Transductive Regression Piloted by Inter-Manifold Relations

Huan Wang HWANG5@IE.CUHK.EDU.HK
IE, The Chinese University of Hong Kong, Hong Kong

Shuicheng Yan SYAN@IFP.UIUC.EDU
Thomas Huang HUANG@IFP.UIUC.EDU
ECE, University of Illinois at Urbana Champaign, USA

Jianzhuang Liu JZLIU@IE.CUHK.EDU.HK
Xiaoou Tang XTANG@IE.CUHK.EDU.HK
IE, The Chinese University of Hong Kong, Hong Kong

Abstract

In this paper, we present a novel semisupervised regression algorithm working on multi-class data that may lie on multiple manifolds. Unlike conventional manifold regression algorithms that do not consider the class distinction of samples, our method introduces the class information to the regression process and tries to exploit the similar configurations shared by the label distribution of multi-class data. To utilize the correlations among data from different classes, we develop a cross-manifold label propagation process and employ labels from different classes to enhance the regression performance. The inter-class relations are coded by a set of inter-manifold graphs and a regularization item is introduced to impose inter-class smoothness on the possible solutions. In addition, the algorithm is further extended with the kernel trick for predicting labels of the out-of-sample data even without class information. Experiments on both synthesized data and real world problems validate the effectiveness of the proposed framework for semisupervised regression.

1. Introduction

Large scale and high dimensional data are ubiquitous in real-world applications, yet the processing and analysis of these data are often difficult due to the curse of dimensionality as well as the high computational cost involved. Usually, these high dimensional data lie approximately on an underlying compact low dimensional manifold, which may turn the problem tractable. Substantive works have been devoted to unveiling the intrinsic structure of the manifold data, among which the popular ones include ISOMAP (Tenenbaum et al., 2000), LLE (Roweis & Saul, 2000), Laplacian Eigenmap (Belkin & Niyogi, 2003), and MVU (Weinberger et al., 2004).

Besides these unsupervised algorithms that purely explore the manifold structure of the data, researchers have also well utilized the latent manifold structure information from both labeled and unlabeled samples to enhance learning algorithms with limited number of labeled samples. Great success has been achieved in various areas, such as classification (Krishnapuram et al., 2005) (Belkin et al., 2005), manifold alignment (Ham et al., 2005) and regression (Belkin et al., 2004) (Zhu & Goldberg, 2005). Recently increasing attention has been drawn to the semisupervised regression problem by considering the manifold structure. (Chapelle et al., 1999) proposed a transductive algorithm minimizing the leave-one-out error of the ridge regression on the joint set composed from both labeled and unlabeled data. To exploit the manifold structure, (Belkin et al., 2004) adds a graph Laplacian regularization item to the regression objective, which imposes extra condition on the smoothness along the data manifold and it has proved to be quite useful in applications. (Cortes & Mohri, 2007) first roughly transduces the function values from the labeled data to the unlabeled ones utilizing local neighborhood relations, and then optimizes the global objective that best fits the

Appearing in *Proceedings of the 24^{th} International Conference on Machine Learning*, Corvallis, OR, 2007.

labels of the training points as well as the estimated labels provided by the first step.

None of these state-of-the-art regression algorithms make use of the class information to guide the regression. In real-world applications, however, multi-class samples are ubiquitous and samples from different classes can be regarded as lying on multiple manifolds. For example, in age estimation, the personal aging process varies differently for different genders, but there still exists consistency between the aging processes of male and female. In the pose estimation from images of multiple persons, each person can be regarded as one class and the images vary similarly with poses from different persons. Moreover, we may also encounter the multi-modality samples whose representations may not be consistent due to the lack of correspondence, while the inner manifold configuration for the label distribution across different modalities may still be similar. Besides, in the function learning stage of the semisupervised regression framework, the class or modality information is usually easy to obtain, thus it is desirable to utilize this information to further boost the regression accuracy. To fully exploit the relations among data manifolds, we develop in this paper a semisupervised regression algorithm called *TRIM* (**T**ransductive **R**egression piloted by **I**nter-**M**anifold Relations). *TRIM* is based on the assumption that the data of different classes share similar configuration for label distributions. In our proposed algorithm, manifolds of different classes are aligned according to the landmark connections constructed from both the sample label distance and manifold structural similarity, and then the labels are transduced across different manifolds based on the alignment output. Besides the intra-manifold smoothness condition within each class, our method introduces an inter-manifold regularization item and employs the transduced labels from various data manifolds to pilot the trend of the function values over all the samples. In addition, the function to be learnt can be approximated by the functionals lying on the Reproducing Kernel Hilbert Space (RKHS), and the regression on the induced RKHS directly leads to an efficient solution for the label prediction of the out-of-sample data even without corresponding class information.

