Semi-Supervised Learning by Locally Linear Embedding in Kernel Space

4 authors, including:

Rujie Liu
Fujitsu Ltd.
32 PUBLICATIONS **278** CITATIONS

SEE PROFILE

Yuehong Wang
Fujitsu Ltd.
9 PUBLICATIONS **122** CITATIONS

SEE PROFILE

Daiki Masumoto
Fujitsu Ltd.
29 PUBLICATIONS **199** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   image processing View project

Semi-Supervised Learning by Locally Linear Embedding in Kernel Space

Rujie Liu[1], Yuehong Wang[1], Takayuki Baba[2], Daiki Masumoto[2]

[1]*Fujitsu Research & Development Center Co. Ltd, Beijing, 100025, China*
[2]*Fujitsu Laboratories, Kawasaki, Japan*
{*rjliu,wangyh*}*@cn.fujitsu.com*, {*Baba-t, masumoto.daiki*}*@jp.fujitsu.com*

Abstract

Graph based semi-supervised learning methods (SSL) implicitly assume that the intrinsic geometry of the data points can be fully specified by an Euclidean distance based local neighborhood graph, however, this assumption may not always be necessarily true. To overcome this problem, we propose to apply locally linear embedding (LLE) method to characterize the geometric structure of the data points; besides this, the embedding process is performed in the kernel induced feature space rather than the original input space. After embedding, the proposed transductive learning method predicts the labels of the unlabeled data within the regularization framework. Experimental results on image retrieval and pattern recognition verify the performance of the proposed approach.

1. Introduction

Semi-supervised learning has received significant attention in recent years due to the lack of sufficient labeled data[1]. The key to semi-supervised learning problems is the prior consistency assumption: (1) nearby points are likely to have the same label; and (2) points on the same structure are prone to have the same label[3], as expected by spectral clustering[4], diffusion kernel[5], and random walks[6].

The main difference between the semi-supervised learning algorithms lies in their way of realizing the consistency property. With the Laplace-Beltrami operator, Belkin[2] builds Laplacian Eigenmaps to realize locality preserving embedding for the manifold, and then the appropriate classifier is estimated on the basis of the labeled examples. Zhou[3] proposes to use manifold ranking method to rank the data points along their underlying manifold by analyzing their relationship in Euclidean space. In [7], the learning problem is formulated in terms of a Gaussian random field over a contin-

uous state space, and a closed form solution is obtained using matrix methods or belief propagation.

In graph based learning methods, it is implicitly assumed that the data is situated on a low dimensional manifold within the ambient space of the data, and that the local linear structure of the data points in the ambient space can approximate a metric structure in corresponding neighborhoods on the manifold[7][12]. This assumption, however, may not always be necessarily true because data drawn from an extrinsically curved manifold are locally nonlinear at any finite scale. Thus, distances in the ambient space are only biased approximation of geodesic arc length distances on the manifold. To alleviate this problem, Hong[8] has proposed to use the path-based similarity measure to exploit the underlying manifold structure.

In this paper, we apply dimensionality reduction techniques in the graph regularization framework. More specifically, locally linear embedding method (LLE)[9] is adopted to characterize the intrinsic geometric structure of the data points instead of the local neighborhood graph. Different from the basic LLE method, we compute the embedding weights in the kernel induced feature space rather than the input space of the data points.

Our transductive learning method consists of two steps: (1) analyze the geometric structure of the data points by reconstructing each point from its neighbors in the kernel space; (2) predict the labels of the unlabeled data within regularization framework. Two constraints are designed in regularization process. The first constraint comes from LLE theory, i.e., the reconstruction weights characterizing the local geometry of the data points should be equally valid for their labels; the second constraint is the fitting term, which requires the prediction results not change too much from the initial label assignment.

The rest of this paper is organized as follows. In section 2, a brief introduction of the locally linear embedding algorithm is provided. Section 3 presents the embedding process in kernel space. In section 4, the

regularization framework is developed to realize prediction. The experimental results are shown in section 5, followed by the conclusion in section 6.

