International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195220

An Efficient Missing Data Imputation Based On Co-Cluster Sparse Matrix Learning

F. Femila*, G. Sridevi, D. Swathi, K. Swetha
Department of Computer Science, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India

ABSTRACT

Missing data padding is an important problem that is faced in real time. This makes the task of data processing challenging. This paper aims to design a solution for this problem which is ways different from traditional approaches. The proposed method is based on co-cluster sparse matrix learning (CCSML) method. This algorithm learns without reference class, and even with data continuous missing rate as high as the existing techniques. This method is based on a tensor optimization model and labeled maximum block. The computational models of sparse recovery learning are based on low-rank matrix and co-clusters of genome-wide association study (GWAS) data matrices, and the performance is better than existing techniques.

Keywords: Data Preprocessing, Missing Value, Co-Cluster Sparse Matrix, Sparse Recovery

I. INTRODUCTION

In the real world data there are few instances where the data will be missing. Missing data are classified into three types as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).It is necessary to implement the imputation of missing values in the stage of data pre processing to reduce errors in further. Even a small percent of missing data occurred it can lead to various problems with analysis and it leads to wrong conclusion at the end. Missing data can be dangerous because it is difficult to identify the problem and can't predict when missing data cause problem because sometimes the results are affected and sometimes they are not.

The missing data are caused by various operations such as equipment failures, erroneous human operations, mismatches of location of operating points, network failures, errors in data transmitting, corrupt data, failure to load the information, or incomplete extraction.

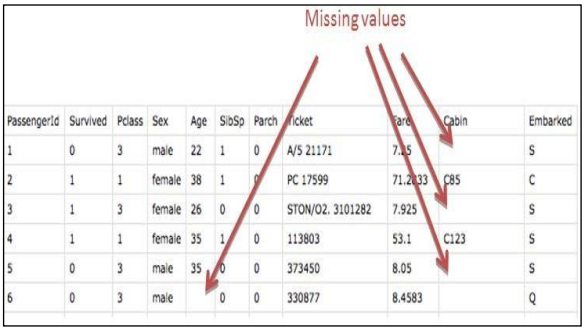

Figure 1. Missing Value Data Set

Due to these missing values it lead to data errors, incomplete results, and inconsistencies. These incomplete data affect the quality of the information and the results even lead to the establishment of the wrong data mining model and it will also deviate from the actual result. Handling these missing values is most challenging task to be done.

The various imputation methods are used such as deleting rows, replacing with mean / median / mode, assigning an unique category Pearson [3] pointed out that missing data that will lead to three major problems: (1) most of the data processing algorithms at this stage cannot process datasets with missing data. Commonly used algorithms or systems are unable to deal with these incomplete datasets; (2) In the data mining process, in order of performing simple operation to save time, the issue of missing records has often been overlooked, which will lead to poor statistical results; (3) mining datasets with missing records. This paper demonstrates some of the popular statistical methods for imputing missing values.

II. RELATED WORKS

According to Xiaolong Xu[20], the existing system of imputation approach of missing values cannot satisfy the analysis requirements because of low accuracy and when the rate of missing values increase the accuracy decreases rapidly. So In this system they propose a novel missing value imputation algorithm which is based on the evidence chain (MIAEC). To extend MIAEC for large-scale data processing, they apply the map reduce programming model to realize the distribution and parallelization of MIAEC. This approach has higher imputation accuracy and also assured with the increasing rate of missing value or the position change of missing value. In this imputation technique only deals with discrete missing data not the continuous missing data.

Md. Geaur Rahman and Md Zahidul Islam[16] described a novel technique called iDM I that imputes missing values of a data set by combining a decision tree algorithm (DT) and an expectation-maximization (EMI) algorithm. We first divide a data set into horizontal segments through applying a DT algorithm and then apply an EMI algorithm on each segment in order to impute the missing values belong to the segment. We evaluate the performance of iDM I over three high quality existing techniques on two real data sets in terms of four evaluation criteria. Our initial experimental results, including several statistical significance analysis, indicate the superiority of iDM I over the existing techniques.

M. Zhu and X. B. Cheng[11],Even though there are some popular imputation methods proposed, these methods perform poorly in the estimation of missing values in the trash pickup logistics management system (TPLMS). The problem in the TPLMS because of missing values is significant and may result in unserviceable decision-making considering the above stated problem, this paper introduces an iterative KNN imputation method which associates with weighted k nearest neighbour (KNN) imputation and the grey relational analysis (GRA).The expected results suggest that the proposed method gets a better performance than the existing methods in terms of imputation accuracy and convergence speed.