2. Background

2.1. Intuitive Motivations

Consider the problem of adult height estimation based on the parents' heights. Dramatic difference exists between male and female adults, while the general configuration of the height distribution within each gender

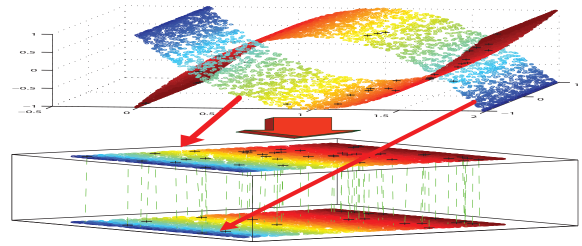


Figure 1. Demonstration of the inter-class relations. Labeled samples are marked by $'+'$ and the inter-manifold relations are indicated by the dashed lines. The manifolds are aligned and the sample labels are propagated across manifolds to pilot the regression.

may still be similar: the taller the parents are, the higher the children are supposed to be. While once we mix up the data from both genders, the height distribution may become complicated. Figure 1 displays a toy example for the obstacles encountered in the multi-class regression problem. We have two classes of samples and the function values within each class manifold share similar configurations, while the number of samples and labels may vary across classes. As is shown in the upper part of Figure 1, it is difficult to get a satisfactory regression if we put all the samples together due to the complex structures produced by the manifold intersections. On the other hand, if the regression is carried out separately on different classes, the accuracy could still be low due to the lack of sufficient labeled samples for each class. Also, for the incoming out-of-sample data, the prediction cannot be done without class information. Moreover, we may face some regression problems with multi-modality samples, e.g., the estimation of human ages from both photos and sketch images. It is meaningless to compose data of different modalities into one manifold since the semantics of the features from different modalities are essentially different, while it is still possible that the intra-modality label configurations are similar. In this paper we focus on the multi-class regression problem, but our framework can also be easily extended to give predictions for the multi-modality data.

2.2. Related Work

Multi-view learning (Blum & Mitchell, 1998) technique can also be applied to the regression problem with multi-class data. Denote the labeled and unlabeled instances as $x^l \in \mathcal{X}^l \subseteq \mathcal{X}$ and $x^u \in \mathcal{X}^u \subseteq \mathcal{X}$ respectively, where $\mathcal{X}^l, \mathcal{X}^u, \mathcal{X}$ are the labeled, unlabeled and whole sample sets. State-of-the-art semisupervised multi-view learning frameworks employ multiple regression functions from the Hilber space $\mathcal{H}_v$ so that the estimation error on the training set and disagreement among the functions on the unlabeled data are minimized (Brefeld et al., 2006), i.e.,

$$\tilde{f}_v|_{v=1}^{M_l} = \min_{f_v \in \mathcal{H}} \sum_{v=1}^{M_l} \Big(\sum_{x \in \mathcal{X}^l} c(f_v(x^l), y(x^l)) + \gamma \|f_v\|^2 \Big) + \lambda \sum_{v_1, v_2 = 1}^{M_l} \sum_{x \in \mathcal{X}^u} c(f_{v_1}(x^u) - f_{v_2}(x^u)), \quad (1)$$

where $y(x)$ is the sample label, $f_v|_{v=1}^{M_l}$ are M_l multiple learners and $c(\cdot)$ is a cost function.

We would like to highlight beforehand some properties of our framework compared to the multi-view learning algorithms:

1. There exists a clear correspondence between the multi-view data from the same instance in the multi-view learning framework, while our algorithm does not require the correspondence among different manifolds. The data of different modalities or classes may be obtained from different instances in our configuration, thus it is much more challenging to give an accurate regression.

2. The class information is utilized in two ways:
a) Sample relations within each class are coded by intra-manifold graphs and a corresponding regularization item is introduced to ensure the within-class smoothness separately.
b) A set of inter-manifold graphs are constructed from the cross manifold label propagation on the aligned manifolds and an inter-manifold regularization item is proposed to fully exploit the information conveyed among different classes.

