ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Spectral clustering and discriminant analysis for unsupervised feature selection

Xiucai Ye, Kaiyang Ji and Tetsuya Sakurai

Department of Computer Science, University of Tsukuba
Tsukuba, Japan

Abstract. In this paper, we propose a novel method for unsupervised feature selection, which utilizes spectral clustering and discriminant analysis to learn the cluster labels of data. During the learning of cluster labels, feature selection is performed simultaneously. By imposing row sparsity on the transformation matrix, the proposed method optimizes for selecting the most discriminative features which better capture both the global and local structure of data. We develop an iterative algorithm to effectively solve the optimization problem in our method. Experimental results on different real-world data demonstrate the effectiveness of the proposed method.

1 Introduction

Feature selection is an efficient technique for data dimension reduction in machine learning and data mining, which brings the immediate effects for applications including: speeding up the algorithms, reducing the risk of over fitting, and improving the accuracy of the predictive results [1].

Unsupervised feature selection has attracted increasing attention in recent years. Without cluster labels, unsupervised feature selection extracts features that effectively maintain the important underlying structure of data, such as the global structure and the local structure. The Maximum Variance method and the global pairwise similarity method [2] select features by preserving the global structure of data. While, the Laplacian Score (i.e., LS) method [3] and the Multi-Cluster Feature Selection (i.e., MCFS) method [4] exploit the local data structure to conduct feature selection.

Since unsupervised feature selection lacks the label information, many methods select features by learning the cluster labels. Spectral clustering is an efficient method which aims to find the optimal partitions among different clusters. Li et al. [5] performed spectral clustering to learn the cluster labels, meanwhile, feature selection was performed to select a better feature subset. Instead of spectral clustering, Yang et al. [6] utilized discriminant analysis to joint learning of the cluster labels with unsupervised feature selection. Discriminant analysis is important to feature selection, which aims to select the discriminative features.

In this paper, we utilize both spectral clustering and discriminant analysis to learn the cluster labels of data, during which feature selection is performed simultaneously. The global structure of data is captured by the discriminant analysis, while the local geometric structure is revealed by spectral clustering.

Our proposed method is referred to as SDFS. SDFS optimizes for selecting the most discriminative features which can better capture both the global and local data structure. We develop an iterative algorithm to effectively solve the optimization problem. Many experimental results are provided for demonstration.

2 The Proposed Method

In this paper, we use x_1 , ..., x_n to denote the n unlabeled data samples, $x_i \in \mathbb{R}^m$ and $X = [x_1, ..., x_n] \in \mathbb{R}^{m\times n}$ is the data matrix. Let $\{f_1, ..., f_m\}$ be the set of features where m is the number of features. Feature selection is to select d features form $f_1, ..., f_m$ to represent the original data, where $d < m$. We use I to denote the identity matrix, and let $1_n \in \mathbb{R}^n$ denote a column vector with all of its elements being 1. The centering matrix is $H_n = I - \frac{1}{n}1_n 1_n^T$.

Consider that x_1 , ..., x_n are sampled from c clusters. Let $Y = [y_1, ..., y_n]^T \in \{0, 1\}^{n\times c}$ denote the label matrix, where $y_i \in \{0, 1\}^{c\times 1}$ is the label vector of x_i. The j^{th} element of y_i is 1 if x_i is in the j^{th} cluster, and 0 otherwise. The scaled cluster indicator matrix F is defined as $F = [F_1, ..., F_n]^T = Y(Y^TY)^{-1/2}$.

2.1 Local structure learning

The proposed SDFS method utilizes spectral clustering to learn the scaled cluster indicator matrix F, which aims to preserve the local data structure. The objective function of spectral clustering can be formulated as

$$\min_F Tr(F^TLF), \qquad s.t. \quad F = Y(Y^TY)^{-1/2}, \tag{1}$$

where $Tr(\cdot)$ is the trace operator and L is a Laplacian matrix constructed based on local data structure using different methods. In this paper, we utilize the k-nearest neighbor graph to construct the normalized Laplacian matrix L. The similarity matrix S with the pairwise similarity S_{ij} as its entries is calculated as $S_{ij} = \begin{cases} \exp(-\frac{\|x_i-x_j\|^2}{2\sigma^2}), & x_i \text{ and } x_j \text{ are } k \text{ nearest neighbors,} \\ 0, & \text{otherwise,} \end{cases}$ where σ is the bandwidth parameter. Let D denote a $n \times n$ diagonal matrix with $D_{ii} = \sum_{j=1}^n S_{ij}$ on the diagonal. The normalized Laplacian matrix L is defined as $L = I - D^{-1/2}SD^{-1/2}$.