According to a survey by A. Karmaker and S. Kwek[9], Missing attribute values in data are quite common in many classification problems. In this paper, we incorporate an Expectation-Maximization (EM) inspired approach for filling up missing values to decision tree learning that is to mainly focus on improving classification accuracy. In this approach each missing attribute-value is iteratively filled using a predictor constructed from the known values and predicted values of the missing attribute-values from the previous iteration. We show that our approach significantly outperforms some standard machine learning methods for handling missing values in classification tasks. This approach is implemented as the default technique of handling incomplete data by many statistical software packages (such as SAS and SPS)

Xiaofeng X. Zhu et., al.[4],proposed various techniques are implemented to deal with missing values in data set with homogeneous attribute .This

paper proposes a system in which imputing missing data in data set with heterogeneous attributes. The first it proposes two consistent estimators for discrete and continuous missing target values. And then, a mixture-kernel-base id iterative estimator is used to impute mixed-attribute data sets. This method is evaluated with extensive experiments compared with some typical algorithms, and the result demonstrates that the proposed approach is better than these existing imputation methods when compared to classification, accuracy and root mean square error (RMSE) at different missing ratios.

III. PROPOSED SYSTEM

The proposed system sparse recovery, for imputing missing genetic data in genome-wide association studies(GWAS), co-cluster sparse matrix learning (CCSML)The models of sparse matrix are designed based on the sparse properties of low-rank, noisy, genetic datasets of matrices with missing data. This proposed system states that the low-rank matrix completion model is similar to Mendel-Impute, but the matrix co-clustering factorization model is completely new. Sparse matrix is easy to use for metadata analysis, easy-to-process input file format and easy-to-interpret output result files. It has better or comparable performance compared to existing system, especially for handling large sample size data with very different sets of SNPs and no reference panels. The overall methodology is described through a brief block diagram in fig .2

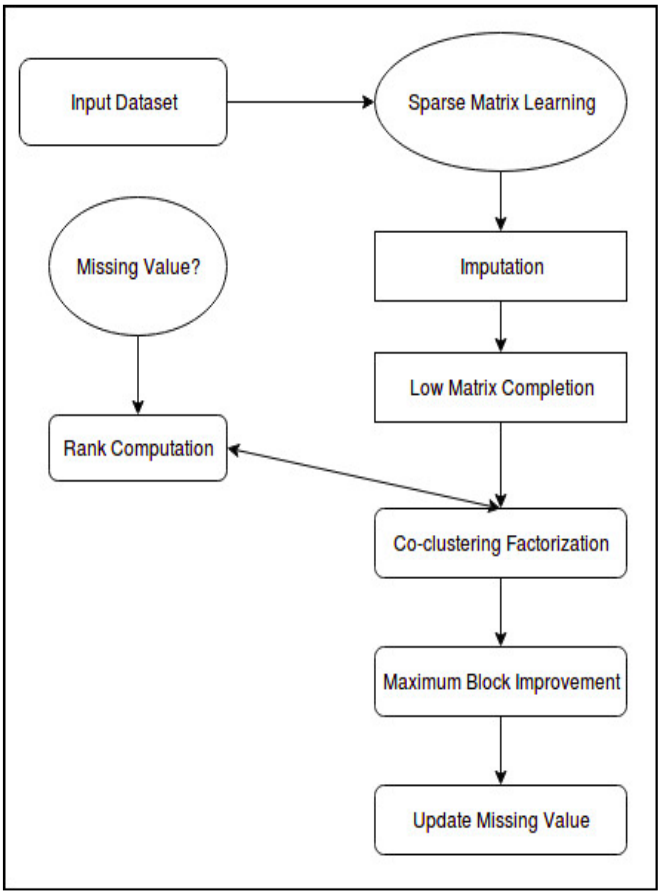

Figure 2. Flow of the Proposed Model

A. FPCA learning

FPCA is designed for solving matrix rank minimization problem and matrix completion problem. SVT is efficient for large matrix completion problems and the algorithm make use of matrix shrinkage. For solving the nuclear norm minimization problem and prove convergence of first of these algorithms we use fixed point iterative. The approximate singular value decomposition procedure, it get a very fast, robust and powerful algorithm, we call it as FPCA (Fixed Point Continuation with Approximate SVD), which can solve very large matrix rank minimization problems. There are many applications in various fields using matrix rank minimization problem such as system identification, optimal control, low-dimensional embedding, etc.

