MATEC Web of Conferences **28**, 06003 (2015)
DOI: 10.1051/matecconf/20152806003

Preserving Global and Local Structures for Supervised Dimensionality Reduction

Yinglei Song[1, a], Yongzhong Li[1], Junfeng Qu[2]

[1]School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, 212003, China
[2]Department of Information Technology, Clayton State University, Morrow, GA 30260, USA

Abstract. In this paper, we develop a new approach for dimensionality reduction of labeled data. This approach integrates both global and local structures of data into a new objective, we show that the objective can be optimized by solving an eigenvalue problem. Testing results on benchmark data sets show that this new approach can effectively capture both the crucial global and local structures of data and thus lead to more accurate results for dimensionality reduction than existing approaches.

1 Introduction

Dimensionality reduction has been extensively used in areas including computer vision, data mining, machine learning and pattern recognition [5, 9]. In general, most dimensionality reduction techniques map the high-dimensional data points in a data set into a set of low-dimensional data points while preserving the features that are most important for recognition and classification.

So far, a large number of approaches have been developed for dimensionality reduction. For example, the Principal Component Analysis (PCA) [8] is a traditional unsupervised dimensionality reduction technique. It reduces the dimensionality of data by seeking for linear projections that can maximize the global variance of the projected data points. PCA thus can preserve the maximum amount of the global information of a data set. However, PCA may not be ideal for classification and thus is seldom used on labeled data.

Linear Discriminant Analysis (LDA) is a classical supervised dimensionality reduction technique [7, 8]. To maximize the extent to which data points in different classes are separated from one another in the space of reduced dimensionality, LDA computes the directions along which the ratio of the between-class distance and the within-class distance is maximized. LDA has been extensively used for dimensionality reduction in a variety of applications, such as microarray data analysis and face recognition [1, 6]. However, LDA is only able to capture the global geometric structure of a data set, the local geometric structure might be lost after the dimensionality of data is reduced [3].

Recently, research has shown that local geometric structure is an important feature of a data set and may affect the recognition accuracy of certain classifiers [2, 3, 4]. A variety of approaches thus have been developed to reduce the dimensionality of a data set while preserving its local geometric structure. For example, Laplacian Eigenmaps (LE) [2] and Locality Preserving Projection (LPP) [10] reduce the dimensionality by minimizing an objective defined in terms of the graph Laplacian matrix. Locally Linear Embedding (LLE) [11] models the local geometric structure with linear dependence and data points are mapped into a low-dimensional manifold while preserving the dependence with minimum error. These dimensionality reduction techniques preserve the local geometric structure of a data set. However, important global features might be lost during the process of dimensionality reduction.

Since global and local structures are both important for recognition and classification, a method that can reduce the dimensionality of a data set while preserving both of its global and local structures is thus highly desirable. In this paper, we develop a new approach that considers both of them while reducing the dimensionality of a data set. Our approach develops a new objective that includes both the global and local structure features of a data set. The optimal direction for projection is thus the optimal solution of the objective. We show that the objective can be optimized by solving an eigenvalue problem. We evaluate the effectiveness of this new approach with six benchmark data sets and compare it with both LDA and LPP. Our results show that this new approach outperforms both LDA and LPP in all benchmark data sets. Testing results also suggest that combining both global and local structures retains more important features in a data set and thus is promising for improving the effectiveness of dimensionality reduction.

2 Algorithms and Methods

[a] Corresponding author: syinglei2013@163.com

2.1 Linear Discriminant Analysis

LDA is a supervised dimensionality reduction technique, it reduces the dimensionality of a data set by projecting the data points in the data set to directions that can maximize the between-class distance while minimizing the within-class distance. Given a l dimensional data set $D = \{d_1, d_2, ..., d_n\}$, LDA computes a linear transformation $W \in R^{r \times l}$ of the data points in D, the projected data point y_i for $d_i \in D$ can thus be computed as follows.

$$y_i = Wd_i \tag{1}$$

We assume D are labelled into k disjoint classes $C_1, C_2, ..., C_k$, where class C_i contains n_i data points. Two scatter matrices are defined as follows

$$S_w = \frac{1}{n}\sum_{i=1}^{k}\sum_{d \in C_i}(d - m_i)(d - m_i)^T \tag{2}$$

$$S_b = \frac{1}{n}\sum_{i=1}^{k} n_i (m - m_i)(m - m_i)^T \tag{3}$$

where the second summation in equation (2) is over all data points in class C_i, m_i denotes the centroid of all data points in C_i. m is the centroid of D.

