Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.000103

TENSOR-BASED ALGEBRA FOR MULTILINEAR STRUCTURE OF MICROARRAY EXPRESSION RECOGNITION

Nguyen Thi Ngoc Anh[1], Nguyen Tran Quoc Vinh[1], Ho Phan Hieu[2], Vo Trung Hung[2]

[1] Khoa Tin học, Trường Đại học Sư Phạm, Đại Học Đà Nẵng, [2] Đại học Đà Nẵng

ngocanhnt@ued.udn.vn, hophanhieu@ac.udn.vn, ntquocvinh@ued.udn.vn, vthung@dut.udn.vn

ABSTRACT: *In this paper, there are two current challenges of microarray expression are investigated, namely missing values recovery and feature extraction for supervised learning. The first axis is focus on how to deal with the main properties of time sequences of microarray, including tensor structure, noise, and temporal dynamic characteristics. This allows discovering latent factors and evolving trends for offering missing imputation. Then, an improvement of orthogonal Tucker decomposition based on discriminant analysis for multilinear structure of microarray expression is presented for dimensionality reduction. Consequently, an exploring a novel type of third-order microarray expression, termed as gene - sample - time (GST), is presented for biological sample classification. The contributions will be distributed along two main thrusts of effectiveness; including latent modeling setting for imputing missing values based on the High-Order Kalman Filter and feature extraction based on Tensor Discriminative Feature Extraction. Those proposal methods are carried out on Interferon beta (INFβ) dataset of GST-microarray expressions to distinguish the patients in the favorable response group and the remaining patients in the problematic-treatment-response group. The experimental performance corroborates the advantages of the proposed approaches upon those of the matrix-based algorithms and recent tensor-based, discriminant-decomposition, in terms of missing values completion, classification accuracy of 90.23% and computation time.*

Keywords: *Tensor Factorization and Decomposition; Microarray; Discriminant Analysis; Missing values; Supervised Learning, Kalman Filter.*

I. INTRODUCTION

Medical and biological sciences are the subjects of important advances, since the last decades, due to new and improved sensors and to the large amount of available data. In this paper, a microarray expression is presented that contains the thousands of genes deriving from expression levels simultaneously [1]. Traditionally, this type of microarray data tend to be formulated a two-dimensional (2D) gene expression matrix so as to facilitate data processing and analysis, represented by gen × sample datasets and gen × time point datasets [1-3]. However, with the availability of large computational power and breakthrough in computing paradigms, new approaches are feasible in bioinformatics with a large scale is necessary instead of standard flat-view matrix-based. This is because the traditional methods are limited in that they risk losing the covariance information among the various modes whereby the underlying information cannot be extracted.

Indeed, with rapid acquisition of biological experiments from different laboratories or studies, many higher-order biological data representing interactions between more than two types of variables can be obtained [2, 3, 4, 7], shown in Figure 1. In this paper, the fundamental question of handling those large datasets of multimodal recordings of microarray scans that is to combine data from conditions of different nature to extract meaningful information is investigated. Consequently, the expression levels of a gene yields a 3-D microarray representation, formulated as gene × sample × time, known as GST is examined. Such three-dimensional microarray, GST structure contains the expression gene with respect to the samples monitored over a series of time points that is used to develop healthcare system for drug-treatment, disease detection or dosage monitoring [9]. Furthermore, the missing data in microarray is encountered frequently due to insufficient resolution, artifacts, systematic error, or incomplete experiments [8]. These missing data can cause distortion, repudiation, and further, reduce the effectiveness of analyzing algorithms. Current methodologies for microarray analysis require a complete set of microarray data matrix as input. Therefore, an accurate and reliable imputation approach for missing values is necessary to avoid incomplete data sets for analyses and further improve the usage of performance techniques.