2.2 Global structure learning

In SDFS, linear discriminant analysis is utilized in the learning process to capture the global data structure. Following [7], the total scatter matrix is defined as $S_t = \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T = \tilde{X}\tilde{X}^T$ and the between-cluster scatter matrix is defined as $S_b = \sum_{i=1}^c n_i(\mu_i - \mu)(\mu_i - \mu)^T = \tilde{X}FF^T\tilde{X}^T$, where μ is the mean of all data, μ_i is the mean of data in the i^{th} cluster, n_i is the number of data in the i^{th} cluster, and $\tilde{X} = XH_n$ is the data matrix after being centered.

The linear discriminant analysis is to find a linear transformation $W \in \mathbb{R}^{m\times q}$ $(q < m)$ that projects X from m-dimensional space to q-dimensional space. In the lower dimensional space, the within-cluster distance is minimized while the between-cluster distance is maximized as [7]

$$\max_{W} Tr((W^T S_t W)^{-1} W^T S_b W). \tag{2}$$

Since (2) has a trivial solution of all zeros when performing sparsity constraint on W for feature selection, Tao et al. [8] presented the nontrivial solution of (2) by setting $W^T S_t W = I$ for supervised feature selection, which also inherits the merit of selecting the most discriminative features. In the proposed method, we consider the nontrivial solution of (2) as [8]. Furthermore, we also consider to preserve the local structure which is not considered in [8].

2.3 The Objective Function

By incorporating spectral clustering, discriminant analysis and $l_{2,1}$-norm regularization into a framework, the proposed SDFS method is formulated as

$$\begin{aligned} \min_{W,F} Tr(F^T L F) + \alpha(-Tr(W^T S_b W) + \beta\|W\|_{2,1}), \\ s.t. FF^T = I_c, F \geq 0, W^T S_t W = I, \end{aligned} \tag{3}$$

where α and β are two balanced parameters. We relax the condition of $F = Y(Y^T Y)^{-1/2}$ to $FF^T = I_c$ as in [6]. Since the nonnegative constraint of F can help to relieve the deviation from the true solution [5], we constrain F to be nonnegative. To avoid the trivial solution [8], we constrain the transformation matrix W to be uncorrelated with respect to S_t , i.e., $W^T S_t W = I$.

Note that in (3) the term $\|W\|_{2,1}$ is introduced to ensure that W is sparse in rows. Let $W = [w_1, ..., w_n]^T \in \mathbb{R}^{m\times q}$, where w_i is the i^{th} row of W. Since w_i corresponds to the weight of feature f_i, the sparsity constraint on rows makes W suitable for feature selection. After the optimal transformation matrix W is obtained, each feature f_i is ranked according to $\|w_i\|_2$ in descending order and the top d features are selected.

2.4 Optimization

We propose an iterative algorithm, which divides the optimization problem into two steps: learning W while fixing F, and learning F while fixing W.

Since $S_t = \tilde{X}\tilde{X}^T$, $S_b = \tilde{X}FF^T\tilde{X}^T$ and $FF^T = I_c$, we rewrite the objective function of SDFS as follows.

$$\begin{aligned} \min_{W,F} Tr(F^T L F) + \alpha(-Tr(W^T \tilde{X}FF^T\tilde{X}^T W) + \beta\|W\|_{2,1}) + \frac{\gamma}{2}\|F^T F - I_c\|_F^2, \\ s.t. F \geq 0, W^T \tilde{X}\tilde{X}^T W = I, \end{aligned} \tag{4}$$

where $\gamma > 0$ is a parameter to ensure the orthogonality.

