ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 5, Issue 7, July 2016

FLEXIBLE FACTORIZATION FOR MULTI ASPECT DATA MINING IN BIG DATA

1. Sneh Chauhan ,**2**.Diksha Negi, **3**.Assoc. Prof. Mr. Manish Sharma
Department of Computer Science and Engineering Graphic Era University Dehradun

ABSTRACT

Data decomposition presents the collective techniques to complete missing data. The proposed methods are based on the Kronecker product (KP) decomposition (PARAFAC). Besides the standard KP decomposition, also a controlled decomposition is utilized. Measures such as normalized error rate and mean square error, bias and variance are used to evaluate the performance of the proposed tensor-based methods in comparison with other widely used approaches, such as mean substitution and *k*-nearest neighbour estimation. The numerical results validate that the proposed methods gives important improvement in comparison to popular methods. The best results are achieved for the normalized decomposition .We use the PARAFAC and SVD to complete the missing data and for finding the distance using KNN. This implementation verified by using MATLAB.

Keywords: Tensor, Big data, KNN, SVD, PARAFAC, MATLAB etc.

I.INTRODUCTION

Missing data can rise in a diversity of settings due to loss of information, errors in the data collection procedure, or costly research. For example, in biomedical signal processing, missing data can be encountered during EEG analysis, where numerous electrodes are used to assemble the electrical activity along the scalp. Many real-world data with missing accesses are ignored because they are deemed unsuitable for analysis, but this work gives to the increasing indication that such data can be examined. Dissimilar most former studies which have only considered matrices, we focus here on the problem of missing data in tensors because it has been exposed progressively that data frequently have more than two methods of difference and are therefore best signified as multi-way arrays (i.e., tensors) [3,7]. For example, in EEG data each signal from an electrode can be characterized as a time-frequency matrix; thus, data from multiple channels is three-dimensional (temporal, spectral, and spatial) and procedures a three-way array [9]. Social network data, network traffic data, and bibliometric data are of importance to numerous applications such as community detection, link mining, and more; these data can have numerous dimensions/modalities, are often massively large, and usually have at least some missing data. Other examples of multi-way arrays with missing entries from different previous work. For in case, [7] shows that, in spectroscopy, intermittent machine failures or different sampling frequencies may result in tensors with missing fibers (i.e., the higher-order equivalents of matrix rows or columns, see Figure 1). Correspondingly, missing fibers are encountered in multidimensional NMR (Nuclear Magnetic Resonance) investigation, where sparse sampling is used in order to decrease the experimental time [8].

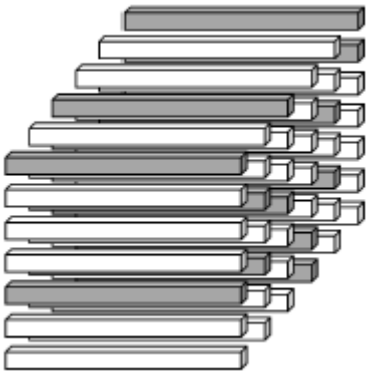

Figure 1: A 3-way tensor with missing row fibers (in gray).

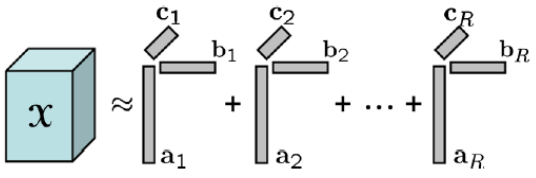

Figure 2: Illustration of an R-component CP model for a third-order tensor X.

Adesign of CP for third-order tensors is given in Figure 2. The CP decomposition is extensible to N-way tensors for N ≥ 3, and there are frequent approaches for computing it [2].

In the case of incomplete data, a standard practice is to attribute the missing values in some method (e.g., replacing the missing entries using average values along a specific mode). Imputation can be useful as long as the quantity of missing data is small; however, performance damages for large amounts of missing data [10, 1].