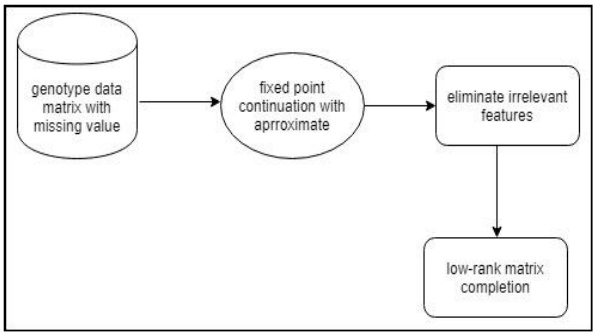

Figure 3. FPCA Learning

B. Sparse low rank matrix algorithm

The sparse low-rank matrix completion model aims to fill in missing data values of a matrix based on the priori information that the matrix under consideration is of low rank. The low-rank matrix completion model can be formulated as the following optimization problem:

$$\min X \ \text{rank}X, \text{s,t.}, Xij=Miji,j \quad (1)$$

Where rank(X) denotes the rank of matrix X and Ω denotes the index set of the known entries of M. That is, it is given a set of known entries of M, and want to fill in the missing entries such that the completed matrix is of low rank. In the genotype missing data imputation problem, each row of the matrix M represents a patient sample, and each column of the matrix M corresponds to a SNP. That is, Mij represents the j th allele of the i th patient sample. It is usually believed that patients can be classified into different categories and patients in the same category should have similar genetic patterns. Therefore, believe that the matrix M is low-rank, or at least numerically low-rank.

The sparse low-rank matrix model has been widely used in online recommendation, collaborative filtering, computer vision and so on. Under certain randomness hypothesis, the model (1) is equivalent to the following convex optimization problem with high probability:

$$\min X \|X\|^*, \text{s,t.}, Xij=Miji,j \quad (2)$$

Where, $||X||^*$ is called the nuclear norm of matrix X and is defined as the sum of singular values of X. The nuclear norm minimization problem (NNM) is numerically easier to solve than the propose model because it is a convex problem. Many efficient numerical algorithms have been suggested to solve the NNM model, use the fixed-point continuation method (FPCA) proposed.

The propose impute method implements Nesterov's accelerated proximal gradient method (APG) to solve (2), while FPCA can be seen as the ordinary version of proximal gradient method for solving (2). Theoretically, APG is faster than FPCA for solving LRMC, because the former attains an -optimal solution in o1 iterations, while the latter one attains an -optimal solution in o1 iterations. Mendel-Impute also implements two important techniques to further accelerate the speed of APG: the sliding window scheme to better balance the trade-offs between accuracy and running time, and the line search technique to find an appropriate step size for the proximal gradient step. From our experiments, we found that the sliding window scheme is quite helpful for missing data imputation. Thus, we incorporated the sliding window scheme to LRMC, denoted as LRMC-s.

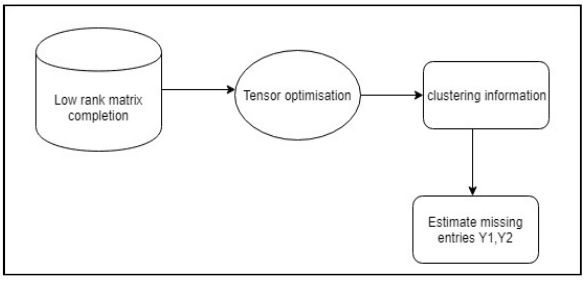

Figure 4. Sparse Low Rank Matrix

C. Co-clustering techniques

The propose a new approach for imputation that is based on matrix co-clustering factorization. Co-clustering model for two-dimensional and higher-dimensional matrix co-clustering, which is based on a

tensor optimization model and an optimization method termed Maximum Block Improvement (MBI) Inspired by the idea of matrix co-clustering for imputation, we develop a basic model as follows.

minA,X,Y1,Y2 fA,X,Y1,Y2:=‖A-Y1XY2‖F2, s,t
Aij=Miji,j Where
A∈Rm×n, Y1Rm×k1,X∈Rk1k2,Y2Rk2×n (3)

In (3), the Frobenius norm of a matrix X is defined as ‖X‖F2=ijXij2 imputation approach, based on the matrix co-clustering factorization, aims to complete matrix M by using a low-rank matrix factorization model. In our framework, A is the data matrix with missing entries; Y1 and Y2 are the artificial row assignment matrix and the artificial column assignment matrix, respectively, and X is the artificial central-point matrix. Note that A is also an unknown decision variable in (3), because only a subset of its entries is known. Moreover, note that (3) requires the input of k1 and k1 , which are closely related to the rank of the matrix to be completed. Therefore, in practice, if we have a good estimation to the rank of the matrix, then (3) is a better model to use than (2), because it also provides us the clustering information of individual samples and SNPs.