The objective of LDA is to maximize the ratio of the between-class distance and the within-class distance after the transformation is applied to the data points in D, which is to solve the following optimization problem.

$$W_p = \arg\max\{tr((WS_wW^T)^{-1}(WS_bW^T))\} \tag{4}$$

It has been shown that, in cases where S_w is nonsingular, the objective in equation (4) can be minimized by solving the following eigenvalue problem [8].

$$S_w^{-1}S_b w = \lambda w \tag{5}$$

Specifically, the r rows of W are the eigenvectors of $S_w^{-1}S_b$ that correspond to the r largest eigenvalues of $S_w^{-1}S_b$. Since the rank of matrix S_b is at most $k-1$, $S_w^{-1}S_b$ has at most $k-1$ nonzero eigenvalues. The dimensionality of the data points generated by LDA is thus at most $k-1$.

2.2 Locally Linearly Embedding

Given a data set $D = \{d_1, d_2, d_3, ..., d_n\}$, LLE asumes each data point and its neighbors lie on a locally linear patch of a manifold. In other words, each data point d_i can be approximately written as a linear combination of its k nearest neighbors in D, as shown in equation (6).

$$d_i \approx \sum_{j \in N_i} W_{ij} d_j \tag{6}$$

where N_i is the set of k nearest neighbors of d_i in D and W_{ij} is the relative weight associated with its neighbor d_j. For each given i, the values of the relative weights W_{ij} must satisfy the constraint as described in equation (7).

$$\sum_{j \in N_i} W_{ij} = 1 \tag{7}$$

W_{ij} 's can be computed by minimizing the total amount of error due to this approximation under the constraint described in (7), the total amount of error is shown in equation (8).

$$e(W) = \sum_{i=1}^{n}(d_i - \sum_{j \in N_i} W_{ij} d_j)^2 \tag{8}$$

The values of can then be computed based on a linear equation set of k equations as shown in equation (9)

$$\sum_{j \in N} C_{ij} W_{ij} = 1 \tag{9}$$

where C_{ij} is the number on the i th row and j th column of the covariance matrix C of the data set. After solving the set of linear equations, the solutions can be normalized to satisfy the constraints described in equation (7). The computed relative weights W_{ij} 's thus collectively describe the local genometric structure of the data set.

Assume that, after the dimension is reduced, we obtain a new set of r dimensional data sets $Y = \{y_1, y_2, ..., y_n\}$, where y_i corresponds to d_i in D. It can be seen that data points in must minimize the following error function to preserve as much local structure information as possible.

$$e_Y(Y) = \sum_{i=1}^{n}(y_i - \sum_{j \in N_i} W_{ij} y_j)^2 \tag{10}$$

The right hand side of equation (10) should be minimized based on the constraint as shown in equation (11).

$$\sum_{i=1}^{n} | y_i |^2 = n \tag{11}$$

The values of y_i 's thus can be computed by solving an eigenvalue problem

$$Mv = \lambda v \tag{12}$$

where the matrix M can be computed from as follows.

$$M = (I - W)^T (I - W) \tag{13}$$

$y_1, y_2, ..., y_n$ can be determined from the eigenvectors that correspond to the r lowest eigenvalues of M [11].

2.3 The New Objective

We observe that, if we assume the mapped data points $y_1, y_2, ..., y_n$ can be obtained by applying a linear transformation T on the original data set D, the relationship between y_i and d_i can be written as follows.

$$y_i = Td_i \tag{14}$$

where T is a $r \times m$ matrix. From equation (14), the error $e_Y(Y)$ in (10) can be written as follows.

$$e_T(T) = \sum_{i=1}^{n} (Td_i - \sum_{j \in N_i} W_{ij} Td_j)^2 \tag{15}$$

It is not difficult to see that equation (14) can be further written as.

$$e_T(T) = tr(TQT^T) \tag{16}$$

where Q is a matrix that can be computed from $d_1, d_2, ..., d_n$ and W.