A. k-nearest neighbour classifier

In this paper, it will use a simple data mining algorithm (k-nearest neighbour classifier) for classification of imperfect data using mathematical data. Goal of the research using KNN

1. Build a simple k-nearest neighbour (KNN) classifier to classify examples in a given dataset.
2. Calculating the consequence of neighbour set size and training set size on the accuracy of the KNN classifier built.

A k-nearest neighbour classifier is a simple data classifier, where the group (class label) of a data item is strong-minded by using a common vote of its neighbours, conveying the data item to the group most common among its *k* nearest neighbours. The input to the algorithm is a set of tuples, where one of the attributes is a class label and the other attributes are the features of the given data. While some of the tuples have known class labels, others have unknown class labels, and the task is to label those tuples with unknown class labels.

B. Singular Value Decomposition (SVD)

Let X be an n × m matrix with n ≥ m and rank (X) = R Then the SVD of X is $X = U S V^T$

Where U is n ×m and $U^TU = Im$
V is m ×m and $V^TV = VV^T = I_m$
S = diag{ s1 , ….., sR ,0,… ,0} is m ×mcolumns of U are equally orthogonal and have length 1 columns of V are equally orthogonal and have length 1 singular values of X are $s_1 \geq \ldots \geq s_R > 0 \geq \ldots \geq 0$rank(X) = R = # positive singular values of X

SVD ➔ $X = s_1 u1 v_1^T + \ldots + s_R u_R v_R^T$ with u_j and v_j the j -th columns of U and Vrank $(u_j v_j^T) = 1$ ➔ SVD decomposes X into R rank-1 matrices**"economy size SVD"** U_R is n ×R and $(U_R)^T U_R = I_R$
V_R is m ×R and $(V_R)^T V_R = I_R$

S_R= diag { S_1 , … ,s_R} is R ×R➔ columns of U_R and V_R are exclusive up to sign if the singular values are all separate.

II. PROBLEM FORMULATION

We train tensors that have orthonormal CANDECOMP/PARAFAC (CP)tensor decomposition with a small number of mechanisms. Furthermore, for simplicity of notation and explanation, we only reflect symmetric third order tensors. We would like to strain that our methods simplifies simply to handle non-symmetric tensors as well as higher-order tensors. Officially, we assume that the true tensor T has the subsequent form:

$$T = \sum_{l=1}^{r} a\sigma(u\sigma \otimes u\sigma \otimes u\sigma) \in R^{n \times n \times n} \quad \textbf{(1)}$$

With r << n, ul∈R^n with IIul II = 1, and ul's are orthogonal to each other. We let U ∈$R^{n \times r}$be a tall-orthogonal matrix where ul's is the l-th column of U and Ui⊥Uj for i ≠ j. We use⊗ to represent the standard outer product such that the (i, j, k)-th element of T is given by: Tijk= ∑a UiaUjaUka. We more accept that the ui's are unstructured, which is formalized by the notion of incoherence usually assumed in matrix completion problems. The incoherence of a symmetric tensor with orthogonal decomposition is

$$\mu(T) \max_{i \in [n], \ell \in [r]} \sqrt{n}\,[U\, i\ell], \quad \textbf{(2)}$$

where [n] = {1, . . . , n} is the set of the first n integers. Tensor completion turn out to be progressively for tensors with larger μ(T), because the 'mass' of the tensor can be focused on a few accesses that might not be exposed. Out of n^3 entries of T, a subset Ω [n] × [n] × [n] is discovered. We use $P_\Omega(\cdot)$ to signify the projection of a matrix onto the discovered set such that

$$P_\Omega(T)_{ijk} = \begin{cases} T_{ijk} \; if (i,j,k) \in \\ 0 \;\; otherwise \end{cases}, \quad \textbf{(3)}$$

We need to improve T precisely using the given entries (PΩ(T)). We undertake that each (i, j, k) for all i ≤ j ≤ k is comprised in with a secure probability p (since T is symmetric, we comprise all permutations of (i, j, k)). This is equivalent to fixing the total number of samples |Ω| and choosing Ω regularly at random over all $\binom{n^3}{|\Omega|}$ choices. The goal is to ensure exact recovery with high probability and for | Ω | that is sub-linear in the number of entries (n^3).

