Published in final edited form as:

Med Image Comput Comput Assist Interv. 2019 October ; 11767: 149–157.

doi:10.1007/978-3-030-32251-9_17.

Deep Granular Feature-Label Distribution Learning for Neuroimaging-based Infant Age Prediction

Dan Hu, Han Zhang, Zhengwang Wu, Weili Lin, Gang Li, Dinggang Shen, UNC/UMN Baby Connectome Project Consortium

Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA

Abstract

Neuroimaging-based infant age prediction is important for brain development analysis but often suffers insufficient data. To address this challenge, we introduce label distribution learning (LDL), a popular machine learning paradigm focusing on the small sample problem, for infant age prediction. As directly applying LDL yields dramatically increased number of day-to-day age labels and also extremely scarce data describing each label, we propose a new strategy, called granular label distribution (GLD). Particularly, by assembling the adjacent labels to granules and designing granular distributions, GLD makes each brain MRI contribute to not only its own age but also its neighboring ages at a *granule* scale, which effectively keeps the information augmentation superiority of LDL and reduces the number of labels. Furthermore, to extremely augment the information supplied by the small data, we propose a novel method named *granular feature distribution* (GFD). GFD leverages the variability of the brain images at the same age, thus significantly increasing the learning effectiveness. Moreover, deep neural network is exploited to approximate the GLD. These strategies constitute a new model: deep granular feature-label distribution learning (DGFLDL). By taking 8 types of cortical morphometric features from structural MRI as predictors, the proposed DGFLDL is validated on infant age prediction using 384 brain MRI scans from 35 to 848 days after birth. Our proposed method, approaching the mean absolute error as 36.1 days, significantly outperforms the baseline methods. Besides, the permutation importance analysis of features based on our method reveals important biomarkers of infant brain development.

Keywords

Infant age prediction; cortical features; label distribution learning

1 Introduction

Neuroimaging-based infant age prediction is important for tracking early brain development [1], identifying potential neurodevelopmental disorders [2], and discerning environmental influences on the brains [3]. Although there are many studies on brain imaging-based age prediction [1, 4], little work has been dedicated to infants, who are in a crucial period with

neurodevelopmental disorders possibly emerging [5]. To this end, this paper focuses on predicting infant age from birth to 2 years of age using brain MR images, as well as identifying some development-related biomarkers. Conventionally, neuroimaging-based age prediction is considered as a regression problem and many machine learning algorithms have been adopted, e.g., support vector regression [6], convolutional neural network, Gaussian process regression [1], and multiple linear regression with elastic-net penalty [7]. On the other hand, age prediction can be taken as a classification problem as well by setting each age as a class label [8]. However, all these methods suffer from insufficient training data issue. Label distribution learning (LDL), as an effective strategy of learning from insufficient data, has become a popular learning paradigm in the past few years [8]. Taking advantage of the similarity of the features at close ages, LDL realizes the information augmentation by utilizing the images at the neighboring ages while learning a particular age. This merit attracts us introducing LDL to handle the data insufficiency issue in infant age prediction. However, there are still two major challenges while applying LDL method to the infant neuroimaging data: (1) high computational burden caused by a large number of “age labels”, since there are nearly 800 different age labels in a day-to-day infant age prediction problem; (2) ineffective learning due to scarce training samples for each age label (384 MR images distributed on over 800 labels).

To overcome these challenges, we propose a deep granular feature-label distribution learning (DGFLDL) method for infant age prediction. First, we exploit the concept of granule, the collection of values that are arranged together due to their adjacency, to assemble the labels with high similarity, and propose a new strategy called granular label distribution (GLD). With GLD, the number of labels is sharply reduced to the number of granules by designing the distribution on granules rather than each single label. Then, by utilizing the inter-subject and noise-caused variability of the brain images at the same age, a novel concept granular feature distributions (GFD) is defined to replace the original single feature vector. The introduction of GFD further augments the information in the small data and effectively increases the robustness of the age prediction model. Furthermore, a deep neural network is employed to fit the granular label distribution by taking its advantage on nonlinear function approximation. Finally, some biomarkers for infant brain development are identified by permutation importance analysis based on our proposed model.

