Facial Age Estimation by Conditional Probability Neural Network

Chao Yin and Xin Geng*

School of Computer Science and Engineering,
Southeast University, Nanjing, 211189, P.R. China
{cyin,xgeng}@seu.edu.cn

Abstract. A new label distribution learning algorithm for facial age estimation, namely the Conditional Probability Neural Network (CPNN), is proposed in this paper. CPNN is based on a three-layer neural network which takes both the target variable (e.g., the age) and the conditional feature vector (e.g., the facial features) as its inputs, and the output is the conditional probability of the target variable given the feature vector. As a label distribution learning algorithm, CPNN can effectively utilize the neighboring ages while learning the real age. Compared with the existing label distribution learning algorithm IIS-LLD, it does not presume the underlying model as the maximum entropy model, but learns it from the training data. Thus CPNN is able to match the real problem better. Experimental results on the FG-NET database show that CPNN performs remarkably better than all the other eight compared methods.

Keywords: Label distribution learning, neural network, age estimation.

1 Introduction

Human face changes remarkably over one's lifetime. From infancy to adulthood, the main changes are due to the craniofacial growth [1], which cause the size of the face gradually getting larger. From adult to elder, the main changes are the texture changes [1], which cause the skin of the face becomes darker and wrinkly. It is these changes that make face an important trait for age estimation.

The age estimation problem can be treated as a multi-class classification problem, a regression problem or a hybrid of the two [1]. In the classification way, for example, Lanitis et al. [7] evaluated several common classifiers for age estimation, including artificial neural networks, nearest neighbor classifier and so on. Selvi et al.[2] used Multilinear Principal Component Analysis to estimate human age. In the regression way, for example, Zhang et al. [3] used MTWGP to estimate the human age. They treated the age estimation problem as a multi-task regression problem in which each learning task refers to the estimation of the age function for each person. In the hybrid way, for example, Guo et al. [6] proposed a method called LARR which achieved a better performance when combining classification and regression together to estimate the human age.

* Corresponding author.

C.-L. Liu, C. Zhang, and L. Wang (Eds.): CCPR 2012, CCIS 321, pp. 243–250, 2012.

Recently, Geng et al. [8] proposed a novel method IIS-LLD based on label distribution learning for facial age estimation. The motivation of this method is to solve the lack of sufficient training data problem [8]. Inspired by the observation that the faces at close ages look quite similar, each face image is associated with a label distribution rather than just the true label (age) [8]. A label distribution of a face image covers multiple ages. For each age there is a real number called description degree indicating the importance of that age. Of course the most important one is the real age. So the description degree of the real age is the highest in the distribution. It gradually decreases with the increase of the distance from the neighboring ages to the real age. The main benefit of this approach is that each image can not only contribute to the learning of its real age, but also to the learning of its adjacent ages, and thus relieves the problem of insufficient training data.

There is a presumption in IIS-LLD, i.e., the conditional probability of an age given the face image is modeled by the maximum entropy model [14]. However, the maximum entropy model may not fit the age estimation problem well. It is better to learn the model based on the training data. This paper proposes a novel neural network learning algorithm which can automatically learns the conditional probability of the age given the face image. Since there is no model presumption like that in IIS-LLD, the proposed method can fit the age estimation problem better and thus can achieve better performance.

The rest of the paper is organized as follows. In Section 2, the learning algorithm of the Conditional Probability Neural Network (CPNN) is proposed. After that, CPNN is tested and compared with several exiting age estimation methods in Section 3. Finally, conclusions are drawn in Section 4.

2 Conditional Probability Neural Network

2.1 The Structure of CPNN

The structure of CPNN is given in Fig. 1. It is composed of an input layer, a hidden layer and an output layer. There are two different kinds of inputs in CPNN. One is the conditional input $\boldsymbol{x}$, the other is the target variable input y. $\boldsymbol{x}$ and y could come from different domains and be within different ranges. For example, in the problem of age estimation, $\boldsymbol{x}$ represents the facial features, each element of which is usually within [0,1]. y represents the age of the face, which is usually within [0,100]. The output of CPNN is the conditional probability $P(y|\boldsymbol{x})$.

