Human Age Estimation by Metric Learning for Regression Problems

Yangjing Long

Max Planck Institute for Mathematics in the Sciences
27 Inselstrasse, Leipzig, Germany
long.yangjing@hotmail.com

Abstract. The estimation of human age from face images is an interesting problem in computer vision. We proposed a general distance metric learning scheme for regression problems, which utilizes not only data themselves, but also their corresponding labels to strengthen the credibility of distances. This metric could be learned by solving an optimization problem. Furthermore, the test data could be projected to this metric by a simple linear transformation and it is feasible to be combined with manifold learning algorithms to improve their performance. Experiments are conducted on the public FG-NET database by Gaussian process regression in the learned metric to validate our framework, which shows that the performance is improved over traditional methods.

Keywords: Age Estimation, Metric Learning, Regression.

1 Introduction

The estimation of human age from face images is an interesting problem in computer vision. As an important hint for human communication, facial images comprehend lots of useful information including gender, expression, age, pose, etc. Unfortunately, compared with other cognition problems, age estimation from face images is still very challenging. This is mainly due to the fact that, aging progress is influenced by not only personal gene but also many external factors. Physical condition, living style etc. may accelerate or slower aging process. Besides, since aging process is slow and with long duration, collecting sufficient data for training is a fairly strenuous work.

[10,17] formulated human ages as a quadratic function. Yan et al. [27,28] modeled the age value as the square norm of a matrix where age labels were treated as a non-negative interval instead of a certain fixed value. However, all of them regarded age estimation as a regression problem without special concern about the own characteristics of aging variation. As Deffenbacher [8] stated, the aging factor has its own essential sequential patterns. For example, aging is irreversible, which is expressed as a trend of growing older along the time axis. Such general evolution of aging course is beneficial to age estimation, especially when training data are limited and distributed unbalanced over each age range.

Geng et al. [13,12] firstly made some pioneer research on seeking for the underlying aging patterns by projecting each face in their aging pattern subspace (AGES).

X. Jiang and N. Petkov (Eds.): CAIP 2009, LNCS 5702, pp. 74–82, 2009.

Guo et al. [16] proposed a scheme based on Orthogonal Locality Preserving Projections (OLPP) [5] for aging manifold learning and get the state-of-art results. In [16], SVR (Support Vector Regression) is used to estimate ages on such a manifold and the result is locally adjusted by SVM. However, they only tested their OLPP-based method on a private large database consisting of only Japanese people, and no dimension reduction work was done to exact the so-called aging trend on the public available FG-NET database [1]. A possible reason is that, FG-NET database may not supply enough samples to recover the intrinsic structure of data. The lack of sufficient data is a prominent barrier in age estimation.

We propose a new framework aiming to learn a special metric for regression problems. Age is predicted based on the learned metric rather than the traditional Euclidean distance. We accomplish this idea by formulating an optimization problem, which approximates a special designed distance that scaled by a factor determined according to the labels of data. In this way, the metric measuring the similarity of samples is strengthened. More importantly, since labels are incorporated to depict the underlying sample distribution tendency, which signifies the inclusion of more information, a smaller amount of training data is required. Unlike the nonlinear manifold learning where it is repeated to find its low dimensional embedding, a merit of our framework is that, a full metric over the input space is learned and expressed as a linear transformation, and it is easy to project a novel data into this metric. Moreover, the proposed framework may also be used as a pre-processing step to assist those unsupervised manifold learning algorithms to find a better solution.

