Hyperedge Prediction using Tensors

Ankit Sharma

September 10, 2013

1 Introduction

$Hypergraphs$ which are a non-trivial generalization of graphs are known to be effective tools for capturing social collaborations like group interactions, author-author collaborations etc. In Hypergraphs each $hyperedge$ can have more than two nodes or vertices unlike the graphs. For example, we model a conference paper written by three authors as an $occurrence$ of a hyperedge containing three vertices one for each author. In this discussion we shall talk in terms of authors and their collaboration to write papers. Therefore, our input is all the previous papers published by different author collaborations of various sizes. We build a hypergraph with this data where each collaboration is a hyperedge and each paper is occurrence of that hyperedge. Our aim is to predict the likeliness that a collaboration (hyperedge) will write a paper (occur) together in future.

2 Approaches and doubts

2.1 Hyperincidence Tensor approach

We model the input hypergraph of all previous collaborations as an array of hypergraph's incidence matrix (hyper-incidence matrix) over several time periods. This forms our hyperincidence tensor $\mathcal{Z}_h$. This intuition is captured in figure 1(a)/(b). Now let us assume that we are considering N_t snapshots or time periods, N_h unique collaborations (hyperedges) and let there be total N_a unique authors. Therefore, tensor $\mathcal{Z}_h = \{H^{(t)}\}_{t=1}^{N_t}$ represents array of incidence matrices $H^{(t)}$. Each of these incidence matrix $H^(t)$ has N_h rows each representing a unique collaboration of authors (hyperedge). Each one of these hyperedges $h_k \forall k \in \{1, 2, ..., N_h\}$ represent a unique collaboration between a subset of authors. Dimension of $\mathcal{Z}_h$ therefore, becomes $N_h \times N_a \times N_t$. Initially this tensor contains all zeros. For each paper in time period t in our input data we find out

the index $k \in \{1, 2, ..., N_h\}$ such that h_k represents the collaboration of this paper. Once we find the index k for the hyperedge we fill the tensor $\mathcal{Z}_h(k, j, t) = 1$ where j is the index of each author (vertex) which is the part of the collaboration (hyperedge) h_k. Thus, in short this tensor is an array of incidence matrices where each $H^{(t)}$ represent the hypergraph formed by the papers published in time period t. And each row of this matrix represents a collaboration with non-zero values at those columns (authors) which are part of this collaboration. Note that the same collaboration of authors (hyperedge) can occur multiple times to write different papers in which case we might add up the values in tensor (or any other way we wish to capture multiplicity).

In order to do future hyperedge prediction we decompose this incidence tensor using CANDECOMP/PARAFAC (CP) [2] tensor decomposition. Given the three dimensional tensor $\mathcal{Z}_{\boldsymbol{h}}$ with size $N_h \times N_a \times N_t$ its decomposition is given by $\mathcal{Z}_{\boldsymbol{h}} \approx \sum_{k=1}^{K} \lambda_k \mathbf{h}_k \circ \mathbf{a}_k \circ \mathbf{t}_k$ where $\lambda_k \in \mathbb{R}^+$, $\mathbf{h}_k \in \mathbb{R}^{N_h}$, $\mathbf{a}_k \in \mathbb{R}^{N_a}$, and $\mathbf{t}_k \in \mathbb{R}^{N_t}$. Then we formulate that $\mathbf{h_k a_k^\top}$ for the component k basically represents the likelihood of i^{th} hyperedge (collaboration) containing a particular j^{th} vertex (author). This results in the final likelihood matrix between each hyperedge-vertex pair as follows:

$$\mathbf{S} = \sum_{k=1}^{K} \gamma_k \lambda_k \mathbf{h_k a_k^\top} \text{ where } \gamma_t = \frac{1}{T_{buf}} \left(\sum_{t=T-T_{buf}+1}^{T} \mathbf{t}_k(t) \right)$$

The compression of the time factors for the past T_{buf} number of years (buffer) similar to heuristic method adopted in [1].