In this paper, there are two folds are investigated, including missing values imputation and feature extraction for supervised modeling. Incomplete microarray is first pre-processed via using the High-Order Kalman Filter (HOKF) method [9] in order to handle missing values of original dataset. In the next phase, feature extraction is applied on the completed dataset so as to extract key set of prominent discriminative features for microarray recognition. With those applied method, the original multilinear representation of dataset is preserved. Particular, the recognition modeling is proposed as following: firstly, the HOKF is applied to solve the missing problems of microarray. Then, the improvement of orthogonal Tucker decomposition [4] is applied to seek new high-order subspace features from tensor-based input of training sample for dimension-reduction purpose. During this phase, the reduced core tensor denotes for training features capturing the correlation among the basic components are discovered. Then, the tensor-based testing inputs are projected onto this feature subspace so as to find testing features. At the last step, the testing features and training features are compared to recognize which class for each input sample belongs to. With this mechanism, the

preservation of tensor-based structure dataset of real time series of microarray expressions is evaluated to favorable and problematic responder recognition. The recognition accuracy corroborates the advantages of the proposed methods with averages of 90.23% that performs an improvement upon those of the matrix-based techniques and discriminant-decomposition algorithms recently.

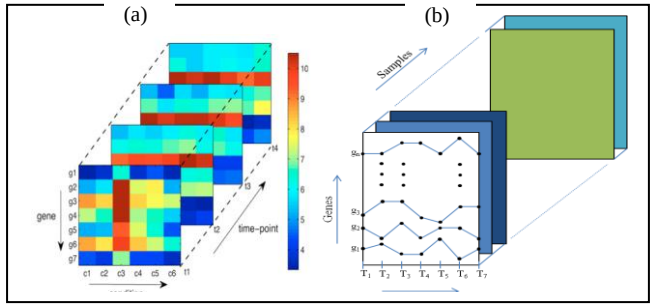

Figure 1. Illustrations of third-order representation of microarray dataset: (a) $gen \times condition \times time$ (b) $gen \times sample \times time$

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the previous works on missing imputation and feature extraction for microarray tensor-based datasets. The preliminary of tensor algebra is is presented in upcoming section. In section 4, the both proposed methods are modeled for microarray dataset in term of high-order missing imputation and feature extraction. Next section will provide the details the experimental performances and discussions of those methods on real microarray dataset. The paper will summarized and point our possible future research in section 6.

II. RELATED WORK

GST data used to propose models for diseases treatment [8]. There are two main problems associated with the analysis of GST data; namely missing values and small samples. Since the microarray is expressed the levels of genes with respect to biological samples in which monitored over a series of time points. Missing values can occurs in gene or samples at some time points due to technical issues in the measuring process. In addition, the expressions of large number of genes are measured from a small number of samples across a small set of time points [9]. Therefore, missing recovery is required as a pre-processed in order to overcome this problem instead of ignore those samples. The simplest way to handle this problem is to replace missing values by appropriate mean values. Interpolation method is also known as the baseline alternative method for missing values imputation in microarray time series dataset [9]. In matrices-based, the well- known techniques related to dimension reduction and latent variables, including Singular Value Decomposition (SVD) and Principal Component Analysis (PCA), could impute missing values through discovering correlations in multiple gene expression or AutoRegression (AR). However, these methods meet a drawback with tensor structure of GST microarray.

Naturally, GST microarray itself is presented as a third-order tensor, thus two-dimensional microarrays of matrices-based method is not suitable to analysis. Additionally, its values could be uncompleted during data recording. Thus, to complete and analysis this kind of microarray data for classification, the tensor-based approaches recently are more suitable over the methods based of unfolding the tensor into the 2D/1D space, known as matrix and vector, respectively. The limitation of matrices/vector-based algorithm is hard to capture and discover the hidden information for such dataset. This is because that unfolding step will lead to break the natural structure of original dataset, and relationship among the modes does not fully employ [9]. Recent researches have been applied tensor-based approaches for microarray, including the methods were derived from the PARAFAC and Tucker. In particular, Y. Li and Ngom presented a method of non-negative tensor factorization (NTF) in order to extract features that maintain non-negative and independent characteristics for the supervised learning of a original GST-microarray sample dataset [2], [6]. Additionally, the authors also reported the techniques of HOSVD and higher-order orthogonal iterations (HOOI) to decompose tensor structure of microarray allowing to extract prominent set of features for GST recognition [7]. Another tensor-based method was proposed by Omberg et al. [5], known as HOSVD, applied for analyzing the integration of DNA microarray data for different studies. Du et al. presented a generalization of ICA algorithm, named as Multilinear Independent Component Analysis (MICA) for high-order gene-expression profiles in order to discover the interactions of the multiple samples for tumor classification [10]. However, these methods do not consider the categories information among the classes. Furthermore, all of these methods considered the kind of this dataset were completed, not maintaining missing values.

