MODELLING OF AGE FROM THE TEETH DEVELOPMENT

Miroslav Čepek[1], Aleš Pilný[1], Radka Kubelková[2], Jana Velemínská[2], Miroslav Šnorek[1]

[1] Department of Computer Science and Engineering
Faculty of Electrical Engineering, Czech Technical University of Prague
Czech Republic
[2] Department of Anthropology and Human Genetics
Faculty of Science, Charles University in Prague
Czech Republic

cepekmir@fel.cvut.cz(Miroslav Cepek)

Abstract

The age estimation is very important in archeology or forensics as well as in human medicine. In the archeology and forensics the age estimation is useful for the examination of the skeletal remains. In the human medicine is important to estimate the development of individual which can be very different from his/her chronological age. In this paper we will process collection of 1393 Czech female and male children between 3 and 20 years of age. To determine development stage of teeth we will utilise method presented by Moorrees, Fanning and Hunt. The aim of this paper is to identify significant teeth by methods well known in data mining field. After this we will present results of several modelling methods and also formulas which may be immediately used. Models will be created from the full set of teeth and then later from several subsets.

Keywords: Dental Age Estimation, Teeth Mineralization, Feature Selection and Ranking, Age Modelling

Presenting Author's Biography

Miroslav Čepek is a PhD student and since 2010 he is Assistant professor at the Department of Computer Science and Engineering of the Czech Technical University in Prague. He graduated in 2006. He is interested in biological data mining and data preprocessing.

1 Introduction

The age estimation is very important in archeology or forensics as well as in human medicine. In the archeology and forensics the age estimation is useful for the examination of the skeletal remains. In the human medicine is important to estimate the development of individual which can be very different from his/her chronological age. One of possibilities how to estimate the age is to examine stages of mineralisation of the teeth. This works well with children and young people. But when the development of teeth is finished in circa twenty years of age, the estimation of age becomes impossible since there are no changes on the teeth.

There are several methodologies to estimate dental age. All are based on X-Ray photography of teeth. The picture is examined by an expert and he or she assigns development stage to each tooth. The most common methodology of development stages is the Demirjian's method. The

One of alternative approaches is methodology presented in this paper was developed by Moorrees, Fanning and Hunt (MFH) in [?]. This methodology contains one more stage and is presented in figure 1.

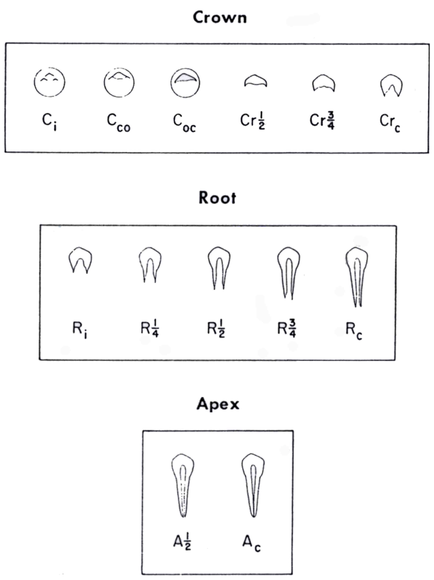

Fig. 1 Development stages of teeth in Moorrees, Fanning and Hunt (FMH) methodology [?]

When development stages are assigned they have to be transformed into age. The classical approach presented by Demirjian is to sum development stages of all teeth and transform the sum into estimation of age. The approach presented by FMH is not so straight forward. The

2 Used Modelling and Feature Selection Methods

In this paper we will compare results of several modelling and feature selection/feature ranking methods. These methods

