Probabilistic Latent Tensor Factorization for 3-way Microarray Data Analysis with Missing Values

Umut Şimşekli Yunus Emre Kara Arzucan Özgür Ali Taylan Cemgil
Department of Computer Engineering
Boğaziçi University, İstanbul, Turkey
{umut.simsekli, yunus.kara, arzucan.ozgur, taylan.cemgil}@boun.edu.tr

Abstract

The recent advances in microarray technology enabled the measurement of gene expression levels of samples over a series of time points. Unlike the traditional 2D microarray data, such experiments generate 3D (gene-sample-time) microarray data, which require specialized methods for analysis. In this study, we propose a novel tensor factorization model for modeling 3D microarray data. The model assumes the existence of certain temporal patterns that are repeated over time. One main advantage of the model is that it handles the missing data implicitly, so that the estimation process is not effected by the existence of missing values, which commonly occur in microarray data. We evaluate our model on classification of the good or bad responders to Interferon beta (INFβ) treatments by using a real gene-sample-time microarray data set and achieve a promising prediction performance.

1 Introduction

With the recent advances in microarray technologies, thousands of genes can be monitored by measuring their relative transcription levels. When combined with the state-of-the-art statistical methods, a lot of useful information such as functions of genes [1] or diagnosis of tumor [2] can be inferred from these data.

Traditionally, there are two types of microarray data: *gene-sample* and *gene-time*. In gene-sample data the expression levels of a set of genes are measured over a set of samples, whereas they are measured with respect to a series of time points in gene-time data. Both types of data can be represented with two dimensional matrices, where the set of genes correspond to the first dimension and the set of samples or time points correspond to the second dimension. Matrix and tensor factorization methods have become useful for analyzing two dimensional microarray data. As a pioneering study, Brunet et al. [3] presented a clustering algorithm based on Non-negative Matrix Factorization (NMF). Schachtner et al. [4] compared the performances of Independent Component Analysis and sparse-NMF on the classification of three gene expression data sets. In [2], the authors presented similar methods for combining NMF with Support Vector Machines (SVM) for tumor and disease classification.

The recent advances in microarray technology made it possible to simultaneously measure the expression levels of different genes over a set of samples and a set of time points, yielding a three dimensional *gene-sample-time* (GST) data set. Matrix and tensor factorization methods for analyzing GST data have not been explored much in the literature yet. Recently, Li and Ngom presented a method for GST microarray data processing that is based on higher-order NMF and reported competitive results to previous works [5]. The authors conducted their experiments on the data set that was presented in [6]. One limitation of this work is that it cannot handle missing data, therefore the authors discarded the genes that contain missing expression measurements and the corresponding

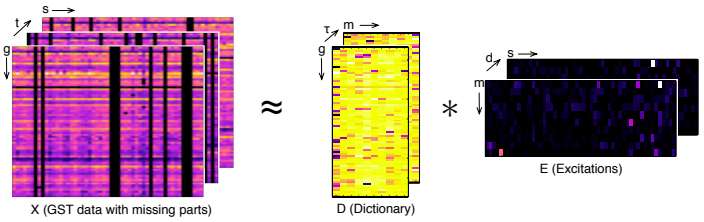

Figure 1: Illustration of the model. The blocks visualize the tensors, '$*$' denotes the convolution operator, and the lower-case letters and arrows represent the indices of a particular tensor.

samples from the data set. Microarray data often contain missing values due to several reasons such as low resolution and incomplete experiments. However, most microarray data analysis methods cannot handle missing values. Therefore, data with missing values are preprocessed by either discarding the instances that contain missing values or estimating the missing values using imputation methods. Several missing value imputation methods have been proposed for 2D microarray data [7]. Imputation methods for 3D microarray data have only recently been proposed in [8]. Nevertheless, many studies prefer to discard the data instances with missing values for simplicity, which results in loss of information.

In this study, we propose a novel tensor factorization model for modeling of clinical GST microarray data. Apart from the previous works, the novel model incorporates temporal information into the factorization process and is able to handle the missing data. We evaluate our model on classification of the good or bad responders to Interferon beta (INFβ) treatments by using a real GST data set, where we form a classification method by combining the novel model with a Support Vector Machine (SVM) classifier.

2 The Model

Factorization based microarray data modeling has become popular in the recent years. In this study, we propose a novel non-negative tensor factorization model for GST microarray data. Let us define the tensor $X(g, s, t)$ as the observed GST data where the indices g, s, and t correspond to *genes*, *samples*, and *time points*, respectively. With an assumption that there are certain temporal patterns that are repeated over time, we define the following novel tensor factorization model as follows:

$$X(g, s, t) \approx \sum_{m,\tau} D(g, m, \tau) E(m, s, t - \tau) \tag{1}$$

where, D is a tensor which contains the so called *metagenes* that encapsulate the temporal and structural information in the microarray data and E is another tensor which determines the metagene expression pattern of the corresponding sample. The index m indicates the metagene index and the index τ indicates the temporal index of a particular metagene. The overall process is explained as the convolution of the metagenes (D) and the excitations (E), where the convolution process ensures the assumption that there are repeated temporal patterns. Figure 1 illustrates this model.