3. The class information is used in the function learning phase but no class attributes are required for the out-of-sample data in the prediction stage.

3. Problem Formulation

3.1. Notations

Assume that the whole sample set $\mathcal{X}$ consists of N samples from M classes, denoted as $X^1, X^2, ..., X^M$. For each class X^k, $N^k = l^k + u^k$ samples are given, i.e.,

$$X^k = \{(x_1^k, y_1^k), ..., (x_{l^k}^k, y_{l^k}^k), (x_{l^k+1}^k, y_{l^k+1}^{*k}), ..., (x_{l^k+u^k}^k, y_{l^k+u^k}^{*k})\},$$

where $Y_l^k = [y_1^k, y_2^k, ..., y_{l^k}^k]^T$ represents the function values with respect to the given labeled samples and $Y_u^k = [y_{l^k+1}^{*k}, ..., y_{l^k+u^k}^{*k}]^T$ corresponds to the function values of the remaining unlabeled data to be estimated in the k^{th} class. In this paper, the 'label' refers to a real value to be regressed.

Let $G^k = (V^k, E^k)$ denote the intra-class graph with vertex set V^k and edge set E^k constructed within the data of the k^{th} class. Here we focus on the undirected graph and it is easy to generalize our algorithm for directed graphs. The edges in E^k reflect the neighborhood relations along the manifold data, which can be defined in terms of k-*nearest* neighbors or an ϵ-*ball* distance criterion in the sample feature space $\mathcal{F}$. One choice for those non-negative weights on the corresponding edges is to use the heat kernel (Belkin & Niyogi, 2003) or the inverse of feature distances (Cortes & Mohri, 2007), i.e.,

$$w_{ij} = e^{\frac{-\|\Phi^k(x_i) - \Phi^k(x_j)\|^2}{t}}$$
$$\text{or} \quad w_{ij} = \|\Phi^k(x_i) - \Phi^k(x_j)\|^{-1},$$

where $t \in \mathbb{R}$ is the parameter for the heat kernel and $\Phi^k(\cdot)$ is a feature mapping from $\mathcal{X}$ to the normed feature vector space $\mathcal{F}$ for the sample of the k^{th} class. For samples from different modalities/manifolds, the feature mappings may be different. The other one is to solve a least-square problem to minimize the reconstruction error and get the weights w_{ij}:

$$w_{ij} = \arg\min_{w_{ij}} \|x_i - \sum_j w_{ij} x_j\|^2,$$
$$s.t. \sum_j w_{ij} = 1, w_{ij} \geq 0$$

(Roweis & Saul, 2000).

To encode the mutual relations among different sample classes, we also introduce the inter-manifold graph, denoted as a triplet $G^{k_i k_j} = (V^{k_i}, V^{k_j}, E^{k_i k_j})$. The inter-manifold graph $G^{k_i k_j}$ is a bipartite graph with one vertex set from the k_i^{th} class and the other from the k_j^{th}. The construction of $G^{k_i k_j}$ will be discussed in the following subsections.

3.2. Regularization along Data Manifold

Now we are given the sample data of multiple classes, denoted as $\mathcal{X}$, including both labeled and unlabeled samples X^l and X^u. The manifold structures within each class data are encoded by the intra-class graph. (Belkin et al., 2004) introduced a manifold regularization item for the semisupervised regression framework, i.e., the graph Laplacian, which is expected to impose smoothness conditions along the manifold on the possible solutions. The final formulation seeks a balance between the fitting item and the smoothness regularization, i.e.,

$$\tilde{f} = \arg\min \frac{1}{l} \sum_i (f_i - y_i)^2 + \gamma f^t L^p f, \tag{2}$$

where $p \in \mathbb{N}$. When $p = 1$, the regularization item is the graph Laplacian and for $p = 2$, the regularization turns out to be the 2-norm of the reconstruction error when the weights w_{ij} are normalized, i.e.,

$$f^T L^T L f = \sum_i \|f_i - \sum_j w_{ij} f_j\|^2$$
$$\text{w.r.t. } \sum_j w_{ij} = 1, w_{ij} \geq 0.$$

3.3. Cross Manifold Label Propagation

In our configuration, there does not exist a clear correspondence among the manifolds of different classes or modalities, and even the representations for different modalities can be distinct. Thus it is rather difficult to construct sample relations directly from the similarity in the sample space $\mathcal{X}$. Alternatively, the function labels still convey some correspondence information, which may be utilized to guide the inter-manifold relations. Moreover, the manifold structure also contains some indications about the correspondence. Thus we first seek a point-to-point correspondence for the labeled data combining the indications provided by both the sample labels and the manifold structures. This is done under two assumptions:

1. Samples with similar labels lie generally in similar relative positions on the corresponding manifold.
2. Corresponding sample sets tend to share similar graph structures on the respective manifold.