2 Method

2.1 Granular Label Distribution and Granular Feature Distribution

Conventional label distribution learning (LDL) assigns distribution for each label to measure the relative importance of different labels in describing an image. This causes great computational burden when the total number of labels is large (and day-to-day age prediction is such a problem). Since a brain MRI contributes to the neighboring labels analogously, it is unnecessary to define the contribution of the image on each label individually. Instead, we assemble the adjacent labels to a *granule* and define the contribution of the image to such a granule rather than many seemingly independent labels, which is a good choice to solve the challenge of a large number of labels. Thus, by transforming the label set to granule set and assigning distributions on the granules, the label

distribution is replaced by granular label distribution (as shown in Fig. 1(a)). Based on this strategy, the large number of labels involved in age learning is substantially reduced to a relatively small number of granules. As such, each granule has more corresponding data samples than a single label has, thus improving the effectiveness of learning as well.

The design of the granular label distribution is important for the whole model. Suppose *all the granules are disjoint intervals*, the distance from a granule to an age label is defined as the absolute difference between the midpoint of the granule and the age. Then, for a brain image x with the age a , the granular label distribution of x should be designed by satisfying the following requirement: the probability that a belongs to the granule decreases with the increase of the distance between a and the granule. A collection of adjacent Gaussian distributions centered at the midpoints of the granules is one of the possible choices satisfying the above requirement.

Except for granular label distribution, the information lying in the data could be further augmented by considering inter-subject and noise-caused variability existing in the brain image features. That is, a certain age can be associated with *not only* its own brain image *but also* many neighboring images with certain probabilities. To this end, we propose a new concept of granular feature distribution (GFD), which assigns distributions to the features on a granular scale. As shown in Fig. 1(b), the information lying in the point $x = (x_1, x_2)$ is augmented to a feature granule based on GFD. Starting from label learning (LL), the data transformation processes in LDL, granular label distribution learning (GLDL), and granular feature-label distribution learning (GFLDL) are illustrated in Fig. 2. Finally, our new model, *deep granular label feature distribution learning* (DGFLDL), is established by exploiting a deep neural network to approximate the granular label distribution and performing the infant age prediction based on the transformed data obtained from granular label distribution and granular feature distribution.

2.2 Problem Formulation

Let $X = \prod_{m=1}^M X_m$ denote a brain image feature space and $Y = \{y_1, y_2, \dots, y_N\}$ denote a set of possible ages. The granular label space is defined as $Y^* = \{g_j = (\mu_j, \sigma_j), j = 1, \dots, c_Y\}$, $\mu_j \in Y$, $\sigma_j \in R$, $c_Y \ll N$. Each (μ_j, σ_j) represents a Gaussian distribution granule with mean μ_j and standard deviation σ_j . The granular feature space is defined as $X^* = \prod_{m=1}^M X_m^*$, where $X_m^* = \{h_m^i = (\phi_m^i, \psi_m^i), i = 1, \dots, c_X^m\}$, $\phi_m^i \in X_m$, $\psi_m^i \in R$. Based on X^* and Y^* , the original training data $D = \{(x, a) | x \in X, a \in Y\}$ is transformed into a new data $D^* = \{(x^*, a^*) | x^* \in \prod_{m=1}^M [0, 1]^{c_X^m}, a^* \in [0, 1]^{c_Y}\}$, where

$$x^* = (P_{11}(x_1), \dots, P_{1c_X^1}(x_1), \dots, P_{M1}(x_M), \dots, P_{Mc_X^M}(x_M)), \sum_{i=1}^{c_X^m} P_{mi}(x_m) = 1,$$

$$a^* = (d_{g_1, x}, \dots, d_{g_j, x}, \dots, d_{g_{c_Y}, x}), \text{ and } \sum_{j=1}^{c_Y} d_{g_j, x} = 1. P_{mi}(x_m) \text{ is the probability that } x_m \text{ belongs to the granule } h_i^m, \text{ and } d_{g_j, x} \text{ is the probability that } a \text{ belongs to the granule } g_j.$$