2 Metric Learning for Regression

Let $S = (\boldsymbol{X}_i, y_i)$ $(1 \le i \le N)$ denotes a training set of N observations with inputs $\boldsymbol{X}_i \in R^d$ and their corresponding non negative labels y_i. Our goal is to rearrange these data in high-dimensional space with a distinct trend as what their labels characterize. In other words, we hope to find a linear transformation T: $R^d \rightarrow R^d$, after applying which, the distances between each pair-wise observation may be measured as:

$$\hat{d}(X_i, X_j) = \| T(X_i - X_j) \|^2 \tag{1}$$

2.1 Problem Formulation

Metrics is a general concept, as a function giving a generalized scalar distance between two argument patterns [11]. Straightforwardly, different distances are also possible to depict the tendency of a data set. Similar to Weinberger et al. [25] and Xing et al. [26], we consider learning a distance metric of the form

$$d_A(X_i, X_j) = \sqrt{(X_i - X_j)^T A (X_i - X_j)} \tag{2}$$

But unlike their works for classification problems, in regression problems, every two observations are of different classes. Better metrics over their inputs are expected and a new metric learning strategy ought to be established.

Suppose given certain well-defined distance $\hat{d}_{ij}=\hat{d}(X_i,X_j)$ ideally delineating the data trend, our target is to approximate $\hat{d}_{ij}$ by $d_A(X_i,X_j)$ minimizing the energy function

$$\varepsilon(A)=\sum_{i,j}\left(d_A(X_i,X_j)^p-(\hat{d}_{ij})^p\right)^2 \tag{3}$$

To promise that A is a metric, A is restricted to be symmetric and positive semi-definite. For simplicity, p is assigned to be 2. This metric learning task is formulated as an optimization problem with the form below

$$\min\sum_{i,j}\left((X_i-X_j)^T A(X_i-X_j)-(\hat{d}_{ij})^2\right)^2 \tag{4}$$

satisfying the matrix A is symmetric and positive semi- definite. And there exists a unique lower triangular L with positive diagonal entries such that $A=LL^T$ [15]. Hence learning the distance metric A is equivalent to finding a linear transform L^T projecting observation data from the original Euclidean metric to a new one by

$$\tilde{X}=L^T X \tag{5}$$

2.2 Distance with Label Information

In practical application, Euclidean distance is not always capable to guarantee the rational relationship among input data. Although manifold learning algorithms may discover the intrinsic low-dimensional parameterizations of the high dimensional data space, at the outset, it also requires Euclidean distance to apply k-Nearest Neighbors to know the local structure of the original space. On the other hand, manifold learning demands a large amount of samples, which is not available in some circumstances.

For many regression and classification problems, it is in fact a waste of information if only data $\boldsymbol{X}_i$ is utilized but with their associated labels y_i ignored in the training stage. Balasubramanian et al. [2] proposed a biased manifold embedding framework to estimate head poses. In their work, the distance between data is modified by a factor of the dissimilarities fetched from labels. The basic form of this modified distance is

$$d'(i,j)=\frac{\beta\times P(i,j)}{\max_{m,n}P(m,n)-P(i,j)}\times d(i,j) \tag{6}$$

where $d(i, j)$ is the Euclidean distance between two samples $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. $P(i, j)$ is the difference of poses between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$.

Through incorporating the label information to adjust Euclidean distance, the modified distances are prone to give rise to the true tendency of data variation i.e. if the distance of two observations is large, then the distance of their labels is also large, vice versa. Hence it is intuitively that the biased distance is a good choice for $\hat{d}_{ij}$ in Eq.(3):

$$\hat{d}(i,j)=\left(\frac{\beta\times|L(i,j)|}{C-L(i,j)}\right)^p\times d(i,j) \tag{7}$$

Analogously, $L(i, j)$ is the label difference between two data. C is a constant greater than any label value in a train set and p is selected to make data easier to discriminate. $d(i, j)$ is the Euclidean distance between two samples $\boldsymbol{X}_i\psi$and $\boldsymbol{X}_j$.

