Learning Nonlinear Distance Functions using Neural Network for Regression with Application to Robust Human Age Estimation

Na Fan
Department of Electronic Engineering
East China Normal University
fanna.cn@gmail.com

Abstract

In this paper, a robust regression method is proposed for human age estimation, in which, outlier samples are corrected by their neighbors, through asymptotically increasing the correlation coefficients between the desired distances and the distances of sample labels. As another extension, we adopt a nonlinear distance function and approximate it by neural network. For fair comparison, we also experiment on the regression problem of age estimation from face images, and the results are very competitive among the state of the art.

1. Introduction

Distance metric learning is a heavily investigated topic (see [1] for a survey), and is frequently applied for computer vision problems, such as [3-4, 6-11] discussed in this paper. Distance functions or dissimilarity measures are pivotal to many models and algorithms in pattern recognition, machine learning, and computer vision, such as the k-Nearest Neighbors (kNN) based algorithms, Radial Basis Function (RBF) networks, Support Vector Machines (SVM), manifold learning, and kernel regression for various classification, clustering or regression problems. It is not only useful for supervised learning, but also supplies semi-supervised or unsupervised learning tasks (such as [2]) with pairwise side information quantifying the degree of similarity or dissimilarity among samples. Even combined with the simplest kNN classifiers, learning a distance metric from labeled examples yields quite competitive results, such as reported in [3].

Recently, researchers found that, effectively taking advantage of label information of data samples is helpful for both classification [4-9] and regression [10, 26] models, especially when the training set is of small size and less structured than the traditional homogeneous one [5, 6, 10, 26]. In classification problems, label information is served as a weight term of energy objective functions to attract samples of the same class to move closer via learning a distance function [4-9]. Without binning[1], these approaches [4-9] cannot be extended directly for regression problems, because the labels in regression problems are not discrete, leading to infinite number of classes.

Like Weinberger et al. [3] and Xing et al. [5], Jin [10] adopted a linear distance function [2] by replacing the inverse covariance matrix in Mahalanobis distance with a symmetric and positive semi-definite matrix, and this matrix is learned via approximating a label-information biased Euclidean distance, according to the least square criterion. Such approximation is resolved by numerically minimizing an energy function using the Newton's method similar to Xing et al. [5]. Based on the learned distance function, Gaussian Process Regression (GPR) is applied for human age estimation and reasonable performance is achieved on the FG-NET Aging Database [14]. The learned distance function is combined with a simple kNN regressor, and state-of-the-art results are reported.

Before Jin et al. [10], Balasubramanian et al. [13] combined an analogous label-information biased Euclidean distance with manifold learning to estimate head pose. However, manifold learning techniques such as Isomap, Locally Linear Embedding (LLE), and Laplacian Eigenmap is not advocated in [10]. The reason is that, opposite to the fundamental assumption of manifold learning, the training data available in the human age estimation problem are usually sparsely and inhomogenously sampled. This is natural considering the difficulty of collecting face images of the same person over lifespan (remember, cameras are luxury and not popular five decades ago).

To incorporate label information in the learning stage, a key issue is how to reconcile the inconsistencies between the feature space and the semantic space. Particularly, it

[1] Here, binning refers to finding appropriate split points to convert continuous age numerical values into a number of age bins. See [15] for a survey and performance evaluation among several popular binning methods. In Section 5, we also show experimentally that such binning conversion is suboptimal.

[2] Ironically, contrary to [10], neither of their distance functions can be referred to as a metric, because the distance functions defined in [10] do not satisfy the triangle inequality, one of metric axioms. Instead, they should be called non-metric distance or semi-metrics, in conformance to most existing literature such as [12]. In Section 4 and Section 5, we will discuss this in detail.

978-1-4577-1102-2/11/$26.00

refers to how to handle the case in which two neighboring points are with disparate labels. In this paper, we further explore such discrepancy between the two spaces. We proposed an iterative framework, in which, outlier samples are prone to be corrected by their neighbors, and thus robust regression is achieved. It is easy to see from this framework that the method of Jin et al. [10] iterate only once, and it is feasible to iterate more times to obtain better results. Besides, we extend the first order correlation coefficients between dimensions of the feature space in [3, 5, 6, 7, 10, 11] to a more generalized polynomial-like form, which is approximated via a variation of Neural Networks (NNs), called Compositional Pattern-Producing Networks (CPPNs) [16]. From these two aspects, the work of Jin et al. [10] may be viewed as a special case of our model.