- Zero Regression Method this is very simple method – from the training data the average output value is calculated. This average value is then output of the model for any data vector [?].
- Multi-Variable Linear Regression – is standard least-square linear regression as described for example [?]. This method works with all input attributes (several teeth in our case).
- Single-Variable Linear Regression – this method uses linear regression but only with one attribute (one tooth). It creates one Linear Regression model for each attribute and then returns the model with the lowest error [?].
- Sequential Minimal Optimizatrion (SMO) – is a variant of Support Vector Machine (SVM) method. The basic idea of the SVM is the same as in linear regression – to find function that approximates the training points. The objetive is to find the function with the minimal error and in the same time with the maximal flatness. Standard way to train SVM is to solve one large Quadratic Programming problem. Instead of solving one large Quadratic Programming problem the SMO breaks the problem into sequence of much smaller Quadratic Programming problems. [http://research.microsoft.com/en-us/um/people/jplatt/smotr.pdf]
- RBF Neural Network – is popular feed-forward neural network. The network have one hidden layer and output layer. The neurons in the hidden layer basicaly represents points in the input space. Activity of each neuron corresponds with the distance of instance from given neuron. The output layer combines outputs from the neurons from the first layer. The output is the weighted sum of the activation of neurons in the hidden layer [?].

We also want to find out if the age may be correctly estimated not just by full set of teeth but also by some subset. Therefore we have created several experiments where we left out teeth we hold less important and we will compare results to the full set of teeth. To determine which teeth are more important we have used several feature selection and feature ranking methods.

3 Data and Experiments

The data were collected among Czech children between 2 and 20 years of age. In this dataset there are data about 736 females and 648 males. The X-Ray images were obtained from S

In our experiments we will assume that there are significant differences between development of teeth between males and females. Therefore we will present two sets of experiments – one for males and the second for females. The data were obtained in form of X-Ray images. These images were scored according to Moorrees, Fanning and Hunt method.

Some individuals have missing teeth. This may be caused by teeth that have not start to develop yet (typically in young children) or individual have lost them (typically in older children). In this case we have imputed the missing value. We have taken the closest older and younger child and averaged development stages of corresponding tooth. After this step we have divided the original data into training and testing set. These sets will be fixed for all experiments.

4 Results

In this part we will present our results. First we will present results of feature selection methods and we will select teeth which we will use un the later part of the paper. Then we will present results of modelling of the age using full set of teeth for both, males and females. And in the last part we will present results for several subsets of teeth.

In many places in this section we will abbreviate names of the teeth to save space. Here are listed all teeth and their abbreviations:

- I1L – First Incisor on the Left Side
- I2L – Second Incisor on the Left Side
- CL – Canine on the Left Side
- P1L – First Premolar on the Left Side
- P2L – Second Premolar on the Left Side
- M1L – First Molar on the Left Side
- M2L – Second Molar on the Left Side
- M3L – Third Molar on the Left Side
- I1P – First Incisor on the Right Side
- I2P – Second Incisor on the Right Side
- CP – Canine on the Right Side
- P1P – First Premolar on the Right Side
- P2P – Second Premolar on the Right Side
- M1P – First Molar on the Right Side
- M2P – Second Molar on the Right Side
- M3P – Third Molar on the Right Side

4.1 Significance of Teeth for Age Modelling

To estimate significance of various teeth we have used several feature selection methods mentioned above. The Table 4.1 shows results of these methods. The

To summarise results from the table 4.1 and to support decision which teeth are important and which we can leave out we have created table 4.1. In this table we

Tab. 3 RMS Errors for female children using all teeth

Name of the Modelling Method	RMS
Zero Regression	3.28
SMO Regression	0.93
RBF Neural Network for Regression	0.92
Multiple-Variable Linear Regression	0.95
Single-Variable Linear Regression	1.15

have shown how many times each tooth was selected in table 4.1.

It shows that the most important teeth are the first and second molar on both sides and left canine. If we will count corresponding teeth on both sides of mandible as one, the most significant teeth will be the first and second molar, canine, the first premolar and with some exception the third molar. Later in this section we will present experiments with teeth selected here.

4.2 Age Modelling from Full Set of Teeth

As we have explained above we assume that teeth development differs between males and females. Therefore here we provide two sets of results – one for males and the other for females.

The table 4.2 shows results for females and lower mandible. The table shows that the most accurate modelling method is the RBF Neural Network. But the differences between the RBF, SMO and Multiple-Variable Regression are very small and therefore it is hard to tell which is the best. The Single-Variable Regression, though significantly less accurate, is not that bad. As we expected the worst modelling method is the Zero Regression method. It shows the worst possible model which is can used for age estimation. The positive thing is that other models are approximately three times more accurate. The another interesting thing about these results is that one tooth is enough to estimate age with reasonable accuracy.