2.3 Optimization Strategy

Since the energy function is not convex, it is a non-convex optimization and consequently it is impossible to find a closed form solution. The metric A is with the property to be symmetric and positive semi-definite, so it is natural to compute a numerical solution to Eq.(4) using the Newton's method. Similar to [26], in each iteration, a gradient descent step is employed to update A. The iteration algorithm is summarized as follows:

1. Initialize A and step length α;
2. Enforce A to be symmetric by $A \leftarrow (A+A^T)/2$;
3. The Singular Value Decomposition of $A=L^T \Delta L$, where the diagonal matrix Δ consists of the eigenvalues $\lambda_1,\ldots,\lambda_n$ of A and columns of L contains the corresponding eigenvectors;
4. Ensure A to be positive semi-definite by $A \leftarrow L^T \Delta' L$, where $\Delta'=\mathrm{diag}(\max(\lambda_1,0),\ldots,\max(\lambda_n,0))$;
5. Update $A' \leftarrow A - \alpha \nabla_A \varepsilon(A)$, where $\nabla_A \varepsilon(A)$ is the gradient of the energy function in Eq.(3) w.r.t. A;
6. Compare the energy function $\varepsilon(A)$ with $\varepsilon(A')$ in Eq.(3), if $\varepsilon(A)<\varepsilon(A')$, then augment the step length α with a momentum to accelerate the optimization process; otherwise, shrink α to assure a local minimum is not overpassed.
7. If A has converged or the maximum iteration times are reached, terminate; otherwise go back to Step 2.

3 Gaussian Process Regression

Given a training set $S = (\boldsymbol{X}_i, y_i)$ $(1 \le i \le N)$ as described in Section II and a sample $\boldsymbol{X}^*$ for query, GPR predicts its output y^* by putting a Gaussian process prior on this function $f(\cdot)$, assuming that all sample points evaluated from the function have a multivariate Gaussian density [20].

Let $\boldsymbol{X}=[\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N]$ and $\boldsymbol{Y}=[y_1,\ldots,y_N]^T$, the Gaussian predictive distribution of y^* is derived of the form

$$p(y^* | \boldsymbol{X}^*, \boldsymbol{X}, \boldsymbol{Y}, \Theta) \sim N(\mu(\boldsymbol{X}^*), V(\boldsymbol{X}^*)) \tag{8}$$

The mean prediction and covariance matrix in Eq.(8) are

$$\mu(\boldsymbol{X}^*) = \mathrm{k}(\boldsymbol{X}^*, \boldsymbol{X})[\mathbf{K}+\sigma^2 \mathrm{I}]^{-1} \boldsymbol{Y} \tag{9}$$

$$V(\boldsymbol{X}^*) = \mathrm{k}(\boldsymbol{X}^*, \boldsymbol{X}^*) - \mathrm{k}(\boldsymbol{X}^*, \boldsymbol{X})^T [\mathbf{K}+\sigma^2 \mathrm{I}]^{-1} \mathrm{k}(\boldsymbol{X}^*, \boldsymbol{X}) \tag{10}$$

where $\mathrm{k}(\cdot,\cdot)$ is the covariance function, $\mathbf{K}$ is the covariance matrix of $\mathbf{X}$ and σ^2 is the variance of noise.

Another way to perceive and thus rewrite Eq.(9) is to treat the mean prediction as a linear combination of N kernel functions:

$$\mu(X^*) = \sum_{c=1}^{N} \alpha_C \mathrm{k}(X^*, x_c) \tag{11}$$

where $\alpha=(\mathbf{K}+\sigma^2\mathrm{I})^{-1}\boldsymbol{Y}$

Gaussian kernel is a good choice for the covariance function

$$k(\boldsymbol{X}_i,\boldsymbol{X}_j)=v^2\exp(-\|\boldsymbol{X}_i-\boldsymbol{X}_j\|^2/2l^2+\sigma^2\sigma_{XiXj}) \tag{12}$$

In respect that the proposed learned metric encodes label information implicitly, it is bestowed as the similarity measure and Eq.(12) becomes

$$k(\boldsymbol{X}_i,\boldsymbol{X}_j)=v^2\exp(-(\boldsymbol{X}_i-\boldsymbol{X}_j)^T A(\boldsymbol{X}_i-\boldsymbol{X}_j)/2l^2+\sigma^2\sigma_{XiXj}) \tag{13}$$

4 Experimental Results

Age estimation is carried on the public FG-NET Aging Database [1] by the regression strategy on the basis of the proposed metric. The database contains totally 1002 color or gray images from 82 people. Each person has around 10 face images with the ranges from 0 to 69 with labeled ground truth. These images are taken under varying lighting condition, poses and expressions. Each image is labeled by 68 points characterizing its shape features. Similar to [13,16,27,28], input features are selected to be the parameters of AAMs [6].