3.3.1. Reinforced Landmark Correspondence

First we search for a set of stable landmarks to guide the manifold alignment. Specifically, we use the ϵ-*ball* distance criterion on the sample labels to initialize the inter-manifold graphs. To give a robust correspondence, we reinforce the inter-manifold connections by iteratively implementing

$$W^{k_i k_j} \Leftarrow W^{k_i} \times W^{k_i k_j} \times W^{k_j}. \tag{3}$$

Similar to the similarity propagation process on the directed graph (Blondel et al., 2004), (3) reinforces the similarity score of sample pairs by the similarity of its neighbor pairs, i.e.,

$$w_{ij}^{k_i k_j} \Leftarrow \sum_{m,n} w_{im}^{k_i} w_{mn}^{k_i k_j} w_{nj}^{k_j}. \tag{4}$$

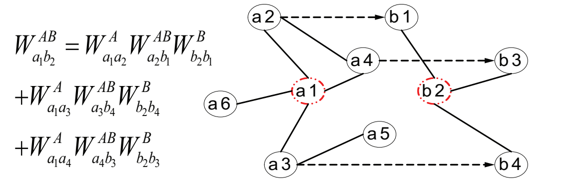


Figure 2. Demonstration of reinforced landmark correspondence. The similarity score of a_1 and b_2 is reinforced by the scores of neighbor pairs.

The assumption here is that two nodes tend to be similar if they have similar neighbors. (4) utilizes the intra-manifold structure information to reinforce the inter-manifold similarity and thus can generate a more robust correspondence. The accuracy of the landmark correspondence is critical for our algorithm. To ensure a robust performance, only the correspondences with the top 20% largest similarity scores are selected as the landmarks and it is common that some classes may miss certain sample labels, so plenty of labeled samples remain unmatched.

3.3.2. Manifold Alignment

To propagate the sample labels to the unlabeled points across manifolds, we 'stretch' all the manifolds with respect to the landmark points obtained from the previous step and this can be realized by the semisupervised manifold alignment (Ham et al., 2005).

In the manifold alignment process, we seek an embedding that minimizes the correspondence error on the landmark points and at the same time keeps the intra-manifold structures, i.e.,

$$f^{k_i}|_{k_i=1}^M = \arg\min\left(\frac{C(f^{k_i}|_{k_i=1}^M)}{\sum_{k_i} f^{k_i T} D^{k_i} f^{k_i}}\right), \tag{5}$$

and

$$C(f^{k_i}|_{k_i=1}^M) = \sum_{k_i k_j} w_{ij}^{k_i k_j} \|f^{k_i}_{x_i^{k_i}} - f^{k_j}_{x_j^{k_j}}\|^2$$
$$+\gamma \sum_{k_i=1}^{M} (f^{k_i T} L^p_{k_i} f^{k_i}) + \beta f^T L^a f, \tag{6}$$

where D^{k_i} is the diagonal degree matrix, L_{k_i} is the graph Laplacian matrix for the k_i^{th} class and $p \in \mathbb{N}$. To ensure the inter-class adjacency, we add a global compactness regularization item $\beta f^T L^a f$ to the cost function C, where L^a is the Laplacian Matrix of W^a with

$$w_{ij}^a = \begin{cases} 1 & \text{if } \mathrm{x_i} \text{ and } \mathrm{x_j} \text{ are of different classes} \\ 0 & o.w. \end{cases}$$

The labels are propagated across manifolds on the derived aligned manifolds using the nearest neighbor ap-

Algorithm 1 Procedure to construct inter-manifold connections

1: Inter-manifold graph initialization

$$w_{ij}^{k_i k_j} = \begin{cases} 1 & \text{if } \|y_i^{k_i} - y_j^{k_j}\|^2 < \epsilon \\ 0 & o.w. \end{cases}$$

2: Correspondence Reinforcement
for Iter = 1 : N_{Iter}

$$W^{k_i k_j} = W^{k_i} \times W^{k_i k_j} \times W^{k_j}$$

end

3: Landmark Selection: Select the sample pairs with the top 20% largest similarity scores as correspondence. Set the corresponding elements in $W^{k_i k_j}$ as 1 and others 0.