For fair comparison with [10], our method is also applied to the regression problem of human age estimation. Literature related to age estimation from face images is too wide to review here. We refer readers to [17] (and also Section 5) for an up-to-date survey.

The rest of the manuscript is organized as follows: Section 2 formulates our distance function model; Section 3 details how to approximate the proposed distance function through NNs; Section 4 presents the iterative framework; Experimental results are discussed in Section 5 and finally, Section 6 draws concluding remarks and points out some possible future work.

2. Problem Formulation

Denote $S = (\boldsymbol{X}_i, y_i)$ $(1 \le i \le N)$ as a training set of N labeled samples with inputs $\boldsymbol{X}_i \in R^d$ and their associated continuous non-negative labels y_i. We use x_{iu} denoting the u^{th} component (i.e. dimension) of $\boldsymbol{X}_i$.

We follow the terminology in [10], and refer to the space constructed by the observed sample data (i.e. input data) $\boldsymbol{X}_i$ as the *feature space*. The *semantic space* refers to a space characterized by the labels y_i. In other words, in the semantic space, data samples are distributed according to what their labels y_i dictate.

Mahalanobis distance is based on correlations between dimensions x_{iu} and x_{jv} $(1 \le i,j \le N, 1 \le u,v \le d)$. With the covariance matrix C, it is defined as

$$d(X_i, X_j) = \sqrt{(X_i - X_j)^T C^{-1} (X_i - X_j)} \quad (1)$$

Mahalanobis distance does not consider the labels y_i. In [3, 5, 6, 7, 10], C^{-1} is replaced by a matrix to be computed, with the objective to satisfy

$$d(X_i, X_j) = \sqrt{(X_i - X_j)^T A (X_i - X_j)} = dd_{ij} \quad (2)$$

Where dd_{ij} is the *desired distance* between sample i and j. In [10], dd_{ij} is selected to be the Euclidean distance between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ biased by labels y_i and y_j. We will discuss how to select dd_{ij} in Section 4. The distance of labels $dd_{ij} = |y_i - y_j|$ is a feasible but not optimal choice.

When solving computer vision problems, attention should be paid to equality constraints. Many equality rarely holds either theoretically (because the problem is over-constraint) or practically (because error is inevitable). In fact, most equality only implies approximate equality. Often, the difference between the left and right side of the equality is measured by a defined norm, and then accumulated (or integrated for continuous case) as an energy objective function for optimization.

Square Eq.(2) on both sides, and using the least square criterion yields an energy objective function

$$E(A) = \sum_{i=1}^{N} \sum_{j=1}^{N} ((X_i - X_j)^T A (X_i - X_j) - dd_{ij}^2)^2 \quad (3)$$

Customarily, the perception and learning mechanism of human beings is complicated and might be simplistic to be model by a linear metric. For better illustration, we rewrite Eq.(2) in component-wise form

$$d(X_i, X_j) = \sum_{u=1}^{d} \sum_{v=1}^{d} a_{uv} (x_{iu} - x_{ju})(x_{iv} - x_{jv}) \approx dd_{ij} \quad (4)$$

A doubt about Eq.(4) is that, why both the power of the factor $(x_{iu}–x_{ju})$ and $(x_{iv}–x_{jv})$ is so coincidental to be 1? To address this question, we generalize Eq.(4) to be

$$d(X_i, X_j) = \sum_{u=1}^{d} \sum_{v=1}^{d} \sum_{k=1}^{m} a_{uv} (x_{iu} - x_{ju})^{p_k} (x_{iv} - x_{jv})^{q_k} \approx dd_{ij} \quad (5)$$

Where $\{p_k\}_{k=1}^{m}$ and $\{q_k\}_{k=1}^{m}$ are two sequences containing the powers of $(x_{iu}–x_{ju})$ and $(x_{iv}–x_{jv})$ respectively. As $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = d(\boldsymbol{X}_i – \boldsymbol{X}_j)$, $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$ is actually a function of d variables $(x_{i1}–x_{j1})$, $(x_{i2}–x_{j2})$,…, $(x_{id}–x_{jd})$.