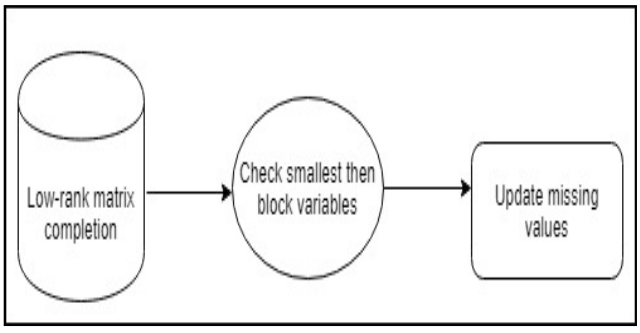

Figure 5. Co-cluster technique

D. Maximum block improvement

The propose model is non-convex, it has some natural block-structure that can be utilized to adopt an efficient solution method. The propose to solve the model (3) using a block coordinate update (BCU) procedure. There are four block variables in the model (3), namely A, X, Y1 and Y2. The basic idea of BCU is, at each iteration, to minimize the function f with respect to one block variable while the other three blocks are fixed at the current known values. This idea is effective because we observed that minimizing f for only one block variable among A, X, Y1 and Y2 is always relatively easy. A naive implementation of the BCU idea is to minimize f in the order of $A \rightarrow$ Y1 $\rightarrow X \rightarrow$ Y2, and in each step only one block variable is updated with the other three blocks being fixed. The propose model, the matrix X actually plays a more important role than the other three blocks. As a result, it is beneficial if we can update the X block more frequently than the other three blocks. Therefore, we implemented the following four different algorithms based on the BCU idea.

"MBI-BL": This is a variant of the MBI algorithm MBI-BL applies MBI algorithm in to minimize f with four blocks variables: X, (Y1 – X), (Y2 – X) and (A – X). In each block, for example, (A – X), we use alternating block minimization scheme to minimize f with respect to A and X alternatively, until the function value ceases to change. After having attempted all four block variables, we update the block variable with maximum improvement.

Algorithm

Maximum Block Improvement

Given initial iterates X0,Y10,Y20, A0, and initial values v0=0, v1=1.
For K=0,1…..run the following until vk-vk+1<ϵ

1. Block Improvement

Xk,1argminX‖AkY1kXY2k‖2 (4)

Y1kXk,2argmin(Y1,X)‖AkY1kXY2k‖2 (5)

Y2kXk,3argminY2,X‖AkY1kXY2k‖2 (6)

Ak,Xk,4argminA,X‖A-Y1kXY2k‖2s,tAij=Miji,j (7)

2. Compute the corresponding objective values

w1=fAk,Xk,1,Y1k,Y2k (8)

w2=fAk,Xk,2,Y1k,Y2k (9)

w3=fAk,Xk,3,Y1k,Y2k (10)

w4=fAk,Xk,4,Y1k,Y2k (11)

3. Maximum Improvement

Compare w1, w2, w3, w4 pick up the smallest value to update the corresponding block variables:

- If w1is the smallest, then

Xk+1Xk, 1w1 (12)

- If w2is the smallest, then

Y1k+1Y1k, Xk+1Xk, 2, vk+1w2 (13)

- If w3is the smallest, then

Y2k+1Y2k, Xk+1Xk, 3, vk+1w3 (14)

- If w4is the smallest, then

Ak+1Ak, Xk+1Xk, 4, vk+1w4 (15)

All the algorithms are terminated when the objective value in the $(k + 1)$-th iteration does not decrease significantly from that in the k-th iteration.

IV.EXPERIMENTAL RESULT

The system is designed such that all the missing values in data set are identified by this system and the result obtain will contain no missing values. The results obtained by CCML method reduce the error rate, decrease the computation time and increase the accuracy when compared with the MIAEC method.

A. Imputation Error Rate

In this graph the comparison is shown between the existing system by using MIAEC and the proposed system by using co cluster sparse matrix. In this graph the x axis represents the missing rate and in y axis it represents the error rate. By implementation of the proposed technique the error rate is decreased and the missing rate is also decreased when compared to MIAC method.