Firstly we hope to testify that the proposed metric is able to disinter some internal patterns of human's aging progression. We randomly choose 300 images out of all the 1002 images in FG-NET Database as training samples, and the rest as test samples. The parameters in Eq.(7) are chosen as C=100, β=1 and p=1. The energy function is converged after 50 iterations or so. Figure 1(a) and 1(b) portrays the positional relationship among training samples in the hyper-space measured by Euclidean distance and the learned metric A. The 2D view is acquired by Multi-Dimensional Scaling (MDS) [7]. Figure 2 plots the relative position of the remaining 702 image samples for test. Contrast to Figure 1, manifold learning algorithms like Isomap, LLE and OLPP fails to predicate the aging trend sometimes. Furthermore, though only 30% of the entire data set is directed for learning the aging trend is effectually set up.

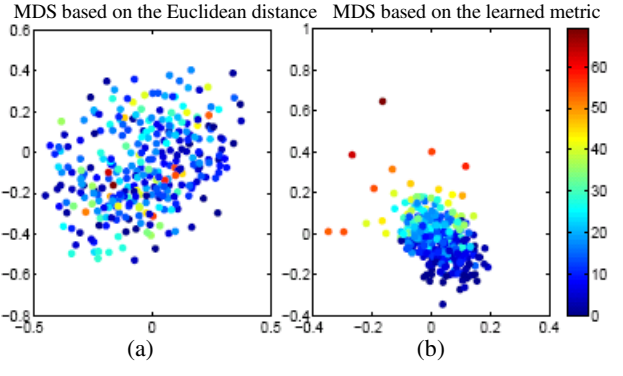

Fig. 1. 2-D view of the clustering effects of the 300 training samples by metric learning. It illustrates the 2 dimensional embedding of the training data sampled from FG-NET Aging Database by MDS. Points of age from 0 to 69 are marked from blue to red. It is seen that, the distance calculated based on our learned metric in Figure (b) preserves local proximity of samples with close labels better than that based on the traditional Euclidean distance in Figure (a).

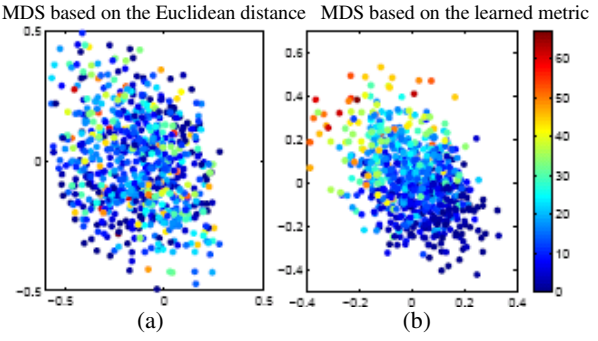

Fig. 2. 2-D view of the clustering effects of the 702 testing samples by metric learning, corresponding to Figure 1. It is obvious that, the actual aging trend is, to some extended, manifested in the hyper-space based on our learned metric.

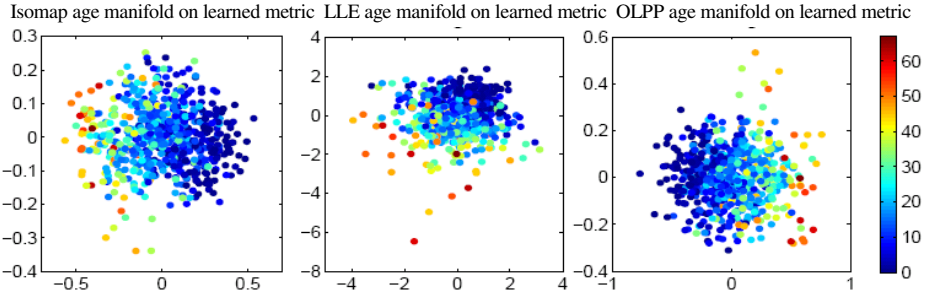

Fig. 3. 2D age manifolds. This figure illustrates the 2 dimensional embedding of FG-NET Aging Database by Isomap, LLE and OLPP algorithms based on our learned metric.