When F is fixed, we need to solve the following problem.

$$\min_{W} -Tr(W^T \tilde{X}FF^T\tilde{X}^T W) + \beta\|W\|_{2,1}, \qquad s.t. \quad W^T \tilde{X}\tilde{X}^T W = I. \tag{5}$$

Since $\frac{\partial \|W\|_{2,1}}{\partial W} = 2UW$ where $U \in \mathbb{R}^{m\times m}$ is a diagonal matrix with the i^{th} diagonal element as $U_{ii} = \frac{1}{2\|w_i\|_2}$, by constructing an auxiliary function, we rewrite (5) as

$$\min_{W} Tr(W^T(-\tilde{X}FF^T\tilde{X}^T + \beta U)W), \qquad s.t. \quad W^T\tilde{X}\tilde{X}^TW = I. \tag{6}$$

The solution of (6) can be obtained by solving the following generalized eigenproblem:

$$(-\tilde{X}FF^T\tilde{X}^T + \beta U)\tilde{w} = \lambda\tilde{X}\tilde{X}^T\tilde{w}. \tag{7}$$

The matrix $W \in \mathbb{R}^{m\times q}$, containing the eigenvectors corresponding to the q smallest eigenvalues as the column vectors, is the solution of (6). Then, we normalize W such that $(W^T\tilde{X}\tilde{X}^TW)_{ii} = 1$, $i = 1, ..., q$.

Next, when W is fixed, we need to solve the following problem.

$$\min_{F} Tr(F^TLF) - \alpha Tr(W^T\tilde{X}FF^T\tilde{X}^TW) + \frac{\gamma}{2}\|F^TF - I_c\|_F^2, \quad s.t.F \geq 0. \tag{8}$$

Since $Tr(W^T\tilde{X}FF^T\tilde{X}^TW) = Tr(F^T\tilde{X}^TWW^T\tilde{X}F)$, let $M = L - \alpha\tilde{X}^TWW^T\tilde{X}$, (8) can be rewritten as

$$\min_{F} Tr(F^TMF) + \frac{\gamma}{2}\|F^TF - I_c\|_F^2, \qquad s.t. \quad s.t.F \geq 0. \tag{9}$$

Following [5], we update F by multiplicative rules, as

$$F_{ij} \leftarrow F_{ij}\frac{(\gamma F)_{ij}}{(MF + \gamma FF^TF)_{ij}}. \tag{10}$$

Then, we normalize F such that $(F^TF)_{ii} = 1$, $i = 1, ..., n$.

In summary, we solve the optimization problem in (4) in an alternative way. We first construct the k-nearest neighbor graph and calculate L. We initialize $F \in \mathbb{R}^{n\times c}$ and set $U \in \mathbb{R}^{m\times m}$ as an identity matrix. W is calculated according to the generalized eigenproblem in (7). Then, F is updated according to (10) and U is updated by setting $U_{ii} = \frac{1}{2\|w_i\|_2}$. After that, W is updated again according to (7). This updating process is continued until (4) is convergent. To optimize the objective function of SDFS, the most time consuming operation is to solve the generalized eigenproblem in (7). The time complexity of the operation is $O(m^3)$ approximately. Empirical results show that the convergence is fast and only several iterations (less than 10 iterations in the presented datasets) are needed to converge. Thus, the proposed method scales well in practice.

3 Experiments

In our experiments, we use a diversity of six public datasets to compare the performance of different unsupervised feature selection methods. Their data properties are summarized in Table 1. We compare the proposed method with several well-known unsupervised feature selection methods, including LS [3], MCFS [4], UDFS [6], and NDFS [5]. We also compare these feature selection methods with the baseline method, which uses all the features for clustering. The number of nearest neighbors is set as $k = 5$, which is similar to [3, 4, 6, 5]. The parameters is tuned from $\{10^{-6}, 10^{-4}, 10^{2}, 1, 10^{2}, 10^{4}, 10^{6}\}$. The number of selected features is ranged from $\{50, 100, 150, 200, 250, 300\}$. We report the best result of all the methods by using different parameters. We first perform each feature selection method to select features and then perform K-means based on the selected features. Two widely used evaluation metrics, i.e., Accuracy (ACC) and Normalized Mutual Information (NMI) [9], are applied to evaluate the clustering results. We repeat the clustering 20 times with random initializations and report the average results.

Dataset	# of samples	# of Features	# of Clusters
UMIST [10]	575	644	20
JAFFE[11]	213	676	10
BA[12]	1404	320	36
MNIST[13]	2000	784	10
Isolet1[13]	1560	617	26
COIL20[13]	1440	1024	20

Table 1: Properties of Datasets.

