Hypergraph Spectra for Semi-supervised Feature Selection

Zhihong Zhang¹, Edwin R. Hancock^{1,*}, and Xiao Bai²

¹ Department of Computer Science,
University of York, UK

² School of Computer Science and Engineering
Beihang University, China

Abstract. In many data analysis tasks, one is often confronted with the problem of selecting features from very high dimensional data. Most existing feature selection methods focus on ranking individual features based on a utility criterion, and select the optimal feature set in a greedy manner. However, the feature combinations found in this way do not give optimal classification performance, since they neglect the correlations among features. While the labeled data required by supervised feature selection can be scarce, there is usually no shortage of unlabeled data. In this paper, we propose a novel hypergraph based semi-supervised feature selection algorithm to select relevant features using both labeled and unlabeled data. There are two main contributions in this paper. The first is that by incorporating multidimensional interaction information (MII) for higher order similarities measure, we establish a novel hypergraph framework which is used for characterizing the multiple relationships within a set of samples. Thus, the structural information latent in the data can be more effectively modeled. Secondly, we derive a hypergraph subspace learning view of feature selection which casting 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. Experimental results demonstrate the effectiveness of our feature selection method on a number of standard face data-sets.

Keywords: Hypergraph representation, Semi-supervised subspace learning.

1 Introduction

In order to render the analysis of high-dimensional data tractable, it is crucial to identify a smaller subset of features that are informative for classification and clustering. Dimensionality reduction aims to reduce the number of variables under consideration, and the process can be divided into feature extraction and feature selection. Feature extraction usually projects the features onto a low-dimensional and distinct feature space, e.g., Locally Linear Embedding (LLE)

* Edwin Hancock is supported by a Royal Society Wolfson Research Merit Award.

[1], kernel PCA [2], Locality preserving Projection (LPP) [3], Neighborhood Preserving Embedding (NPE) [4] and Laplacian eigenmap [5]. Unlike feature extraction, feature selection identifies the optimal feature subset in the original feature space. By maintaining the original features, feature selection improves the interpretability of the data, which is preferred in many real world applications, such as face recognition [28]. Feature selection algorithms can be roughly classified into three groups, namely a) supervised feature selection, b) unsupervised feature selection and c) semi-supervised feature selection.

Supervised feature selection algorithms usually evaluate the importance of features using the correlation between the features and the class label. An important line of research in this area is the use of methods based on mutual information. For example, Battiti [6] has developed the Mutual Information-Based Feature Selection (MIFS) criterion, where the features are selected in a greedy manner. Given a set of existing selected features S , at each step it locates the feature x_i that maximizes the relevance to the class i.e. $I(x_i; C)$. The selection is regulated by a proportional term $\beta I(x_i; S)$ that measures the overlap information between the candidate feature and the existing features. The parameter β may significantly affect the features selected, and its control remains an open problem. Peng et al. [7] on the other hand, use the so-called Maximum-Relevance Minimum-Redundancy criterion (MRMR), which is equivalent to MIFS with $\beta = \frac{1}{n-1}$.

Supervised spectral feature selection algorithms provide powerful alternatives to those discussed above. Examples include the Fisher score [8], the trace ratio [9] and ReliefF [10]. Given d features, and a similarity matrix S for the samples, the idea underpinning spectral feature selection algorithms is to identify features that align well with the leading eigenvectors of S . The leading eigenvectors of S contain information of concerning the structure of the sample distribution and group similar samples into compact clusters. Consequently, features that align closely to the eigenvectors will better preserve sample similarity [11].

While the labeled data required by supervised feature selection can be scarce, there is usually no shortage of unlabeled data. Hence, there are obvious attractions in developing unsupervised feature selection algorithms which can utilize this data. The typical examples in unsupervised learning are graph-based spectral learning algorithms. Examples include the Laplacian score [12], SPEC [11], Multi-Cluster Feature Selection (MCFS) [22] and Unsupervised Discriminative Feature Selection (UDFS) [23]. A frequently used criterion in unsupervised graph-based spectral learning is to select the features which best preserve the structure of the data similarity or a manifold structure derived from the entire feature set. For example, the Laplacian score [12] uses a nearest neighbor graph to model the local geometric structure of the data, using the pairwise similarities between features calculated using the heat kernel. In this framework, the features are evaluated individually and are selected one by one. Another unsupervised spectral feature selection algorithm is SPEC [11], which is an extension of the Laplacian score aimed at making it more robust to noise. The method selects

the features that are most consistent with the graph structure. Note that SPEC also evaluates features independently.