III.PROPOSED METHOD

Algorithm 1: ALS for 3-way PARAFAC decomposition.

Involve: Tensor $\chi \in R^{I \times J \times K}$, rank R, and maximum iterations T

Confirm: PARAFAC decomposition $\lambda \in R^{Rx1}$, A C R^{1xR}, B $\in R^{J \times R}$, C $\in R^{K \times R}$
1: Reset A, B, C;
2: for t = 1,..., T do
3: A $\leftarrow X_{(1)}$ (C Θ B) $(C^T C * B^T B)^{\dagger}$:
4:Measure the distance of the tuple *t* to each of the labelled tuples in the dataset using KNN.
5. Find the set *S* of the k-nearest neighbours of *t* (i.e. *k* tuples with the minimum distance to *t*).
6. Discover the popular class label in the set *S*. If two or more classes have the common (i.e. the same number of incidences in the set), decrease *k* by 1 till you find the majority class label (Note that you decrease k only for the classification of this occurrence, not the entire dataset).

7. Standardize columns of A (storing norms in vector λ);
8. B $\leftarrow X_{(2)}$ (C Θ A) $(C^T C * A^T A)^{\dagger}$:
9. Standardize columns of B (storing norms in vector λ);

10. SVD is used for dimension reduction, we use it both as a graph dividing and as a way to observe the strangest user of numerical random data, and hence most motivating user in a tensor mining.

11. SVD changes data based on association, and so can extract structure that is imperfect; it does not need pre-specification of the constructions of concern.
12: C $\leftarrow X_{(3)}$ (B Θ A) $(B^T B * A^T A)^{\dagger}$:
13: Stabilize columns of C (storing norms in vector λ);
14: if meeting criterion is met then
15: break for loop;
16: end if
17: end for
18: return λ, A, B, C:

The Parallel Factor analysis (PARAFAC)

PARAFAC decays a matrix **D** into a produce of three matrices allowing to three modes. In its place of having a score and a loading matrix as in SVD, each constituent contains of a score matrix meant**A** and two loading matrices signified **B**and**C**. In PARAFAC, it is common not to differentiate the score and the loading matrices. In other words, the PARAFAC model of a three-dimensional 3-D matrix is assumed by three loading matrices**A**, **B** and **C** with elementsa_{if}, b_{jf} and c_{kf} as follows:

$$d_{ijk} = \sum_{f=1}^{F} a_{if}\, b_{jf}\, c_{kf} + e_{ijk} \qquad (4)$$

The elements of **D** having a size $I \times J \times K$ are denoted d_{ijk}. The trilinear model is found to minimize the sum of squares of the residuals denoted e_{ijk}. $F$is the number of factors extracted in each mode, which describes the maximum of information contained in the matrix **D**. The model use a cost function as follows:

$$L(\mathbf{A},\mathbf{B},\mathbf{C}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(d_{ijk} - \sum_{f=1}^{F} a_{if}\, b_{jf}\, c_{kf}\right)^2 \qquad (5)$$

The benefit of this procedure is that it arranges for simple and strong models that can be simply interpreted [Harshman, 1970]. Also, the solution of the PARAFAC model is exclusive [Kruskal, 1976]. Kruskal (1977) planned even less preventive circumstances in cases where single solutions can be predictable. This latter author usages the k-rank of the loading matrices, showing that if $k_A + k_B + k_c \geq 2F + 2$ then the PARAFAC solution is unique, with k_A being the k-rank of matrix**A**, k_B the k-rank of **B** and $k_C$the k-rank of**C**. $F$is the probable number of factors or mechanisms.