Given observed GST data X, the aim is to estimate the tensors D and E. The estimated tensors can be used for various applications including clustering, classification, and missing value imputation. In the next section, we describe how the estimation process is carried out.

2.1 Probabilistic Latent Tensor Factorization

Probabilistic Latent Tensor Factorization (PLTF) is a recently proposed algorithmic framework for modeling multiway data, where any arbitrary tensor factorization model can be realized by this framework [9]. This framework makes use of the practical aspects of graphical modeling of machine learning and combines them with tensor factorization models. Once a factorization model is defined in the PLTF framework notation, the estimation algorithm is immediately available. The PLTF framework is defined as

$$X(v_0) \approx \hat{X}(v_0) = \sum_{\vec{v_0}} \prod_{\alpha} Z_{\alpha}(v_{\alpha}), \tag{2}$$

where X is the observed tensor and Z_α are the individual factors. In this framework, each tensor is represented by an index set: we define V as the set of all indices in a model, V_0 as the set of observed indices, V_α as the set of indices in Z_α, and $\bar{V}_\alpha = V - V_\alpha$ as the set of all indices not in Z_α. We use small letters as v_α to refer to a particular setting of indices in V_α.

PLTF framework aims to find the optimal factors $Z_{1:N}$ such that $Z_{1:N} = \arg\min_{Z_{1:N}} d(X, \hat{X})$ where we select $d(\cdot)$ as the β-*divergence* that generalizes Euclidean, Kullback-Leibler, and Itakura-Saito divergences and enables a unified treatment. It is shown that the estimation of the latent factors Z_α can be achieved via iterative optimization [9]. One can obtain the following compact fixed point equation where each Z_α is updated in an alternating fashion fixing the other factors $Z_{\alpha'}$ for $\alpha' \neq \alpha$:

$$Z_\alpha \leftarrow Z_\alpha \circ \frac{\Delta_\alpha(M \circ X \circ \hat{X}^{\beta-2})}{\Delta_\alpha\left(M \circ \hat{X}^{\beta-1}\right)} \qquad \Delta_\alpha(A)(v_\alpha) \equiv \sum_{\bar{v}_\alpha} \left(A(v_0) \prod_{\alpha' \neq \alpha} Z_{\alpha'}(v_{\alpha'}) \right), \quad (3)$$

where $\circ$ is the element-wise product and β determines the cost function to be used: $\beta = \{0, 1, 2\}$ correspond to Itakura-Saito, Kullback-Leibler, and Euclidean divergences, respectively. Besides, M is a tensor of size X where $M(v_0) = 1$ (0) if $X(v_0)$ is observed (missing). By adjusting the tensor M, missing values in a data set can easily be handled. In this iterative method, the key quantity is the $\Delta_\alpha(\cdot)$ function and we need to compute this function twice for arguments $A = M \circ X \circ \hat{X}^{\beta-2}$ and $A = M \circ \hat{X}^{\beta-1}$ while updating Z_α.

In order to take benefit from the PLTF framework, we should define our model in the framework. However, due to the convolution operator, the tensor E has the index $(t - \tau)$, which violates the PLTF notation. Therefore we define a dummy index and update the model as follows:

$$\hat{X}(g, s, t) = \sum_{m,\tau,d} D(g, m, \tau) E(s, m, \overbrace{d}^{t-\tau}) \overbrace{Z(d, t, \tau)}^{\delta(d-t+\tau)} \quad (4)$$

where $\delta(\cdot)$ is the dirac-delta function where $\delta(x)$ equals to 1 (0) if x equals to 0 (otherwise). The ultimate model can now be expressed in the PLTF notation. After defining the index sets for each tensor, we can obtain the following update equations:

$$\begin{aligned} D &\leftarrow D \circ \frac{\Delta_D(M \circ X \circ \hat{X}^{\beta-2})}{\Delta_D\left(M \circ \hat{X}^{\beta-1}\right)} & \Delta_D(A)(g, m, \tau) &\equiv \textstyle\sum_{s,t} A(g, s, t) E(s, m, t - \tau) \\ E &\leftarrow E \circ \frac{\Delta_E(M \circ X \circ \hat{X}^{\beta-2})}{\Delta_E\left(M \circ \hat{X}^{\beta-1}\right)} & \Delta_E(A)(s, m, d) &\equiv \textstyle\sum_{g,t} A(g, s, t) D(g, m, t - d) \end{aligned} \quad (5)$$

By iteratively applying these equations, the factors D and E can be estimated.