2. Locally linear embedding

Denote a set of n data points in a D dimensional input space as $\chi = \{x_1, x_2, \ldots, x_n\}$. For each point x_i, find its K nearest neighbors in the dataset, $x_1^i, \ldots, x_K^i$. Suppose the data points are sampled from some smooth underlying manifold, each x_i can be linearly reconstructed from its neighbors, and the reconstruction error are then measured by the cost function:

$$\varepsilon(W) = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{K} w_{ij}x_j^i\|^2 \qquad (1)$$

which adds up the squared distances between the data points and their reconstructions. The weights w_{ij} indicates the contribution of the j-th neighbor to x_i's reconstruction.

Under the constraint that $\sum_j w_{ij} = 1$ for each i, the cost function is minimized by solving a constrained least squares problem, and the optimal weights are:

$$w_{ij} = \frac{\sum_m (C_{jm}^i)^{-1}}{\sum_{mn}(C_{mn}^i)^{-1}} \qquad (2)$$

where C^i represents a $K \times K$ local covariance matrix associated with x_i, and the element value is computed by:

$$C_{mn}^i = (x_i - x_m^i) \cdot (x_i - x_n^i)$$

The reconstruction coefficients, which reflect local geometry of the data points in the original space, are expected to be equally valid for local patches on the manifold. In particular, weights w_{ij} should also reconstruct the i-th data point in the embedded space.

3   Geometry analysis in kernel space

3.1. Graph Laplacian kernel

Let the data points χ be described by a connected graph $G = (V, E)$, with nodes V corresponding to the n data points and edge set $E \subset V \times V$. For each edge $(i, j) \in E, i \neq j$, a non-negative weight a_{ij} is assigned to represent the pairwise similarity between vertices i and j. With this manner, an adjacency matrix, also called affinity matrix, $A = [a_{ij}]$, is obtained.

The graph Laplacian L is defined in terms of the adjacency matrix as: $L = D - A$, where D is a diagonal matrix with $D_{ii} = \sum_j a_{ij}$, and the normalized graph Laplacian is defined as $\widetilde{L} = D^{-\frac{1}{2}}LD^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$.

The regularized Laplacian is defined as:

$$P = r(\widetilde{L}) = \widetilde{L} + \varepsilon \cdot I \qquad (3)$$

With the regularization matrix P, a Hilbert space H can be defined on graph G via $< f, f >_H = < f, Pf >$, and it can be proved that the inverse of P is the reproducing kernel of $H(G)$[10], i.e.,

$$k(i, j) = [P^{-1}]_{ij} \qquad (4)$$

3.2. LLE in kernel space

The graph Laplacian kernel K implicitly defines a mapping ϕ from input space to a high dimensional feature space F such that k corresponds to a dot product in F by $k(x, x') = < \phi(x), \phi(x') >$. With this trick, we may describe the geometric structure of the data points in the kernel induced feature space F.

For each $\phi(x_i)$, denote its K nearest neighbors in feature space F as $N(\phi(x_i))$, and the reconstruction error associated with $\phi(x_i)$ is:

$$\varepsilon(W_i) = \|\phi(x_i) - \sum_{\phi(x_j) \in N(\phi(x_i))} w_{ij}\phi(x_j)\|^2 \qquad (5)$$

With the constraint that $\sum_j w_{ij} = 1$, equation 5 can be rewritten as:

$$\varepsilon(W_i) = \| \sum_{\phi(x_j) \in N(\phi(x_i))} w_{ij}(\phi(x_i) - \phi(x_j))\|^2 = \mathbf{W}_i^T \mathbf{C}_i \mathbf{W}_i$$
$$(6)$$

where, C_i is the local Gram matrix of $\phi(x_i)$ in F, and:

$$C_i(j, k) = (\phi(x_i) - \phi(x_j))^T((\phi(x_i) - \phi(x_k)) \qquad (7)$$

The closed form solution for this optimization problem is given by:

$$W_i = (C_i^{-1}\overline{\mathbf{1}})/(\overline{\mathbf{1}}^T C_i^{-1}\overline{\mathbf{1}}) \qquad (8)$$

Instead of performing matrix inversion, a more efficient way of finding the solution is to solve the linear system of equations, $C_i W_i = 1$, and then to scale the weights so that they sum to one.

In this algorithm, the distances of the data points in kernel space F are required to obtain their nearest neighbors, which are defined as follows:

$$\|\phi(x_m) - \phi(x_n)\|^2 = k(m, m) + k(n, n) - 2k(m, n)$$

4. Regularization

The above obtained reconstruction weights W, which characterize the local geometry of the data points, are expected to be equally valid for their labels.

