Yan *Applied Informatics* (2015) 2:8
DOI 10.1186/s40535-015-0009-3

Applied Informatics
a SpringerOpen Journal

RESEARCH Open Access

Sparsity preserving score for feature selection

Hui Yan

Correspondence:
yanhui@njust.edu.cn
School of Computer Science and Engineering, Nanjing University of Science and Technology, 210094, China

Abstract

Compared with supervised feature selection, selecting features in unsupervised learning scenarios is a much harder problem due to the lack of label information. In this paper, we propose sparsity preserving score (SPS) for unsupervised feature selection based on recent advances in sparse representation technique. SPS evaluates the importance of a feature by its power of sparse reconstructive relationship preserving. Specially, SPS selects features that minimize reconstruction residual based on sparse representation in the space of selected features. SPS aims to jointly select features by transforming data from a high-dimensional space of original features to a low-dimensional space of selected features through a special binary feature selection matrix. When the sparse representation is fixed, our searching strategy is an essentially discrete optimization and our theoretical analysis guarantees our objective function can be easily solved with a closed-form solution. The experimental results on two face data sets demonstrate the effectiveness and efficiency of our algorithm.

Keywords: Feature selection; Sparse representation; Binary matrix

Introduction

In many areas, such as text processing, biological information analysis, and combinatorial chemistry, data are often represented as high-dimensional feature vectors, but often only a small subset of features is necessary for subsequent learning and classification tasks. Thus, dimensionality reduction is preferred, which can be achieved by either feature selection or feature extraction (Guyon & Elisseeff 2003) to a low dimensional space. In contrast to feature extraction, feature selection aims at finding out the most representative or discriminative subset of the original feature spaces according to some criteria and maintains the original representation of features. During recent years, feature selection has attracted much research attention and widely used in a variety of applications (Yu et al. 2014; Ma et al. 2012b).

According to the availability of labels of training data, feature selection can be classified into supervised feature selection (Kira et al. 1992; Nie et al. 2008; Zhao et al. 2010) and unsupervised feature selection (He et al. 2005; Zhao & Liu 2007), (Yang et al. 2011; Peng et al. 2005). Supervised feature selection selects features according to label information of each training data. Unsupervised methods, however, are not able to obtain label information directly, and they frequently select the features which best preserve the data similarity or manifold structure of data.

Feature selection mainly focuses on search strategies and measurement criteria. The search strategies for feature selection can be divided into three categories: exhaustive search, sequential search, and random search. The exhaustive search aims to find out the optimal solution from all possible subsets. However, it is NP-hard and thus it is impractical to run. Sequential search methods, such as sequential forward selection and sequential backward elimination (Kohavi & John 1997), start from an empty set or the set of all candidates as the initial subset selected and successively add features to the selected feature or eliminate features from a subset one by one. The major drawback of the traditional sequential search methods relies heavily on search routes. Although the sequential methods do not guarantee the global optimality of selected subset, they have been widely used because of their simplicity and relatively low computational cost even for large-scale data. Plus-l-minus-r (l-r) (Devijver 1982), a slightly more reliable sequential search method, considers deleting features that were previously selected and selecting features that were previously deleted. However, it only partially solves the limit of search routes and brings in additional parameters. The random search methods, such as the random hill climbing and its extension sequential floating search (Jain & Zongker 1997), take advantage of randomized steps of the search and select features from all candidates with a chance probability per feature.

Measurement criterion is also an important research direction in feature selection. Data variance (Duda et al. 2001) ranks the score of each feature by the variance along a dimension. The measurement criterion of data variance finds features that are useful for representing data; however, these features may not be useful for preserving discriminative information. Laplacian score (He et al. 2005) is a recent locality graph-based unsupervised feature selection algorithm. Laplacian score reflects locality preserving power of each feature.

Recently, Wright et al. present a Sparse Representation-based Classification (SRC) (Wright et al. 2009) method. Afterwards, sparse representation-based feature extraction becomes an active direction. Qiao et al. (2010) present a Sparsity Preserving Projections (SPP) method, which aims to preserve the sparse reconstructive relationship of the data. Zhang et al. (2012) recently present a graph optimization for dimensionality reduction with sparsity constrains, which can be viewed as an extension of SPP. Clemmensen et al. (2011) provide a sparse linear discriminant analysis with a sparseness constraint on projection vectors.

As we know, feature selection with direct connection to SRC has not emerged. In this paper, we use SRC as a measurement criterion to design an unsupervised feature selection algorithm called sparsity preserving score (SPS). The formulated objective function, which is an essentially discrete optimization, aims to seek a binary linear transformation such that in a low-dimensional space the sparse representation coefficients are preserved. As the sparse representation is fixed, our theoretical analysis guarantees our objective function can be easily solved with a closed form, which is optimal solution. SPS simply ranks the score of each feature by Frobenius norm of sparse linear reconstruction residual in the space of selected features.

