Hybrid Clustering of Multiple Information Sources via HOSVD

Xinhai Liu[1,3], Lieven De Lathauwer[1,2], Frizo Janssens[1], and Bart De Moor[1]

[1] *K.U. Leuven, ESAT-SCD , Kasteelpark Arenberg 10, Leuven 3001, Belgium*
[2] *K.U.Leuven, Group Science, Engineering and Technology, Campus Kortrijk*
[3] *WUST, CISE and ERCMAMT , Wuhan 430081, China*

Abstract

We present a hybrid clustering algorithm of multiple information sources via tensor decomposition, which can be regarded an extension of the spectral clustering based on modularity maximization. This hybrid clustering can be solved by the truncated higher-order singular value decomposition (HOSVD). Experimental results conducted on the synthetic data have demonstrated the effectiveness.

keywords: hybrid clustering, HOSVD, spectral clustering, tensor

1 Introduction

Hybrid clustering of multiple information sources means the clustering of the same class of entities that can be described by different representations from various information sources. The need for clustering multiple information sources is almost ubiquitous and applications abound in all fields, including market research, social network analysis and many scientific disciplines. As an example in social network analysis, with the pervasive availability of Web 2.0, people can interact with each other easily through various social media. For instance, popular sites like Del.icio.us, Flickr, and YouTube allow users to comment shared content and users can tag their favorite content [1]. These diverse individual activities result in a multi-dimensional network among users. An interesting research problem that arises here is how to unify heterogeneous data sources from different point views to facilitate clustering.

Intuitively, multiple information sources can facilitate inferring more accurate latent cluster structures among entities. Nevertheless, due to the heterogeneous property of different information sources, it becomes a challenge to identify clusters in multiple information sources as we have to fuse the information from all data sources for joint analysis.

While most clustering algorithms are conceived for clustering data from a single source, the need to develop general theories or frameworks for clustering multiple heterogeneous information sources that share dependency has become more and more crucial. Unleashing the full power of multiple information sources is, however, a very challenging task, for example, the scheme of different data collections might be very different (data heterogeneity). Although

several approaches for utilizing multiple information sources have been proposed [2, 3, 1], these methods seem ad-hoc.

Increasingly, tensors are becoming common in modern applications dealing with complex heterogeneous data, which provide novel tools for joint analysis on multiple information sources. Tensors have been successfully applied to several domains, such as chemometrics [4], signal processing [5] and Web search [6]. Tensor clustering is a recent generalization to the basic one-dimensional clustering problem, and it seeks to decompose a n-order input tensor into coherent sub-tensors while minimizing some cluster quality measures [7]. Higher-order singular value decomposition (HOSVD) is a convincing generalization of the matrix SVD to tensor decomposition [6]. Meanwhile, multiple information sources can be easily modeled as a tensor and the inner relationship among them can be naturally investigated by tensor decomposition analysis.

In this work, we first review modularity maximization, a recently developed measure for clustering. We discuss its application on single information source and then extend it to multiple information sources. Since multiple matrices factorization is involved in our hybrid clustering of multiple information sources, we formulate our problem within the framework of tensor decomposition and propose a novel algorithm: hybrid clustering based on HOSVD (HC-HOSVD). Our experiments on synthetic data validate the superiority of our proposed approach.

2 Related Work

Some hybrid clustering methods to integrate multiple information sources have emerged: clustering ensemble [3], multi-view clustering [2] and kernel fusion [8]. Recently, Tang et al. [1] propose a method named principle modularity maximization (PMM) to detect the cluster in multi-dimensional networks. They also compare PMM with average modularity maximization (AMM), which combines the multiple information sources averagely and then maximizes the modularity. Although above methods are effective for certain application scenarios, they seem to be restricted that they lack an effective scheme to investigate the inner relationship among diverse information sources.

