BEST: International Journal of Management, Information Technology and Engineering (BEST: IJMITE)
ISSN (P): 2348-0513, ISSN (E): 2454-471X
Vol. 5, Issue 02, Feb 2017, 17-22

MATHEMATICAL SEGMENT INITIALIZATION MODEL USED IN MAIN STREAM FOR FINDING MULTIPLE DATA IMPUTATION

P. LOGESHWARI[1] & ANTONY SELVADOSS THANAMAI[2]

[1]Research scholar, Department of Computer Science, NGM College Pollachi, Tamil Nadu, India

[2]Associate Professor and Head, Department of Computer Science, NGM College Pollachi, Tamil Nadu, India

ABSTRACT

With a continuous source of data relating to transactions, the data may be segmented and processed in a data flow arrangement, optionally in parallel, and the data may be processed without storing the data in an intermediate database. Data from multiple sources may be processed in parallel. The segmentation also may define points at which aggregate outputs may be provided, and where checkpoints may be established. In this paper using the Mathematical Segment Initialization Model used to find a multiple data imputation in main stream.

KEYWORDS: Segment, Missing Data, Multiple Imputation, Data Stream, Transaction, Memory

INTRODUCTION

Segmentation means to divide the marketplace into parts, or segments, which are definable, accessible, actionable, and profitable and have a growth potential. In other words, a company would find it impossible to target the entire market, because of time, cost and effort restrictions. It needs to have a 'definable' segment - a mass of people who can be identified and targeted with reasonable effort, cost and time. With a continuous source of data relating to transactions, the data may be segmented and processed in a data flow arrangement, optionally in parallel, and the data may be processed without storing the data in an intermediate database. Data from multiple sources may be processed in parallel. The segmentation also may define points at which aggregate outputs may be provided, and where checkpoints may be established. In this paper using the Mathematical Segment Initialization Model used to find a multiple data imputation in main stream. In addition, since the transit of a data stream is usually at a high speed, and the impact of one single transaction to the entire set of transactions in the current data stream is very negligible, making it reasonable to handle the data stream Imputation in a wider magnitude. *Segment-oriented data stream Imputation* has taken for handling this problem.

"One size doesn't fit all….. Meaning…. Segmentation is a must in the online world"

SEGMENTATION

Segmentation Process

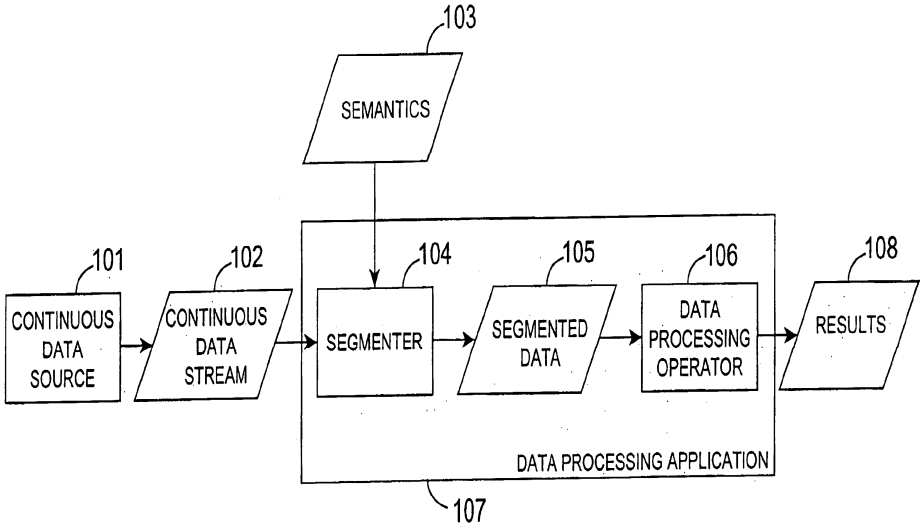

Figure 1: Segmentation Process

This is the segmentation process getting the data from continuous data stream after that segmenter getting the streaming datas and datas are segmented data processing operator is ready to getting the segmented datas and produce the results.

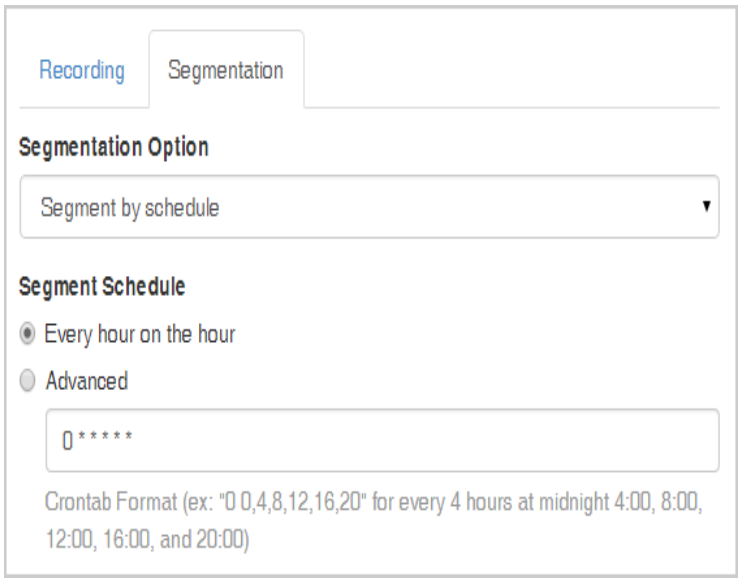

Figure 2: Segmentation options

*Figure 2.*Shows the options of the segmentation we have to choose an option regarding that option the streaming datas are segmented.