Background

Unsupervised feature selection criterion

Let $x_i \in R^{m \times 1}$ be the i th training sample and $X = [x_1, x_2, \ldots, x_N] \in R^{m \times N}$ be a matrix composed of entire training samples. The unsupervised criterion to select m' $(m' < m)$ features is defined as

$$\min_A \text{loss}\left(X, XU^A\right) + \mu\Omega\left(U^A\right)$$

where A is the set of the indices of selected features, U^A is the corresponding $m \times m$-sized feature selection matrix, and XU^A is reconstruction of the reduced space in $R^{m' \times N}$ to the original space in $R^{m \times N}$. loss$(\cdot)$ is the loss function, and $\mu\Omega(U^A)$ is the regularization with μ as its parameter.

Sparse representation

Given a test sample y, we represent y in an overcomplete dictionary whose basis vectors are training sample themselves, i.e., $y = X\beta$. If the system of linear equation is underdetermined, this representation is naturally sparse. The sparsest solution can be sought by solving the following l_1 optimization problem (Donoho 2006; Cands et al. 2006):

$$\hat{\beta} = \arg\min_\beta ||\beta||_1, s.t., y = X\beta \tag{1}$$

This problem can be solved in polynomial time by standard linear programming algorithms (Chen et al. 2001).

Methods

We formulate our strategy to select $n(n < m)$ features as follows: given a set of unlabeled training samples $x_i \in R^{m \times 1}$, $i = 1,.., N$, learn a feature selection matrix $P \in R^{m \times n}$ such that P is optimal according to our objective function. For the task of feature selection, P is required to be a special 0–1 binary matrix which satisfies two constraints: (1) each row of P has one and only one non-zero entry of 1 and (2) each column of P has at most one non-zero entry. Accordingly, the sum of entries in each row equals 1 and the sum of entries in each column less than or equals 1. For test, $x_i' = U^T x_i$ is the new representation of χ_i where $x'_i(k) = x_i(k)$ if the kth feature is selected, and otherwise $x_i'(k) = 0$.

We define the following objective function to minimize the sparse linear reconstruction residual and measure the sparsity by the l_1 -×norm of coefficients.

$$\min_{P,\{\beta_i, i=1,\dots,N\}} \begin{array}{ll} J(P,\beta_i) := & \sum_{i=1}^N ||Px_i - PD_i\beta_i||_F^2 + \lambda||\beta_i||_1 \\ s.t., & \sum_{j=1}^m P(i,j) = 1 \\ & \sum_{i=1}^n P(i,j) \leq 1, \\ & P(i,j) = 0 \text{ or } 1 \end{array} \tag{2}$$

Here, $D_i = [x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N] \in R^{m \times (N-1)}$ is the collection of training samples without the ith sample, β_i is the sparse representation coefficient vector of χ_i over D_i, and λ is a scalar parameter. The items in line 1 of (2) are approximation and sparse constraints in the features selected space, respectively. (2) is a joint optimization of P and β_i ($i = 1, \dots, N$).

Since P and β_i ($i = 1,..,N$) are dependent on each other, this problem cannot be solved directly. We update the variables alternately with others fixed.

By fixing β_i ($i = 1,..,N$), removing terms irrelevant to P and rewriting the first term in (2) in a matrix form, the optimization problem (2) is reduced to

$$\begin{aligned} &\min_P \operatorname{trace}\left\{P\Gamma\Gamma^T P^T\right\} \\ &s.t., \ \sum\nolimits_{j=1}^{m} P(i,j) = 1 \\ &\qquad \sum\nolimits_{i=1}^{n} P(i,j) \leq 1 \\ &\qquad P(i,j) = 0 \ \text{ or} \, 1 \end{aligned} \tag{3}$$

where $\Gamma = [\gamma_1, \ldots, \gamma_N]$, and $\gamma_i = x_i - D_i\beta_i$.