To illustrate errors made by models we have created figures 4.2 and 4.2. The second figure 4.2 shows the error for each individual. The each dark dot in the figure represent actual age of one individual. The corresponding (in the same column) light dot represent estimated age. These figures shows errors for the RBF Neural Network model. The figure 4.2 shows the histogram of distribution of difference between actual and modelled age. The histogram shows that age of about 78% of children in the testing set are estimated with error between cca -1 – +1.2 years.

The remaining 20% of female children whose age was predicted with higher error presents mix of children whose teeth development is ahead or behind their chronological age and errors of the models. Errors are particularly visible at the right end of the figure 4.2 where the model and actual age differs considerably. The reason is simple – the development of teeth which are used for prediction stop and therefore the model

Tab. 1 Results of the feature selection and feature ranking methods. In case of feature ranking methods in the second column the first eight most significant teeth are shown (rows marked by FR). When using feature selection method teeth are in random order (methods marked FS).

Name of feature selection/ranking method	Selected Teeth
InfoGain Method (FR)	M2L M2P P1P P1L M1L M1P CL P2P
Gain Ratio Method (FR)	I1L M2L M1P I1P M1L M2P CP CL
χ^2 Method (FR)	M2P M2L P1P P1L M1L M1P P2P CL
Cfs Subset Evaluation (FS)	M1P M2P M3P CL P1L M2L M3L
Wrapper Subset Evaluation with Linear Regression (FS)	M1P M2P M3P CL
Wrapper Subset Evaluation with RBF Neural Network (FS)	M2P M3P M1L M1P

Tab. 2

Tooth	I1	I2	C	P1	P2	M1	M2	M3
Left side of mandible	1	0	5	3	0	4	4	1
Right side of mandible	1	0	1	2	2	6	6	3
Both sides combined	2	0	6	5	2	10	10	4

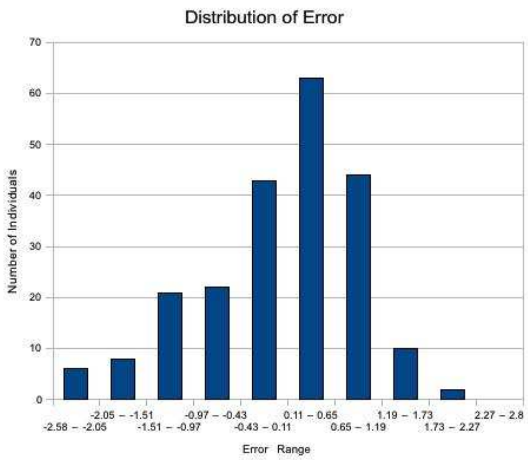

Fig. 2 Histogram showing distribution of errors of age in female children and full set of teeth

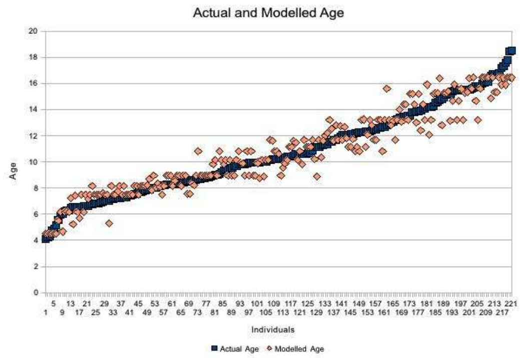

Fig. 3 Actual and predicted age in female children and full set of teeth

reaches the maximum predictable age and the gap between modelled and actual age begins to appear. This

Some modelling methods provides directly usable equations which one may directly use to estimate the age. We will begin with the equation produced by Single-Variable Linear Regression.

$$Age = 1.04 * M2P + 0.9 \tag{1}$$

.

The Multiple-Variable Linear Regression provides much more sophisticated equation

$$\begin{aligned} Age = {} & 0.88 * I1P + 0.91 * M1P \\ & + 0.46 * M2P + 0.45 * M3P - 0.96 * I1L \\ & + 0.34 * P1L - 0.82 * M1L + 0.65 \end{aligned} \tag{2}$$

Meaning of I1P, M1P, M2P, M3P, I1L, P1L and M1L is explained at the beginning of this section.