Tensor decomposition, more especially HOSVD, is a basic data analysis task with growing importance in the application of data mining. J.-T. Sun et al. [9] use HOSVD to analyze web site click-through data. Liu et al. [10] apply HOSVD to create a tensor space model, analogous to the well-known vector space model in text analysis. J. Sun et al. [11] have written two papers on dynamically updating a HOSVD approximation, with applications ranging from text analysis to environmental and network modeling. Based on tensor decomposition, Kolda et al. [6] propose TOPHITS algorithm for Web analysis by incorporating text information and link structure. Huang et al. [12] present a kind of HOSVD based clustering and employ it to image recognition. We call that method data vector clustering based on HOSVD (DVC-HOSVD). Our algorithm has the similar flavor but is formulated by modularity tens instead of data vector.

Collectively, multiple information analysis requires a flexible and scalable framework that exploits the inner correlation among different information sources, while tensor decomposition can fit in with this requirement. To the best of our knowledge, our work is the first unified attempt to address the modularity max-

imization based hybrid clustering via tensor decomposition.

3 Modularity-based Spectral Clustering

3.1 Modularity Optimization on Single Data Source

Modularity is a benefit function used in the analysis of networks or graphs. Its most common use is as a basis for optimization methods for detecting cluster structure in networks [13]. Consider a network composed of N nodes or vertices connected by m links or edges, modularity of this network is defined as follows

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{1}$$

where A_{ij} represents the weight of the edge between i and j, $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to vertex i, c_i is the cluster to which vertex i belongs, the δ function $\delta(u, v)$ is 1 if $u = v$, and 0 otherwise. The value of the modularity lies in the range [-1,1]. It is positive if the number of edges within groups exceeds the number expected on the basis of chance. In general, one aims to find a cluster structure such that Q is maximized. While maximizing the modularity over hard clustering is proved to be NP hard, a relaxing of the problem can be solved efficiently [14]. Let $d \in Z_+^N$ be the degree of each node, $S \in \{0, 1\}^{N \times C}$ (C is the number of clusters in the network) be a cluster indicator matrix defined below

$$S_{ij} = \begin{cases} 1 & \text{if vertex } i \text{ belongs to cluster } j \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

The modularity matrix is defined as

$$B = A - \frac{dd^T}{2m}. \tag{3}$$

Here we define $\mathrm{tr}(\cdot)$ as trace operation, so modularity can be reformulated as

$$Q = \frac{1}{2m} \mathrm{tr}(U^T B U). \tag{4}$$

Relaxing U to be continuous, it can be inferred that the optimal U is composed of the top k eigenvectors of the modularity matrix [13]. Given a modularity matrix B, the objective function of this spectral clustering is

$$\begin{aligned} &\max_S \ \mathrm{tr}(S^T B S), \\ &\text{s.t. } S^T S = I. \end{aligned} \tag{5}$$

3.2 Modularity Optimization on Multiple Data Sources

By matrix decomposition, we can easily obtain U in (5), whereas it is hard to directly get the optimal solution of multiple extension of (5). Therefore, we turn to tensor methods based on Frobenius norm (F-norm) optimization. Preliminarily, we need to formulate the spectral clustering with F-norm optimization.

The Frobenius norm (or the Hilbert-Schmidt norm) of a modularity matrix A can be defined in various ways [15]

$$\|B\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|b_{ij}|^2} = \sqrt{tr(B^*B)} = \sqrt{\sum_{i=1}^{\min\{m,n\}}\sigma_i^2}, \tag{6}$$

where B^* denotes the conjugate transpose of B and δ_i are the singular values of B.

Considering the following F-norm maximization

$$\begin{aligned} &\max_U \|U^TBU\|_F^2, \\ &\text{s.t. } U^TU = I, \end{aligned} \tag{7}$$

if B is positive (semi)definite, the objective functions in (5) and (7) are different but happen to have their optima under the same matrix U, whose columns span the dominant eigenspace of U [15]. Regarding the positive (semi)definite property of modularity matrix, we can regularize the modularity matrix to guarantee that it is positive (semi)definite [14]. Thus the spectral clustering defined in (5) can be alternatively formulated by F-norm optimization in (7).

From various (K types of) information sources, we can generate multi-dimensional modularity matrices $B^{(i)}$ ($i \in 1, 2, ..., K$) accordingly. Then by linear combination, we formulate the multi-dimensional spectral clustering as follows

$$\begin{aligned} &\max_U \sum_{i=1}^{K}\|U^TB^{(i)}U\|_F^2, \\ &\text{s.t. } U^TU = I, \end{aligned} \tag{8}$$

which is also hard to solve directly, so we will represent it by a 3-order tensor method in the next Section.