As in Eq.(5), the original parameters from AAMs can be linearly transformed into a hyper-space based on our learned metric, by multiplying L^T satisfying $A=LL^T$. Figure 3 draws the 2D aging manifold inputted with the transformed data. Compared to Figure 1, the linear transform L^T is salutary for other manifold algorithms to find an improved aging trend.

Then, age estimation of our methodology is compared with the performance of some state-of-art approaches. The Leave- One-Person-Out mode [13,16,27,28] is the mechanism for experimentation, i.e. each time we choose one person for testing and all others for training. The same as in [13,16,27, 28], two criteria are adopted for performance evaluation. One is the Mean Absolute Value (MAE), which is defined as

$$MAE = \sum_{i=1}^{N} | \widetilde{age}_i - age_i | / N \tag{14}$$

where for each $\boldsymbol{X}_i$, $\widetilde{age}_i$ is its labeled ground truth and age_i is the estimated age. N is the number of testing images.

Another widely acknowledged criterion is the cumulative score at error level l [13]

$$CumScore(l)=N_{error \le l}/\mathrm{N}\times 100\% \tag{15}$$

In respect that, when a face image is labeled as O years old, the person is customarily thought to be $[O,O+1)$ years old [27], thus the error less than a specified number of years is by and large neglectable in practical application. Eq.(15) is an indicator of the algorithmic correct rate.

The parameters in Eq.(7) are rectified to be C=80, β=1 and p=0.6. Table 1 lists the MAE of different approaches. The MAE of the proposed method is almost the same as the best one [16]. However, unlike their LARR, we simply predict ages in a new metric by regression without any local refinement. LARR slides the estimated age up and down by checking different age values to see if it can come up with a better prediction [16]. The parameters defining the search range is determined manually, which is at least not convenient and automatic enough, and may be laborious and not feasible in some real-world applications. Table 2 details Table 1 with separate MAEs over different age range. The MAE of our method in younger people is slightly higher than other recent methods. As compensation, an outstanding improvement is achieved in the larger age range. This trait is fairly attractive considering the fact that, people over 30 years old account for less than 15% of the whole FG-NET database. Even if there are only a few samples (for example, there are only 8 images out of 1002 over 60 years old), a relatively acceptable age prediction can be obtained.

Table 1. MAE comparison of different methods

Reference	Method	MAE
[13]	AGES	6.77
[12]	KAGES	6.18
[27]	RUN1	5.78
[28]	RUN2	5.33
[16]	LARR	5.07
Proposed	Metric learning+GPR	5.08

Table 2. MAEs over various age ranges on FG-NET Database for the proposed method, GPR and RUN. In the first column, the value in the parenthesis stands for the proportion (percentage) for each age group out of the whole database.

Age Range	Proposed	GPR	RUN[27]
0-9(37.0%)	2.99	3.55	2.51
10-19(33.8%)	4.19	4.34	3.76
20-29(14.4%)	5.34	5.09	6.38
30-39(7.9%)	9.28	9.04	12.51
40-49(4.6%)	13.52	14.65	20.09
50-59(1.5%)	17.79	19.77	28.07
60-69(0.8%)	22.68	31.76	42.50
Average	5.08	5.45	5.78

5 Conclusions

In this paper, a new metric learning framework is proposed to resolve regression problems. It is feasible to be applied to many other problems in machine learning or computer vision. No assumptions about the structure or distribution of the samples are made, and a relatively small quantity of training samples is required to learn their underlying variation trend. Experiments shows the effectiveness of the learned metric to restore the intrinsic infrastructure of input sample data and encouraging performance is acquired on a widely used public face aging database.