In other words, we hope that the label of a data point can be reconstructed by its neighbors in the kernel space.

Without the loss of generality, we assume that the former l points in the data set χ are labeled while the rest points are to be classified. Let F denote the space of functions defined on χ, and $\forall f \in F$ assigns a real value f_i to each x_i. Thus, following cost function can be defined to measure the reconstruction property of f:

$$\zeta(f) = \sum_i ||f_i - \sum_{\phi(x_j) \in N(\phi(x_i))} w_{ij} f_j||^2 \quad (9)$$

Besides this, another constraint, i.e. fitting term, is imposed to prevent the prediction results from changing too much from the initial label assignment.

Define a $n \times 1$ vector $y = [y_1, \ldots, y_n]^T$ with $y_i = 1$ or -1 if x_i is labeled as positive or negative, and $y_i = 0$ for $i = l+1, \ldots, n$. The fitting constraint is then represented by a square loss function, as follows:

$$\tau(f) = \sum_i (f_i - y_i)^2 \quad (10)$$

Above two items are linearly combined by a trade-off parameter μ, and the integrated cost function is:

$$\varepsilon(f) = \zeta(f) + \mu\tau(f) \quad (11)$$

The optimal classification function f^* is thus obtained by solving following optimization problem,

$$f^* = \arg\min_{f \in F} \varepsilon(f) \quad (12)$$

We can rewrite $\varepsilon(f)$ as

$$\varepsilon(f) = \sum_i ||f_i - \sum_j w_{ij} f_j||^2 + \mu \sum_i (f_i - y_i)^2$$
$$= f^T M f + \mu(f - y)^T(f - y) \quad (13)$$

where $M = (I - W)^T(I - W)$.

Differentiating $\varepsilon(f)$ with respect to f, we have

$$f^* = \mu(M + \mu I)^{-1} y \quad (14)$$

It is obvious that this result is well consistent with other spectral graph theory based learning methods [3][7][10], except the difference of characterizing the geometric property of the data points. In previous methods, the graph Laplacian matrix is directly adopted, based on which, graph Laplacian kernel is built and the prediction is then made. This manner, however, usually leads to distortion because distances in the input space are only biased approximation of geodesic arc lengths on the manifold. In our method, the coefficients of locally linear embedding are applied to describe the intrinsic geometry of the data points. Moreover, the embedding is performed in Laplacian kernel space to make the result more reasonable.

5. Experiments

The performance of the proposed method is empirically evaluated by object recognition and content based image retrieval. In the experiments, we compare our method to Zhou's manifold ranking method[3], moreover, we include an ordinary k-NN nearest neighbor classifier for baseline comparison.

Firstly, the digits recognition task is performed using the USPS handwritten 16x16 digits dataset. We randomly select 300 examples for each digit, thus build a database of total 3000 examples.

In the experiment, we let all algorithms use their respective optimal parameters, and it is found that 0.01 and 5e-4 are best suitable for parameter μ in manifold ranking and our method respectively. The number of neighbors, i.e. parameter K in section 3, for kernel construction and locally linear embedding in our method is set to be 10. Besides this, the k in k-NN classifier is set to 1.

We consider the 10-way problems of classifying digits "0" through "9". In classification, same numbers of digits are randomly selected for each class as the labeled data, based on which, 10 predictions are made for the remaining data. Each unlabeled data is then assigned to the class with maximum prediction value. This process is repeated 100 times, and the average classification precision is calculated. The experimental results with different number of labeled points are shown in figure 1.

It is clear that our method is consistently superior to the manifold ranking method. The performance gain is particularly impressive when only a small number of labeled points are used, which is often the true scenario of real world applications with training data scarcity problems. As the increase of the number of labeled points, the performance of our method and manifold ranking becomes very comparable.

Next, our algorithm is further evaluated by content based image retrieval task, where a general purpose image database consisting of 4,000 Corel images is used. These images are categorized into 40 groups. Each of the categories contains 100 images of essentially the same semantic concept, which serves as the ground-truth. In retrieval, all these images are used in turn as queries, and the average results over 4,000 queries are presented. The precision vs. scope is used to evaluate the performance of various methods.

The features that are used to represent the images include color histogram and wavelet texture feature. Color histogram is obtained by quantizing the HSV color space into 64 bins. To calculate the wavelet feature, three level Daubechies wavelet transform is firstly performed to the image, and then the first three mo-

ments of the coefficients in High/High, High/Low, and Low/High bands at each decomposition level are used to construct the feature vector[11].

