Comparison of MSMIA and Moment Algorithm Using Streaming Dataset of Barclays

P.Logeshwari
Research scholar, Department of Computer Science, NGM College Pollachi Email: tppselvalogu@gmail.com

Dr.AntonySelvadossThanamani
Associate Professor and Head, Department of Computer Science, NGM College Pollachi Email: selvdoss@gmail.com

Abstract--- Main Stream Data Multiple Imputation is one of the main models for Missing Data Imputation in data stream mining, in which a fixed length of recently arrived data is considered. In a Main Stream Data Multiple Imputation over a transactional data stream, by the arrival of a new transaction, the oldest transaction is removed from the Data Stream and the new transaction is inserted into the Data Stream. The MSMIA algorithm compared with the Moment, a state-of-the-art incremental mining algorithm. Real-world dataset of Barclays Bank has been used and now fix the data stream size to 10K transactions

Index Terms—:Missing Data, Multiple Imputation, Data Stream, transaction, MSMIA algorithm, Moment, Main Stream,Delay Time.

1.INTRODUCTION

The MSMIA algorithm relies mainly on a verifier function and it is an exact and efficient algorithm for mining very large main stream Data Multiple Imputation s over data streams. The performance of the MSMIA improves when small delays are allowed in reporting new Missing Data; however this delay can be set to 0 with a small performance overhead.Main Stream Data Multiple Imputation is one of the main models for Missing Data Imputation in data stream mining, in which a fixed length of recently arrived data is considered. In a Main Stream Data Multiple Imputation over a transactional data stream, by the arrival of a new transaction, the oldest transaction is removed from the Data Stream and the new transaction is inserted into the Data Stream. The MSMIA algorithm compared with the Moment, a state-of-the-art incremental mining algorithm. Real-world dataset of Barclays Bank has been used and now fix the data stream size to 10K transactions.

2. COMPARISON OF MSMIA AND MOMENT

2.1 Problem Statement and Notations

Let D be the dataset to be mined (a data stream in our case); now then D contains several transactions, where each transaction contains one or more items. Let $I = i_1, i_2, ..., i_n$ be the set of all such distinct items in D. Each subset of I is called an itemset, and by k-Data we mean an Data containing k different items. The *frequency* of an Data $s$is the number of transactions in D that contain Data s, and is denoted as Count(s,D). The support of s, sup(s,D), is defined as its frequency divided by the total number of transactions in D. Therefore, $0 \leq sup(s,D) \leq 1$ for each Data s. The goal of Missing Data mining is to find all such Data s, whose support is greater than (or equal to) some given minimum support threshold α. The set of Missing Data in D is denoted as $\sigma_\alpha(D)$.

Here now Missing Data mining is considered over a data stream, thus D is defined as a main stream Data Multiple Imputation over the continuous stream. D moves forward by a certain amount δ by adding the new Impute $\delta+$ and dropping the expired one $\delta-$. Therefore, the successive instances of D are shown as W_1, W_2, ….W_n. The number of transactions that are added to (and removed from) each data stream is called its Impute size.

For the purpose of simplicity, it is assumed that all Imputes have the same size, and also each data stream consists of the same number of Imputes. Thus, $n = |W| \cdot |S|$ is the number of Imputes in each data stream, where $|W|$ denotes the data stream size and $|S|$ denotes the size of the Imputes.

2.2 The MSMIA Algorithm

The Main stream Data Multiple Imputation Algorithm (MSMIA) always maintains a union of the Missing Data of all Imputes in the current data stream W, called Segment(S), which is guaranteed to be a superset of the Missing Data over W. Upon arrival of a new Impute and expiration of an old one, we update the true count of each segment in S, by considering its frequency in both the expired Impute and the new slide. To assure that S contains all Data that are frequent in at least one of the Imputes of the current data stream $\cup_i(\sigma_{\alpha}(S_i))$, we must also mine the new Impute and add its Missing Data to S. The difficulty is that when a new segment is added to S for the first time, its true frequency in the whole data stream is not known, mostly since this segment wasn't frequent in the previous n -1 Imputes. To address this problem, MSMIA uses an auxiliary array, *aux array*, for each new segment in the new slide.