IMPUTATION

Imputation is the process of replacing missing data with substituted values. When substituting for a data point, it

is known as "unit imputation"; when substituting for a component of a data point, it is known as "item imputation". Because missing data can create problems for analyzing data, imputation is seen as a way to avoid pitfalls involved with list wise deletion of cases that have missing values. That is to say, when one or more values are missing for a case, most statistical packages default to discarding any case that has a missing value, which may introduce bias or affect the representativeness of the results. Imputation preserves all cases by replacing missing data with an estimated value based on other available information. Once all missing values have been imputed, the data set can then be analyzed using standard techniques for complete data.

MULTIPLE IMPUTATIONS

In order to deal with the problem of increased noise due to imputation, Rubin (1987) developed a method for averaging the outcomes across multiple imputed data sets to account for this. The way this works is that imputation processes similar to stochastic regression are run on the same data set multiple times and the imputed data sets are saved for later analysis. Each imputed data set is analyzed separately and the results are averaged except for the standard error term (SE). The SE is constructed by the within variance of each data set as well as the variance between imputed items on each data set. These two variances are added together and the square root of them determines the SE, thus the noises due to imputation as well as the residual variance are introduced to the regression model.

Multiple imputations involve drawing values of the parameters from a posterior distribution. The posterior distribution reflects the noise associated with the uncertainty surrounding the parameters of the distribution that generates the data. Therefore the multiple imputations simulate both the process generating the data and the uncertainty associated with the parameters of the probability distribution of the data. More traditional methods like hot-deck imputation and Maximum-likelihood-based imputation fail to give a complete simulation of the uncertainty associated with missing data.

In machine learning, it is sometimes possible to train a classifier directly over the original data without imputing it first. That was shown to yield better performance in cases where the missing data is structurally absent, rather than missing due to measurement noise.

SEGMENT INITIALIZATION

Mining Missing Data over data streams using main stream Data Multiple Imputation model handled the data streams transaction by transaction. Unlike the landmark data stream model, transactions in the main stream Data Multiple Imputation model will be both inserted into and dropped out from the data stream. The transaction-by-transaction Imputation of a data stream leads to excessively high frequency of processing. In addition, since the transit of a data stream is usually at a high speed, and the impact of one single transaction to the entire set of transactions in the current data stream is very negligible, making it reasonable to handle the data stream Imputation in a wider magnitude. *Segment-oriented data stream Imputation* has taken for handling this problem.

A main stream Data Multiple Imputation in the stream is a data stream of n number of most recent w transactions which Imputes forward for every transaction or every segment of transactions. The notation I_l to denote all the Data of length l together with their respective counts in a set of transaction. In addition, T_n and S_n are used to denote the latest transaction and segment in the current data stream, respectively. Thus, the current data stream is either $W \leq T_{n-w+1}, \ldots, T_n>$ or $W \leq S_{n-m+1}, \ldots, S_n>$, where w and n denote the size of W and the number of segments in W, respectively. The main stream Data Multiple Imputation will be divided into m segments. Each of the m segments contains a set of successive equal

number of *s* transactions. Also in each segment, the summary of transactions belonging to that segment is stored for further analysis.

By taking this segment based manner of Imputation, each time when a new transaction is entered in to the segments, the earliest segment deleted based on the summary of transactions. As a result, no need to maintain the whole transactions within the current data stream in all along to support data stream Imputation. The parameter *m* directly affects the consumption of memory and it is recorded simultaneously. A large volume of *m* means the data stream will Impute and update the transactions more frequently.

THE MATHEMATICAL MODEL SEGMENT INITIALIZATION

The Main stream Data Multiple Imputation 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, it's 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, 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 an Impute expires and then compute the frequency of such new Data over this Impute and update the *aux array*s accordingly.

For Each New Impute S

1: For each segment $s \in S$
updates.*freq* over S
2: Mine S to compute $\sigma_\alpha(S)$
3: For each existing segment $s \in \sigma_\alpha(S) \cap S$
Remember S as the last Impute in which s is frequent
4: For each new segment $s \in \sigma_\alpha(S) \setminus S$
$S \leftarrow S \in \{s\}$
Remember S as the first Impute in which s is frequent
create auxiliary array for s and start monitoring it

For Each Expiring Impute S

5: For each segment $s \in S$
updates.*freq*, if S has been counted in
updates.*aux array*, if applicable
reports as delayed, if frequent but not reported
at query time
deletes.*aux array*, if s has existed since arrival of S
deletes, if s no longer frequent in any of the current slides

Figure A1: Mathematical Segment Initialization Pseudo Code

Max Delay: The maximum delay allowed by the $n-1$ Imputes. Indeed, after expiration of $n-1$ Imputes, 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.

CONCLUSIONS

In this paper I have to use Mathematical Segment Initialization Model for Finding Multiple Data Imputation in Main Streams. Mining Missing Data over data streams using main stream Data Multiple Imputation model handled the data streams transaction by transaction. Unlike the landmark data stream model, transactions in the main stream Data Multiple Imputation model will be both inserted into and dropped out from the data stream. *Segment-oriented data stream Imputation* has taken for handling this problem.

REFERENCE

1. *Tensor Voting Techniques and Applications in 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, Jukka Miettinen, 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 Leering Approach, 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 Missing Value Imputation Journal: International Journal for Research in Science & Advanced Technologies, Vol 1.Issue-2, 2013, Ms.R. Malarvizhi and Dr.Antony SelvadossThanamani..*