IV. RESULT

Table 1 and no of figures summarize our results. As there is no straightforward algorithm is known to determine the rank of a tensor (i.e. smallest number of components in CP decomposition of the tensor, we used cross-validation to estimate the rank by trying different number of components during training. The proposed method significantly improves the imputation accuracy in terms of error rate compared to existing methods. Imputing missing data should not strongly affect the variance of the data (responses in medical questionnaires). Obviously, mean substitution significantly reduces the variances, since it replaces all missing data by the mean value.

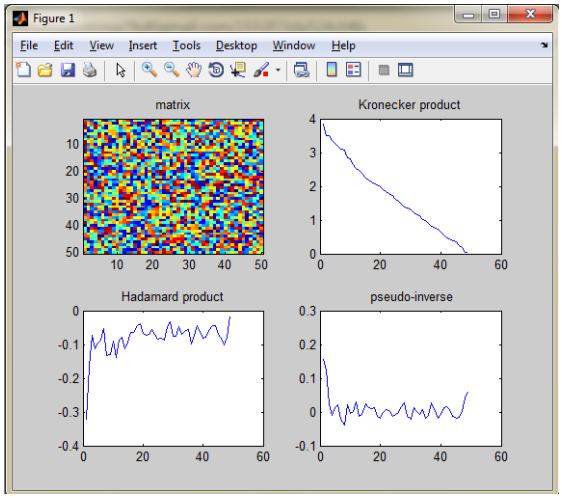

Figure: 4.1 A graph of matrix, kronecker, Hadamard and Pseudo inverse for tensor

In figure 4.1, we mentioned the result categorisation in four parts namely matrix, kronecker, Hadamard and Pseudo inverse for tensor.
Here matrix composition using SVD and other result created by ALS- PARAFAC technique.

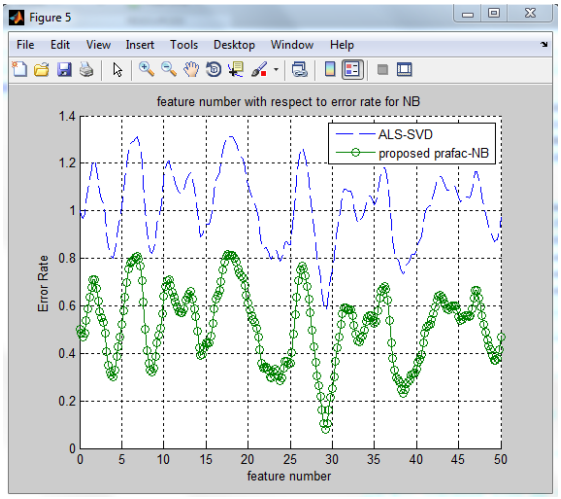

Figure 4.2: feature number with respect to error rate for NB

In this figure we can see that the error rate value of Proposed ALS-NB is less than ALS-SVD.