In this paper, in order to deal with two aforementioned challenges of microarray analysis while preserving tensor structure of dataset, High-Order Kalman Filter is first applied missing imputation contribution. Then, the second thrust is contribution in the area of tensor decomposition and factorization by present an improvement of Tucker decomposition to extract discriminant subspaces from tensor-based input of microarray sequences. Different from the original tensor models of Tucker whereby do not employ discriminative-information; the proposed method discover categories information in the tensor-based inputs of microarray datasets so as to seek the compact discriminative

subspace for classification phase. To verify the effectiveness of the proposed approaches, the comparison is also carried out on the two recent tensor methods, named as linear Laplacian discriminant analysis (TLLDA) [11] and the local tensor discriminant analysis (LTDA) [12] those also explore the categories information retrieval as the prior step for supervised setting scenarios.

III. PRELIMINARIES and PROPOSED METHODS

A. Tensor Algebra: Mathematical Fundamentals

Mathematically, tensors is a generalized of vector and matrices, known as "multiway arrays" or "higher-order tensor" [13]. An *Nth-order* tensor (*N*-way) is denoted by boldface underline letters, denoted as $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, matrix is denoted bold capital letter **A.**

Vectorization: Given *N*-order tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, the vectorization of tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, denoted by $\text{vec}\ \underline{\mathbf{X}} \in \mathbb{R}^{I_1 I_2 \ldots I_N}$ is obtained by iterating elements of tensor $\underline{\mathbf{X}}$ into a vector. Specifically, the *k*th element of $\text{vec}\ \underline{\mathbf{X}}$ is given by [9], [16]:

$$\text{vec}\ \underline{\mathbf{X}}_{\ k} = \underline{\mathbf{X}}_{i_1 i_2 \ldots i_N} \ \text{where}\ k = 1 + \sum_{m=1}^{M} \prod_{n=1}^{m-1} I_n \ \ i_m - 1 \tag{1}$$

Matricization: Given a tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N \times J_1 \times J_2 \times \cdots \times J_N}$, which can be written in a shorthand $\underline{\mathbf{X}} \in \mathbb{R}^{I\,J}$ the matricization of tensor $\underline{\mathbf{X}}$, denoted as $\text{mat}\ \underline{\mathbf{X}} \in \mathbb{R}^{I_1 \cdots I_N \times J_1 \cdots J_N}$ is obtained by flatting the tensor into a matrix. The element of matricization $\text{mat}\ \underline{\mathbf{X}}$ is given as following:

$$\text{mat}\ \underline{\mathbf{X}}_{\ kl} = \underline{\mathbf{X}}_{i_1 \cdots i_N j_1 \cdots j_N} \ \text{where} \begin{cases} k = 1 + \sum_{n=1}^{N} \prod_{m=1}^{n-1} I_m \ \ i_n - 1 \\ l = 1 + \sum_{n=1}^{N} \prod_{m=1}^{n-1} J_m \ \ j_n - 1 \end{cases} \tag{2}$$

Mode-*n* tensor-matrix product: The mode-*n* matrix product of the tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, with the matrix $\mathbf{A} \in \mathbb{R}^{J_n \times I_n}$ that is denoted by $\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_n \mathbf{A}$ or $\underline{\mathbf{Y}}_{\ n} = \mathbf{A}\underline{\mathbf{X}}_{\ n}$ is the tensor $\underline{\mathbf{Y}} \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N}$. Element-wise, $\underline{\mathbf{Y}} \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N}$ comprises the following [15]:

$$\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_n \mathbf{A} = (\underline{\mathbf{X}} \times_n \mathbf{A})_{i_1 \ldots i_{n-1} j_n i_{n+1} \ldots i_N} = \sum_{i_n = 1}^{I_N} x_{i_1 i_2 I i_{n-1} i_n i_{n+1} \ldots i_N} a_{j_n i_n} \tag{3}$$