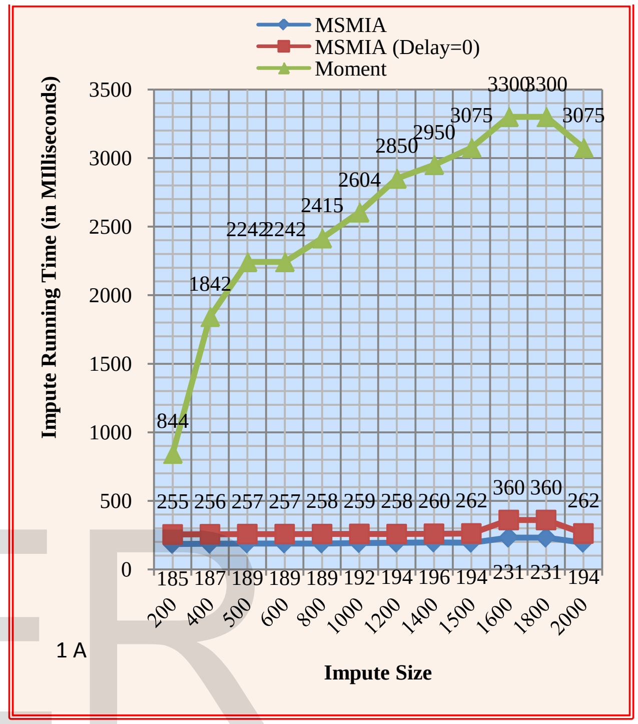
The *aux array now* stores the frequency of a segment in each data stream starting at a particular Impute in the current data stream. In other words, the *aux array* stores the frequency of a segment for each data stream, for which the frequency is not known. The key point in this is that this counting can either be done eagerly or lazily. Under the laziest approach, we wait until a Impute expires and then compute the frequency of such new Data over this Impute and update the *aux arrays* accordingly.

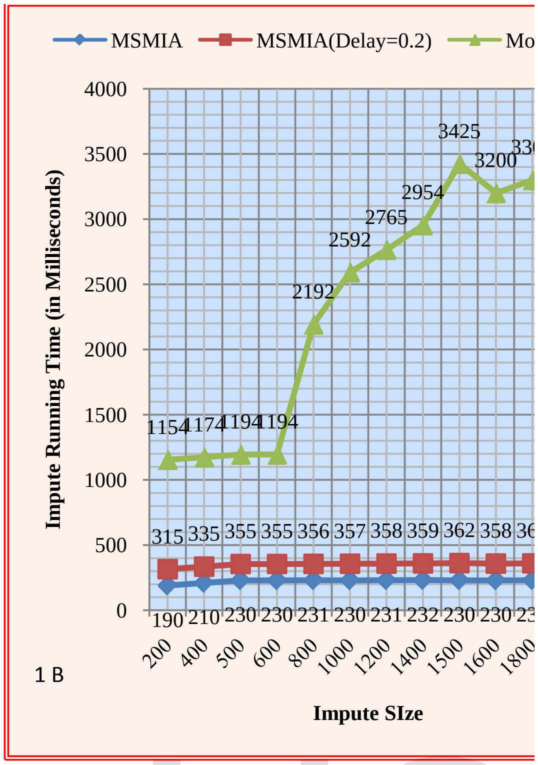
At the end of each slide, MSMIA outputs all Data in S whose frequency at that time is $\geq \alpha \cdot n \cdot |S|$. However few Data will be missed due to the lack of knowledge at the time of the output, but it will then be reported as delayed when other Imputes expire.

3.Max Delay Time Calculation

Max Delay: The maximum delay allowed by the MSMIA is $n-1$ Imputes. Indeed, after expiration of $n-1$ Imputes, MSMIA will have a complete history of the frequency of all Missing Data of W and can report them. Moreover, the case in which a segment is reported after $(n-1)$ Imputes of time, is quite rare. For this to happen, segment's support in all previous $n-1$ Imputes must be less than α but very close to it, say $\alpha \cdot |S| - 1$, and suddenly its occurrence goes up in the next Impute to say β, causing the total frequency over the whole data stream to be greater than the support threshold.

Formally, this requires that, $(n-1)\cdot(\alpha \cdot |S| - 1) + \beta \geq \alpha \cdot n \cdot |S|$ which implies $\beta \geq n + \alpha \cdot |S| - 1$. This is not impossible, but in however the real-world such events are very rare, especially when n is a large number (i.e., a large data stream spanning many slides). While MSMIA (Delay=L) represents an efficient incremental mining algorithm, counting frequencies of Data over a given dataset in $n - L + 1$ Imputes in this case.