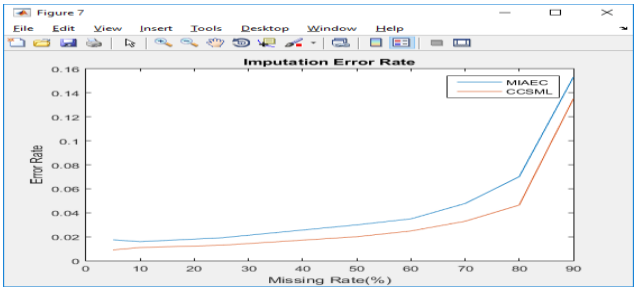

Figure 6. Comparison of Imputation Error Rate Between MIAEC and CCSML

B. Imputation Accuracy

The result obtained by CCSML method shows higher accuracy when compared with MIAEC. When there is a increase rate in accuracy the missing value problems are being reduced.

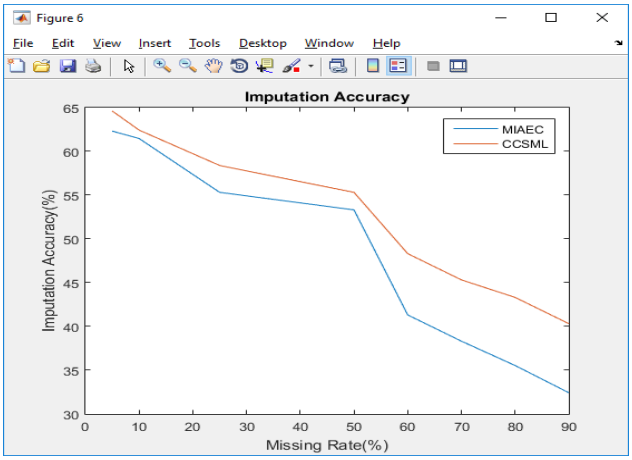

Figure 7. Comparison of Imputation Accuracy Between MIAEC and CCSML

C. Imputation computation time

The resultant graph represents that the CCML method results in decrease in the computation time

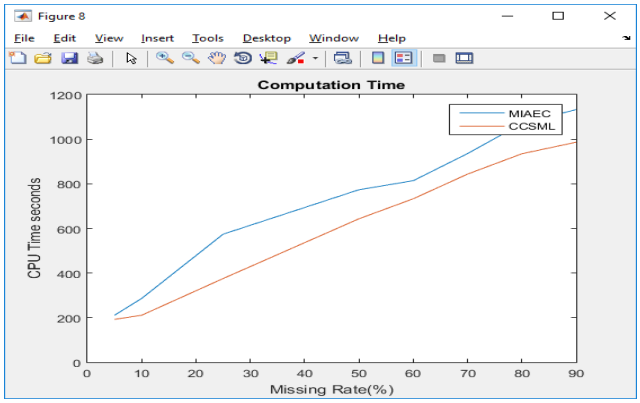

Figure 8. Comparison of imputation computation time between MIAEC and CCSML

V.CONCLUSION

The propose system sparse recovery, for imputing missing genetic data in genome-wide association studies. co-cluster sparse matrix learning (CCSML)The models of sparse matrix are designed based on the sparse properties of low-rank and low numbers of co-clusters of the large, noisy, genetic datasets of matrices with missing data. would like to point out that the low-rank matrix completion model is similar to Mendel-Impute, but the matrix co-clustering factorization model is completely new. The propose approach is able to effectively find patterns for imputation within study data, both with and without reference panels, and even with data missing rate as high as 90%. The performance of our approach with several other main stream approaches for genotype imputation. Sparse matrix is easy to use for metadata analysis, and it requires very simple, easy-to-process input file format and easy-to-interpret output result files. It has better or comparable performance compared to current state-of-the-art methods, especially for handling large sample size data with very different sets of SNPs and no reference panels.

II. REFERENCES

[1]. R. K. Pearson, "The problem of disguised missing data," ACM SIGKDD Explorations News. Lett., vol. 8, no. 1, pp. 83-92, 2006.

[2]. R. J. A. Little and D. B. Rubin, Statistical Analysis With Missing Data, 2nd ed. Hoboken, NJ, USA: Wiley, 2002, pp. 200-220.