Under the constraints in (3), we suppose
$P(i, k_i) = 1$, then

$$\operatorname{trace}\left\{P\Gamma\Gamma^T P^T\right\} \begin{aligned} &= \sum\nolimits_{i=1}^{m} P(i,:)\Gamma\Gamma^T P^T(i,:) \\ &= \sum\nolimits_{i=1}^{m} \{P(i,:)\Gamma\}\{P(i,:)\Gamma\}^T\} \\ &= \sum\nolimits_{i=1}^{m} \sum\nolimits_{j=1}^{N} \{\Gamma(k_i,j)\}^2 \end{aligned} \tag{4}$$

The optimization problem in (3) is converted into computing the sparsity preserving score of each feature, which is defined as

$$\operatorname{Score}(i) = \sum\nolimits_{j=1}^{N} \{\Gamma(k_i,j)\}^2, i = 1, \ldots, m \tag{5}$$

And then we rank and select the n smallest ones from Score(i), $i = 1, \ldots, m$. Without loss of generality, suppose the n selected features are indexed by $k_i^*, i = 1, \ldots, n$. We can construct the matrix P as

$$P(i,j) = \begin{cases} 1, & j = k_i^* \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

By fixing P, removing terms irrelevant to β_i ($i = 1, \ldots,N$), the optimization problem (3) is reduced to the following l_1 optimization problem

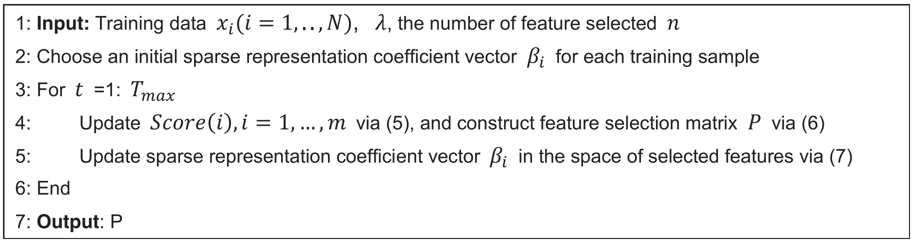
$$\min_{\{\beta_i, i=1,\ldots,N\}} \sum\nolimits_{i=1}^{N} ||Px_i - PD_i\beta_i|| \ {}_F^2 + \lambda ||\beta_i|| \ {}_1 \tag{7}$$

The iterative procedure is given in Algorithm 1. The initial solution of β_i can be calculated directly in the original space of selected features, and it can be used as a good initial solution of the iterative algorithm (Yang et al. 2013).

Note that since the P obtained via the first iteration is 0–1 matrix, some values of features (corresponding to $j \neq k_i^*$) are equal to zero in the second iteration. Thus, it is meaningless to compute the coefficient vector β_i for features whose values are equal to zero. In other words, P becomes a stable value after the first iteration. Thus, we give non-iterative version of Algorithm 1, i.e., Algorithm 2, where we compute β_i in the original space as

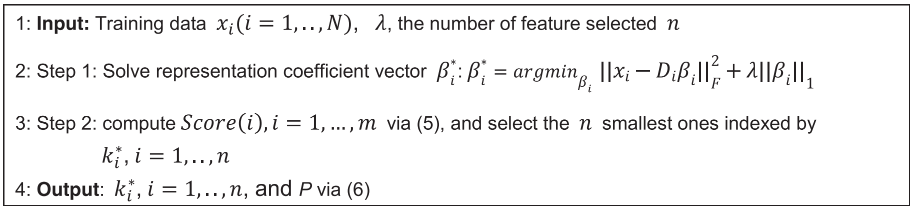
$$\min_{\beta_i} ||x_i - D_i\beta_i|| \ {}_F^2 + \lambda ||\beta_i|| \ {}_1 \tag{8}$$

Some standard convex optimization techniques or TNIPM in (Kim et al. 2007) can be used to solve β_i. In our experiments, we directly use source code provided by authors in (Kim et al. 2007).



```
Input: Training data $x_i(i = 1,..,N)$, $\lambda$, the number of feature selected $n$
Choose an initial sparse representation coefficient vector $\beta_i$ for each training sample
For $t$ =1: $T_{max}$
    Update $Score(i), i = 1, \ldots, m$ via (5), and construct feature selection matrix $P$ via (6)
    Update sparse representation coefficient vector $\beta_i$ in the space of selected features via (7)
End
Output: P
```

Algorithm 1: Iterative procedure for sparsity preserving score



```
Input: Training data $x_i(i = 1,..,N)$, $\lambda$, the number of feature selected $n$
Step 1: Solve representation coefficient vector $\beta_i^*$: $\beta_i^* = argmin_{\beta_i} ||x_i - D_i\beta_i||_F^2 + \lambda||\beta_i||_1$
Step 2: compute $Score(i), i = 1, \ldots, m$ via (5), and select the $n$ smallest ones indexed by $k_i^*, i = 1,..,n$
Output: $k_i^*, i = 1,..,n$, and $P$ via (6)
```

Algorithm 2: Non-iterative procedure for sparsity preserving score

Results and discussion

Several experiments on Yale and ORL face datasets are carried out to demonstrate the efficiency and effectiveness of our algorithm. In our experiments, all samples are not pre-processed. Our algorithm is an unsupervised method, and thus, we compare our Algorithm 2 with other four representative unsupervised feature selection algorithms including data variance, Laplacian score, feature selection for multi-cluster data (MCFS) (Cai et al. 2010), and spectral feature selection (SPEC) (Zhao & Liu 2007) with all the eigenvectors of the graph Laplacian. In all the tests, the number of the nearest neighbors in Laplacian score, MCFS, and SPEC is taken to be half of the number of training images per person.