As suggested in [1], being continuous is a reasonable assumption for distance functions. In 1957, Kolmogorov [18] proved a theorem stating that, any continuous multivariate function can be exactly represented by superpositioning several continuous univariate functions. To make it easier to understand, we present the Kahane's representation [19] of this Kolmogorov theorem below.

$$f(x_1, x_2, ..., x_n) = \sum_{q=1}^{2n+1} g\left(\sum_{p=1}^{n} \lambda_p h_q(x_p) \right) \quad (6)$$

Nielsen [20] pointed out that, this theorem can be interpreted in the context of Neural Network (NN). Specifically, any continuous multivariate function is equivalent to an NN with two hidden layers (i.e. four-layer NN) whose transfer functions are all continuous univariate functions.

However, Kolmogorov's proof is not completely algorithmic as it does not describe how to derive $g(\cdot)$ and $h_q(\cdot)$ from $f(\cdot)$. To solve this, many algorithmic solutions have been proposed, such as the one by Kurkova [21]. Nevertheless, they do not solve our problem of approximating $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$, because $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$ is unknown. What we know is only N^2 input-output pairs i.e. $\{\boldsymbol{X}_i – \boldsymbol{X}_j, dd_{ij}\}$, as given by the training set.

3. Neural Networks

As we are not able to solve the above formulated model mathematically, we resort to CPPNs [16]. CPPNs are a variation of NNs with a different set of transfer functions. While NNs typically only contain a certain type of sigmoid or radial basis functions, CPPNs can comprise a mixed configuration of various types of functions. CPPNs were originally invented to mimic the rationale of biological neurons, and it has been successfully applied to several problems in the area of cognitive science.

The topology of our four-layer feed-forward back propagation NN is specifically designed for learning distance functions. The input, the 1st hidden, the 2nd hidden and the output layers has 200, 226, 3 and 1 neuron respectively (i.e. a 200-226-3-1 topology [3]) for the case of the FG-NET Aging Database. For the sake of adjustability and flexibilities, the 1st hidden layer has multiple transfer functions, each of which is assigned a specific number of neurons as listed in Table 1. Note that the selected 9 transfer functions are commonly used to approximate general high-order functions. The 2nd hidden layer has 3 transfer functions, namely, $f(x)=x$, $logsig(x)$ and $radial\_basis(x)$, for maintaining both linearity and non-linearity passed in from the 1st layer.

Table 1. Transfer functions in the 1st hidden layer

Transfer function definition	No. of neurons assigned
$f(x)=x$	1
$f(x)=exp(x)$	15
$tan\_sig(x)=2/(1+exp(-2x))-1$	30
$log\_sig(x)=1/(1+exp(-x))$	30
$double\_sig(x)=exp(-x)/(1+exp(-x))^2$	30
$double\_log\_sig(x)=sgn^4(x)(1-exp(-x^2))$	30
$f(x)=cos(x)$	30
$radial\_basis(x)=exp(-x^2)$	30
$f(x)=1-\|x\|, -1\le x\le 1, f(x)=0$ otherwise	30

We use the Scaled Conjugate Gradient method proposed in [22] as the training algorithm. Similar to Eq.(3), the training of NNs i.e. weights and bias values adjustment is also according to the least square criterion [5].

4. The Proposed Iterative Framework

In this section, we will discuss how to select the desired distances dd_{ij} in Eq.(5) for each iteration. For expression convenience, we call the 1-D distance of labels as *ideal distance* $id_{ij}=|y_i-y_j|$, and the results of approximating dd_{ij} via a distance function as *achieved distance* $ad_{ij}=d(\boldsymbol{X}_i,\boldsymbol{X}_j)$.