Modha et al. [9] used neural network to estimate joint probability density function. Conditional probability can be indirectly obtained via this method. But it needs to build two neural networks, one is for the joint probability of the target variable and the conditional variable, the other is for the marginal probability of the conditional variable. Then the conditional probability can be calculated by dividing the former by the later.

Based on Modha's neural network, Sarajedini et al. [10] proposed a method to directly estimate the conditional probability. One of the main differences between

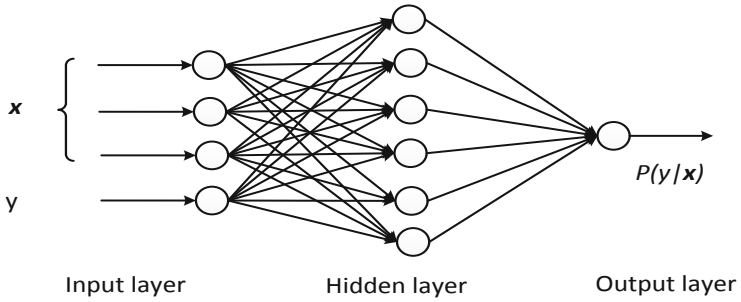

Fig. 1. The Structure of CPNN

Sarajedini's work and the method proposed in this paper is that their algorithm is unsupervised while the algorithm proposed in this paper is supervised by the label distributions of the training data.

2.2 The Learning Rule of CPNN

Let $\boldsymbol{t} = [x_1, x_2, ..., x_n, y]^T$ be the input of CPNN. Then the net activation of the output unit can be represented as

$$f(\boldsymbol{x}, y, \boldsymbol{w}) = f(\boldsymbol{t}, \boldsymbol{w}) = \sum_{i=1}^{M_2} G(\sum_{j=0}^{M_1} \boldsymbol{t}_j w_{1ij}) w_{2i}, \tag{1}$$

where G is a sigmoid function, M_1 is the number of the units in the input layer, M_2 is the number of units in the hidden layer. The output of CPNN can be treated as the joint probability density function using the following formula

$$p(\boldsymbol{x}, y) = exp(c(\boldsymbol{w}) + f(\boldsymbol{x}, y, \boldsymbol{w})), \tag{2}$$

where

$$c(\boldsymbol{w}) = -ln(\sum_{y}(\int exp(f(\boldsymbol{x}, y, \boldsymbol{w}))d\boldsymbol{x})) \tag{3}$$

is the bias of the output unit, i.e.,

$$p(\boldsymbol{x}, y) = \frac{exp(f(\boldsymbol{x}, y, \boldsymbol{w}))}{\sum_y(\int exp(f(\boldsymbol{x}, y, \boldsymbol{w}))d\boldsymbol{x})}. \tag{4}$$

The marginal probability can be calculated by

$$p(\boldsymbol{x}) = \sum_{y} p(\boldsymbol{x}, y) = \frac{\sum_y exp(f(\boldsymbol{x}, y, \boldsymbol{w}))}{\sum_y(\int exp(f(\boldsymbol{x}, y, \boldsymbol{w}))d\boldsymbol{x})}. \tag{5}$$

The conditional probability is then

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})} = \frac{exp(f(\boldsymbol{x}, y, \boldsymbol{w}))}{\sum_y exp(f(\boldsymbol{x}, y, \boldsymbol{w}))}. \tag{6}$$

Eq. (6) can be rewritten as

$$p(y|\boldsymbol{x}) = exp(b(\boldsymbol{x}, \boldsymbol{w}) + f(\boldsymbol{x}, y, \boldsymbol{w})), \tag{7}$$

where $b(\boldsymbol{x}, \boldsymbol{w})$ is

$$b(\boldsymbol{x}, \boldsymbol{w}) = -ln(\sum_y exp(f(\boldsymbol{x}, y, \boldsymbol{w}))). \tag{8}$$

After getting the conditional probability, the Kullback-Leibler divergence can be used to measure the distance between the estimated probability distribution and the true probability distribution, i.e.,

$$G(\boldsymbol{w}) = \sum_{\boldsymbol{x}} r(\boldsymbol{x}) log \frac{r(\boldsymbol{x})}{e(\boldsymbol{x})}. \tag{9}$$