We summarize the clustering results on the six datasets in Table 2 and Table 3. Most of the unsupervised feature selection methods performs better than the baseline method. The LS method can not improve the accuracy of clustering results for many datasets. On most of the datasets, NDFS, UDFS and SDFS perform better than MCFS. Both NDFS and SDFS apply spectral clustering for feature selection, which results in more accurate clustering than other methods on most of the data sets. The proposed SDFS method obtains best performance on all the six datasets. That is because SDFS utilizes spectral clustering and discriminant analysis simultaneously, which is able to select the most discriminative features to better capture both the global and local structure of data.

4 Conclusion

In this paper, we propose a novel unsupervised feature selection method, which incorporates spectral clustering, discriminant analysis and $l_{2,1}$- norm regularization into a joint framework. We derive an efficient algorithm to solve the optimization problem of the proposed method. Experiments on various types of datasets demonstrate the advantages of the proposed method.

Acknowledgments

This work was supported in part by JST/CREST and MEXT KAKENHI (Grant No.25286097).

Dataset	UMIST	JAFFE	BA	MNIST	Isolet1	COIL20
Baseline	64.1±5.2	73.5±5.8	56.0±2.0	47.7±2.5	71.3±3.0	73.0±4.3
LS	60.2±4.8	72.8±6.2	56.3±1.8	48.2±2.8	61.7±2.8	65.9±4.4
MCFS	64.8±4.6	76.2±4.4	56.5±1.8	50.8± 2.3	70.5±3.2	68.2±4.5
UDFS	65.0±4.9	75.3±4.6	57.7±1.5	51.2± 2.1	66.8±3.7	72.4±4.1
NDFS	65.2±4.8	77.4±4.0	58.0±1.8	51.8± 2.0	71.7±2.8	72.6±4.4
SDFS	65.6±5.0	78.2±4.2	58.9±2.0	52.6± 2.2	72.9±2.5	73.6±4.0

Table 2: Clustering Results (NMI % ± std) of Different Feature Selection Methods.

Dataset	UMIST	JAFFE	BA	MNIST	Isolet1	COIL20
Baseline	43.0±3.7	68.2±6.5	38.5±3.1	52.4±5.0	55.4±3.5	57.5±3.2
LS	40.2±3.8	69.5±6.4	40.6±2.9	55.2±4.8	48.3±3.6	45.8±6.2
MCFS	42.8±3.6	72.4±5.8	41.2±2.8	56.8± 4.3	53.8±4.2	52.2±5.2
UDFS	44.5±3.5	71.2±6.2	42.2±2.6	57.6± 4.0	50.7±5.0	56.2±3.8
NDFS	45.1±3.2	72.5±5.2	42.9±2.8	58.2± 3.5	56.5±3.4	57.6±4.0
SDFS	45.6±3.0	73.2±5.1	43.8±2.5	59.1± 3.8	57.3±3.0	58.2±3.6

Table 3: Clustering Results (ACC % ± std) of Different Feature Selection Methods.

References

[1] J. G. Dy and C. E. Brodley. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5:845–889, 2004.

[2] J. Zhang J. Yin X. Liu, L. Wang and H. Liu. Global and local structure preservation for feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, 25:1083–1095, 2014.

[3] D. Cai X. He and P. Niyogi. Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514, 2006.

[4] C. Zhang D. Cai and X. He. Unsupervised feature selection for multi-cluster data. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 333–342, 2010.

[5] J. Liu X. Zhou Z. Li, Y. Yang and H. Lu. Unsupervised feature selection using nonnegative spectral analysis. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1026–1032, 2012.

[6] Z. Ma Z. Huang andX. Zhou Y. Yang, H. Shen. L21-norm regularized discriminative feature selection for unsupervised learning. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1589–1594, 2011.

[7] K. Fukunaga. Introduction to statistical pattern recognition. *Academic press*, 2013.

[8] F. Nie Y. Jiao H. Tao, C. Hou and D. Yi. Effective discriminative feature selection with nontrivial solution. *IEEE Transactions on Neural Networks and Learning Systems*, 2015.

[9] A. Strehl and J. Ghosh. Cluster ensembles–a knowledge reuse framework for combining multiple partitions,. *Journal of Machine Learning Research*, 3:583?617, 2002.

[10] http://www.sheffield.ac.uk/eee/research/iel/research/face.

[11] http://www.kasrl.org/jaffe.html.

[12] http://www.cs.nyu.edu/ roweis/data.html.

[13] http://www.cad.zju.edu.cn/home/dengcai/data/facedata.html.