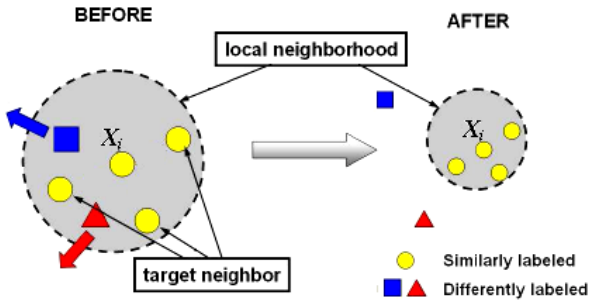
[3] Other NN topologies with similar size only lead to a slight performance difference. Investigation of the optimal topology is a pure machine learning problem, which is out of the scope of this work. Here we only present a good network configuration, but its optimality is not guaranteed.

[4] $sgn(x)$ denotes the sign function. $sgn(x)=0$ for $x=0$; $sgn(x)=x/|x|$ for $x\neq 0$.

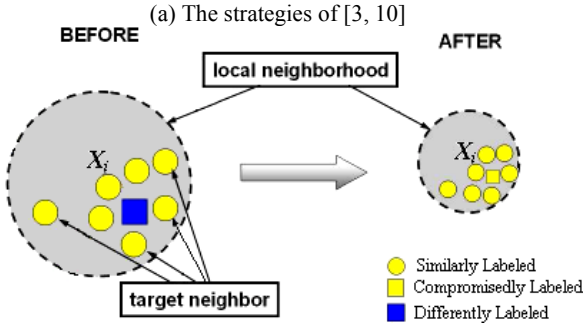
[5] It is referred to as Mean Square Error (MSE) in the context of NN.

By reviewing Section 2, the ideal distance id_{ij} corresponds to the semantic space, and each achieved distance ad_{ij} corresponds to a learned space. Like the labels y_i, the ideal distances id_{ij} is fixed and do not change in the learning stage.

He et al. [23] argued that, it is better to apply learning algorithms in the semantic space than directly in the feature space, because the former is more consistent with human perception. But this is just a good will, considering the fact that, the ideal distance is known only among the N points of the training set, as denoted by id_{ij}. This is exactly the reason that, we have to approximate id_{ij} via $d(\boldsymbol{X}_i,\boldsymbol{X}_j)$ according to the feature space, with the objective that ad_{ij} is as close to id_{ij} as possible.



(a) The strategies of [3, 10]



(b) Our strategy to trade off the contradiction between id_{ij} and ad_{ij} so that the label of the blue squared sample (that turns yellow after training) is compromised to be dd_{ij}. This is the situation that [3, 10] fails to tackle.

Figure 1. Schematic diagrams of the strategies in [3, 10] and this paper. Here we only take the neighborhood of one sample $(\boldsymbol{X}_i, y_i)$ as an example. In both Figure (a) and (b), the left part is the situation of neighborhood before training and the right part delineates the situation after training. Data with similar labels are marked in the same shape and color.

Attempts have been made to reduce the gap between the feature space and its corresponding semantic space, such as manifold learning [23]. As plotted in Figure 1(a), a critical situation considered by [3, 10] is that, two neighboring points in the feature space are with large distance in the semantic space. Note that, unlike [3, 10], in regression problems it is not necessary to leave a margin among points of different labels, as the transition between labels are

continuous and smooth.

But what if the blue squared point shown in Figure 1(b) is impossible to be detached away? There exist three explanations for such situation: 1) The blue squared point is an outlier; 2) The dimensionality of the distance function learning algorithm [6] or the feature space is not high enough; 3) Under the defined optimality criterion, it does not worth to detach this "heretical" point away. The actual situation might be a combination of the above three. But whatever reason it is, our strategy is to compromise by lowering our expectation on the desired distance dd_{ij} and making it closer to the currently achieved distance ad_{ij}.

Our observation is that, an appropriate desired distance dd_{ij} is of crucial importance, and it should reconcile the inconsistencies between the ideal distance id_{ij} and a previously achieved distance ad_{ij}. The chosen dd_{ij} should correspond to a transitional space between the semantic space and the currently learned space, so that the distance function learning process is essentially a process of transiting from the feature space to the semantic space.