$$P_{mi}(x_m) = \frac{1}{\psi_m^i \sqrt{2\pi} Z} \exp\left(-\frac{(x_m - \phi_m^i)^2}{2(\psi_m^i)^2}\right), Z = \frac{1}{\psi_m^i \sqrt{2\pi}} \sum_{i=1}^m \exp\left(-\frac{(x_m - \phi_m^i)^2}{2(\psi_m^i)^2}\right) \quad (1)$$

$$d_{gj, \mathbf{x}} = \frac{1}{\sigma_j \sqrt{2\pi} V} \exp\left(-\frac{(y - \mu_j)^2}{2\sigma_j^2}\right), V = \frac{1}{\sigma_j \sqrt{2\pi}} \sum_{j=1}^{c_Y} \exp\left(-\frac{(y - \mu_j)^2}{2\sigma_j^2}\right) \quad (2)$$

Then, the problem of granular feature-label distribution learning for infant age prediction can be formulated. Given a training set $S = \{(\mathbf{x}_1, a_1), (\mathbf{x}_2, a_2), \dots, (\mathbf{x}_K, a_K)\}$, where $\mathbf{x}_k \in X$, $k = 1, \dots, K$, represents the cortical feature vector of a brain MRI and $a_k \in Y$ is the chronological age of $\mathbf{x}_k$. Based on granular feature space X^* and granular label space Y^* , S is transformed to $S^* = \{(\mathbf{x}_1^*, a_1^*), (\mathbf{x}_2^*, a_2^*), \dots, (\mathbf{x}_K^*, a_K^*)\}$. After learning a conditional probability mass function (p.m.f.) $P(g|\mathbf{x}^*)$ from S^* , for $g \in Y^*$, the age could be estimated as

$$\hat{a} = E(g | \mathbf{x}^*) = \sum_{j=1}^{c_Y} \frac{\mu_j P(g | \mathbf{x}^*)}{\sum_{j=1}^{c_Y} P(g | \mathbf{x}^*)} \quad (3)$$

Thus, DGFLDL aims to learn the parametric model $\hat{a} = f(\mathbf{x}; \boldsymbol{\theta}, \mathbf{W})$, where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\phi}, \boldsymbol{\psi})$ is the parameter vector in distributions defined by Eqs. 1 and 2, which corresponds to granular feature and label distributions; $\mathbf{W}$ is the parameter vector of a deep neural network (DNN), $P(g|\mathbf{x}^*; \mathbf{W})$, which generates a mapping similar to S^* .

2.3 DGFLDL Model

The frame work of DGFLDL is depicted in Fig. 3 and also detailed as follows.

1) Initialization: $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$ are first initialized by the equal-distance partition of $\{X_m | m = 1, \dots, M\}$ and Y , respectively. For each feature, $\boldsymbol{\psi}$ is designed as the standard deviation of its values on the training set. $\boldsymbol{\sigma}$ is designed as the same value (set as 20 in our experiment) for all granules of the label. To simplify the learning, only $\boldsymbol{\sigma}$ will be learned, while keeping $\boldsymbol{\mu}$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$ fixed. Suppose $(\boldsymbol{\mu}, \boldsymbol{\sigma}_0, \boldsymbol{\phi}, \boldsymbol{\psi})$ are the predefined initial values, each training image $\mathbf{x}$ is transformed to its feature distribution $\mathbf{x}^*$ and associated with an initial label distribution $\mathbf{a}_0^*$.