However, there are some drawbacks with the above graph-based spectral learning methods when they are used to deal with computer vision problems. One of these is that they evaluate features individually and hence cannot handle redundant features. Redundant features increase the dimensionality unnecessarily, and degrade learning performance when faced with a shortage of data. It is also shown empirically that removing redundant features can result in significant performance improvement. Another problem is that their graph representations are very sensitive to the topological structure and lack sufficient robustness in real-world learning tasks. This is because in real world problems the objects and their features tend to exhibit multiple relationships rather than simple pairwise ones. This factor will considerably compromise the performance of learning approaches. For example, consider the problem of grouping images of five different persons based on identity, each of which is imaged in the same pose, but under different lighting conditions. See Fig. 1 for an illustration. Recent studies on illumination have shown that images of the same objects may look drastically different under different lighting conditions while different objects may appear similar under different illumination conditions [24]. Furthermore, it is well known that the set of images of a Lambertian surface under arbitrary lighting (without shadowing) lies on a 3-D linear subspace in the image space [13]. As any three images span a 3-D subspace, one needs to consider at least four images at a time to define a measure of similarity. Therefore, suitable for graph construction and similarity measure are needed.

Fig. 1. Shown above are images of five persons under varying illumination conditions

A natural way of remedying the information loss described above is to represent the data set as a hypergraph instead of a graph. Hypergraph representations allow vertices to be multiply connected by hyperedges. They can hence be both accurate and complete in describing feature relations and structures. Due to their effectiveness in representing multiple relationships, in this paper, we propose a hypergraph based semi-supervised feature selection method which can be used to classify face images under varying illumination conditions. This method jointly evaluates the utility for 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 effective capture the higher order relations among samples. Secondly, inspired from the recent works on mutual

information [26], [27], 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 semi-supervised feature selection strategy through hypergraph subspace learning, 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, thus it is able to handle feature redundancy.

2 Hypergraph Construction

In this section, we establish a novel hypergraph framework which is used for characterizing the multiple relationships within a set of samples. To this end, we commence by introducing a new method for measuring higher order similarities among samples based on information theory. According to Shannon's study, the uncertainty of a random variable X can be measured by the entropy $H(X)$. For two random variables X and Y , the conditional entropy $H(Y|X)$ measures the remaining uncertainty about Y when X is known. The mutual information $I(X;Y)$ of X and Y quantifies the information gain about Y provided by X . The relationship between $H(Y)$, $H(Y|X)$ and $I(X;Y)$ is $I(X;Y) = H(Y) - H(Y|X)$. As defined by Shannon, the initial uncertainty for X is $H(X) = -\sum_{x \in X} P(x) \log P(x)$, where $P(x)$ is the prior probability density function over $x \in X$. The remaining uncertainty for Y if X is known is defined by the conditional entropy $H(Y|X) = -\int_x p(x) \{ \sum_{y \in Y} p(y|x) \log p(y|x) \} dx$, where $p(y|x)$ denotes the posterior probability for $y \in Y$ given $x \in X$. After observing x , the amount of additional information gain is given by the mutual information

$$I(X;Y) = \sum_{y \in Y} \int_x p(y,x) \log \frac{p(y,x)}{p(y)p(x)} dx . \quad (1)$$

The mutual information (1) quantifies the information which is shared by X and Y . When the $I(X;Y)$ is large, it implies that x and y are closely related. Otherwise, when $I(X;Y)$ is equal to 0, it means that two variables are totally unrelated. Analogically, the conditional mutual information of X and Y given Z , denoted as $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$, represents the quantity of information shared by X and Y when Z is known. The conditioning on a third random variable may either increase or decrease the original mutual information. In this context, the Interaction Information $I(X;Y;Z)$ is defined as the difference between the conditional mutual information and the simple mutual information, i.e.