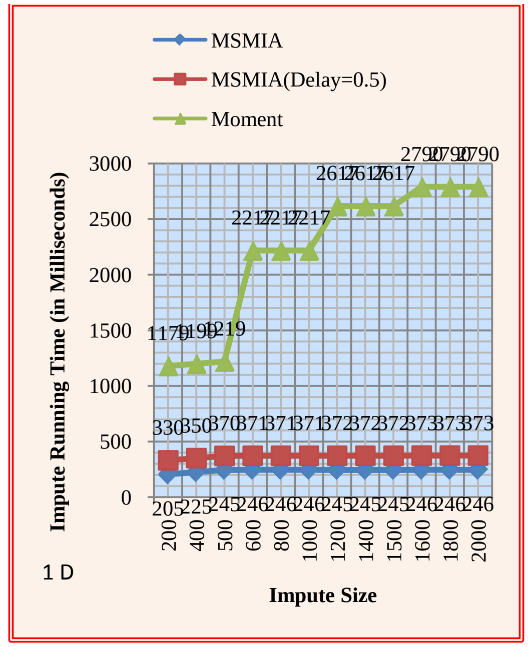
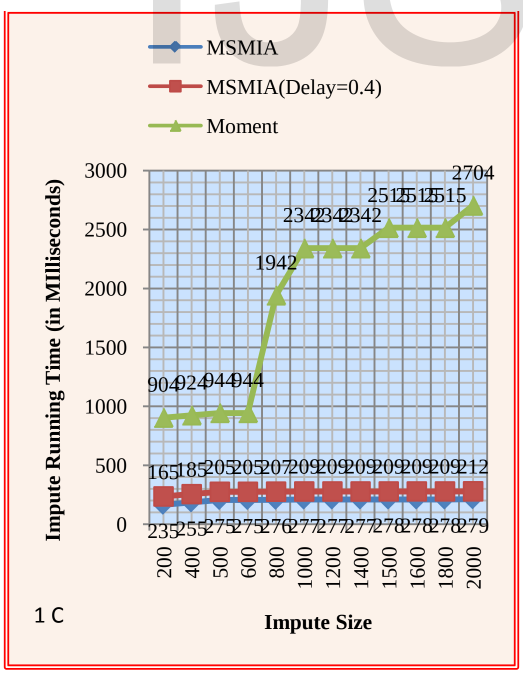


Fig. 1 Comparison of MSMIA and Moment



The MSMIA algorithm compared with the Moment, a state-of-the-art incremental mining algorithm. Real-world dataset of Barclays Bank has been used and now fix the data stream size to 10K transactions. Furthermore, the support thresholds set to 1% and vary the Impute size to measure the scalability of these algorithms. As shown in Figure 1 (a), (b), (c) and (d) MSMIA is much more scalable compared to the Moment algorithm. In fact, both versions MSMIA and MSMIA (Delay) algorithms, one with maximum data stream size delay and the other one without any delay, are much faster than Moment. The Moment algorithm is intended for incremental maintenance of Missing Data, but is not suitable for batch processing of thousands of transactions. The proposed algorithm however is aimed at maintaining Missing Data over large main stream Data Multiple Imputation s. In fact, the proposed algorithm handles a Impute size of up to 1 million transactions.

Impute sizes	**200**	**400**	**500**	**600**	**800**	**1000**	**1200**	**1400**	**1500**	**1600**	**1800**	**2000**
MSMIA	185	187	189	189	189	192	194	196	194	231	231	194
MSMIA (Delay=0)	255	256	257	257	258	259	258	260	262	360	360	262
Moment	844	1842	2242	2242	2415	2604	2850	2950	3075	3300	3300	3075
MSMIA	190	210	230	230	231	230	231	232	230	230	231	230
MSMIA(Delay =0.2)	315	335	355	355	356	357	358	359	362	358	360	362

Moment	1154	1174	1194	1194	2192	2592	2765	2954	3425	3200	3300	3425
MSMIA	165	185	205	205	207	209	209	209	209	209	209	212
MSMIA(Delay =0.4)	235	255	275	275	276	277	277	277	278	278	278	279
Moment	904	924	944	944	1942	2342	2342	2342	2515	2515	2515	2704
MSMIA	205	225	245	246	246	246	245	245	245	246	246	246
MSMIA(Delay =0.5)	330	350	370	371	371	371	372	372	372	373	373	373
Moment	1179	1199	1219	2217	2217	2217	2617	2617	2617	2790	2790	2790

Table T1 Comparison of MSMIA, MSMIA (Delay) and Moment with various Impute sizes

Real-world dataset of Barclays Bank with 21567 instances and 21 attributes and Normalized Boeing Data set with 3600 instances and 129 attributes were used for comparing the MSMIA and the Moment. The dataset names, describe the data characteristics, where T is average transaction length, I is average segment length, and D signifies the number of transactions. Table T1 lists out the values for running time comparison betweenMSMIA, MSMIA (Delay) and Moment with different delay rate.