2) GLD adaption: In the t^{th} iteration, the target is to find $\mathbf{W}_t$ that can generate the granular label distribution most similar to $\mathbf{a}_{t-1}^*$. As Kullback-Leibler (K-L) divergence is used to measure the dissimilarity between distributions, the best $\mathbf{W}_t$ is learned by solving the following minimization problem:

$$\mathbf{W}_t = \operatorname{argmin}_{\mathbf{W}} \sum_{j=1}^{c_Y} P(g_j | \mathbf{x}^*; \mathbf{W}) \ln \frac{P(g_j | \mathbf{x}^*; \mathbf{W})}{d_{g_j, \mathbf{x}}^{t-1}} \quad (4)$$

3) Granular error-based sample selection: After the GLD adaption, the conditional p.m.f. $P(g|\mathbf{x}^*; W_t)$ is obtained. Thus, the age of each training image $\mathbf{x}$ can be estimated as the expectation of $P(g|\mathbf{x}^*; W_t)$, i.e.,

$$\hat{a} = E(g | \mathbf{x}^*) = \sum_{j=1}^{c_Y} \frac{\mu_j P(g | \mathbf{x}^*)}{\sum_{j=1}^{c_Y} P(g | \mathbf{x}^*)} \quad (5)$$

A *granular based mean absolute error strategy* is designed to select images from the training set to fit the granule parameter vector σ . For each granule $g_j = (\mu_j, \sigma_j)$, the granule induced interval $[\alpha_j^-, \alpha_j^+]$ is obtained by the upper percentile of the corresponding Gaussian distribution, where $P(y \geq \alpha_j^-) = 1 - \beta$, $P(y \geq \alpha_j^+) = \beta$, β is a threshold and set as 0.05 in our experiment. In total, the selected images are

$$B = \bigcup_{j=1}^{c_Y} \{(\mathbf{x}, a) | |a - \hat{a}| \leq \text{MAE}_j, a \in [\alpha_j^-, \alpha_j^+]\} \quad (6)$$

$$\text{MAE}_j = \text{mean}(|a - \hat{a}|, a \in [\alpha_j^-, \alpha_j^+]) \quad (7)$$

4) Granules parameters adaption: Based on the selected image set B , σ is learned to be the best parameter vector that can generate the granular label distribution most similar to the predicted granular label distributions in B . Here, K-L divergence is again chosen to measure the dissimilarity between the two distributions. Together with K-L divergence, the prediction loss is included to estimate the effectiveness of σ as well. These two losses are weighted-combined with a regularization parameter $0 \leq \lambda \leq 1$.

$$\sigma_t = \text{argmin}_{\sigma} (\lambda \mathcal{L}_{Exp} + (1 - \lambda) \mathcal{L}_{KL}) \quad (8)$$

$$\mathcal{L}_{Exp} = \sum_{x_b \in B} (a_b - \sum_{i=1}^C \frac{\mu_j d_{g_i, x_b}}{\sum_i d_{g_i, x_b}})^2, \mathcal{L}_{KL} = \sum_{x_b \in B} \sum_{j=1}^{c_Y} P(g_i | x_b; W_t) \ln \frac{P(g_i | x_b; W_t)}{d_{g_j, x_b}} \quad (9)$$

Substituting Eq. 9 into Eq. 8 yields a nonlinear programming problem, which can be effectively solved by the sequential least squares programming method. W in the DNN and the parameters of feature-label distributions are optimized alternatively until the training error stop decreasing or the specified maximum number of epochs reaches.

Furthermore, the most contributive features in the age prediction model could be taken as the important infant brain development biomarkers, since they have high discriminative capability for age estimation. Based on the well-trained DGFLDL model, permutation importance was calculated for measuring the contribution of each feature for age prediction.