The target is to minimize the Kullback-Leibler divergence between the true probability distribution $r(\boldsymbol{x})$ and the estimated probability distribution $e(\boldsymbol{x})$. It is equivalent to minimizing $L(\boldsymbol{w})$ given by Eq. (10)

$$L(\boldsymbol{w}) = -\sum_x r(\boldsymbol{x}) log(e(\boldsymbol{x})). \tag{10}$$

So the target function of the age estimation problem can be written as

$$J(\boldsymbol{w}) = -\sum_{k=1}^{K} \sum_{age=0}^{A} (r_{k,age} * log(p(y_{age}|\boldsymbol{x}_k))), \tag{11}$$

where the label distribution of the image $\boldsymbol{x}_k$ is r_k, $r_{k,age}$ is the description degree of age in the distribution r_k , K is the total number of images in the database. A is the maximum age in the databases. The partial derivative of $J(\boldsymbol{w})$ is

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = \sum_{k=1}^{K} \sum_{age=0}^{A} -1 * r_{k,age} * (\frac{\partial b(\boldsymbol{x}_k, \boldsymbol{w})}{\partial \boldsymbol{w}} + \frac{\partial f(\boldsymbol{x}_k, y_{age}, \boldsymbol{w})}{\partial \boldsymbol{w}}), \tag{12}$$

where $\frac{\partial b(\boldsymbol{x}_k, \boldsymbol{w})}{\partial \boldsymbol{w}}$ is

$$\frac{\partial b(\boldsymbol{x}_k, \boldsymbol{w})}{\partial \boldsymbol{w}} = -1 * \frac{\sum_y (\exp(f(\boldsymbol{x}_k, y_{age}, \boldsymbol{w})) \times \frac{\partial f(\boldsymbol{x}_k, y_{age}, \boldsymbol{w})}{\partial \boldsymbol{w}})}{\sum_y \exp(f(\boldsymbol{x}_k, y_{age}, \boldsymbol{w}))}. \tag{13}$$

The calculation of $\frac{\partial f(\boldsymbol{x}_k, y_{age}, \boldsymbol{w})}{\partial w_{lij}}$ is similar to Modha's method [9], i.e.,

$$\frac{\partial f(\boldsymbol{x}_k, y_{age}, \boldsymbol{w})}{\partial w_{lij}} = z_{(l-1)j} \delta_{li}, \tag{14}$$

where $z_{(l-1)j}$ is the output of the j-th unit before the current layer. When l is the output layer, $\delta_{li} = 1$. When l is the hidden layer, δ_{li} is calculated as

$$\delta_{li} = G'(I_{li}) \sum_{p=1}^{n} \delta_{(l+1)p} w_{(l+1)pi}. \tag{15}$$

Finally, the weights are updated by the RPROP algorithm [11]. One property of RPROP algorithm is that the size of the updating step depends on the sequence of signs rather than the magnitude of the derivative [11]. There are two steps in the RPROP algorithm: 1. calculate the sign of the update-value; 2. calculate the size of update-value. The sign of the update-value is determined by the following formula

$$\Delta w_{lij}^{(t)} = \begin{cases} -\Delta_{lij}^{(t)} & , if \frac{\partial J}{\partial w_{lij}}^{(t)} > 0 \\ +\Delta_{lij}^{(t)} & , if \frac{\partial J}{\partial w_{lij}}^{(t)} < 0, \\ 0 & , else \end{cases} \tag{16}$$

where $\frac{\partial J}{\partial w_{lij}}^{(t)}$ is the derivation in the current step. The size of update-value $\Delta_{lij}^{(t)}$ is defined as

$$\Delta_{lij}^{(t)} = \begin{cases} \eta^+ * \Delta_{lij}^{(t-1)} & , if \frac{\partial J}{\partial w_{lij}}^{(t-1)} * \frac{\partial J}{\partial w_{lij}}^{(t)} > 0 \\ \eta^- * \Delta_{lij}^{(t-1)} & , if \frac{\partial J}{\partial w_{lij}}^{(t-1)} * \frac{\partial J}{\partial w_{lij}}^{(t)} < 0, \\ \Delta_{lij}^{(t-1)} & , else \end{cases} \tag{17}$$