To guarantee that both the global and local structures are preserved to some extent, we propose to find a linear transformation T that can maximize the between-class distance while minimizing both the within-class distance and the error term $e_T(T)$. First, since the rank of the between-class scatter matrix S_b used in traditional LDA is at most $k-1$, the dimensionality of the data points generated by LDA is at most $k-1$. To resolve this issue, we propose a new scatter matrix S_b^n to represent the between-class distance. S_b^n can be computed as follows.

$$S_b^n = \sum_{i=1}^{k} \sum_{j \in C_i} \frac{1}{n_i} (d_j - m)(d_j - m)^T \tag{17}$$

Based on the within-class scatter matrix S_w in equation (2) and $e_T(T)$ in equation (15), we propose to maximize the following objective.

$$O(T) = \frac{tr(TS_b^n T^T)}{\lambda_1 tr(TS_w T^T) + \lambda_2 tr(TQT^T)} \tag{18}$$

Where λ_1 and λ_2 are positive constants that determine the relative weights of the within-class distance and the error that arises from the transformation. Similar to the approximation used in LDA, we can approximate $O(T)$ as follows.

$$O(T) \approx tr((T(\lambda_1 S_w + \lambda_2 Q)T^T)^{-1}(TS_b^n T^T)) \tag{19}$$

It is not difficult to see that the right hand side of equation (18) can be optimized by solving the following eigenvalue problem.

$$(\lambda_1 S_w + \lambda_2 Q)^{-1} S_b^n t = \lambda t \tag{20}$$

3 Testing Results

We have implemented this new dimensionality reduction approach and evaluated its performance with six benchmark data sets. The effectiveness of a dimensionality reduction result is evaluated based on the classification accuracy obtained on it. In our experiments, the constants λ_1 and λ_2 are computed as follows.

$$\lambda_1 = \sqrt{\frac{tr(Q)}{tr(S_w)}} \tag{21}$$

$$\lambda_2 = \frac{1}{\lambda_1} \tag{22}$$

The reason for selecting above values for λ_1 and λ_2 is that the contributions from within-class distance and the error $e_T(T)$ can be well balanced when the denominator of equation (18) is approximately $\sqrt{tr(TS_w T^T)tr(TQT^T)}$. If we assume:

$$\sqrt{\frac{tr(TQT^T)}{tr(TS_w T^T)}} \approx \sqrt{\frac{tr(Q)}{tr(S_w)}} \tag{23}$$

It is not difficult to see that the denominator in equation (18) is approximately $\sqrt{tr(TS_w T^T)tr(TQT^T)}$ when the values of λ_1 and

λ_2 are as above. In addition, for each data point, five of its nearest neighbors are used to compute matrix Q.

Table 1 shows the information on the benchmark data sets we have used in our experiments. These data sets include image documents, such as USPS, and text documents, such as 20Newsgroups. The rest of data sets are from UCI Machine Learning Repository. These data sets include Satimage, Waveform, Soybean, and Letter. We randomly partition each data set into a training set and a test set. The sizes of both training and testing sets are also listed in table 1.

Table 1. Information on the test benchmark data sets

Data Set	Sample Size	Dimensionality	# of Classes	Training	Test
20Newsgroups	1200	8298	4	300	900
Waveform	1200	40	3	300	900
Satimage	3600	36	6	900	2700
Soybean	562	35	15	150	412
Letter	3900	16	13	800	3100
USPS	3000	256	10	700	2300

We then compare the effectiveness of our approach with that of LDA and LLE, two approaches that only consider global or local structures in dimensionality reduction. The effectiveness of an approach is evaluated by the classification accuracy measured with the Nearest-Neighbor (NN) approach. To estimate the classification accuracy more accurately, we repeat the partition of each data set for 50 times and compute the average accuracy and the standard deviation. Table 2 shows the average accuracy and the standard deviation obtained with our approach, LDA, and LLE respectively on all six benchmark data sets.

Table 2. The average classification accuracy and standard deviation of our approach, LDA and LLE in percentage.