3 Experiments

We verified the effectiveness of our proposed DGFLDL method on the brain MRI based infant age prediction by using a high-quality dataset obtained from the UNC/UMN Baby Connectome Project [12]. In experiments, we used 185 subjects (97 females) with 384 longitudinal MRI scans acquired at different ages ranging from 35 to 848 days. All infant MR images were preprocessed by a well-established infant-specific computational pipeline [13]. Briefly, the pipeline includes intensity inhomogeneity correction, skull stripping, cerebellum removal, tissue segmentation, separation of left/right hemispheres, cortical topology correction, and inner and outer surface reconstruction. Since using the morphological features of each vertex for prediction is redundant and computationally expensive, a region-based analysis is used. Each individual cortical surface of each hemisphere was parcellated into 180 cortical regions of interest (ROIs) [9] by aligning onto the public 4D Infant Cortical Surface Atlas (<https://www.nitrc.org/projects/infantsurfatlas/>) [14]. With the ROIs on each surface of each hemisphere, 6 types of morphological features (i.e., sulcal depth measured with Euclidean distance (SDE), myelin content, local gyrification index (LGI), average convexity, mean curvature, and cortical thickness) were obtained by averaging the corresponding values of all vertices inside each ROI; and 2 types of features (i.e., surface area and cortical volume) were obtained by summing the corresponding values of all vertices inside each ROI. Therefore, for each scan, 2880 imaging features ($8 \text{ feature types} \times 180 \text{ ROIs} \times 2 \text{ hemispheres}$) were acquired.

Four state-of-the-art regression methods for age prediction and 2 label distribution-related methods are chosen as the baseline methods, which include support vector regression (SVR), Gaussian process regression (GPR), random forest (RF), deep neural network (DNN) applying to the original data, deep label distribution learning applying to the data with ordinary label distribution (DLDL), and deep granular label distribution applying to the data with our granular label distribution (DGLDL). These age prediction algorithms were assessed by 10 times of 10-fold cross validation with 3 criteria, i.e., the mean absolute error (MAE), the mean relative absolute error (MRAE), and the correlation value r between the predicted ages and the chronological ages. For r , the 95% confidence interval was computed by the 2.5 and 97.5 percentiles of the correlation values obtained from 1,000 bootstrap samples. During the cross-validation, for the longitudinal scans from the same subject, we guaranteed that the scans in the training data were scanned earlier than the scans in the testing data. The DNN employed in our DGLDL and DGFLDL model consists of 10 fully-connected layers with ReLU as the activation function of the first 9 layers and softmax as the activation function of the last layer.

The prediction results were presented in Table 1. Scatter plots of the predicted ages and the chronological ages are shown in Fig. 4, where r is the average of r_L and r_R . Our DGFLDL model outperformed the other 6 models and reduced the mean absolute error to 36.1 days. The correlation coefficient of the predicted ages and the chronological ages reached 0.963. These comparison results demonstrate that replacing label distribution learning (LDL) with the proposed granular label distribution method (GLDL) improved the prediction accuracy. In addition, when the granular feature distribution was adopted, the performance was further significantly improved.

The permutation importance of each feature type (or ROI) was obtained by summing the permutation importance of its corresponding 360 ROIs (or 8 features), respectively, as shown in Fig. 5. We found that cortical thickness, surface area, and cortical volume are the most important feature types in the prediction. This possibly results from the most dramatic development of these features in the first 2 years after birth. Furthermore, we revealed the top 5 important ROIs in age prediction, which includes the bilateral medial prefrontal cortex, left frontal opercular cortex, left posterior opercular cortex, and the left primary visual cortex.

4 Conclusion

In this paper, we proposed deep granular feature-label distribution learning to address the insufficient data challenge of neuroimaging-based infant age prediction. Granular label distribution was introduced, which *not only* solved the problem of dramatically increased number of labels *but also* improved the learning efficiency. We further proposed granular feature-label distribution to realize *dual information augmentation* of the data, which utilizes the inter-subject and noise-caused variability of the brain images and has been validated as an effective method with high prediction accuracy. Moreover, the importance analysis of features based on our proposed model revealed the importance of cortical thickness in infant age prediction. Some brain regions were also identified as important infant brain development biomarkers. Our proposed DFGLDL serves as a model for prediction with insufficient data and a potentially powerful method for studying normal and abnormal infant brain development.

Acknowledgments.

This work was partially supported by NIH grants (MH107815, MH116225, and MH117943). This work also utilizes approaches developed by an NIH grant (1U01MH110274) and the efforts of the UNC/UMN Baby Connectome Project Consortium.