4 Tensor Decomposition for Hybrid Clustering

This section provides notation and minimal background on tensors and tensor decomposition used in this research. We refer readers to [16, 4] for more comprehensive review on tensors.

Tensor is a mathematical representation of a multi-way array. The order of a tensor is the number of modes (or ways). A first-order tensor is a vector, a second order tensor is a matrix and a tensor of order three or higher is called a higher-order tensor. We only investigate 3-order tensor decomposition that is relevant to our problem.

4.1 Basic Conceptions of Tensor Decomposition [17, 18]

The n-mode matrix unfolding: Matrix unfolding is the process of reordering the elements of a 3-way array into a matrix. The n-mode ($n = 1, 2, 3$) matrix unfoldings of a tensor $\mathcal{A} \in \mathbb{R}^{I\times J\times K}$ are denoted by $A_{(1)}, A_{(3)}$ and $A_{(3)}$ separately. For example, the matrix unfolding $A_{(1)}$ is a matrix with the number of

rows I and the number of its columns is the product of dimensionalities of all other modes, that is, $J \times K$.

The n-mode product: For instance, the 1-mode product of a tensor $\mathcal{A} \in \mathbb{R}^{I\times J\times K}$ by a matrix $H \in \mathbb{R}^{I\times P}$, denoted by $\mathcal{A} \times_1 H$, is a $(P \times J \times K)$-tensor of which the entries are given by

$$(\mathcal{A} \times_1 H)_{pjk} = \sum_p a_{pjk}. \tag{9}$$

The analogous definitions are for 2-mode and 3-mode products.

Higher-order Singular Value Decomposition(HOSVD): HOSVD is a form of higher-order principle component analysis. It decomposes a tensor into a core tensor multiplied by a matrix along each mode. In the three-way case where $\mathcal{A} \in \mathbb{R}^{I\times J\times K}$, we have

$$\mathcal{A} = \mathcal{S} \times_1 U \times_2 V \times_3 W, \tag{10}$$

where $U \in \mathbb{R}^{I\times I}$, $V \in \mathbb{R}^{J\times J}$ and $W \in \mathbb{R}^{K\times K}$ are called factor matrices or factors and can be thought of as the principle components of the original tensor along each mode. The factor matrices U, V and W are assumed column-wise orthonormal. The tensor $S \in \mathbb{R}^{I\times J\times K}$ is called the core tensor and its elements show the level of interaction between different components. According to [17], given a tensor $\mathcal{A}$, its matrix factors U, V and W as are defined in (10) can be calculated as the left singular vectors of its matrix unfoldings $A_{(1)}, A_{(2)}$ and $A_{(3)}$ respectively.

Truncated HOSVD [17, 16]: An approximation of a tensor $\mathcal{A}$ can be obtained by truncating the decomposition, for instance, the matrix factors U, V and W can be obtained by only considering the first left singular vectors of the corresponding matrix unfoldings. This approximate decomposition is named truncated HOSVD.

4.2 Hybrid Clustering via HOSVD(HC-HOSVD)

A tensor $\mathcal{A}$ can be built from several modularity matrices $\{B^{(1)}, B^{(2)}, \cdots, B^{(K)}\}$: the first and the second dimensions I and J of the tensor $\mathcal{A}$ are equal to the dimensions of the modularity matrices $B^{(i)}$ $(i = 1, \ldots, K)$, and its third dimension K equals the number of several information sources (different modularity matrices). According to the definition of F-norm of tensors [17],

$$\sum_{i=1}^{K} \|U^T B^{(i)} U\|_F^2 = \|\mathcal{A} \times_1 U^T \times_2 U^T\|_F^2. \tag{11}$$

So the optimization of (6) can be formulated equivalently as

$$\begin{aligned} \max_U & \|\mathcal{A} \times_1 U^T \times_2 U^T\|_F^2, \\ & \text{s.t. } U^T U = I. \end{aligned} \tag{12}$$