$$I(X;Y;Z) = I(X;Y|Z) - I(X;Y) . \quad (2)$$

The interaction information $I(X; Y; Z)$ measures the influence of the variable Z on the amount of information shared between variables X and Y . Its value can be positive, negative, or zero. Zero valued Interaction Information $I(X; Y; Z)$ implies that the relation between X and Y entirely depends on Z . A positive value of $I(X; Y; Z)$ implies that X and Y are independent of each other themselves, but are correlated with each other when combined with Z . A negative value of $I(X; Y; Z)$ indicates that Z can account for or explain the correlation between X and Y . The generalization of Interaction Information to K variables is defined recursively as follow

$$I(\{X_1, \dots, X_K\}) = I(\{X_2, \dots, X_K\} | X_1) - I(\{X_2, \dots, X_K\}) . \quad (3)$$

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 $\mathbf{X} = [x_1, \dots, x_N]^T$ 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 K th order relations among a K -tuple of participating samples. A K -uniform hypergraph can be represented in terms of K th order matrix, i.e. a tensor $\mathcal{W}$ of order K , whose element $W_{i_1, \dots, i_K}$ is the hyperedge weight associated with the K -tuple of participating vertices $\{v_{i_1}, \dots, v_{i_K}\}$. In our work, the hyperedge weight associating with $\{x_{i_1}, x_{i_2}, \dots, x_{i_K}\}$ is computed as follows

$$W_{i_1, \dots, i_K} = K \frac{I(x_{i_1}, x_{i_2}, \dots, x_{i_K})}{H(x_{i_1}) + H(x_{i_2}) + \dots + H(x_{i_K})} . \quad (4)$$

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

3 Hypergraph Representation

Unlike matrix eigen-decomposition, there has not yet been a widely accepted method for spanning a rationale eigen-space for a tensor [30]. Therefore, it is hard to directly embed a hypergraph into a feature space spanned by its tensor representation through eigen-decomposition. 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 $\mathbf{A}$, 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 $W_{i_1, \dots, i_{K-2}, i, j}$ associated with the edge weight $A_{i, j}$

$$\tilde{A}_{i, j} = \sum_{i_1=1}^{|V|} \cdots \sum_{i_{K-2}=1}^{|V|} W_{i_1, \dots, i_{K-2}, i, j} \quad (5)$$

The edge weight $\tilde{A}_{i, j}$ for edge ij is generated by a uniformly weighted sum of hyperedge weights $W_{i_1, \dots, i_{K-2}, i, j}$. However, the form appearing in (5) 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 [15]. We characterize the relationships between $\mathbf{A}$ and $\mathcal{W}$ as follows

$$W_{i_1, \dots, i_K} = \sum_{\{i, j\} \subseteq \{i_1, \dots, i_K\}} A_{i, j} \quad (6)$$

There are $\binom{|V|}{2}$ variables and $\binom{|V|}{K}$ equations in the system of equations described in (5). When $K > 2$, the linear system (5) is over-determined and cannot be solved analytically. We thus approximate the solution to (5) by minimizing the least squares error

$$\hat{\mathbf{A}} = \underset{\mathbf{A}}{\operatorname{argmax}} \sum_{i_1, \dots, i_K} \left(\sum_{\{i, j\} \subseteq \{i_1, \dots, i_K\}} A_{i, j} - W_{i_1, \dots, i_K} \right)^2 \quad (7)$$

In practical computation, we normalize the compatibility tensor $\mathcal{W}$ by using the extended Sinkhorn normalization scheme [31], and constrain the element of $\mathbf{A}$ to be in the interval $[0, 1]$ to avoid unexpected infinities. Effective iterative numerical methods are used to compute the approximated solutions [32].

The adjacency matrix $\mathbf{A}$ computed through (7) is one effective representation for a K -uniform hypergraph, because it naturally avoids the operation of arithmetic average and thus to a certain degree overcomes the low pass information loss arising in (5). Furthermore, the $\mathbf{D}$ is the diagonal matrix with its i th diagonal element being $A_{ii} = \sum_j A_{ij}$. In this context, a hypergraph can be easily embedded into a feature space spanned by a semi-supervised subspace learning, which will be explained in detail in the next Section.