Parameter values in this experiment are kept the same as those in USPS digits recognition, and the results are shown in Figure 2. Similar conclusions are obtained once again from this experiment. The proposed algorithm is superior to manifold ranking method, and these two semi-supervised learning schemes outperform k-NN technique.

It is noticed that all the superiorities are very slight. This is because the images in Corel database are all semantically classified, which results in huge gaps between low level features and high level semantics. The semi-supervised learning schemes, no matter what kind of methods are used wherein to characterize the intrinsic geometry of the data points, are all stemmed from the distance of the low level features. In other words, the semantic gap in image retrieval is not solved by using unlabeled data.

6. Conclusion

Graph based semi-supervised learning methods assume that the data is situated on a low dimensional manifold within the ambient space of the data and that this manifold can be approximated by a local similarity graph. However, this assumption usually leads to systematic distortions. To overcome this problem, this paper proposes to adopt dimensionality reduction techniques in the graph regularization framework. That is to say, locally linear embedding in kernel space is applied to characterize the local geometry of the data points. The performance of this method is verified by pattern recognition and image retrieval experiments.

References

[1] X. Zhu. Semi-supervised learning literature survey. *Technical report 1530*, Department of computer sciences, University of Wisconsin at Madison, 2006.

[2] M. Belkin, P. Niyogi. Semi-Supervised learning on Riemannian manifolds. *Machine Learning*, vol.56, pp. 209-239, 2004.

[3] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Scholkopf. Learning with local and global consistency. *In Advances in neural information processing systems*, vol.16, 2003.

[4] O. Chapelle, J. Weston, B. Scholkopf. Cluster kernels for semi-supervised learning. *In Advances in neural information processing systems*, vol.15, 2002.

[5] R.I. Kondor, J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. *In Proc. 19th Int'l Conf. on machine learning*, 2002.
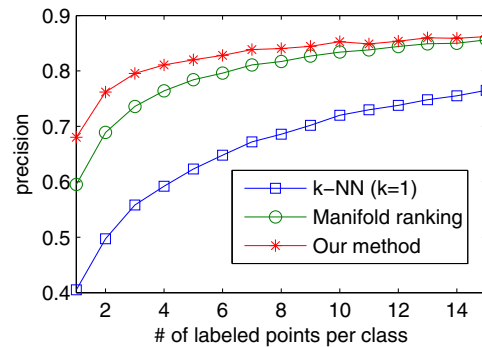
Figure 1. USPS handwritten digits recognition
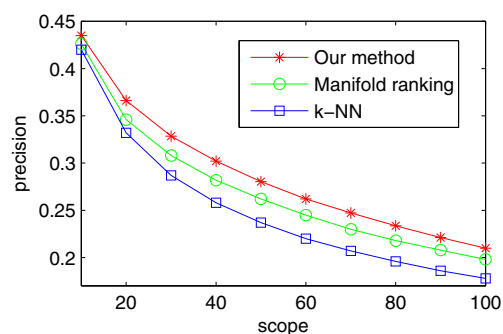
Figure 2. Retrieval precision on Corel dataset

[6] M. Szummer, T. Jaakkola. Partially labeled classification with Markov random walks. *In Advances in neural information processing systems*, vol.14, 2001.

[7] X. Zhu, Z. Ghahramani, J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. *In Proc. 20th Int'l Conf. on machine learning*, 2003.

[8] H. Chang, D.-Y. Yeung. Graph Laplacian kernels for object classification from a single example. *In Proc. IEEE Int'l Conf. CVPR*, 2006.

[9] S.T. Roweis, L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, vol.290, pp. 2323-2326, 2000.

[10] A.J. Smola, R. Kondor. Kernels and regularization on graphs. *In Conference on Learning Theory*, COLT 2003.

[11] W.Y. Ma, B.S. Manjunath. A Comparison of Wavelet Transform Features for Texture Image Annotation. *in Proc. of the IEEE int'l Conf. on Image Processing*, pp. 256-259, 1995.

[12] J.-J. Verbeek, N. Vlassis. Gaussian fields for semi-supervised regression and correspondence learning. *Pattern Recognition*, vol.39, pp. 1864-1875, 2006.