Since the modularity matrices $B^{(i)}$ $(i = 1, \ldots, K)$ are symmetric, the matrix unfoldings $A_{(1)}$ and $A_{(2)}$ are identical. Consequently, the matrices U and V in (10) are the same. In (11), there is no compression along the third mode. In

this case, we may take W equal to any orthogonal matrix, without affecting the cost function. Hence, we might take $W = I$, in (11). As is explained in [17], projection on the dominant higher-order singular vectors usually gives good approximation of the given tensor. Consequently, taking the columns of U equal to the dominate 1-model singular vectors is expected to yield a large value of the objective function in (11). The dominant 1-model singular vectors of U are equal to the dominant left singular vectors of $A_{(1)}$. The truncated HOSVD that is obtained this way, does not maximize (11) in general. However, the result is usually pretty good, the algorithm is simple to implement and quite efficient. Moreover, there exists an upper bound on the approximation error [17]. The pseudo code of this hybrid clustering algorithm based on HOSVD (HC-HOSVD) is presented in as follows.

Algorithm 4.1: HC-HOSVD($B^{(1)}, B^{(2)}, ..., B^{(R)}, C$)

comment: C is the number of clusters

1. *Build a modularity − based tensor* $\mathcal{A}$
2. *Compute* U *from the subspace spanned by the dominant left* $(C-1)$ *singular vectors of* $A_{(1)}$
3. *Normalize the rows of* U *to unit length*
4. *Calculate the cluster idx with* $k-$*means on* U

return (*idx* : *the clustering label*)

5 Experiment on Synthetic data [1]

Generally, real-world data does not provide the ground truth information of cluster membership, so we turn to synthetic data with multiple information sources to conduct some controlled experiments. In this section, we evaluate and compare different strategies applied to multi-dimensional networks. The synthetic data[1] has 3 clusters, with each having 50, 100, 200 members respectively. There are 4 kinds of interactions among these 350 nodes, that is, we have four different information sources. For each dimension, cluster members connect with each other following a random generated within-cluster interaction probability. The interaction probability differs with respect to groups at distinct dimensions. After that, we add some noise to the network by randomly connecting any two nodes with low probability. Normalized Mutual Information(NMI) [3] is adopted to measure the clustering performance.

We cross compare the four types of hybrid clustering on multiple information sources with clusterings on each single information source. The four types of hybrid clustering methods are averagely modularity maximization(AMM) [1], PMM, DVC-HOSVD, HC-HOSVD.

We regenerate 100 different synthetic data sets and report the average performance of each method plus its standard deviation in Table 1. Clearly, hybrid clustering on multiple information source outperforms clustering on single information source with lower variance. Due to the randomness of each run, it is not surprising that single information source method shows large variance. Among the four hybrid clustering strategies, HC-HOSVD obviously outperforms the other three.

[1] The data was offered by Lei Tang in Arizona University [1]

.

	Strategy	Performance
Single Information Source	A1	0.6029 ± 0.1798
	A2	0.6158 ± 0.1727
	A3	0.5939 ± 0.1904
	A4	0.6114 ± 0.2065
Multiple Information Source	AMM	0.8939 ± 0.0945
	PMM	0.8414 ± 0.1201
	DVC-HOSVD	0.8975 ± 0.1109
	HC-HOSVD	0.9264 ± 0.1431

Table 1: Clustering on multiple synthetic networks

6 Conclusion and Further Direction

Our main contributions are two-fold:

Based on tensor decomposition, we proposed a kind of hybrid clustering algorithm named HC-HOSVD to integrate multiple information sources.

We applied our method to synthetic data and cross compared our method with other clustering methods. The clustering performance demonstrated that our method is superior to other methods.

In later research, we will deeply explore the inner relationship among multiple information sources via tensor decomposition and carry out our algorithm to tackle large-scale and real databases.

7 Acknowledgments

Research supported by (1)China Scholarship Council(CSC, No. 2006153005); (2) Research Council K.U.Leuven: GOA-Ambiorics, GOA-MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), CIF1, STRT1/08/023; (3) F.W.O.: (a) project G.0321.06, (b) Research Communities ICCoS, ANMMM and MLDM; (4) the Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, "Dynamical systems, control and optimization", 2007–2011); (5) EU: ERNSI; (6) Wuhan university of science and technolgy(WUST), college of information science and engineering (CISE).