Eq.(1) in [13] and Eq.(7) in [10] composed two metrics in a multiplicative way, and the resulting distance function does not satisfy the triangle inequality (An obvious counterexample is to combine two three-point metrics both with $d(a,b)=1$, $d(b,c)=1$, $d(a,c)=2$).

Although Tan et al. [12] argued that, non-metric distances might fit practical problems better than metrics. In their face image matching problem, their observation is based on a rationale intuitively explained as, both a human and a horse may be similar to a centaure, but a human and a horse is not similar to each other. This motivates them to use min{} function to define the similarity measure as the most similar sub-block between face images, and consequently the triangle inequality property is lost. But the situation is not analogous in either head pose estimation or age estimation: If age 31 is similar to age 32, and age 32 is similar to age 33, it is improper to claim that, age 31 is not similar to age 33.

In this paper, we set the desired distance in the current iteration to be a weighted sum of the ideal distance and the achieved distance in the previous iteration:

$$dd_{ij} = \alpha\, id_{ij} + (1-\alpha) ad_{ij} \qquad 0 < \alpha < 1 \qquad (7)$$

Where α is a weight adjusting the importance of id_{ij} and ad_{ij}. It is not difficult to prove that, if id_{ij} and ad_{ij} are both metrics, dd_{ij} is also a metric.

Normally, it is not feasible to add two values of different scales and units, but id_{ij} and ad_{ij} is supposed be close, since ad_{ij} tries to approximate id_{ij}. Note that, the weighted sum here is component-wise: if ad_{ij} is closer to dd_{ij} (and also id_{ij} very probably) for certain i and j in the previous iteration, the component dd_{ij} will be updated to be closer toward id_{ij} in the current iteration independently, regardless of other component e.g. $dd_{i'j'}$ ($i' \neq i$ or $j' \neq j$).

Initially, ad_{ij} is set to be proportional to the Euclidean distance [7] of the feature space. The pseudo-code of the proposed iterative framework is outlined as follows:

1. Initialize ad_{ij} according to the Euclidean distance between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$
2. $dd_{ij} = \alpha\, id_{ij} + (1-\alpha)\, ad_{ij}$
3. $ad_{ij} = d(\boldsymbol{X}_i, \boldsymbol{X}_j)$, where $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$ is the output of the trained NN with dd_{ij} as target value
4. If the stopping criterion is not met, goto step 2

The stopping criterion is met when the average update of ad_{ij} in the current iteration is less than 0.01. A trick to save computing cost is that, we simply leave the weight and bias values of the NN in the previous iteration as the initial weight and bias values in the current iteration. From the perspective of NN, our algorithm may be viewed as a weight and bias value updating algorithm, in which, once the learned NN converges to target value dd_{ij}, the target value dd_{ij} is modified in terms of Eq.(7), and then weight and bias adjustment is needed again, so such process iterates. Figure 2 is a visualization of samples in the space based on iteratively learned distance functions. It shows that, after 3 iterations the aging trend is gradually clear.

5. Experimental Results

The final learned distance $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$ is combined with the common regression technique kNN with k=10. Our method is tested on the FG-NET database that actually serve as benchmark for human age estimation methods [10, 26-33].

Each face image in FG-NET has 68 labeled points characterizing shape features, which are combined with appearance features to form a face representation of 200 parameters [26-28,30,31], called Active Appearance Model (AAM) [25]. Figure 3 gives some typical AAM labeled face images. We follow the popular test scheme, namely Leave-One-Person- Out (LOPO), which was usually taken for the FG-NET database, as suggested in [10, 26, 27, 30-33].

For performance evaluation, two widely used criteria in [10, 26, 27, 30-33] are adopted: Mean Absolute Error (MAE) and Cumulative Score (CS). The MAE is defined as the average of the absolute errors between the estimated and the ground truth ages, $MAE = \sum_{i=1}^{M} | y_i^* - y_i | / M$, where y_i is the ground truth age for the test sample, y_i^* is the estimated age, and M is the total number of test samples. The CS is defined as $CS(w) = M_{e \le w}/M \times 100\%$, where $M_{e \le w}$ is the number of test samples on which the absolute error of estimated age is no higher than w years.