References

1. FG-NET Aging Database, http://www.fgnet.rsunit.com
2. Balasubramanian, V.N., Ye, J., Panchanathan, S.: Biased manifold embedding: A framework for person-independent head pose estimation. In: IEEE Conf. CVPR, pp. 1–7 (2007)
3. Bar-Hillel, A., Weinshall, D.: Learning distance function by coding similarity. In: Proc. ICML, pp. 65–72 (2007)
4. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: a geometric framework for learing from labeled and unlabeled examples. Journal of Machine Learning Research 7, 2399–2434 (2006)
5. Cai, D., He, X., Han, J., Zhang, H.J.: Orthogonal laplacianfaces for face recognition. IEEE Trans. Image Processing 15, 3608–3614 (2006)
6. Cootes, T., Edwards, G., Taylar, C.: Active appearance models. IEEE Trans. Pattern Analysis & Machine Intelligence 23(6), 681–685 (2001)
7. Cox, T., Cox, M.: Multidimensional Scaling. Chapman & Hall, Lodon (1994)
8. Deffenbacher, K.A., Vetter, T., Johanson, J., O'Toole, A.J.: Facial aging, attractiveness, and aistinctiveness. Perception 27 (1998)
9. Donoho, D.L., Grimes, C.E.: When does geodesic distance recover the true hidden parametrization of families of articulated images? In: Proc. European Symposium on Artificial Neural Networks (2002)
10. Draganova, A.L.C., Christodoulou, C.: Comparing different classifiers for automatic age estimation. IEEE Trans. Systems, Man, and Cybernetics 34(1), 621–628 (2004)
11. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. John Wiley & Sons, Inc., New York (2001)
12. Geng, X., Smith-Miles, K., Zhou, Z.-Z.: Facial age estimation by nonlinear aging pattern subspace. In: Proc. ACM Conf. Multimedia (2008)
13. Geng, X., Zhou, Z.H., Zhang, Y., Li, G., Dai, H.: Learning from facial aging patterns for automatic age estimation. In: Proc. ACM Conf. Multimedia, pp. 307–316 (2006)
14. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: NIPS (2005)
15. Golub, G.H., Loan, C.F.V.: Matrix Computations. Johns Hopkins Univ. Press (1996)
16. Guo, G., Fu, Y., Dyer, C., Huang, T.S.: Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Trans. on Image Processing 17, 1178–1188 (2008)
17. Lanitis, A., Taylor, C.J., Cootes, T.: Toward automatic simulation of aging effects on face images. IEEE Trans. Pattern Analysis and Machine Intelligence 24(4), 442–455 (2002)
18. Neal, R.M.: Monte carlo implementation of gaussian process models for bayesian regression and classification. Technical Report CRG-TR-97-2

19. Nilsson, J., Sha, F., Jordan, M.I.: Regression on manifolds using kernel dimension reduction. In: IEEE Conf. ICML, pp. 265–272 (2007)
20. Raumussen, C.E., Williams, C.K.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
21. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
22. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
23. Sugiyama, M., Hachiya, H., Towell, C., Vijayakumar, S.: Geodesic gaussian kernels for value function approximation. Autonomous Robots 25, 287–304 (2008)
24. Tenebaum, J.B., de. Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionally reduction. Science 290(5500), 2319–2323 (2000)
25. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. In: Proc. NIPS, pp. 1475–1482 (2006)
26. Xing, E., Ng, A., Jordan, M.I., Russell, S.: Distance metric learning with application to clustering with side-information. In: Proc. NIPS (2002)
27. Yan, S., Wang, H., Huang, T.S., Tang, X.: Ranking with uncertain labels. In: IEEE Conf. Mulitimedia and Expo, pp. 96–99 (2007)
28. Yan, S., Wang, H., Tang, X., Huang, T.S.: Learning autostructured regressor from uncertain nonnegative labels. In: IEEE Conf. ICCV, pp. 1–8 (2007)