Briefly, multiple mode-*n* matrix products in can be formulated as the following equations [4]:

$$\underline{\mathbf{X}} \times_n \mathbf{A} \ \times_m \mathbf{B} = \ \underline{\mathbf{X}} \times_m \mathbf{B} \ \times_n \mathbf{A} \qquad m \neq n \tag{4}$$

On the whole, for all possible modes (n = 1, 2,…, N) of a tensor $\underline{\mathbf{X}}$ and a set of matrices $\mathbf{A}^{\,n}$, we can present the multiplication notation as following:

$$\underline{\mathbf{X}} \times \ \mathbf{A} \ = \underline{\mathbf{X}} \times_1 \mathbf{A}^{\,1} \ \times_2 \mathbf{A}^{\,2} \ \cdots \times_N \mathbf{A}^{\,N} \tag{5}$$

The multiplication of a tensor, $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ with all except from mode-n is given:

$$\underline{\mathbf{X}} \times_{-n} \ \mathbf{A} \ = \underline{\mathbf{X}} \times_1 \mathbf{A}^{\,1} \ \times_2 \mathbf{A}^{\,2} \ \cdots \times_{n-1} \mathbf{A}^{\,n-1} \ \times_{n+1} \mathbf{A}^{\,n+1} \ \cdots \times_N \mathbf{A}^{\,N} \tag{6}$$

Tensor factorization: Given a tensor $\underline{\mathbf{U}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N \times J_1 \times J_2 \times \cdots \times J_N}$, the factorization of $\underline{\mathbf{U}}$ is to decompose it into N factor matrices $\underline{\mathbf{U}}^{\,n} \in \mathbb{R}^{I_n \times J_n}{}_{n=1}^{\ N}$, so that $\underline{\mathbf{U}}_{i_1, i_2 \cdots i_N, j_1, j_2, \cdots, j_N} = \prod_{n=1}^{N} \mathbf{U}^{\,n}_{i_n j_n}$. It can be written by the Kronecker product of N matrices as: $\text{mat}(\underline{\mathbf{U}}) = \mathbf{U}^{\,M} \otimes \mathbf{U}^{\,M-1} \otimes \cdots \otimes \mathbf{U}^{\,1}$ [16].

The contracted product of tensor $\underline{\mathbf{U}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N \times J_1 \times J_2 \times \cdots \times J_N}$ and $\underline{\mathbf{Z}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is given by $\underline{\mathbf{X}} = \underline{\mathbf{U}} \circledast \underline{\mathbf{Z}}$, then vec $\underline{\mathbf{X}}$ = mat $\underline{\mathbf{U}}$ $\circledast$ vec $\underline{\mathbf{Z}}$. If $\underline{\mathbf{U}}$ is factorizable with matrices $\underline{\mathbf{U}}^{n} \in \mathbb{R}^{I_n \times J_n}{}_{n=1}^{N}$ then:

$$\text{vec } \underline{\mathbf{X}} = \text{vec } \underline{\mathbf{U}} \circledast \underline{\mathbf{Z}} = \left[\mathbf{U}^{M} \otimes \mathbf{U}^{M-1} \otimes \cdots \otimes \mathbf{U}^{1} \right] \text{vec } \underline{\mathbf{Z}} \tag{7}$$

Where $\text{mat}(\underline{\mathbf{U}}) = \mathbf{U}^{M} \otimes \mathbf{U}^{M-1} \otimes \cdots \otimes \mathbf{U}^{1}$

B. Proposal Methods Setup for Microarray Recognition

1. Missing value imputation

2. Given a high-order time sequences of microarray with missing values, at this step HOKF [9] is applied to recovery missing entries to obtain completed dataset for supervised learning. In particular, the given incomplete dataset is first filled by linear interpolation for missing entries only along with the indication tensor sequences that has the same size with the input dataset. The indication tensor contains value of 1 if this observation is not missing and 0 otherwise. Then, a probabilities modeling is setup to estimate the conditional expectation. The nature of the method for missing recovery is the iterative with three steps: latent factor estimation, parameters maximization, missing values estimation, and iterating until convergence.