Table 2 and Figure 4 is the MAE over different age intervals and the CS of our method respectively. Compared

[6] Here, the dimensionality of a regressor refers to the Vapnik–Chervonenkis dimension based complexity, see [24] for details.

[7] It is scaled so that the mean of such Euclidean distance equals to the mean of id_{ij}. Note that, distance itself is first order derivative and we do not need to scale according to its variance.

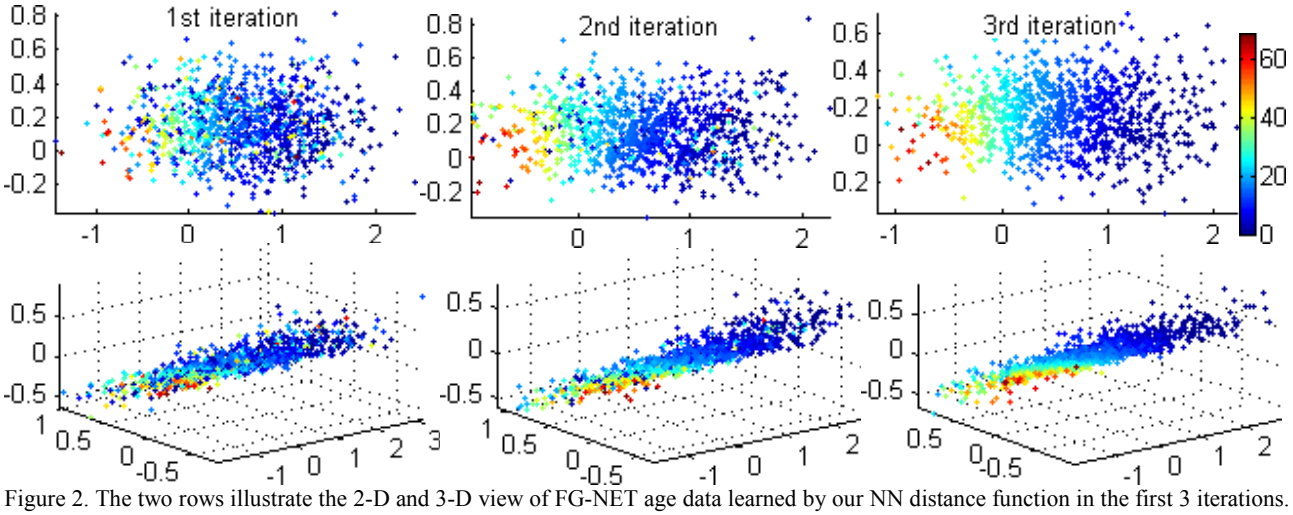


Figure 2. The two rows illustrate the 2-D and 3-D view of FG-NET age data learned by our NN distance function in the first 3 iterations. The data points of age from 0 to 69 are colored from blue to red.

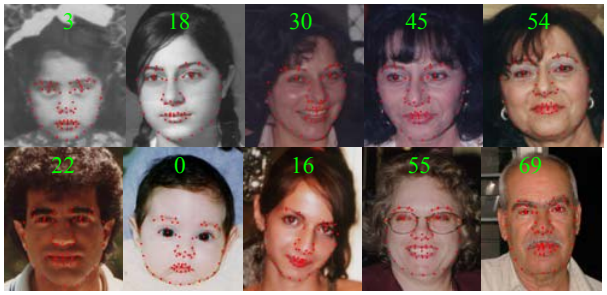


Figure 3: Some typical sample images of FG-NET Aging Database. The 68 red spots are landmark points labeled by AAM. The first row shows collected aging images from the same person and the last row is from different persons. Ground truth age is displayed at the top of each image. Because original images in the database are not of the same size, they are properly scaled for displaying purpose.