[3]. F. Z. Poleto, J. M. Singer, and C. D. Paulino, "Missing data mechanisms and their implications on the analysis of categorical data," Stat. Comput., vol. 21, no. 1, pp. 3143, Jan. 2011.

[4]. X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu, "Missing value estimation for mixed-attribute data sets," IEEE Trans. Knowl. Data Eng., vol. 23, no. 1, pp. 110-121, Jan. 2011.

[5]. Y. Qin, S. Zhang, X. Zhu, J. Zhang, and C. Zhang, "Semi-parametric optimization for missing data imputation," Appl. Intell., vol. 27, no. 1, pp. 79-88, 2007.

[6]. U. Dick, P. Haider, and T. Scheffer, "Learning from incomplete data with innate imputations," in Proc. 25th Int. Conf. Mach. Learn., Jul. 2008, pp. 232-239.

[7]. Z. Shan, D. Zhao, and Y. Xia, "Urban road traffic speed estimation for missing probe vehicle data based on multiple linear regression model," in Proc. 16th Int. IEEE Conf. Intel. Transp. Syst. (ITSC), The Hague, The Netherlands, Oct. 2013, pp. 118-123.

[8]. F. Bashir and H.-L. Wei, "Parametric and non-parametric methods to enhance prediction performance in the presence of missing data," in Proc. 19th Int. Conf. Syst. Theory, Control Compute. (ICSTCC), Cheile Gradistei, Romania, 2015, pp. 337-342.

[9]. A. Karmaker and S. Kwek, "Incorporating an EM-approach for handling missing attribute-values in decision tree induction," in Proc. 5th Int. Conf. Hybrid Intell. Syst. (HIS), 2005, p. 6.

[10]. D.-H. Yang, N.-N. Li, H.-Z.Wang, J.-Z. Zhao, and H. Gao, "The optimization of the big data cleaning based on task merging," Chin. J. Comput., vol. 39, no. 1, pp. 97-108, 2016.

[11]. M. Zhu and X. B. Cheng, "Iterative KNN imputation based on GRA for missing values in TPLMS," in Proc. 4th Int. Conf. Comput. Sci. Netw. Technol. (ICCSNT), Harbin, China, 2015, pp. 94-99.

[12]. P. Keerin, W. Kurutach, and T. Boongoen, "Cluster-based KNN missing value imputation for DNA microarray data," in Proc. IEEE Int. Conf. Syst., Man, (SMC), Seoul, South Korea, Oct. 2012, pp. 445-450.

[13]. L. Jin, H. Wang, S. Huang, and H. Gao, ``Missing value imputation in big data based-on map-reduce,'' J. Comput. Res. Develop., vol. 50, no. S1, pp. 312-321, 2013.

[14]. M. G. Rahman and M. Z. Islam, "iDMI: A novel technique for missing value imputation using a decision tree and expectation-maximization algorithm," in Proc. 16th Int. Conf. Comput. Inf. Technol. (ICCIT), Khulna, Bangladesh, 2014, pp. 496-501.

[15]. S. Wu, X.-D. Feng and Z.-G. Shan, "Missing data imputation approach based on incomplete data clustering," Chin. J. Comput., vol. 35, no. 28, pp. 1726-1738, Aug. 2012.

[16]. Md. Geaur Rahman and Md Zahidul Islam"iDMI: A Novel Technique for Missing Value Imputation using a Decision Tree and Expectation-Maximization Algorithm ",16th Int'l Conf. Computer and Information Technology,Khulna, Bangladesh, 8-10 March 2014.

[17]. M. G. Rahman and M. Z. Islam, "kdmi: A novel method for missing values imputation using two levels of horizontal partitioning in a data set," in The 9th International Conference on Advanced Data Mining and Applications (ADMA 13). in press, Hangzhou, China: Springer, 2013.

[18]. K. Cheng, N. Law, and W. Siu, "Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data," Pattern Recognition, vol. 45, no. 4, pp. 1281-1289, 2012.

[19]. Browning, S. R. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet. 124, 439-450, doi: 10.1007/s00439-008-0568-7 (2008).

[20]. Xiaolong xu, weizhi chong, shancang li, abdullahi arabo and jianyu xiao. "MIAEC: Missing Data Imputation Based on the Evidence Chain",

Cite this article as :

F. Femila, G. Sridevi, D. Swathi, K. Swetha, "An Efficient Missing Data Imputation Based On Co-Cluster Sparse Matrix Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 215-222, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT195220
Journal URL : http://ijsrcseit.com/CSEIT195220