4: Manifold Alignment using the Inter-manifold Graphs $W^{k_i k_j}$.

5: Find the corresponding points from different classes for the unmatched labeled samples using the nearest neighbor approach in the aligned space and update the inter-manifold graphs $W^{k_i k_j}$.

proach, i.e., we connect the labeled samples with the nearest points from other classes on the aligned space. The derived inter-manifold graphs are concatenated to form

$$W^r = \begin{pmatrix} \boldsymbol{O} & W^{12} & ... & W^{1M} \\ W^{21} & \boldsymbol{O} & ... & W^{2M} \\ ... & ... & ... & ... \\ W^{M1} & W^{M2} & ... & \boldsymbol{O} \end{pmatrix}, \tag{7}$$

which is then symmetrized and employed in the following Inter-Manifold regularization item.

3.4. Inter-Manifold Regularization

We rearrange the samples to place the data from the same class together and put the labeled points first for each class, that is,

$$\mathcal{X} = \{x_1^1, x_2^1, ..., x_{l^1}^1, x_{l^1+1}^1, ..., x_{l^1+u^1}^1, x_1^2, x_2^2, ..., x_{l^2}^2, x_{l^2+1}^2, ..., x_{l^2+u^2}^2, ..., x_1^M, x_2^M, ..., x_{l^M}^M, x_{l^M+1}^M, ..., x_{l^M+u^M}^M\}.$$

Denote the corresponding function values as f, which is a concatenation of $f^k = [f_{x_1^k}^k, ..., f_{x_{l^k+u^k}^k}^k]^T$ from all the classes. Our regression objective is defined as:

$$\tilde{f} = \arg\min_f \sum_k \frac{1}{l^k} \sum_{x_i^k \in \mathcal{X}^l} \|f_{x_i^k}^k - y_i^k\|^2 + \beta \sum_k \frac{1}{(N^k)^2} (f^k)^T L_k^p f^k + \frac{\lambda}{N^2} f^T L^r f, \tag{8}$$

where L^r is the Laplacian matrix of the symmetrized W^r.

The minimization of the objective is achieved when

$$\tilde{f} = R^{-1} \sum_k \frac{1}{l^k} (S_{l^k}^k S^k)^T Y_l^k \tag{9}$$

where

$$S_{l^k}^k = \begin{pmatrix} I_{l^k \times l^k} & \boldsymbol{O}_{l^k \times u^k} \end{pmatrix},$$

$$S^k = \begin{pmatrix} \boldsymbol{O}_{N^k \times \sum_{k_i=1}^{k-1} N^{k_i}} & I_{N^k \times N^k} & \boldsymbol{O}_{N^k \times \sum_{k_i=k+1}^{M} N^{k_i}} \end{pmatrix}$$

are label and class selection matrices respectively and

$$R = \sum_k \frac{1}{l^k} (S_{l^k}^k S^k)^T (S_{l^k}^k S^k) + \beta \sum_k \frac{1}{(N^k)^2} S^{kT} L_k^p S^k + \frac{\lambda}{N^2} L^r.$$

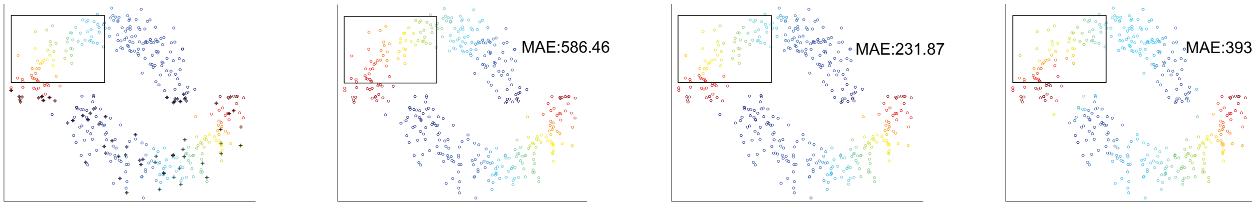
4. Regression on Reproducing Kernel Hilbert Space (RKHS)

A series of algorithms such as SVM, Ridge regression and LapRLS (Belkin et al., 2005) employ different regularization items and empirical cost measures to the objective and solve the optimization problem in an appropriately chosen Reproducing Kernel Hilbert Space (RKHS). One merit for the regression on RKHS is the ability to predict out-of-sample labels.