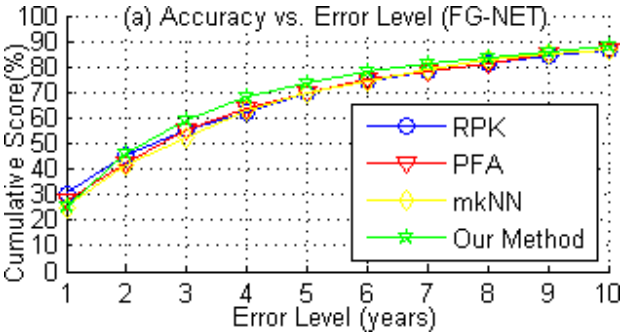


Figure 4. Cumulative scores on the FG-NET databases of the methods in [10, 32, 33] and this paper

to recent methods in [32, 33], 1) our method has smaller MAE over almost all age intervals; 2) the CS(w) of our method is evidently higher when $3 \le w \le 6$, which manifests the robustness of our method, because the absolute error of our estimated age tends to fall between 3 and 6 years rather than a larger range over 10 years.

As a performance baseline for regression algorithms, two popular discretization algorithms, the entropy method and the density method in [15], are used to compute split points, so that every numerical age label y_i falls into an age interval (i.e. an age bin) and then classification algorithms such as [3] can be applied in a brute force way. In order to get lower MAE, we manually tune the number of splits and find that, 38 split points are the best for the FG-NET dataset. We also tried a scheme that directly sets $\boldsymbol{X}_i$ as the inputs and y_i as the target values to the NN illustrated in Section 3 and report the MAE under the name "direct NN" in Table 3. Table 3 shows that, neither direct NN nor binning plus classification can give satisfactory regression accuracy. An explanation is that, turning the problem to classifying age bins demands more samples to depict the infrastructure of each age bin class. This is exactly a prominent challenge of FG-NET, as it does not provide sufficient samples over some age intervals. As a series of previous literature, we also present an up-to-date performance comparison among the state of the art in Table 4.

Table 2: MAEs over different age intervals of our method

Age Interval	#img.	FG-NET
0-9	371	2.04
10-19	339	4.69
20-29	144	3.84
30-39	79	6.73
40-49	46	14.30
50-59	15	23.60
60-69	8	29.22
70-93	0	–
Overall	1002	4.67

Table 3. MAE comparison of suboptimal methods as baseline

Method	FG-NET
Entropy Binning[15]+Classifier[3]	18.06
Density Binning[15]+Classifier[3]	17.42
direct NN	7.13

Table 4. MAE comparison of different methods

Method	FG-NET
WAS[27]	8.06
AGES[27]	6.77
KAGES[28]	6.18
QM[29]	6.55
MLPs[29]	6.98
RUN[30]	5.78
BM[31]	5.33
LARR[26]	5.07
PFA[32]	4.97
RPK[33]	4.95
Linear metric+GPR[10]	5.08
Proposed	4.67

6. Concluding Remarks

In this paper we propose an iterative framework for distance function learning in regression problems, of which, as an example, the single iteration module is chosen to be a CPPN and the assembled algorithm is applied for human age estimation. We expect to see more applications of this framework to other challenging regression problems.

References

[1] L. Yang, R. Jin, Distance metric learning: a comprehensive survey, Technical report, Michigan State University, 2006. http://www.cs.cmu.edu/~liuy/frame_survey_v2.pdf

[2] D. Y. Yeung, H. Chang, A kernel approach for semi-supervised metric learning, IEEE Transactions on Neural Networks, 18(1): 141-149, 2007.

[3] K. Weinberger, J. Blitzer, L. Saul, Distance metric learning for large margin nearest neighbor classification, Proc. NIPS, pp.1475-1482, 2006.

[4] A. Bar-Hillel, D. Weinshall, Learning distance function by coding similarity, Proc. ICML, pp.65-72, 2007.

[5] E. Xing, A. Ng, M. I. Jordan, S. Russell, Distance metric learning with application to clustering with side-information, Proc. NIPS, pp.505-512, 2002.

[6] J. Goldberger, S. Roweis, G. Hinton, R. Salakhutdinov, Neighbourhood components analysis, Proc. NIPS, pp.513-520, 2005.

[7] S. Shalev-Shwartz, Y. Singer, A. Y. Ng, Online and batch learning of pseudo-metrics, Proc. ICML, pp.743-750, 2004.

[8] N. Shental, T. Hertz, D. Weinshall, M. Pavel, Adjustment learning and relevant component analysis, Proc. ECCV, pp.776-792, 2002.