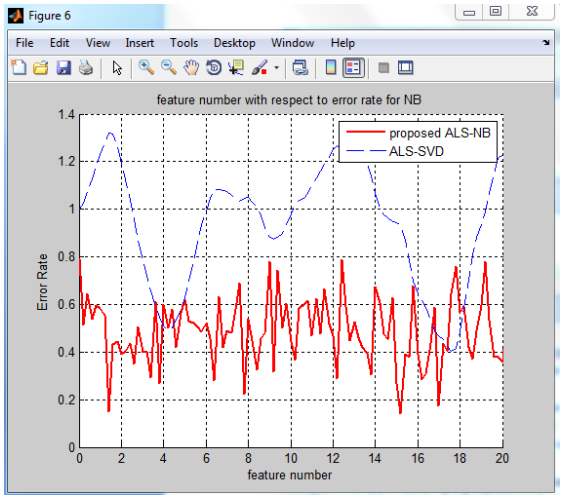

Figure 4.3: feature number with respect to error rate for NB

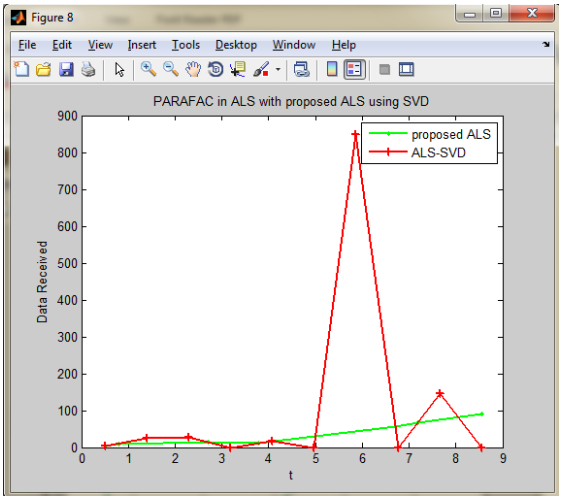

Figure 4.4: PARAFAC in ALS with proposed ALS using SVD

Time	ALS SVD (Data received)	Proposed ALS (Data received)
0	0	--
1	20	10
2	30	10
3	2	10
4	19	10
5	0	40
6	850	50
7	29	70
8	101	90
9	--	--

Table 1: PARAFAC in ALS with proposed ALS using SVD

V. CONCLUSION

We have seen a change in data mining in current years: from models converging on matrices to those learning higher-order tensors and now we are in need of replicas to discover and excerpt the unstructured data from many sources. One method is to express this problem as a coupled matrix and tensor factorization problem. In this paper, we addressed the problem of solving coupled matrix and tensor factorizations when we have squared Euclidean distance as the loss function and introduce a k-nearest neighbor algorithm, which resolves for all distance factors in all data sets instantaneously. We have also protracted our algorithm to data with missing entrances by addressing the case where we have an incomplete higher-order tensor coupled with matrices. The PARAFAC and SVD algorithm can simply be removed to multiple incomplete data sets.

REFEENCES

[1] LeeSaela, InahJeonb, UKangb, "Scalable Tensor Mining" Big Data Research 2 (2015) 82–86.

[2]E.E. Papalexakis, U. Kang, C. Faloutsos, N.D. Sidiropoulos, A. Harpale, Large scale tensor decompositions: algorithmic developments and applications, IEEE Data Eng. Bull. 36(3) (2013) 59–66 http://sites.computer.org/debull/A13sept/p59.pdf.

[3]E.E. Papalexakis, C. Faloutsos, N.D. Sidiropoulos, Parcube: sparse paralleliz-able tensor decompositions, in: Machine Learning and Knowledge Discovery in Databases – European Conference, ECML PKDD 2012, Proceedings, Part I, September 24–28, 2012, Bristol, UK, 2012, pp.521–536.

[4] D. Tao, X. Li, X. Wu, and S.J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, 2007.

[5] J. D. Carroll and J. J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young" decomposition. Psychometrika, 35:283{319, 1970.

[6] P. Comon, X. Luciani and A. L. F. de Almeida, Tensor Decompositions, Alternating Least Squares and other Tales, Journal of Chemometrics, 23 (2009), pp. 393-405.

[7] Z. Cai, M. Heydari, G. Lin, Iterated Local Least Squares Imputation for Microarray Missing Values. Journal of Bioinformatics and Computational Biology, vol. 4 (5), 2006, pp. 935-957.

[8] M. M. Subasi, E. Subasi, M. Anthony, P. L. Hammer, A new imputation method for incomplete binary data, Discrete Applied Mathematics, vol. 159 (10), 2011, pp. 1040-1047.

[9] L. De Lathauwer, B. De Moor, and J. Vandewalle, On the best rank-1 and rank-(R1;R2; : : : ;RN) approximation of higher-order tensors, SIAM J. matrix Anal. Appl., 21 (2000), pp. 1324 1342.

[10] A. Ruhe, Numerical computation of principal components when several observations are missing, Tech. Rep. UMINF-48-74, Department of Infor-mation Processing, Institute of Mathematics and Statistics, University of Umea, Umea, Sweden (1974).