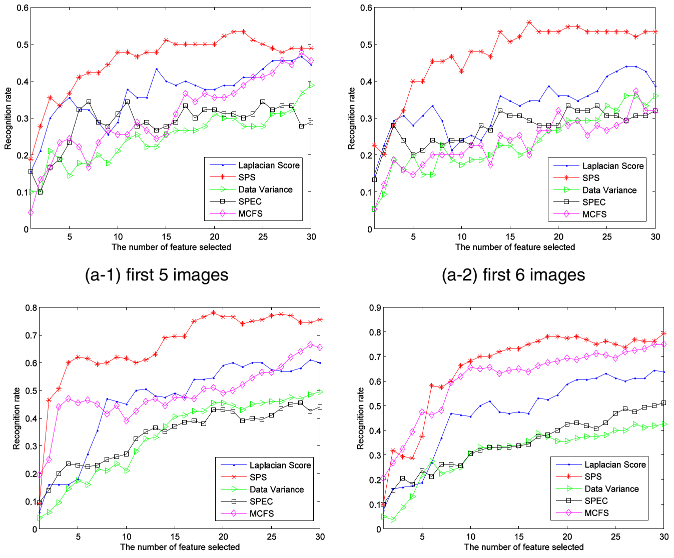
For both datasets, we choose the first five and six images, respectively, per person for training and the rest for testing. After feature selection, the recognition is performed by the "L2"-distance based 1-nearest neighbor classifier. Table 1 reports the top performance as well as the corresponding number of features selected, and Fig. 1 illustrates the recognition rate as a function of the number of features selected. As shown in Table 1, our algorithm reaches the highest or comparable recognition rate at the lowest dimension of feature selected space. From Fig. 1, we can see that with only a very small

Table 1 The comparison of the top recognition rates and the corresponding number of features selected

Methods	Training date			
	Yale		ORL	
	5	6	5	6
Data variance	0.6889 (704)	0.6800 (829)	0.9450 (2503)	0.9563 (2112)
Laplacian score	0.7111 (434)	0.7067 (952)	0.9450 (2390)	0.9563 (1901)
MCFS	0.6556 (974)	0.6933 (825)	0.9250 (1593)	0.9500 (588)
SPEC	0.7111 (836)	0.7200 (780)	0.9150 (2563)	0.9500 (2350)
SPS	0.7333 (551)	0.7333 (569)	0.9450 (2355)	0.9563 (1823)

(b) Yale

Methods	5	6	7	8
Data Variance	0.636	0.706	0.790	0.717
Laplacian Score	0.646	0.712	0.789	0.683
MCFS	0.602	0.684	0.783	0.745
SPEC	0.621	0.685	0.762	0.735
SPS	0.669	0.728	0.808	0.756



(a-1) first 5 images

(a-2) first 6 images

(b-1) first 5 images

(b-2) first 6 images

Fig. 1 Recognition results of the feature selection methods with respect to the number of selected features on (*a-1*, *a-2*) Yale and (*b-1*, *b-2*) ORL

number of features, SPS can achieve significant better recognition rates than the other methods. It can be interpreted from two aspects: (1) SPS jointly selects features and obtain the optimal solution of a binary transformation matrix, while the other methods only add features one by one. Thus, SPS considers the interaction and dependency among features. (2) Features selected with sparse reconstructive relationship preserving are capable of enhancing recognition performance.

Table 2 The comparison of average top recognition rates

Methods	5	6	7	8
(a) ORL				
Data variance	0.970	0.978	0.989	0.980
Laplacian score	0.960	0.976	0.981	0.984
MCFS	0.950	0.958	0.960	0.955
SPEC	0.940	0.947	0.958	0.950
SPS	0.985	0.989	0.993	0.991
Data variance	0.636	0.706	0.790	0.717
Laplacian score	0.646	0.712	0.789	0.683
MCFS	0.602	0.684	0.783	0.745
SPEC	0.621	0.685	0.762	0.735
SPS	0.669	0.728	0.808	0.756

We randomly choose five and six images, respectively, per person for training and the rest for testing. Since the training set is randomly chosen, we repeat this experiment ten times and calculate the average result. The average top performances obtained are reported in Table 2. The results further verify that SPS can select more informative preserving feature subset.