References

[1] L. Tang, X. Wang, H. Liu.: Uncovering Groups via Heterogeneous Interaction Analysis. In *ICDM'09: Proceedings of 9th IEEE International Conference on Data Mining* pages 143–152. New York:ACM Press (2009)

[2] S. Bickel, T. Scheffer.: Multi-view Clustering. In *ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining* pages 19–26 (2004)

[3] A. Strehl, J. Ghosh.: Cluster Ensembles-a Knowledge Reuse Framework for Combining Multiple Partitions. *JMLR* 3:583–617 (2002)

[4] A. Smilde, R. Bro, P. Geladi.: *Multi-way Analysis: Applications in the Chemical Sciences*. Wiley, West Sussex, England (2004)

[5] L. De Lathauwer, J. Vandewalle.: Dimensionality Reduction in Higher-order Signal Processing and Rank-$(r_1, r_2, \ldots, r_n)$ Reduction in Multilinear Algebra. *Lin. Alg. Appl.* 391:31–55 (2004)

[6] T. Kolda, B. Bader.: The TOPHITS Model for Higher-order Web Link Analysis. In *Proceedings of the SIAM Data Mining Conference Workshop on Link Analysis, Counterterrorism and Security* (2006)

[7] S.Jegelka, S. Sra, A. Banerjee.: Approximation Algorithms for Tensor Clustering. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory*, pages 822–833. Springer (2009)

[8] X. Liu, S. Yu, Y. Moreau, B. De Moor, W. Glänzel, F. Janssens.: Hybrid clustering of text mining and bibliometrics applied to journal sets. In *SDM'09: Proceedings of the 2009 SIAM International Conference on Data Mining*. SIAM (2009)

[9] J. tao Sun, H.-J. Zeng, H. Liu, Y. Lu.: Cubesvd: A Novel Approach to Personalized Web Search. In *WWW'05: Proceedings of the 14 th International World Wide Web Conference* pages 382–390 (2005)

[10] N. Liu, B. Zhang, J. Yan, Z. Chen, W. Liu, F. Bai, L. Chien.: Text Representation: From Vector to Tensor. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining* (2005)

[11] J. Sun, D. Tao, C. Faloutsos.: Beyond Streams and Graphs: Dynamic Tensor Analysis. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (2006)

[12] H. Huang, C. Ding, D. Luo, T. Li.: Simultaneous Tensor Subspace Selection and Clustering: the Equivalence of High Order SVD and k-means Clustering. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (2008)

[13] M. E. J. Newman.: Modularity and Community Structure in Networks. *PNAS* 103(23):8577–8582 (2006)

[14] M. E. J. Newman.: Finding Community Structure in Networks using the Eigenvectors of Matrices. *Physical Review E* 74(3):036104 (2006)

[15] D. C. Lay.: *Linear Algebra and its Applications (3rd Edition).* Addition Wesley (2003)

[16] T. G. Kolda, B. W. Bader.: Tensor Decompositions and Applications. *SIAM Review* 51(3):455–500 (2009)

[17] L. De Lathauwer, B. D. Moor, J. Vandewalle.: A Multilinear Singular Value Decomposition. *SIAM J. Matrix Anal. Appl.* 21(4):1253–1278 (2000)

[18] L. De Lathauwer, B. D. Moor, J. Vandewalle.: On the Best Rank-1 and Rank-$(r_1, r_2, \ldots, r_n)$ Approximation of Higher-order Tensors. *SIAM J. Matrix Anal. Appl.* 21(4):1324–1342 (2000)

NOTICE: this is the author's version of a work that was accepted for publication in Advances in Neural Networks - ISSN 2010, Lecture Notes in Computer Science (LNCS). Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was published as

X. Liu, L. De Lathauwer, F. Janssens, B. De Moor, "Hybrid Clustering on Multiple Information Sources via HOSVD", *Advances in Neural Networks - ISSN 2010*, June 6–9, 2010, Shanghai, China, Lecture Notes in Computer Science, Vol. 6064/2010, Springer, pp. 337–345, DOI: 10.1007/978-3-642-13318-3_42.

The original publication is available at www.springerlink.com.