3. Feature extraction for supervised recognition

After missing imputation step, we obtain complete dataset for feature extraction and recognition purpose. Three steps for this phase are illustrated in generality of *N*-order tensor dataset, as following:

Step 1: Tensor-based Structured Representation: All the given *K* sample set for training consists of an *N*-order tensor that is composed as formed $\underline{\mathbf{X}}^{k} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ for *k = 1, 2, 3, ..., K.* Herein, each training sample $\underline{\mathbf{X}}^{k}$ is assigned to the class with the label c_k. The formed of the sample training set then can be represented as the mode *(N+1)*-th-order sample tensor $\underline{\mathbf{X}}^{train} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N \times K}$. The testing dataset is organized as a same way with structure $\underline{\mathbf{X}}^{\text{test}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N \times K'}$

Step 2: *N*-basic factor matrices Estimation: The training dataset is decomposed into *N*- basic factor matrices via applying the improvement orthogonal Tucker decomposition employing categories information [], as given:

$$\underline{\mathbf{X}}^{\text{train}} = \underline{\mathbf{G}} \times_1 \mathbf{A}^{1} \times_2 \mathbf{A}^{2} \cdots \times_N \mathbf{A}^{N}$$

The core tensor $\underline{\mathbf{G}}^{k} \in \mathbb{R}^{J_1 \times J_2 \times \cdots \times J_N}$ representing the training features. This tensor contain the reduced features of $\underline{\mathbf{X}}^{k}$ that are much lower dimension than the original data tensor training $\underline{\mathbf{X}}^{k}$. Additionally, the basic factors components $\mathbf{A}^{1} \in \mathbb{R}^{I_1 \times J_1}$, $\mathbf{A}^{2} \in \mathbb{R}^{I_2 \times J_2}$, …, and $\mathbf{A}^{N} \in \mathbb{R}^{I_N \times J_N}$ utilized as bases of the features space spanned by factor $\mathbf{A}^{n}$.

Our target is to estimate the set of factor matrices $\mathbf{A}^{n}$ by employing the categories information so that the core tensor $\underline{\mathbf{G}}^{k}$ maximizes the differences between the two classes. On the other words, the proposed method tries to find the optimization of the basic matrices so as to ensure that the reduced core tensor $\underline{\mathbf{G}}^{k}$ of the same classes is as similar as possible. In general, we can fine the maximization of the cost function in equation (8) is to determine the orthogonal basic factor $\mathbf{A}^{n}$,*n = 1,2,3,....,N* of the feature subspace of the training set $\underline{\mathbf{X}}^{k}$:

$$\varphi = \arg \max_{\mathbf{A}^{1},\cdots,\mathbf{A}^{N}} \frac{\sum_{c=1}^{C} K_c \left\| \overline{\underline{\mathbf{G}}}^{c} - \overline{\overline{\underline{\mathbf{G}}}} \right\|_F^2}{\sum_{k=1}^{K} \left\| \underline{\mathbf{G}}^{k} - \overline{\underline{\mathbf{G}}}^{c_k} \right\|_F^2} \tag{8}$$

where, $\overline{\underline{\mathbf{G}}}^{c}$ indicates the mean tensor of the *c*-th class that contains the *Kc* training samples, whereby c = 1,…, C denotes the class to which each training sample $\underline{\mathbf{X}}^{k}$ belongs. The class category to which the k-th training-tensor sample $\underline{\mathbf{X}}^{k}$ belongs is denoted by c_k. $\overline{\overline{\underline{\mathbf{G}}}}$ is the mean tensor of all training features. Via dealing with the optimization of equation (8) will lead to the learning rule to estimate discriminant bases [9].

