

With sincere regards

Thanking you profoundly

Academically yours

Sd/-

Co-ordinator

ABOUT THE JOURNAL

In this age of Commerce, Economics, Computer, I.T. & Management and cut throat competition, a group of intellectuals felt the need to have some platform, where young and budding managers and academicians could express their views and discuss the problems among their peers. This journal was conceived with this noble intention in view. This journal has been introduced to give an opportunity for expressing refined and innovative ideas in this field. It is our humble endeavour to provide a springboard to the upcoming specialists and give a chance to know about the latest in the sphere of research and knowledge. We have taken a small step and we hope that with the active co-operation of like-minded scholars, we shall be able to serve the society with our humble efforts.

Our Other Journals