References

1. Cole JH, Franke K: Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends in neurosciences*, 40(12), 681–690 (2017). [PubMed: 29074032]
2. Nenadic I, Dietzek M, Langbein K, et al.: BrainAGE score indicates accelerated brain aging in schizophrenia, but not bipolar disorder. *Psychiatry Research: Neuroimaging*, 266, 86–89 (2017). [PubMed: 28628780]
3. Jason S, Christian H: Differences between chronological and brain age are related to education and self-reported physical activity. *Neurobiology of aging*, 40, 138–144 (2016). [PubMed: 26973113]
4. Toews M, Wells WM, Zollei L: A feature-based developmental model of the infant brain in structural MRI. In: *MICCAI 2012*, 15(2), 204–211(2012).
5. Gilmore JH, Kang C, Evans DD, et al.: Prenatal and neonatal brain structure and white matter maturation in children at high risk for schizophrenia. *American Journal of Psychiatry* 167, 1083–1091 (2010). [PubMed: 20516153]
6. Liem F, Varoquaux G, Kynast J, et al. Predicting brain-age from multimodal imaging data captures cognitive impairment. *Neuroimage*, 148, 179–188 (2017). [PubMed: 27890805]
7. Lewis JD, Evans AC, Tohka J, and Grp B. D. C.: T1 white/gray contrast as a predictor of chronological age, and an index of cognitive performance. *Neuroimage*, 173, 341–350 (2018). [PubMed: 29501876]

8. Antipov G, Baccouche M, Berrani SA, Dugelay JL: Effective training of convolutional neural networks for face-based gender and age prediction. *Pattern Recognition*, 72, 15–26 (2017).
9. Glasser M, Coalson TS, Robinson EC, et al.: A multi-modal parcellation of human cerebral cortex. *Nature*, 536, 171–178 (2016). [PubMed: 27437579]
10. Geng X: Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7), 1734–1748 (2016).
11. Gao BB, Xing C, Xie CW, Wu J, Geng X: Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6), 2825–2838 (2017). [PubMed: 28371776]
12. Brittany RH, Martin AS, Wei G et al.: The UNC/UMN Baby Connectome Project (BCP): An Overview of the Study Design and Protocol Development *Neuroimage*, 185, 891–905 (2019). [PubMed: 29578031]
13. Gang L, Li W, Pew-Thian Y, et al.: Computational neuroanatomy of baby brains: A review. *Neuroimage*, 185, 906–925 (2019). [PubMed: 29574033]
14. Gang L, Wang L, Shi F et al.: Construction of 4D high-definition cortical surface atlases of infants: Methods and applications. *Medical image analysis*, 25 (1), 22–36 (2015) [PubMed: 25980388]

Fig. 1.

Illustration of the granular label distribution and granular feature distribution.

Fig. 2.
An illustration of the data transformation in LL, LDL, GLDL, and GFLDL.

Fig. 3.
The framework of our proposed DGFLDL method.

Fig. 4. Scatter plots of the chronological ages and predicted ages obtained by DGFLDL (our method) and GPR (one of the competing methods with the best performance).

Fig. 5.
The permutation importance of 360 ROIs and 8 feature types.

Table 1.

The comparison of the DGFLDL with SVR, GPR, RF, DNN, DLDL, and DGLDL. r_L and r_R are the left and right endpoints of the 95% confidence interval of r .

	MAE (days)	MRAE (%)	r_L	r_R
SVR	60.3±3.5	23.6±2.3	0.902±0.005	0.928±0.004
GPR	55.9±2.8	18.5±1.7	0.925±0.007	0.949±0.005
RF	65.3±3.6	26.1±2.1	0.872±0.006	0.921±0.004
DNN	66.1±2.5	19.2±1.5	0.889±0.004	0.933±0.002
DLDL	41.5±2.7	12.2±1.4	0.934±0.004	0.966±0.003
DGLDL (ours)	39.2±2.6	11.9±0.9	0.943±0.005	0.971±0.003
DGFLDL (ours)	36.1±2.3	11.1±0.5	0.951±0.003	0.976±0.002