where $0 < \eta^- < 1 < \eta^+$. When the partial derivative of the weight changes its sign, it means that the last update was too big and the algorithm has jumped over a local minimum. So the update-value Δ_{lij} is decreased by the factor η^- [11]. If the partial derivative retains its sign, it indicates that the last update was in the right direction and the update-value can be increased by multiplying a factor η^+. The ultimate update equation is:

$$w_{lij}^{(t+1)} = w_{lij}^{(t)} + \Delta w_{lij}^{(t)}. \tag{18}$$

3 Experiment

3.1 Methodology

The database used in the experiments is the FG-NET aging database [12]. There are 1002 images from 82 subjects with age ranges from 0 to 69 in FG-NET. The facial feature extractor is the Appearance Model [5]. 200 model parameters are extracted as the feature vector to explain about 95% of the variation in the training data.

The algorithms are tested through the LOPO (Leave-One-Person-Out) mode [4], i.e., in each fold, the images of one person are used as the test set and those of the others are used as the training set. After 82 folds, each subject has been used as test set once, and the final results are calculated from all of the estimates.

The performance of the age estimators is measured by both MAE(Mean Absolute Error) and CS (Cumulative Score). MAE is defined as $MAE = \sum_{k=1}^{N} |\hat{e}_k - e_k|/N$, where e_k is the real age , $\hat{e}_k$ is the estimated age, and N is the number of test images. CS is defined as $CS(t) = N_{e \leq t}/N \times 100\%$, where $N_{e \leq t}$ is the number of test images which have an absolute error no higher than t.

Two kinds of label distributions are generated for each face image. The first is the Gaussian distribution (represented by G) and the second is the Triangle distribution (represented by T), both centered at the real age. Several parameters for the distributions are tested and the best results are reported. In detail, the standard deviation of the Gaussian distribution varies within the values 1, 2, and 3, and the bottom length of the Triangle distribution varies within the values 4, 6, and 8.

3.2 Results

The compared methods include several existing facial age estimation methods IIS-LLD [8], AGES [4], WAS [12], and AAS [7], and some general-purpose classification methods BP (Backpropagation neural network), C4.5 (C4.5 decision tree), KNN, and SVM (Support Vector Machine).

Table 1 shows the MAE results of CPNN and all the compared methods. As can be seen, the two label distribution learning algorithms (CPNN and IIS-LLD) perform remarkably better than the other compared methods. This proves the effectiveness of using label distribution learning to relieve the lack of training samples problem. In general, CPNN performs significantly better than IIS-LLD. The main advantage of CPNN comes from that the relationship between the given face image and the corresponding label distribution is learned from the training data, rather than based on any presumptions, such as the maximum entropy model used in IIS-LLD.

The cumulative scores of CPNN and other compared methods on the FG-NET database are shown in Fig. 2. As can be seen, the cumulative scores of the label distribution learning algorithms (CPNN and IIS-LLD) are remarkably better than all other compared methods. The comparison between CPNN and IIS-LLD shows clear advantage of CPNN, which is more apparent with higher error levels. In summary, the experimented results reveal that the label distribution learning algorithms can effectively solve the problem of insufficient training data in age estimation. Among the label distribution learning algorithms, CPNN performs better than IIS-LLD due to that it learns the underlying model from the training data, rather than making any presumptions.