2.1.1 Doubt 1

Currently we are only restricting ourselves to the problem of finding recurring hyperedge prediction i.e. predicting only the collaborations (hyperedges) that have been observed in past. Now, $\mathbf{S}(i, j)$ for a given i represents the likeliness of the event that hyperedge i occurs and contains vertex j. Now we assume all these events for different values of j are independent. Therefore, we take the likelihood of a hyperedge i occuring as a whole with likeliness $P(h_i) = \prod_{j \in V} \mathbf{S}(i, j)$ were V is the set of author indices who are a part of the collaboration represented by hyperedge h_i. Is this a right way to think ?

2.1.2 Doubt 2

In contrast to the method in Doubt 1 how good is to assume that the probability or likelihood of hyperedge i occurring as just dependent upon the hyperedge factors

alone: $P(h_i) = \sum_{k=1}^{K} \lambda_k \mathbf{h_k}(i)$ or by may be including time factors also: $P(h_i) = \sum_{k=1}^{K} \lambda_k \gamma_k \mathbf{h_k}(i)$? Here, $\mathbf{h_k}$ and $\gamma_\mathbf{k}$ are the factors from CP decomposition.

2.1.3 Doubt 3

In continuation of Doubt 1 if we try to extend our prediction to new hyperedges by observing that fact that each row $\mathbf{S}(i,:)$, which represents a hyperedge, will have non-zero entries (after decomposition) other than just the column indices in the set V. Then is it safe to assume that each of the $\mathbf{S}(i,j)$ for all $j \in \{\{1,....,N_a\} - V\}$ represents the likelihood of new vertex j becoming a part of the hyperedge i which originally had only the vertices in V. Again here V is the set of author indices who are a part of the collaboration represented by hyperedge h_i.

2.2 K-way tensor Approach

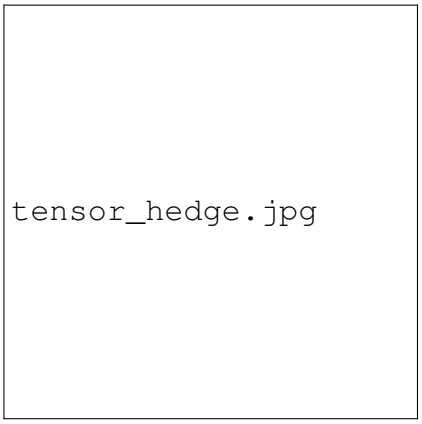
2.2.1 Doubt 4

If the approach of incidence matrix as in the section 2.1 above is somewhat wrong then our doubt is that what is then the cleanest way to capture a hypergraph. Is storing hypergraph in form of a K-way Tensor a better approach. By that we mean that if let us suppose or collaborations are of maximum size S then we use a tensor of dimension S which captures each collaboration as entries in this high dimensional tensor. Some of the literature like [3][4] are mentioning these ways but no direct application is provided. Is this a possible direction ? Given that these methodologies are quiet new to us will it be possible for some redirection from your end ?

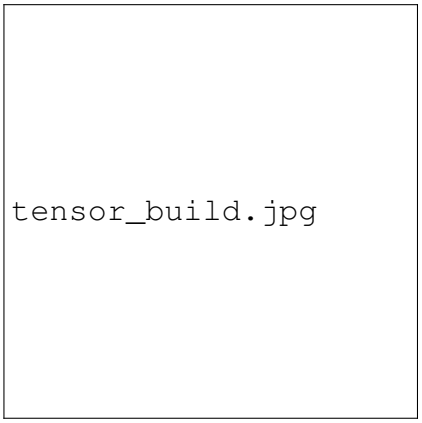
References

[1] Daniel M. Dunlavy, Tamara G. Kolda, and Evrim Acar., "Temporal Link Prediction Using Matrix and Tensor Factorizations.", ACM Trans. Knowl. Discov., Feb'11.

[2] T. G. Kolda and B. W. Bader., "Tensor Decompositions and Applications", SIAM Review 51(3):455-500, September'09.



(a) Temporal hyperedges captured in Tensor



(b) Loading data into Tensor

Figure 1

[3] Liqun Qi,"The Spectral Theory of Tensors (Rough Version)", eprint arXiv:1201.3424, ArXiv e-prints, Jan'12.

[4] Joshua Cooper and Aaron Dutle, "Spectra of Uniform Hypergraphs", eprint arXiv:1106.4856, ArXiv e-prints, June'11.