Unsupervised Feature Selection via Hypergraph Embedding

Zhihong Zhang[1]
zhihong@cs.york.ac.uk
Peng Ren[2]
pengren@upc.edu.cn
Edwin R. Hancock [1]
erh@cs.york.ac.uk

[1] Department of Computer Science,
The University of York,
York, UK.

[2] College of Information and Control Engineering,
China University of Petroleum,
Qingdao, China.

For the task of feature selection addressed in this paper, we introduce a hypergraph embedding view of feature selection by subspace learning. The method jointly evaluates the utility sets of features rather than individual features. There are three novel ingredients. The first is that by incorporating hypergraph representation into feature selection, we can be more effectively capture the higher order relations among samples. Secondly, inspired from the recent works on mutual information [1], we determine the weight of a hyperedge using an information measure referred to as multidimensional interaction information (MII) which precisely preserves the higher order relations captured by the hypergraph. The advantage of MII is that it is sensitive to the relations between sample combinations, and as a result can be used to seek third or even higher order dependencies among the relevant samples. Thus, the structural information latent in the data can be more effectively modeled. Finally, we describe a new feature selection strategy through hypergraph embedding, which casts the feature discriminant analysis into a regression framework that considers the correlations among features. As a result, we can evaluate joint feature combinations, rather than being confined to consider them individually.

Hypergraph Construction: we establish a novel hypergraph framework which is used for characterizing the multiple relationships within a set of samples. Based on the higher order similarity measure, we establish a hypergraph framework for characterizing a set of high dimensional samples. A hypergraph is defined as a triplet $H = (V,E,w)$. Here V denotes the vertex set, E denotes the hyperedge set in which each hyperedge $e \in E$ represents a subset of V, and w is a weight function which assigns a real value $w(e)$ to each hyperedge $e \in E$. We only consider K-uniform hypergraphs (i.e. those for which the hyperedges have identical cardinality K) in our work. Given a set of high dimensional samples $\mathbf{X} = [x_1, \cdots x_N]^T$ where $x_i \in \mathbb{R}^d$, we establish a K-uniform hypergraph, with each hypergraph vertex representing an individual sample and each hyperedge representing the Kth order relations among a K-tuple of participating samples. A K-uniform hypergraph can be represented in terms of Kth order matrix, i.e. a tensor $\mathcal{W}$ of order K, whose element $W_{i_1,\cdots,i_K}$ is the hyperedge weight associated with the K-tuple of participating vertices $\{v_{i_1}, \cdots, v_{i_K}\}$. In our work, the hyperedge weight associating with $\{x_{i_1}, x_{i_2}, \cdots, x_{i_K}\}$ is computed as follows

$$W_{i_1,\cdots,i_K} = K \frac{I(x_{i_1}, x_{i_2}, \cdots, x_{i_K})}{H(x_{i_1}) + H(x_{i_2}) + \cdots H(x_{i_K})} . \tag{1}$$

It is clear that $W_{i_1,\cdots,i_K}$ is a normalized version of K-th order Interaction Information. The greater the value of $W_{i_1,\cdots,i_K}$ is, the more relevant the K samples are. On the other hand, if $W_{i_1,\cdots,i_K} = 0$, the K samples are totally unrelated.

Hypergraph Representation: In our work, we consider the transformation of a K-uniform hypergraph into a graph. Accordingly, the associated hypergraph tensor $\mathcal{W}$ is transformed to a graph adjacency matrix $\boldsymbol{A}$, and the higher order information exhibited in the original hypergraph can be encoded in an embedding space spanned by the related matrix representation. In this scenario, one straightforward way for the transformation is marginalization which computes the arithmetical average over all the hyperedge weights $W_{i_1,\cdots,i_{K-2},i,j}$ associated with the edge weight $A_{i,j}$

$$\tilde{A}_{i,j} = \sum_{i_1=1}^{|V|} \cdots \sum_{i_{K-2}=1}^{|V|} W_{i_1,\cdots,i_{K-2},i,j} \tag{2}$$

The edge weight $\tilde{A}_{i,j}$ for edge ij is generated by a uniformly weighted sum of hyperedge weights $W_{i_1,\cdots,i_{K-2},i,j}$. However, the form appearing in (2) behaves as a low pass filter, and thus results in information loss through marginalization.