Let K denote a Mercer kernel: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\mathcal{H}_K$ denote the induced RKHS of functions $\mathcal{X} \to \mathbb{R}$ with norm $\| \ \|_K$. The regression with inter-manifold regularization on the RKHS is defined as

$$\tilde{f} = \arg\min_{f \in \mathcal{H}_K} \sum_k \frac{1}{l^k} \sum_{x_i^k \in \mathcal{X}^l} \|f_{x_i^k}^k - y_i^k\|^2 + \gamma \|f\|_K^2 + \beta \sum_k \frac{1}{(N^k)^2} (f^k)^T L_k^p f^k + \frac{\lambda}{N^2} f^T L^r f. \tag{10}$$

Similar to the Tikhonov regularization, we add an RKHS norm penalization item to the *TRIM* algorithm as a smoothness condition.



(a) (b) (c) (d)

Figure 3. Regression on the nonlinear Two Moons Dataset. (a) Original Function Value Distribution. (b) Traditional Graph Laplacian Regularized Regression (separate regressors for different classes). (c) Two Class *TRIM*. (d) Two Class *TRIM* on RKHS. Note the difference in the area indicated by the rectangle.

For the multi-class data under the same modality, we have the following theorem:

Theorem-1. The solution to the minimization problem (10) admits an expansion

$$\tilde{f}(x) = \sum_{i=1}^{N=\sum_k (l^k+u^k)} \alpha_i K(x_i, x) \qquad (11)$$

Theorem-1 is a special version of the **Generalized Representer Theorem** (Schlkopf et al., 2001) and the proof is omitted here. It says that the minimizer of (10) can be expressed in terms of the linear expansion of $K(x_i, x)$ on both labeled and unlabeled data over all the sample classes. Thus the minimization over Hilbert space boils down to minimizing the coefficient vector $\alpha = [\alpha_1^1, ..., \alpha_{l^1}^1, ..., \alpha_{l^1+u^1}^1, ..., \alpha_1^M, ..., \alpha_{l^M}^M, ..., \alpha_{l^M+u^M}^M]^T$ over $\mathbb{R}^N$ and the minimizer is given by:

$$\tilde{\alpha} = J^{-1} \sum_k \frac{1}{l^k} (S_{l^k}^k S^k)^T Y_l^k, \qquad (12)$$

where

$$J = \sum_k \frac{1}{l^k} (S_{l^k}^k S^k)^T (S_{l^k}^k S^k) K + \gamma I + \beta \sum_k \frac{1}{(N^k)^2} S^{kT} L_k^p S^k K + \frac{\lambda}{N^2} L^r K.$$

and K is the $N \times N$ Gram matrix of labeled and unlabeled points over all the sample classes.

For the out-of-sample data, the estimated labels can be obtained using:

$$y_{new} = \sum_{i=1}^{N=\sum_k (l^k+u^k)} \tilde{\alpha}_i K(x_i, x_{new}). \qquad (13)$$

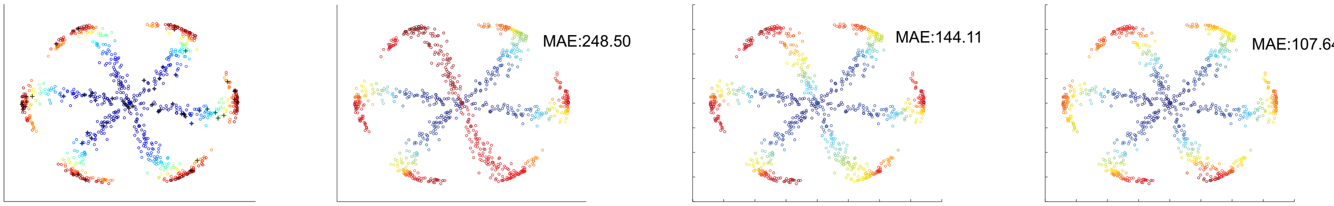
Note here in this framework the class information for the incoming sample is not required in the prediction stage.