Step 3: Feature Extraction: In order to seek the extracted features, the testing input is projected onto the feature subspaces that are simultaneously spanned by every basic factor that has been discovered in the previous step. The formulation is given:

$$\underline{\mathbf{G}}^{\text{test}} = \underline{\mathbf{X}}^{\text{test}} \times_1 \mathbf{A}^{1} \times_2 \mathbf{A}^{2} \cdots \times_N \mathbf{A}^{N}$$

Step 4: Microarray Recognition: The core training features and testing features is compared to decide which class sample is assigned via classifier of KNN model.

IV. EXPERIMENTAL RESULTS

The proposed method is evaluated for corroborating its effectiveness on a real Interferon beta (INFβ) recognition which taken from a third-order GST microarray dataset. The source of dataset is distributed online and also is available on the material of [12], [19]. In this dataset, there are two groups, namely patients who responded favorably to treatments and those whose response is problematic. The original dataset contains the expression measurements for 76 genes at 7 time points (0, 3, 6, 9, 12, 18 and 24 months) for each patient, with 31 patients responding well and the remaining 22 responding bad to the treatment. This dataset contains genes with missing expression measurements at some time points.

In this paper, instead of ignore the samples containing missing values; we applied HOKF to impute these missing values first. Then, the proposed method of feature extraction is applied directly on the tensor-based structure of microarray to find out the discriminative subspaces of features that is known as core tensors. In our experiment, the form training set of microarray training data comprising as 3-D: $\underline{\mathbf{X}} \approx \underline{\mathbf{G}} \times_1 \mathbf{A}^{1} \times_2 \mathbf{A}^{2}$ with $\mathbf{A}^{1}$ denoted for gene and $\mathbf{A}^{2}$ implied for time points, respectively. As such, the training data will be decomposed into two factor matrices $\mathbf{A}^{1} \in \mathbb{R}^{76 \times J_1}$, and $\mathbf{A}^{2} \in \mathbb{R}^{7 \times J_2}$ that are the projected filters for the gene and time point components, respectively. The reduced core tensor $\underline{\mathbf{X}} \in \mathbb{R}^{J_1 \times J_2}$ with $J_1 \leq 76$ and $J_2 \leq 7$ denotes for the reduced features of the original training set of $\underline{\mathbf{X}}$ that has smaller dimension than the original data tensor $\underline{\mathbf{X}}$. Therefore, in order to reduce the dimension, a projection is performed by mapping the original tensor samples X onto the compressed core tensor $\underline{\mathbf{G}}$ with the proper dimensions of J_1 and J_2. As the results, the current training feature of tensor $\underline{\mathbf{G}} \in \mathbb{R}^{J_1 \times J_2}$ will be vectorized into the feature vector with a length of $J_1 \times J_2$. The feature extraction is performed by projecting the testing data onto the subspace features obtaining from the training phase with the basic factors $\mathbf{A}^{1}$, and $\mathbf{A}^{2}$. Our target is to discover the discriminative features to classify the patient who belongs to favorable INFβ responders or problematic one. In this paper, to decide the number of components for each basic factor, we use the energy of 98% of the original dataset. Our proposed method discover successful discriminative features of the raw dataset and then could transform a high-dimensional sample tensor-based structure data into a new subspace features with low-dimensionality.

To verify the effectiveness of our proposed method, we compare our proposed approaches with the traditional matrices-based method, known as PCA. Furthermore, two recent tensor-based method of discriminant analysis, namely Laplacian discrimination analysis (TLLDA) and local tensor discriminant analysis (LTDA), are also implemented for our comparison. The PCA algorithm performed the limitation on capturing the covariance among modes in tensor structure of data since it requires the original dataset unfold to matrices formulation. As the result, it shows of only 75.37% classification accuracy. While, the two recent tensor methods show higher performances than the matrix-based method with achievement of accuracy accuracies of 89.14% and 78.61% with respect to TLLDA and LTDA, respectively. However, our proposed method deal with the missing observations first and then with the extracted discriminative features, the recognition gets the highest accuracy of 90.23%. While the original Tucker methodology performs 84.17% accuracy. The advantage of our proposed method not only presents on recognition accuracy but also on processing speed issue with its processing time of 0.0816s. Its computation time is faster than the TLLDA and LTDA method with given values of 0.0889s and 0.1495s, respectively.