4 Feature Selection through Semi-supervised Subspace Learning

In this section, we formulate the procedure of semi-supervised feature selection on a basis of hypergraph subspace learning. Feature selection can be seen as a special subspace learning task, where the projection matrix is constrained to be selection matrix.

4.1 Hypergraph Semi-supervised Subspace Learning

One goal of hypergraph subspace learning 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 $\mathbf{X} = [x_1, \dots, x_N]^T$ and $\mathbf{Y} = [y_1, \dots, y_k, \dots, y_C]$, where y_k is a N -dimensional vector and its N elements represent the N samples $x_1, \dots, x_N$ separately in the k th dimension of the low dimensional feature space. Based on the hypergraph transformation described in Section 3, the semi-supervised hypergraph subspace learning can be easily conducted as the following two steps, a) label propagation, b) label regression:

Label Propagation: First, we obtain the soft labels of unlabeled data using a new label propagation method [25], in which an additional class labeled $C + 1$ is introduced to accommodate the outliers data. Give a data set $\mathbf{X} = \{x_1, \dots, x_l, \dots, x_N\} \subset \mathbb{R}^d$ and the first l data are labeled and the subsequent $u = N - l$ data are unlabeled. Define the initial label matrix $L = [(l_1)^T, (l_2)^T, \dots, (l_N)^T] \in \mathbb{R}^{N \times (C+1)}$, for the labeled data, $L_{ij} = 1$ if x_i is labeled as $l_i = j$ and $L_{ij} = 0$ otherwise. For the unlabeled data, $L_{ij} = 1$ if $j = C+1$ and $L_{ij} = 0$ otherwise. Denote a stochastic matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$, to assign labels to the unlabeled data, an iteration function $\mathbf{F}(t+1) = \alpha\mathbf{P}\mathbf{F}(t) + (1-\alpha)L$ is computed on the nodes of graph until convergence so as to satisfy two constraints, where α is a parameter in $(0, 1)$. During each iteration, each point receives information from its neighbors (first term), and also retains its initial information (second term). $\mathbf{F} = [\mathbf{F}_1^T, \dots, \mathbf{F}_N^T]^T \in \mathbb{R}^{N \times (C+1)}$ is the predicted soft label matrix, where $\mathbf{F}_{ij}$ reflects the posterior probability of data point x_i belonging to class j [29]. When $j = C + 1$, $\mathbf{F}_{i,C+1}$ represents the probability of x_i belonging to outlier. The iteration process will converge to the fixed point

$$\mathbf{F} = \lim_{t \rightarrow \infty} \mathbf{F}(t) = (1 - \alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}L. \quad (8)$$

Using this iterative label propagation scheme, the outlier in data can be detected and the label for each unlabeled point is assigned to the class from which it has received the most information during the iteration process.

Label Regression: After we obtain the soft label $\mathbf{F}_i$ for each data x_i , i.e., the probability $\mathbf{F}_{ij}$ of x_i belonging to class j , we can learn the subspace for data using the soft labels. Assuming the low-dimensional data can be obtained from the linear projection $\mathbf{Y} = \mathbf{X}\mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{d \times C}$ is the projection vector, a regression function can be defined as

$$\arg \min_{\mathbf{W}, b} \gamma \|\mathbf{W}\|^2 + \sum_{i=1}^N \sum_{j=1}^C \mathbf{F}_{ij} \|\mathbf{W}^T x_i + b - t_j\|^2, \quad (9)$$

where the bias term $b \in \mathbb{R}^{C \times 1}$, $t_j = [0, \dots, 0, 1, 0, \dots, 0]^T$ is the class indicator vector for the j -th class and γ is the regularization parameter. It can be easily

verified that for larger values of $\mathbf{F}_{i,C+1}$, which indicates x_i is belonging to the outlier, then the values of $\mathbf{F}_{i,j}$ ($1 \leq i \leq C$) should be smaller to reduce the effect of the outlier data point x_i in the regression model according to Equation (9). Differentiating the matrix form in Equation (9) w.r.t $\mathbf{W}$ and b , we have the solution to the regression problem is

$$\mathbf{W} = (\mathcal{X}\mathbf{L}_s\mathcal{X}^T + \gamma\mathbf{I})^{-1}\mathcal{X}\mathbf{C}_s\mathbf{F}_C, \quad (10)$$