5. Experiments

We performed experiments on two synthetic datasets and one real-world regression problem of human age estimation. Comparisons are made with the traditional graph Laplacian regularized semisupervised regression (Belkin & Niyogi, 2003). We also evaluate the generalization performance of the multi-class regression on RKHS. In the experiments, the intra-manifold graphs are constructed using 10 nearest neighbors and the inter-manifold graphs are constructed by following the procedure described in Section 3. For all the configurations, the parameter β for the intra-manifold graph is empirically set as 0.001 and the λ for the inter-manifold regularization is set as 0.1. For the kernelized algorithms, the coefficient for the RKHS norm $\gamma = 0.001$ and the gaussian kernel $K(x, y) = \exp\{-\|x - y\|^2/\delta_o^2\}$ with parameters $\delta_o = 2^{1/2.5}\delta$ is applied, where δ is the standard deviation of the sample data. We use the Mean Absolute Error (MAE) criterion to measure the regression accuracy and it is defined as an average of the absolute errors between the estimated labels and the ground truth labels.

5.1. Synthetic Data: Nonlinear Two Moons

The nonlinear two moons dataset is shown in Figure 3. The colors in the figure are associated with the function values and the variation of those sample labels along the manifold is not uniform. The labeled samples are marked by $'+'$ and their distributions are quite different across classes. As we can see, the sample labels for the class lying on the upper part of the figure are not enough to give an accurate guidance for the regression on the nonlinear label distribution. The traditional graph Laplacian regularized regression algorithm does not make use of the information conveyed by the inter-class similarity and the prediction result is not satisfactory, while in our algorithm the sample labels from different classes can be utilized to guide the regression and thus estimation accuracy is much higher.



(a) (b) (c) (d)

Figure 4. Regression on Cyclone Dataset: (a) Original Function Values. (b) Traditional Graph Laplacian Regularized Regression (separate regressors for different classes). (c) Three Class *TRIM*. (d) Three Class *TRIM* on RKHS.

5.2. Synthetic Data: Three-class Cyclones

One merit of our algorithm is that the regression accuracy may be boosted as the class number increases. This is because the cross manifold information that could be utilized grows rapidly as the class number increases. The Cyclone dataset consists of three classes of samples and the class distribution is demonstrated in Figure 5. The label distributions among different classes are quite similar, while the labeled samples scatter differently from class to class. As we may observe from Figure 4, without the inter-class regularization, the regression for certain class may fail due to the lack of sufficient labeled samples while our algorithm still gives a satisfying performance.

5.3. Human Age Estimation

In this experiment we consider as a regression problem the age estimation from facial images. In real applications, we often cannot obtain enough age labels though we may easily get plenty of facial images. Those unlabeled data can be used to guide our regression in the semisupervised framework. One challenge for this estimation is caused by the gender difference. Although the male and female may share some similar configurations in the age distribution, just as demonstrated in Figure 1, mixing the two may complicate the regression problem. On the other hand, the gender information is usually easy to obtain and it is thus desirable to use both the age labels and gender information to guide the regression.

The aging data set used in this experiment is the Yamaha database, which contains 8000 Japanese facial images of 1600 persons with ages ranging from 0 to 93. Each person has 5 images and the Yamaha database is divided into two parts with 4000 images from 800 males and another 4000 images for 800 females. The ages distribute evenly from 0 to 69 with 3500 male images and 3500 female images, and the rest are distributed in the age gap from 71 to 93. We randomly sampled 1000 photos from the male and female

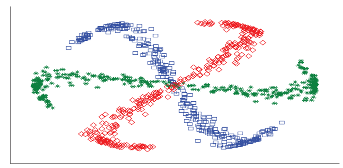


Figure 5. Class Distribution of the Cyclone Dataset

subset respectively and thus altogether 2000 images are used in our experiments. Before the regression, the input image data is preprocessed with Principal Component Analysis and the first 20 dimensions are used for data projection. To fully evaluate the regression performance for both close set samples and those out-of-sample data, we design two experimental configurations for this dataset.

Configuration 1: Close Set Evaluation. In this configuration, we do not have the out-of-sample data and the close set performance of *TRIM* is evaluated in comparison with the traditional graph Laplacian regularized regression (Belkin et al., 2004). The close set contains altogether 2000 samples with 1000 images from males and 1000 images from females respectively. We vary the number of randomly selected labeled samples and examine the performance of different regression algorithms. The comparison results between *TRIM* and the traditional single class Laplacian Regularized regression are shown in Figure 6. We can observe that our algorithm generally gives a higher regression accuracy and the performance improvement is remarkable especially when the number of labeled samples is small. As the number of sample labels increases, the difference between the two algorithms becomes smaller. This may be caused by the fact that as the labels are sparse, the label guidance is far less enough and thus the class information and inter-manifold relations will have a greater influence on the regression accuracy while when the labels are abundant enough to guide the regression, the class information as well as those inter-manifold connections becomes less important.