Data Set	Our Approach		LDA		LLE	
	average	stand. dev.	average	stand. dev.	average	stand. dev.
20Newsgroups	81.32	1.78	75.11	2.32	53.25	3.61
Waveform	80.62	2.32	71.73	3.12	63.32	2.55
Satimage	84.52	1.82	74.32	1.57	81.72	1.23
Soybean	85.63	1.37	84.73	1.62	82.52	1.92
Letter	81.35	1.53	73.21	1.64	78.33	2.23
USPS	83.72	0.78	80.62	1.22	75.52	1.44

It can be seen clearly from table 2 that our approach outperforms both LDA and LLE on all six benchmark data sets, while LDA outperforms LLE on 20Newsgroups, Waveform, Soybean, and USPS and LLE outperforms LDA on Satimage and Letter. The results suggest that for Satimage and Letter, local structures are more important than global ones while global structures are more important than local ones for the rest four data sets. Table 2 also suggests that preserving both global and local structures while reducing the dimensionality can effectively improve the classification accuracy.

To evaluate the computational efficiency of our approach, we measure the computation time needed by our approach on data sets of different sizes. Specifically, we select a few candidate data sets of different sizes from 20Newsgroup, which has the largest dimensionality of all six benchmark data sets. We then measure the computation time needed by our approach on these candidate data sets. Table 3 shows the computation time (in seconds) needed by our approach on candidate data sets of different sizes. It is clear from the table that our approach is able to efficiently process high-dimensional data.

Table 3. Computation time(in seconds) of our approach on candidate sets of different sizes from 20Newsgroups.

Size	100	200	300	400	500	600	700	800	1000
Time	10.72	13.63	17.67	24.62	29.55	36.21	42.34	47.61	57.34

4 Conclusions

In this paper, we develop a new approach for supervised dimensionality reduction. Our approach considers both the global and local geometric structures of a data set and develops a new objective that includes contributions from both of them. We show that this new objective can be optimized by solving an eigenvalue problem. Our testing results show that our approach outperforms both LDA and LLE on six benchmark data sets, which consider only global or local structures of data.

Currently, the parameters used in the objective are estimated from the traces of two matrices, which may not be optimal for the effectiveness of dimensionality reduction. Future work will focus on the development of approaches to determining the optimal parameters. In addition, this new approach can probably be combined with our previous work [12-14] to solve a few important bioinformatics problems.

Acknowledgments

Y. Song's work is fully supported by the University Fund of Jiangsu University of Science and Technology, under the numbers 635301202 and 633301301.

References

1. P. Belhumeour, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces, "Recognition using class specific linear projection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.
2. M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering", *Advances in Neural Information Processing Systems*, volume 15, 2001.

3. J. Chen, J. Ye, and Q. Li, “Integrating global and local structures: a least squares framework for dimensionality reduction”, *Proceedings of 24th International Conference on Machine Learning*, 2007.
4. V. de Silva and J. Tenenbaum, “Global versus local methods in nonlinear dimensionality reduction”, *Advances in Neural Information Processing Systems*, pages 705–712, 2002.
5. R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley Interscience, 2nd edition, 2000.
6. S. Dudoit, J. Fridlyand, and T. P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data”, *Journal of the American Statistical Association*, 97(457):77–87, 2002.
7. R. Fisher, “The use of multiple measurements in taxonomic problems”, *Annals of Eugenics*, 7:179–188, 1936.
8. K. Fukunaga, *Introduction to Statistical Pattern Classification*. Academic Press, San Diego, California, USA, 1990.
9. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning : Data mining, Inference, and Prediction*. Springer, 2001.
10. X. He and P. Niyogi, “Locality preserving projection”, *Advances in Neural Information Processing Systems*, 2003.
11. S. Roweis and L. Saul, “Nonlinear dimensionality reduction by locally linear embedding”, Science, 290(5500): 2323-2326, 2000.
12. Y. Song, “A New Parameterized Algorithm for Rapid Peptide Sequencing”, *PLoS ONE* 9(2): e87476, 2014.
13. Y. Song and A. Y. Chi, “A new approach for parameter estimation in the sequence-structure alignment of non-coding RNAs”, *Journal of Information Science and Engineering*, 2014, in press.
14. Y. Song, “An improved parameterized algorithm for the independent feedback vertex setprobem”,*Theoretical Computer Science*, 535(22): 25-30, 2014.