# Study on tensor calculus and CP-decomposition

Nguyen Vu Linh (1350016)

School of Knowledge Science,
Japan Advanced Institute of Science and Technology

February, 2015 [1]

**Keywords:** Tensor, CP-decomposition, Multi-way array, Temporal link prediction, Spectral clustering.

Tensor have been widely studied in mathematics and physics for along time and increasingly applied in many areas of data mining. There are two ways to think about tensors: tensors are representations of multilinear maps; tensors are elements of a tensor product of two or more vector spaces. For our purpose, "a $N^{th}$-order tensor" is defined as "an element of tensor product of $N$ real vector spaces $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_N$, denoted by $\mathcal{V}_1 \otimes \mathcal{V}_2 \otimes \ldots \otimes \mathcal{V}_N$. When fixing the bases of $\mathcal{V}_1$, $\mathcal{V}_2$ and $\mathcal{V}_N$, a tensor can be represented by a $N$"-way array in the vector space $\mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ and elements in $\mathcal{V}_n$ can be represented as vectors in $\mathbb{R}^{d_n}$, where $d_n$ is the dimension of $\mathcal{V}_n$.

The equivalence between two vector space $\mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2} \otimes \ldots \otimes \mathbb{R}^{d_N}$ and $\mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ allows us to not distinguish these two spaces. Such equivalence provides great advantages for data mining applications because $N$-way array may provide nature and compact representation for numerous complex kinds of data that the integrated result of several inter-related variables or they are combinations of underlying latent components or factors. Furthermore, when considering $N$-way array as element of $\mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$, powerful results related to tensor can be employed to construct tensor based methods to solve challenging problems. Especially, when working with challenging problems for big data related to capture, manage, search, visualize, cluster, classify, assimilate, merge, and process the data within a tolerable elapsed time. When working on tensor data, the large required storage

memory and the inter-relation between variables often make the problems become more complicated. Many tensor-based models known as multiway models have been constructed to deal with those challenges by exploring the meaningful hidden structure and to finding low-rank representation of data. Also, each kind of model have its own advantages and disadvantages which should be carefully considered based on the application context. For example, using CP-decomposition may lead to losing important data structure while Tucker decomposition may be problematic in high-dimensions with many irrelevant features.

One interesting tensor-based method is temporal link prediction method based on CP-decomposition constructed to do temporal link prediction task on bipartite networks whose links evolve over time and node set consists vertices of two types such that only vertices of diffent types can be linked. Such problem is important and has been studied in many researches because prediction is crucial tasks in real applications and bipartite networks can be used to represent various kinds of structures, dynamics, and interaction patterns found in social activities. Temporal link prediction method based on CP-decomposition have shown advantages comparing with others, such as its power in exploring the structure of data, requiring less memory and giving outperformed experimental results. Also, tensor based-methods can predict the links for times $(T+1)^{th}$, ... , $(T+L)^{th}$ while other models are limited to temporal prediction for a single time step. Motivating by these advantages, we extend tensor-based method on bipartite networks to do temporal link prediction problem on specific class of bipartite networks in which new verticies of one type may join networks at concerning time and may link to other vertices in the next time point. The key ideas of the proposed methods are to employ CP-decomposition to decompose weight data into factors of three separated kinds, each fluctuates independently from others and collect additional information of vertices of open type and learn a function to predict values of open type vertex factors from the additional information and use to predict values of those factors corresponding to new vertices.

Clustering plays an outstanding role in numerous data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical

diagnostics, computational biology, and many others. Intuitively, clustering aims to identify groups of "similar behavior" data considered as a first impression on data when dealing with the empirical data. Since tensor data have became popular in data mining, we focus on constructing a versatile clustering for tensor data. Considering clustering methods based on vector space model, spectral clustering have several attractive advantage such that it is versatile, easy to implement, often provide better performance comparing with traditional methods like $K$-mean. Furthermore, spectral clustering is easy to extend for tensor data since it work with only requirement about similarity while other methods often require more additional information. To handling the clustering task on tensor data, we construct a CP-decomposition based spectral clustering by constructing appropriate similarities and employing CP-decomposition to exploring the hidden structure of data and reducing the storage memory. The empirical results provide the evidence to conclude that the proposed models can give the acceptable accuracy and CP-decomposition may help to reduce the storage memory and improve the clustering accuracy by exploring the hidden structure of data.

Concerning temporal link prediction and clustering problems, we discuss about the achievements and provide suggestions to and make plan to complete these objective as future works. For temporal link prediction problem, we plan to implement the method and evaluate using empirical results and extend method to do more general problem when vertices of two type join the concerning networks at the same time. Considering the clustering problem, we plan to construct several similarity measures and extend the available vector space model based multi-view spectral clustering for tensor data. We also give opportunity and suggestions to construct a tensor space model based clustering method which tensor data is transformed in to vector space data and spectral clustering methods are applied on the transformed data in order to cluster the data.