4.CONCLUTION

MSMIA and MSMIA (Delay) algorithms, one with maximum data stream size delay and the other one without any delay, are much faster than Moment. The Moment algorithm is intended for incremental maintenance of Missing Data, but is not suitable for batch processing of thousands of transactions. The proposed algorithm however is aimed at maintaining Missing Data over large main stream Data Multiple Imputation s. In fact, the proposed algorithm handles a Impute size of up to 1 million transactions.

ACKNOWLEDGMENT

Mrs.P.Logeswarireceived her MCA., degree in computer Science from SreeSaraswathiThiyagaraja College of arts and science, Pollachi, India in 2010. She completed her M.Phil., degree in computer Science from SreeSaraswathiThiyagaraja College of arts and science, Pollachi, India on 2012 . Presently she is pursuing PhD (Full Time) degree in Computer Science in NGM College (Autonomous), Pollachi under Bharathiar University, Coimbatore. She served as a Faculty of Computer Science at Government Arts College Udumalpet, from 2012 to 2013 and she served as a Faculty of Computer Science at SreeRamu College of Arts and Science, NM Sunggam,Pollachi,India.from April 2013 to August 2014. She has presented papers in International/National conferences and published two papers in International journal. Her research focuses on Data Mining.

Dr. Antony SelvadossThanamaniis presently working as Professor and Head, Dept of Computer Science, NGM College, Coimbatore, India (affiliated to Bharathiar University, Coimbatore). He has published more than 100 papers in international/ national journals and conferences. He has authored many books on recent trends in Information Technology. His areas of interest include E-Learning, Knowledge Management, Data Mining, Networking, Parallel and Distributed Computing. He has to his credit 24 years of teaching and research experience. He is a senior member of International Association of Computer Science and Information Technology, Singapore and Active member of Computer Science Society of India, Computer Science Teachers Association, New York.

REFERENCES

[1]*Tensor Voting Techniques and Applicationsin Mobile Trace Inference,*

IEEE Access SPECIAL SECTION ON ARTIFICIAL INTELLIGENCE ENABLE NETWORKING,VOLUME 3, 2015 Received October 30, 2015, accepted November 16, 2015, date of publication December 24, 2015, date of current version January 7, 2016.ERTE PAN, (Student Member, IEEE), MIAO PAN, (Member, IEEE), AND ZHU HAN, (Fellow, IEEE)

[2]*Cluster Based Mean Imputation International Journal of Research and Reviews in Applicable Mathematics& Computer Science.Vol 2.No.1,2012,Ms.R.Malarvizhi and Dr.AntonySelvadossThanamani*

[3]*Bayesian Learning of Noisy Markov Decision Processes,ACM Transactions on Modeling and Computer Simulation Vol. 23, No. 1, Article 4, Publication date: January 2013.SUMEETPAL S. SINGH, University of Cambridge*

[4]*Estimating Burned Area in MatoGrosso, Brazil,Using an Object-Based Classification Method on a Systematic Sample of Medium Resolution Satellite Images,IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, VOL. 8, NO. 9, SEPTEMBER 2015,Yosio EdemirShimabukuro, JukkaMiettinen, René Beuchle, Rosana Cristina Grecchi,DarioSimonetti, and FrédéricAchard*

[5]*On-Line PMU-Based Transmission Line Parameter Identification,CSEE JOURNAL OF POWER AND ENERGY*

SYSTEMS,VOL. 1, NO. 2, JUNE 2015,Xuanyu Zhao, Huafeng Zhou, Di Shi, Huashi Zhao, Chaoyang Jing, Chris Jones

[6]*Cluster Based Mean Imputation,International Journal of Research and Reviews in Applicable Mathematics& Computer Science.Vol 2.No.1,2012,Ms.R.Malarvizhi and Dr.AntonySelvadossThanamani*

[7]*K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation,International Journal for Research in Science & Advanced Technologies,Vol 1.Issue-2,2013,Ms.R.Malarvizhi and Dr.AntonySelvadossThanamani.*

[8]*Classification of Efficient Imputation Method for Analyzing Missing Values,International Journal of Computer Trends and Technology(IJCTT),Vol 12.No.4-Jun 2014 ,S.Kanchana and Dr.AntonySelvadossThanamani.*

[9]*Multible Imputation of Missing Data Using Efficient Machine LearingApproach,International Journal of Applied Engineering Research,Vol 1.No.1 ,2015,S.Kanchana and Dr.AntonySelvadossThanamani*

[10]*K-NN Classifier Performs Better Than K-Means Clustering in MissingValue Imputation*
Journal: International Journal for Research in Science & Advanced Technologies,Vol 1.Issue-2,2013,Ms.R.Malarvizhi and Dr.AntonySelvadossThanamani..