V. CONCLUSION

In this paper, we present two proposed methods that deal with two current problems in microarray tensor-based inputs, named as HOKF for missing values imputation and the improved TUCKER-decomposition method based

on higher-order discriminant analysis for feature learning. In particular, the HOKF helps to recover missing values, then the discriminant Tucker decomposition employs on an exploitation of categorical information to discover the power of the discriminative features for microarray recognition while the tensor structure is not broken. The effectiveness of the proposed method can be classified in term of: (i) effectiveness of classification accuracy, missing values imputation; (ii) interpretability of successfully identify of the underlying discriminative characteristics among classes; and (iii) computation time by reduce the processing time. With the promising missing imputation and discriminative features extracted, our future work will be extended for unsupervised learning and compression tasks in microarray mining.

ACKNOWLEDGEMENTS

This research is funded by Funds for Science and Technology Development of the University of Danang under project number B2017-ĐN03-07.

REFERENCES

[1] D. Jiang, J. Pei, M. Ramanathan, C. Lin, C. Tang, A. Zhang, "Mining gene–sample–time microarray data: a coherent gene cluster discovery approach", Knowledge and Information Systems, Vol. 13, pp. 305-335, 2006.

[2] Y. Li, A. Ngom, "Classification of Clinical Gene-Sample-Time Microarray Expression Data via Tensor Decomposition Methods", Springer Berlin Heidelberg, Berlin, pp. 275-286, 2011.

[3] M. G. Du, S. W. Zhang, H. Wang, "Tumor Classification Using High-Order Gene Expression Profiles Based on Multilinear ICA", Advances in Bioinformatics, pp. 926- 945, 2009.

[4] N. A. T. Nguyen, H. J. Yang and S. H. Kim, "Hidden Discriminative feature Extraction for Supervised High-Order Time Series Modeling", Computer in Biology and Medicine, vol. 78, pp. 81-90, 2016.

[5] Y. Li, A. Ngom, "Non-negative matrix and tensor factorization based classification of clinical microarray gene expression data", IEEE Bioinformatics and Biomedicine (BIBM), pp. 438-443, 2010.

[6] Y. E. K. U. Şimşekli, A. Ozgur, A. T. Cemgil, "Probabilistic Latent Tensor Factorization for 3-way Microarray Data Analysis with Missing Values", Machine Learning in Computational Biology Workshop (MLCB) in Neural Information Processing Systems Conference, 2012.

[7] L. Omberg, G. H. Golub, O. Alter, "A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies", Proceedings of the National Academy of Sciences, 104, pp. 18371-18376, 2007.

[8] Y. Li, A. Ngom, L. Rueda, "Missing value imputation methods for gene sample time microarray data analysis", IEEE, In: Proc. IEEE Symposium on Computational Intelligence in Bioin- formatics and Computational Biology, IEEE Press, New York, pp. 183–189, 2010.

[9] N. A. T. Nguyen, H. J. Yang, S. H. Kim, "HOKF: High Order Kalman Filter for Epilepsy Forecasting Modeling", BioSystems, 2017.

[10] M. G. Du, S. W. Zhang, H. Wang, "Tumor Classification Using High-Order Gene Expression Profiles Based on Multilinear ICA, Advances in Bioinformatics", 2009.

[11] Z. L. W. Zhang, and T. Xiaoou, "Tensor linear Laplacian discrimination (TLLD) for feature extraction", Pattern Recognition, 42, pp. 1941-1948, 2009.

[12] F. Nie, S. Xiang, Y. Song, C. Zhang, "Extracting the optimal dimensionality for local tensor discriminant analysis", Pattern Recognition, 42, pp. 105-114, 2009.