where $\mathbf{I}$ denotes an identity matrix with proper size, $\mathbf{L}_s = \mathbf{S} - \frac{1}{\mathbf{1}_N^T\mathbf{S}\mathbf{1}_N}\mathbf{S}\mathbf{1}_N\mathbf{1}_N^T\mathbf{S}$, $\mathbf{S}_{ii} = \sum_{j=1}^C \mathbf{F}_{ij}$, $\mathbf{F}_C \in \mathbb{R}^{N \times C}$ is formed by the first C columns of $\mathbf{F}$, $\mathbf{C}_s = \mathbf{I} - \frac{1}{\mathbf{1}_N^T\mathbf{S}\mathbf{1}_N}\mathbf{S}\mathbf{1}_N\mathbf{1}_N^T$ and $\mathbf{1}_N \in \mathbb{R}^{N \times 1}$.

4.2 Robust Feature Selection Based on ℓ_1 -Norms

The hypergraph subspace learning procedure can be viewed as feature extraction, and can be expressed as $\mathbf{Y} = \mathbf{X}\mathbf{W}$ where $\mathbf{W} \in \mathbb{R}^{d \times C}$ 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 $\mathbf{W} = [w_1, \dots, w_C]$ can be constrained to be a selection matrix $\Phi = [\Phi_1, \dots, \Phi_C]$ which contains the combination coefficients for different features in approximating $\mathbf{Y} = [y_1, \dots, y_C]$. That is, given the k th 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

$$\hat{\Phi} = \underset{\Phi}{\operatorname{argmin}} \sum_{k=1}^C \|y_k - X\Phi_k\|^2. \quad (11)$$

where $\Phi = [\Phi_1, \dots, \Phi_k, \dots, \Phi_C]$ and Φ_k is a d dimensional vector that contains the combination coefficients required to compute for different features in approximating y_k . However, feature selection requires to locate a optimal subset of features that are close to y_k . This is a combinatorial problem which is NP-hard. Thus we approximate the problem in (11) subject to the constraint

$$|\Phi_k| \leq \gamma \quad (12)$$

where $|\Phi_k|$ is the ℓ_1 -norm and $|\Phi_k| = \sum_{j=1}^d |\Phi_{j,k}|$. When applied in regression, the ℓ_1 -norm constraint is equivalent to applying a Laplace prior [18] on Φ_k . This tends to force some entries in Φ_k to be zero, resulting in a sparse solution. Therefore, the representation $\mathbf{Y}$ is generated by using only a small set of selected features in $\mathbf{X}$.

In order to efficiently solve the optimization problem in Equations (11) and (12), we use the Least Angle Regression (LARs) algorithm [20]. Instead of setting the parameter γ , LARs allow us to control the sparseness of Φ_k . This is done by specifying the cardinality of the number of nonzero subset of Φ_k , which is particularly convenient for feature selection.

We consider selecting m features from the d feature candidates. For a dataset containing C clusters, we can compute C selection vectors $\{\Phi_k\}_{k=1}^C \in R^d$. The cardinality of each Φ_k is m and each entry in Φ_k corresponds to a feature. Here, we use the following computationally effective method for selecting exactly m features based on the C selection vectors. For every feature j , we define the HG score for the feature as

$$HGscore(j) = \max_k |\Phi_{j,k}|. \quad (13)$$

where $\Phi_{j,k}$ is the j th element of vector Φ_k . We then sort the features in descending order according to their HG scores, and then select the top m features.

5 Feature Evaluation Indices

Our proposed semi-supervised feature selection method utilizes hypergraph based label regression and the Least Angle Regression (LARs) algorithm for semi-supervised feature selection. It involves applying label regression to embed the data into a new space and then uses LARs to select features that align well to the embedded data resulting from label regression. In order to examine the performance of our proposed method (referred to as the HG+Semi), we need to assess the data transformation obtained and its useful information content it. In view of this, we would like to measure the performance of our proposed algorithm using three different indices, namely, (1) **data transformation**, (2) **classification accuracy** and (3) **redundancy rate**. Assume $\mathbf{F}$ is the set of selected features, the redundancy rate can be defined as:

$$RED(F) = \frac{1}{m(m-1)} \sum_{f_i, f_j \in \mathbf{F}, i > j} \rho_{i,j}. \quad (14)$$

where $\rho_{i,j}$ returns the Pearson correlation score between two features f_i and f_j . The measurement assesses the averaged correlation among all feature pairs, and a large value indicates that many selected features are strongly correlated, and thus redundancy is expected to exist in $\mathbf{F}$.