[9] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, Proc. CVPR, pp.539-546, 2005.

[10] C. Jin, Y. J. Long, On label information incorporated metric learning for regressions, International Journal of Computational Intelligence and Applications, 9(4): 339 - 351, 2010.

[11] Y.J.Long, and Y.Z.Huang, Image based source camera identification using demosaicking, Proceedings of the 8th International Workshop on Multimedia Signal Processing, pp. 419-424, 2006.

[12] X. Tan, S. Chen, J. Li, Z. Zhou, Learning non-metric partial similarity based on maximal margin criterion, Proc. CVPR, pp.138-145, 2006

[13] V. N. Balasubramanian, J. Ye, S. Panchanathan, Biased manifold embedding: A framework for person-independent head pose estimation, Proc. CVPR, pp.1-7, 2007.

[14] FG-NET Aging Database, http://www.fgnet.rsunit.com

[15] S. A. Macskassy, H. Hirsh, A. Banerjee, A. A. Dayanik, Converting numerical classification into text classification, Artificial Intelligence, 143(1):51-77, 2003.

[16] K. O. Stanley, Compositional pattern producing networks: A novel abstraction of development, Genetic Programming and Evolvable Machines, 8(2): 131-162, 2007.

[17] N. Ramanathan, R. Chellappa, S. Biswas, Age progression in human faces: A survey, Journal of Visual Languages and Computing, In Press, 2009.

[18] A. N. Kolmogorov, On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition, Doklady Akademii Nauk SSSR, 114: 953-956, 1957.

[19] J. P. Kahane, Sur le théorème de superposition de Kolmogorov, Journal of Approximation Theory, 13(3): 229-234, 1975.

[20] R.H.Nielsen, Kolmogorov's mapping neural network existence theorem, Proceedings of IEEE International Conference on Neural Networks 1987, pp.11-14.

[21] V.Kurkova, Kolmogorov's theorem and multilayer neural networks, Neural Networks, 5(3): 501-506, 1992.

[22] A. F. Moller, A scaled conjugate gradient algorithm for fast supervised learning, Neural Networks, 6(4): 525-533, 1993.

[23] X. He, W. Y. Ma, H. J. Zhang. Learning an image manifold for retrieval, Proc. ACM Multimedia, pp.17-23, 2004.

[24] V. Cherkassky, X. Shao, F. M. Mulier, V. N. Vapnik, Model complexity control for regression using VC generalization bounds, IEEE Transactions on Neural Networks, 10(5): 1075-1089, 1999.

[25] T. F. Cootes, G. J. Edwards, C. J. Taylor, Active appearance models, IEEE Transactions on PAMI, 23(6):681-685, 2001.

[26] G. D. Guo, Y. Fu, C. Dyer, T. Huang, Image-based human age estimation by manifold learning and locally adjusted robust regression, IEEE Transactions on Image Processing, 17(7): 1178-1188, 2008.

[27] X. Geng, Z. H. Zhou, K. S. Miles, Automatic age estimation based on facial aging patterns, IEEE Transactions on PAMI, 29(12):2234-2240, 2007.

[28] X. Geng, K. S. Miles, Z. Z. Zhou, Facial age estimation by nonlinear aging pattern subspace, Proc. ACM Multimedia, pp.721-724, 2008.

[29] A. Lanitis, C. Draganova, C. Christodoulou, Comparing different classifiers for automatic age estimation, IEEE Transactions on SMC-B, 34(1):621–628, 2004.

[30] S. Yan, H. Wang, X. Tang, T. Huang. Learning auto-structured regressor from uncertain nonnegative labels, Proc. ICCV, pp.1-8, 2007.

[31] S. Yan, H. Wang, T. S. Huang, X. Tang, Ranking with uncertain labels, Proc. ICME, pp.96-99, 2007.

[32] G. D. Guo, Y. Fu, T. S. Huang, C. Dyer, A probabilistic fusion approach to human age prediction, Proc. CVPR-SLAM Workshop, pp.1-6, 2008.

[33] S. Yan, X. Zhou, M. Liu, M. H. Johnson, T. Huang, Regression from patch-kernel, Proc. CVPR, pp.1-8, 2008.