3 Experimental Results

In order to evaluate the performance of our model, we have conducted several experiments. We combine our method with an SVM classifier and apply the resulting method to the prediction of the responder types to Interferon beta (INFβ) treatments. We use the data set of [6] that consists of the GST microarray data of the patients afflicted with multiple-sclerosis (MS). The patients are classified into two groups: the ones that respond to treatments *good* and the ones that respond *bad*. There are 53 patients in the data set (33 of them respond good and 20 of them respond bad) and 76 genes that are sampled at 7 different time points. Therefore, we have the observed tensor X of size $76 \times 53 \times 7$, where some of the genes and time points are missing for some patients.

The experiments are held on a 5-fold cross validation basis. During training, we estimate the tensors D_{trn} and E_{trn} and save D_{trn} for further use. We train the SVM by using the excitation tensor E_{trn} as features. For the test case, we use the previously learned tensor D_{trn}, and estimate the excitation tensor for the test sample E_{val}. Finally, we use E_{val} as the feature set of the test sample and estimate the sample's label by feeding it to the SVM. Besides, we use an RBF kernel in the SVM and investigate different parameter settings.

In our second experiment, we use the same settings as in the first experiment, except in this experiment, we impute the missing parts in the gene-sample-time data by using the imputation method of [8]. For the imputation method, we use the parameter setting that is reported as the best in the related study. Table 1 shows the average results for Accuracy, Sensitivity, Specificity, and F-Measure values that are obtained through the experiments.

Table 1: Average results that are obtained through the experiments. Here K is the number of metagenes (indexed with m) and T is the number of time slices (indexed with τ).

Exp.	K	T	β	SVM params	Accuracy	Sensitivity	Specificity	F-Measure
1	16	4	2	$C = 100, \gamma = 10^{-5}$	0.87	0.75	0.88	0.81
2	17	3	1	$C = 100, \gamma = 10^{-5}$	0.93	0.95	0.92	0.91

By using the raw GST data with missing values, we obtain promising results (81% F-measure and 87% accuracy). After imputing the missing parts in the data set, the performance is further increased, where all the metrics are over 90%. Li and Ngom used the same data set in their recent study and reported 82% accuracy [5]. However, as discussed in the Introduction section, the instances with missing values were discarded in their experiments due to the limitations of the proposed higher-order NMF method.

4 Conclusion

In this study, we presented a novel method for modeling 3D microarray data where the main assumption in the model is the existence of certain temporal patterns that are repeated over time. In order to make inference on the model, we made use of the Probabilistic Latent Tensor Factorization framework, which helped us to obtain the update equations for the optimization algorithm and to handle the missing data.

In order to evaluate the performance of our model, we combined it with a classifier and evaluated the resulting method on classification of the good or bad responders to Interferon beta treatments by using a real GST data set. We achieved promising results by using the raw data set containing missing values. We showed that the performance of the method can be further improved by incorporating a missing value imputation approach.

Acknowledgement: The authors would like to thank Yifeng Li for providing the software for missing data imputation.

References

[1] T. Puelma, R. A. Gutiérrez, and A. Soto, "Discriminative local subspaces in gene expression data for effective gene function prediction," *Bioinformatics*, vol. 28, pp. 2256–2264, 2012.

[2] P. Zhang, C.-H. Zheng, B. Li, and C.-G. Wen, "Tumor classification using non-negative matrix factorization," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Contemporary Intelligent Computing Techniques*. Springer Berlin Heidelberg, 2008.

[3] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *PNAS*, vol. 101, pp. 4164–4169, 2004.

[4] R. Schachtner, D. Lutter, A. Tome, E. Lang, and P. Vilda, "Exploring matrix factorization techniques for classification of gene expression profiles," in *WISP*, 2007.

[5] Y. Li and A. Ngom, "A new kernel non-negative matrix factorization and its application in microarray data analysis," in *CIBCB*, 2012, pp. 371–378.

[6] S. E. Baranzini, P. Mousavi, J. Rio, S. J. Caillier, A. Stillman, P. Villoslada, M. M. Wyatt, M. Comabella, L. D. Greller, R. Somogyi, and et al., "Transcription-based prediction of response to ifnbeta using supervised computational methods," *PLoS Biology*, vol. 3, p. e2, 2005.

[7] M. Celton, A. Malpertuy, G. Lelandais, and A. G. de Brevern, "Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments," *BMC Genomics*, vol. 11, no. 15, 2010.

[8] Y. Li, A. Ngom, and L. Rueda, "Missing value imputation methods for gene-sample-time microarray data analysis," in *CIBCB*, 2010.

[9] Y. K. Yilmaz and A. T. Cemgil, "Probabilistic latent tensor factorization," in *LVA/ICA*, 2010.