6 Experiments and Comparisons

Data sets: The data sets used to test the performance of our proposed algorithm are publicly available face-recognition benchmarks. Table. 1 summarizes the coverage and properties of the three data-sets. The original images were normalized (in scale and orientation) such that the two eyes were aligned at the same position. Then, the facial areas were cropped to give images for matching. In Fig. 2, we show the closely cropped images and these all contain facial structure. In our experiments, we only examine the performance of our proposed method with 50% data labeled.

Table 1. Summary of benchmark data sets

Data-set	Examples	Features	Classes
ORL	400	1024	40
CMU PIE	1428	1024	68
AR	130	2400	10

(a) ORL dataset

(b) CMU PIE dataset

(c) AR dataset

Fig. 2. The sample cropped face images form three face datasets

The ORL dataset contains 40 distinct individuals with ten images per person. The images are taken at different time instances, and include variations in facial expression and facial detail (glasses/no glasses). All images are resized to 32×32 pixels.

The CMU PIE dataset is a multiview face dataset, consisting of 41,368 images of 68 people. The views cover a wide range of poses from profile to frontal views, varying illumination and expression. In this experiment, we fixed the pose and expression. Thus, for each person, we have 21 images obtained under different lighting conditions. All images are resized to 32×32 pixels.

The AR dataset contains over 4000 face images from 126 people (70 men and 56 women). A subset of 10 subjects and 13 images per subject are selected in our experiments and the images are resized to 60×40 pixels. All the images are frontal view faces with different facial expressions, illumination conditions, and occlusions (sunglasses and scarf).

Data Transformation: we compare the data transformation performance of our proposed method using label regression with alternative methods, including kernel PCA [2], the Laplacian eigenmap [5] and LPP [3]. In order to visualize the results, we have used five randomly selected subjects from each dataset, and these are shown in Fig. 3, Fig. 4 and Fig. 5. In each figure, we have shown the projections onto the leading two most significant eigenmodes from different

Fig. 3. Distribution of samples of five subjects in ORL dataset

spectral embedding methods, ordered according to their eigenvalues. This provides a low-dimensional representation for the images. From the above figures, it is clear that our hypergraph based label regression method demonstrates much more clear cluster structure than that by traditional spectral clustering method. This implies that the hypergraph representation is both more appropriate and more complete in describing feature relations and structures.

Table 2. The best result of all methods and their corresponding size of selected feature subset on the three face datasets

Dataset	MRMR	FS	LS	SPEC	UDFS	HG+Semi
ORL	83.5%(95)	80%(99)	65.25%(99)	64.5%(95)	76.5%(99)	93%(88)
PIE	99.15%(99)	99.17%(100)	71.43%(99)	89.64%(100)	96%(98)	99.19%(58)
AR	88.15%(509)	87.69%(548)	60.77%(591)	86.15%(598)	82.31%(562)	95.38%(427)

Classification Accuracy: In order to explore the discriminative capabilities of the information captured by our method, we use the selected features for further classification. We compare the classification results from our proposed method HG+Semi with five representative feature selection algorithms. For unsupervised

Fig. 4. Distribution of samples of five subjects in CMU PIE dataset

learning, three alternative feature selection algorithms are selected as baselines. These methods are the Laplacian score (referred to as LS) [12], SPEC [11] and UDFS [23]. We also compare our results with two state-of-the-art supervised feature selection methods, namely a) the Fisher score (referred to as FS) [9] and b) the MRMR algorithm [7]. We use 5-fold cross-validation for the SVM classifier on the feature subsets obtained by the feature selection algorithms to verify their classification performance. Here we use the linear SVM with LIBSVM [21].