To make the process of marginalization more comprehensive, we use marginalization to constrain the sum of edge weights and then estimate their values through solving an over-constrained system of linear equations. Our idea is motivated by the so called *clique average* introduced in the higher order clustering literature [4]. We characterize the relationships between $\boldsymbol{A}$ and $\mathcal{W}$ as follows

$$W_{i_1,\cdots,i_K} = \sum_{\{i,j\} \subseteq \{i_1,\cdots,i_K\}} A_{i,j} \tag{3}$$

There are $\binom{|V|}{2}$ variables and $\binom{|V|}{K}$ equations in the system of equations described in (2). When $K > 2$, the linear system (2) is over-determined and cannot be solved analytically. We thus approximate the solution to (2) by minimizing the least squares error

$$\widehat{\boldsymbol{A}} = \underset{\boldsymbol{A}}{\operatorname{argmax}} \sum_{i_1,\cdots,i_K} \left(\sum_{\{i,j\} \subseteq \{i_1,\cdots,i_K\}} A_{i,j} - W_{i_1,\cdots,i_K} \right)^2 \tag{4}$$

In practical computation, we normalize the compatibility tensor $\mathcal{W}$ by using the extended Sinkhorn normalization scheme [2], and constrain the element of $\boldsymbol{A}$ to be in the interval $[0,1]$ to avoid unexpected infinities. Effective iterative numerical methods are used to compute the approximated solutions [3].

Feature Selection through Hypergraph Embedding: we formulate the procedure of feature extraction on a basis of hypergraph spectral embedding. One goal of spectral embedding is to represent the high dimensional data $\mathbf{X} \in \mathbb{R}^{N \times d}$ by a low dimensional representation $\mathbf{Y} \in \mathbb{R}^{N \times C}$ ($C \ll d$) in the low dimensional feature space such that the structural characteristics of the high dimensional data are well preserved or are more "obvious". Here we use the representations $\mathbf{X} = [x_1, \cdots x_N]^T$ and $\mathbf{Y} = [y_1, \cdots, y_k, \cdots, y_C]$, where y_k is a N-dimensional vector and its N elements represent the N samples $x_1, \cdots x_N$ separately in the kth dimension of the low dimensional feature space.

The hypergraph embedding procedure can be viewed as feature extraction, and can be expressed as $\mathbf{Y} = \mathbf{X}\Phi$ where $\Phi \in \mathbb{R}^{d \times C}$ is a column-full-rank projection matrix. However, unlike feature extraction, feature selection attempts to select the optimal feature subset in the original feature space. Therefore, for the task of feature selection, the projection matrix $\Phi = [\Phi_1, \ldots \Phi_C]$ can be constrained to be a selection matrix which contains the combination coefficients for different features in approximating $\mathbf{Y} = [y_1, \ldots, y_C]$. That is, given the kth column of $\mathbf{Y}$, i.e y_k, we aim to find a subset of features, such that their linear span is close to y_k. This idea can be formulated as the minimization problem

$$\widehat{\Phi} = \underset{\Phi}{\operatorname{argmin}} \sum_{k=1}^{C} \|y_k - X\Phi_k\|^2 . \tag{5}$$

where $\Phi = [\Phi_1, \cdots, \Phi_k, \cdots, \Phi_C]$ and Φ_k is a d dimensional vector that contains the combination coefficients required to compute for different features in approximating y_k.

[1] Z. Zhang and E. R. Hancock. Hypergraph based Information-theoretic Feature Selection. *Pattern Recognition Letters*, 2012.

[2] A. Shashua, R. Zass and T. Hazan. Multi-way clustering using super-symmetric non-negative tensor factorization. *In Proc. ECCV.*, (4): 595-608, 2006.

[3] A. Björck. Numberical methods for least squares problems. *In Proc. SIAM.*, 1996.

[4] S. Agarwal, J. Lim, L. Zelnik-Manor, P. Perona, D. Kriegman, and S. Belongie. Beyond pairwise clustering. *In Proc. CVPR.*, pages 838-845, 2005.