Configuration 2: Open Set Evaluation. Now we examine the out-of-sample prediction performance for

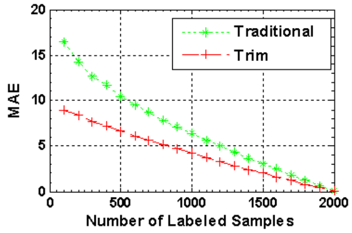
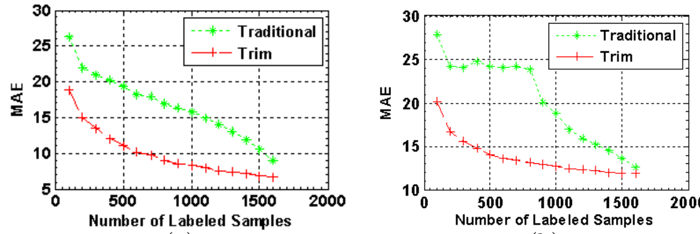


Figure 6. TRIM vs traditional graph Laplacian regularized regression for the close set evaluation on YAMAHA database.



(a) (b)

Figure 7. Open set evaluation for the kernelized regression on the YAMAHA database. (a) Regression on the training set. (b) Regression on out-of-sample data.

the kernelized *TRIM* compared with kernelized version of the traditional graph Laplacian regularized regression (Belkin et al., 2005). In this configuration, the sample set is divided into two subsets: one for the regression function training and the other for the evaluation of aging estimation performance on those out-of-sample data. The training set contains randomly selected 800 male and 800 female images and the remaining ones are used for out-of-sample evaluation. In the testing phase we do not input any gender information. As is demonstrated in Figure 7, our algorithm achieves a lower MAE in both training and testing sets. Another observation similar as in the case for close set configuration is that the regression accuracy improvement grows as the sample labels become sparser.

6. Conclusions

In this paper, we have presented a novel algorithm dedicated for the regression problem on multi-class data. Manifolds constructed from different classes are regularized separately and to utilize the inter-manifold relations, we developed an efficient cross manifold label propagation method and the labels from different classes can thus be employed to pilot the regression. Moreover, the regression function is further extended by the kernel trick to predict the labels of the out-of-sample data without class information. Both synthesized experiments and real world applications demonstrated the superiority of our proposed framework over the state-of-the-art semisupervised regression algorithms. To the best of our knowledge, this is the first work to discuss the semisupervised regression problem on multi-class data.

7. Acknowledgement

The work described in this paper was supported by the Research Grants Council of the Hong Kong SAR (Project No. CUHK 414306) and DTO Contract NBCHC060160 of USA.

References

Belkin, M., Matveeva, I., & Niyogi, P. (2004). Regularization and semi-supervised learning on large graphs. *COLT*.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*

Belkin, M., Niyogi, P., & Sindhwani, V. (2005). On manifold regularization. *AISTATS*.

Blondel, V. D., Gajardo, A., Heymans, M., Senellart, P., & Dooren, P. V. (2004). A measure of similarity between graph vertices: Applications to synonym extraction and web searching. *SIAM Rev.*

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *COLT*.

Brefeld, U., Gartner, T., Scheffer, T., & Wrobel, S. (2006). Efficient co-regularised least squares regression. *ICML*.

Chapelle, O., Vapnik, V., & Weston, J. (1999). Transductive inference for estimating values of functions. *NIPS*.

Cortes, C., & Mohri, M. (2007). On transductive regression. In *Nips*.

Ham, Lee, & Saul (2005). Semisupervised alignment of manifolds. *AISTATS*.

Krishnapuram, B., DavidWilliams, Xue, Y., Hartemink, A., & Carin, L. (2005). On semi-supervised classification. *NIPS*.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*.

Schlkopf, B., Herbrich, R., & Smola, A. J. (2001). A generalized representer theorem. *COLT/EuroCOLT*.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*.

Weinberger, K. Q., Sha, F., & Saul, L. K. (2004). Learning a kernel matrix for nonlinear dimensionality reduction. *ICML*.

Zhu, X., & Goldberg, A. B. (2005). Semi-supervised regression with order preferences. *Computer Sciences TR*.