The classification accuracies obtained with different feature subsets are shown in Fig. 6. From the figure, it is clear that our proposed method HG+Semi is, by and large, superior to the alternative feature selection methods. Specifically, it selects both a smaller and better performing (in terms of classification accuracy) set of discriminative features on the three face data sets. Moreover, HG+Semi rapidly converges, with typically around 30 features (see Fig. 6(a)-(b)). Each of the alternative unsupervised methods, usually require more than 100 features to achieve a comparable result. There are two reasons for this improvement in performance. First, the hypergraph representation is effective in capturing the high-order relations among samples rather than approximating them in terms of pairwise interactions which can lead to a substantial loss of information. Thus the structural information latent in the data can be effectively preserved. Second, the LARs algorithm is applied to select features that align well to the embedded

Fig. 5. Distribution of samples of five subjects in AR dataset

data resulting from label regression. As a result the optimal feature combinations can be located so as to remove redundant features.

Compared with the two state-of-art supervised feature selection algorithms, our proposed semi-supervised method (HG+Semi) outperforms the MRMR algorithm and FS in all cases. On the CMU PIE dataset (see Fig. 6(b)), even though MRMR and FS can give good classification performance when more than 100 features are selected, HG+Semi achieves a comparable result with a much smaller number of features, i.e., less than 20 features. This implies that our proposed method is able to locate both the optimal size of the feature subset and perform accurate classification of the samples based on just a few of the most important features.

The best result for each method together with the corresponding size of the selected feature subset are shown in Table. 2. In the table, the classification accuracy is shown first and the optimal number of features selected is reported in brackets. Overall, HG+Semi achieves the highest degree of dimensionality reduction, i.e. it selects a smaller feature subset compared with those obtained by the alternative methods. For example, in the ORL data set, the best result obtained by the alternative feature selection methods is 83.5% with the MRMR algorithm and 95 features. However, our proposed method (HG+Semi) gives a better accuracy of 93% when only 88 features are used. The results further verify that our feature selection method can guarantee the optimal size of the feature subset, as it not only achieves a higher degree of dimensionality reduction but it

Fig. 6. Accuracy rate vs. the number of selected features on three face datasets

also gives better discriminability. We also observe that the UDFS algorithm gives a better result than the alternative unsupervised methods (i.e. the Laplacian score and the SPEC). The reason for this is that unlike traditional methods which treat each feature individually and which hence are suboptimal, the UDFS method directly optimizes the score over the entire selected feature subset. As a result, a better feature subset can be obtained.

Table 3. Averaged Redundancy rate of Subsets Selected Using Different Algorithms

Dataset	MRMR	FS	LS	SPEC	UDFS	HG+Semi
ORL	1.47	1.72	1.68	1.65	1.62	1.38
PIE	0.51	0.53	0.67	0.66	0.63	0.45
AR	0.56	0.68	0.76	0.70	0.72	0.55

Redundancy Rate: Table. 3 shows a comparison of results from our proposed method to the alternative feature selection methods using the top n features, where n is the number of training data. We chose n , since when the number of selected features is larger than n , any feature can be expressed by a linear combination of the remaining ones, which will introduce unnecessary redundancy in the evaluation stage. In the table, the boldfaced values are the lowest redundancy rates. The subset obtained by our proposed scheme has the least redundancy. This further verifies that our propose algorithm is able to remove redundant features.

The accuracy rate (Table. 2) and redundancy rate (Table. 3) together indicate that HG+Semi both gives the least redundancy, and results in the highest accuracy. They also underline the necessity of removing redundant features for improving learning performance. It should also be observed that the MRMR algorithm also produces low redundancy rates. However, it does not perform as well in terms of classification accuracy. This can be explained by the observation that in MRMR, feature contributions to the classification process are considered individually by evaluating the correlation between each feature and the class label. However, the class label may be jointly determined by a set of features. This interaction among features is not considered by MRMR.

7 Conclusion

In this paper, we have presented an semi-supervised feature selection method based on hypergraph subspace learning. The proposed feature selection method offers two major advantages. The first is that by incorporating MII for higher order similarities measure, we establish a novel hypergraph framework which is used for characterizing the multiple relationships within a set of samples. Thus, the structural information latent in the data can be more effectively modeled. Secondly, we derive a hypergraph subspace learning view of semi-supervised feature selection which casting 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. These properties enable our method to be able to handle feature redundancy effectively.