Conclusions

This paper addresses the problem on how to select features with power of sparse reconstructive relationship preserving. In theory, we prove our feature subset is the optimal solution in closed form if the sparse representation vectors are fixed. Experiments are done on the ORL and Yale face image databases, and results demonstrate our proposed sparsity preserving score is more effective than data variance, Laplacian score, MCFS, and SPEC.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements
The authors would like to thank the anonymous reviewers for their constructive advice. This work is supported by the National Natural Science Foundation of China (Grant No. 61202134), Jiangsu Planned Projects for Postdoctoral Research Funds, China Planned Projects for Postdoctoral Research Funds, and National Science Fund for Distinguished Young Scholars (Grant No. 61125305).

Received: 7 April 2015 Accepted: 29 May 2015
Published online: 16 July 2015

References

Cai, D., Zhang, C., He, X.: Unsupervised feature selection for multi-cluster data. In: International Conference on Knowledge Discovery and Data Mining. ACM, Washington, DC, USA (2010)

Cands E, Romberg J, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. Commun Pure Appl Math 59(8):1207–1223

Chen S, Donoho D, Saunders M (2001) Atomic decomposition by basis pursuit. SIAM Rev 43(1):129–159

Clemmensen L, Hastie T, Witten D, Ersboll B (2011) Sparse discriminant analysis. Technometrics 53(4):406–413

Devijver, P. A., Kittler, J (1982) Pattern recognition: a statistical approach. Prentice-Hall, Englewood Cliffs, London

Donoho D (2006) For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. Commun Pure Appl Math 59(6):797–829

Duda R, Hart P, Stork D (2001) Pattern classification. John Wiley Sons, New York

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

He, X., Cai, D., Niyogi, P. : Laplacian score for feature selection. In: Advances in neural information processing systems. MIT Press, Cambridge, MA (2005)

Jain A, Zongker D (1997) Feature selection: evaluation, application, and small sample performance. IEEE J Pattern Analys Machine Intell 19:153–158

Kim SJ, Koh K, Lustig M, Boyd S, Gorinevsky D (2007) A method for largescale l1-regularized least squares. IEEE J Selected Topics Signal Process 1(4):606–617

Kira K, Rendell L (1992) A practical approach to feature selection. In: 9th International Workshop on Machine Learning, San Francisco, Morgan Kaufmann 249-256.

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 92(12):273–324

Ma Z, Nie F, Yang Y, Sebe N (2012b) Web image annotation via subspace-sparsity collaborated feature selection. IEEE Transsact Multimedia 14(4):1021–1030

Nie, F. P., Huang, H., Cai, X., Ding, C. : Efficient and robust feature selection via joint $l_{2,1}$ -norms minimization. In: Advances in neural information processing systems, Vancouver, BC, Canada (2010)

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transact Pattern Analys Machine Intell 27(8):1226–1238

Qiao LS, Chen SC, Tan XY (2010) Sparsity preserving projections with applications to face recognition. Pattern Recogn 43(1):331–341

Wright J, Yang A, Ganesh A, Sastry S, Ma Y (2009) Robust face recognition via sparse representation. IEEE J Pattern Analys Machine Intell 31:210–227

Yang, Y., Shen, H., Ma, Z., Huang, Z., Zhou, X.: $l_{2,1}$ -Norm regularized discriminative feature selection for unsupervised learning. In: International Joint Conferences on Artificial Intelligence. Morgan Kaufmann, San Francisco, USA (2011)

Yang J, Chu D, Zhang L, Xu Y, Yang JY (2013) Sparse representation classifier steered discriminative projection with applications to face recognition. IEEE Transsact Neural Networks Learn Syst 24(7):1023–1035

Yu D, Hu J, Yan H, Yang X, Yang J, Shen H (2014) Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace SVMs ensemble. BMC Bioinformatics 15:297

Zhang LM, Chen S, Qiao L (2012) Graph optimization for dimensionality reduction with sparsity constraints. Pattern Recogn 45(3):1205–1210

Zhao Z, Liu H (2007) Spectral feature selection for supervised and unsupervised learning. In: Proceedings of international conference on machine learning. ACM, New York

Zhao, Z., Wang, L., Liu, H.: Efficient spectral feature selection with minimum redundancy. In: International Joint Conferences on Artificial Intelligence. Morgan Kaufmann, Georgia, USA (2010)