Table 1. Mean Absolute Error on FG-NET Database

CPNN		IIS-LLD		AGES	WAS	AAS	BP	C4.5	KNN	SVM
G	T	G	T							
4.76	5.07	5.77	5.90	6.77	8.06	14.83	11.85	9.34	8.24	7.25

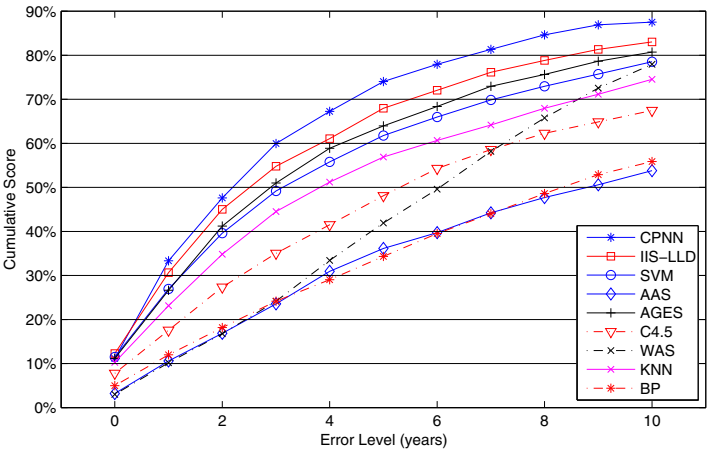

Fig. 2. Cumulative Scores on the FG-NET Database

4 Conclusion

A novel label distribution learning algorithm named CPNN is proposed in this paper for facial age estimation. CPNN relies on the label distribution rather than the real age of each face image to relieve the common lack of training samples problem in age estimation. It learns the relationship between the face image and the corresponding label distribution via a neural network from the training data. This gains advantages for CPNN over the existing label distribution learning algorithm IIS-LLD, which assumes that relationship to be modeled by the maximum entropy model. CPNN is compared with eight other methods on the FG-NET database. Experimental results show that CPNN performs significantly better than all the compared methods in the task of facial age estimation.

Acknowledgment. This work is supported by ARC (DP0987421), NSFC (60905031), JiangsuSF (BK2009269), SRF for ROCS, SEM, the Excellent Young Teachers Program of SEU, the Open Projects Program of NLPR, and the Key Lab of Computer Network and Information Integration of Ministry of Education of China.

References

1. Fu, Y., Guo, G.D., Huang, T.S.: Age Synthesis and Estimation via Faces: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(11), 1955–1976 (2010)
2. Selvi, V.T., Vani, K.: Age estimation system using MPCA. In: Proceedings of the International Conference on Recent Trends in Information Technology, Dubai, United Arab Emirates, pp. 1055–1060 (2011)

3. Zhang, Y., Yeung, D.Y.: Multi-task warped Gaussian process for personalized age estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hong Kong, China, pp. 2622–2629 (2011)
4. Geng, X., Zhou, Z.-H., Li, G., Dai, H.: Learning from facial aging patterns for automatic age estimation. In: Proceedings of the ACM International Conference on Multimedia, Santa Barbara, CA, pp. 307–316 (2006)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
6. Guo, G.D., Fu, Y., Dyer, C.R.: Image-Based Human Age Estimation by Manifold Learning and Locally Adjusted Robust Regression. IEEE Transactions on Image Processing 17(7), 1178–1188 (2008)
7. Lanitis, A., Draganova, C., Christodoulou, C.: Comparing different classifiers for automatic age estimation. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 34(1), 621–628 (2004)
8. Geng, X., Smith-Miles, K.A., Zhou, Z.-H.: Facial Age Estimation by Learing from Label Distributions. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence, Atlanta, GA, pp. 451–456 (2010)
9. Modha, D.S., Fainman, Y.: A learning law for density estimation. IEEE Transactions on Neural Networks 5(3), 519–523 (1994)
10. Sarajedini, A., Hecht-Nielsen, R., Chau, P.M.: Conditional probability density function estimation with sigmoidal neural networks. IEEE Transactions on Neural Networks 10(2), 231–238 (1999)
11. Riedmiller, M., Braun, H.: A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In: Proceedings of the IEEE International Conference on Neural Networks, San Francisco, CA, pp. 586–591 (1993)
12. Lanitis, A., Taylor, C.J., Cootes, T.F.: Toward automatic simulation of aging effects on face images. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(4), 442–455 (2002)
13. Guo, G.D., Mu, G.W.: Human Age Estimation: what is the Influence Across Race and Gender? In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, pp. 71–78 (2010)
14. Berger, A., Pietra, S.D., Pietra, V.J.D.: A maximum entropy approach to natural language processing. Computational Linguistics 22(1), 39–71 (1996)