References

1. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. In: Proc. Syst. (1993)
2. Scholkopf, B., Smola, A., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. In: Neural Computation, pp. 1299–1319 (1998)
3. He, X., Niyogi, P.: Locality preserving projections. Advances in Neural Information Processing Systems (2004)
4. He, X., Cai, D., Yan, S., Zhang, H.J.: Neighborhood preserving embedding. In: Tenth IEEE International Conference on Computer Vision, pp. 1208–1213 (2005)
5. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in Neural Information Processing Systems, pp. 585–592 (2002)
6. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks, 537–550 (2002)
7. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1226–1238 (2005)
8. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. Wiley, New York (2001)
9. Nie, F., Xiang, S., Jia, Y., Zhang, C., Yan, S.: Trace ratio criterion for feature selection. In: Proceedings of the 23rd National Conference on Artificial Intelligence, pp. 671–676 (2008)
10. Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. In: Machine Learning, pp. 23–69 (2003)
11. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: Proceedings of the 24th International Conference on Machine Learning, pp. 1151–1157 (2007)
12. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: Advances in Neural Information Processing Systems (2005)
13. Belhumeur, P.N., Kriegman, D.J.: What is the set of images of an object under all possible illumination conditions? International Journal of Computer Vision, 245–260 (1998)

14. Agarwal, S., Branson, K., Belongie, S.: Higher order learning with graphs. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 17–24 (2006)
15. Agarwal, S., Lim, J., Zelnik-Manor, L., Perona, P., Kriegman, D., Belongie, S.: Beyond pairwise clustering. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 838–845 (2005)
16. Chung, F.: The Laplacian of a hypergraph. AMS DIMACS Series in Discrete Mathematics and Theoretical Computer Science, pp. 21–36 (1993)
17. Li, W.C.W., Solé, P.: Spectra of regular graphs and hypergraphs and orthogonal polynomials. *European Journal of Combinatorics*, 461–477 (1996)
18. Seeger, M.W.: Bayesian inference and optimal design for the sparse linear model. *The Journal of Machine Learning Research*, 759–813 (2008)
19. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. In: *Machine Learning*, pp. 243–272 (2008)
20. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. In: *The Annals of Statistics*, pp. 407–499 (2004)
21. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001)
22. Cai, D., Zhang, C., He, X.: Unsupervised feature selection for multi-cluster data. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 333–342 (2010)
23. Yang, Y., Shen, H.T., Ma, Z., Huang, Z., Zhou, X.: L21-norm regularized discriminative feature selection for unsupervised learning. In: *International Joint Conferences on Artificial Intelligence*, pp. 1589–1594 (2011)
24. Jacobs, D.W., Belhumeur, P.N., Basri, R.: Comparing images under variable illumination. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 610–617 (1998)
25. Nie, F., Xiang, S., Liu, Y., Zhang, C.: A general graph-based semi-supervised learning with novel class discovery. In: *Neural Computing & Applications*, pp. 549–555 (2010)
26. Zhang, Z., Hancock, E.R.: Feature selection for gender classification. In: *5th Iberian Conference on Pattern Recognition and Image Analysis*, pp. 76–83 (2011)
27. Zhang, Z., Hancock, E.R.: Hypergraph based Information-theoretic Feature Selection. *Pattern Recognition Letters* (2012)
28. Yang, A.Y., Wright, J., Ma, Y., Sastry, S.S.: Feature selection in face recognition: A sparse representation perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2007)
29. Nie, F., Xu, D., Li, X., Xiang, S.: Semi-supervised dimensionality reduction and classification through virtual label regression. *IEEE Transactions on Systems, Man, and Cybernetics*, 1–11 (2011)
30. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM Review*, 455–500 (2009)
31. Shashua, A., Zass, R., Hazan, T.: Multi-way Clustering Using Super-Symmetric Non-negative Tensor Factorization. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006*. LNCS, vol. 3954, pp. 595–608. Springer, Heidelberg (2006)
32. Björck, A.: Numerical methods for least squares problems. In: *Proc. SIAM* (1996)