[13] T. G. Kolda, B. W. Bader, Tensor Decompositions and Applications, SIAM Review, 51, pp. 455-500, 2009.

[14] Y. Matsubara, Y. Sakurai, C. Faloutsos, T. Iwata, M. Yoshikawa, Fast mining and forecasting of complex time-stamped events, the 18th ACM SIGKDD international conference on Knowl. Discov. Data mining, ACM, China, pp. 271-279, 2012.

[15] A. H. Phan, A. Cichocki, "Tensor decompositions for feature extraction and classification of high dimensional datasets", Nonlinear Theory and Its Applications, IEICE, 37-68, 2010.

[16] Y. Cai, H. Tong, W. Fan, P. Ji, Q. He, "Facets: Fast Comprehensive Mining of Coevolving High-order Time Series", 1st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2015.

[17] A. Jukić, M. Filipović, "Supervised feature extraction for tensor objects based on maximization of mutual information, Pattern Recognition Letters", 34, pp. 1476-1484, 2013.

[18] H. Wang, S. Yan, D. Xu, X. Tang, T. Huang, "Trace Ratio vs. Ratio Trace for Dimensionality Reduction, Computer Vision and Pattern Recognition", CVPR '07. IEEE Conference on, 2007, pp. 1-8, 2007.

[19] S. E. Villoslada, M. M. Wyatt, M. Comabella, L. D. Greller, R. Somogyi, X. Montalban, and J. R. Oksenberg, "Transcription-based prediction of response to INF-beta using supervised computational methods", PLOS Biology, vol. 3, no. 1, pp. 166-176, 2005.

NHẬN DẠNG BIỂU HIỆN MICROARRAY VỚI CẤU TRÚC ĐA TUYẾN TÍNH DỰA TRÊN ĐẠI SỐ TENSOR

Nguyễn Thị Ngọc Anh, Nguyễn Trần Quốc Vinh, Hồ Phan Hiếu, Võ Trung Hùng

TÓM TẮT: *Trong bài báo này, hai thách thức hiện nay về biểu hiện của microarray được đầu tư, đó là sự phục hồi các giá trị bị mất và sự trích xuất các thuộc tính cho mô hình học có giám sát. Trục đầu tiên của bài báo tập trung vào việc làm thế nào để xử lý các thuộc tính chính của chuỗi thời gian microarray: bao gồm cấu trúc tensor, xử lý nhiễu, và các đặc tính động về thời gian. Điều này cho phép phát hiện ra các yếu tố tiềm ẩn và các chiều hướng liên quan nhau để phục vụ cho việc phục hồi dữ liệu. Sau đó, một sự cải tiến của sự phân hủy trực giao Tucker dựa trên việc phân tích phân biệt đối với cấu trúc đa tuyến tính của biểu hiện microarray được trình bày trong việc giảm số chiều. Do đó, một khám phá cấu trúc bậc ba của biểu hiện gen, được cấu thành với thuật ngữ gen - sample - time (GST), được trình bày để phân loại mẫu sinh học. Đóng góp của bài báo được phân bổ theo hai động lực chính; bao gồm việc thiết lập mô hình tiềm ẩn để phục hồi các giá trị bị mất dựa trên thuật toán bậc cao của Kalman Filter và việc trích xuất thuộc dựa trên thuật toán Tensor Discriminative Feature Extraction. Các phương pháp đề xuất này được hiện thực trên dữ liệu của Interferon beta (INFβ) thuộc biểu hiện của GST-microarray trong việc phân loại hiệu ứng điều trị với việc phân biệt bệnh nhân trong nhóm phản ứng tích cực và những bệnh nhân còn lại trong nhóm thuộc vấn đề điều trị. Hiệu năng thực nghiệm trên dữ liệu thực tế đã chứng minh được ưu điểm vượt trội của các phương pháp đề xuất so với với các thuật toán dựa trên ma trận và phương pháp dựa trên việc phân tách phân biệt gần đây trên các khía cạnh phục hồi dữ liệu, độ chính xác trong việc phân loại là 90,23% và thời gian tính toán.*