The last equation was produced by SMO regression model. This model works with the values normalised between $0 - 1$. This means that before application of the following equation one have to divide the development stage of all teeth by 13 (there are 14 development stages numbered $0, 1, \ldots 13$). The result is again number between 0 and 1 which represents age between 2.5 and 20.5 years.

$$\begin{aligned} Age = {} & 0.06 * I1P + 0.03 * I2P \\ & - 0.05 * CP - 0.06 * P1P + 0.17 * P2P \\ & + 0.08 * M1P + 0.20 * M2P + 0.19 * M3P \\ & - 0.01 * I1L - 0.07 * I2L - 0.01 * CL \\ & + 0.20 * P1L - 0.08 * P2L - 0.06 * M1L \\ & + 0.04 * M2L + 0.19 * M3L + 0.05 \end{aligned} \tag{3}$$

The results for the male children is very similar to the

Tab. 4 RMS Errors for male children using all teeth

Name of the Modelling Method	RMS
Zero Regression	3.47
SMO Regression	1.13
RBF Neural Network for Regression	1.08
Multiple-Variable Linear Regression	1.11
Single-Variable Linear Regression	1.38

results for female children. In table 4.2 we present modelling results for male children with all teeth as inputs. The accuracy of the resulting model is quite close to results for females and even though the results are slightly worse the difference is not significant.

To illustrate errors of the SMO model we will present the histogram of errors on figure 4.2 and difference between modelled and actual age on the figure 4.2. Results are similar to results from other models. The histogram shows that difference between actual and modelled age of 76% of individuals is between -0.82 and 1.43 years. To compare the SMO model for male children with the RBF model for female children we also have calculated portion of individuals with -1 – +1.2 difference between modelled and actual age. For the male children it is 72%. This is 8% less than in case of females. The same situation shows the figure 4.2. If you compare it with the same figure for females (figure 4.2) you will see the higher distance between modelled and actual age.

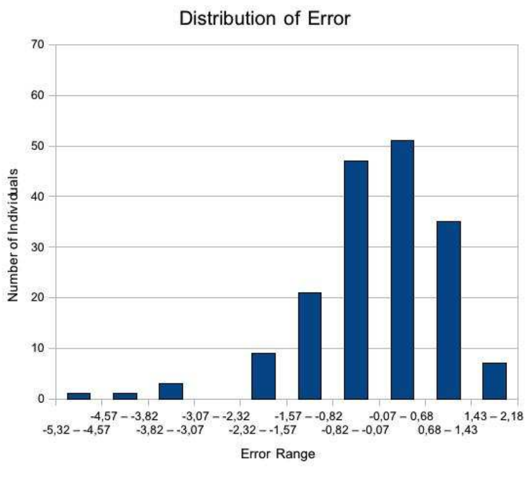

Fig. 4 Histogram showing distribution of errors of age in male children and full set of teeth

There are two possible explanations of higher error achieved for age prediction for male children – the first is that it have a biological background and development of teeth in male children is more variable and less correlated with the chronological age. The second explanation is that there is something wrong with the data for male children. The data for males are from another source and for example the development stages may be

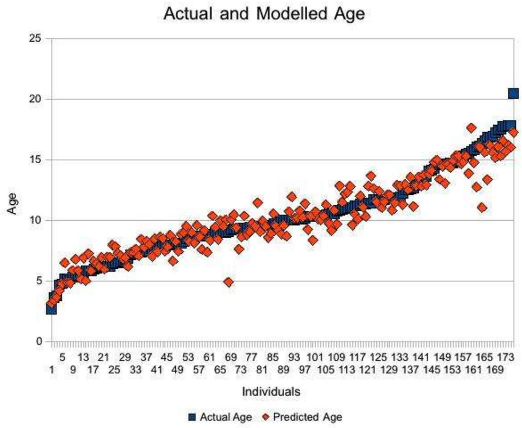

Fig. 5 Actual and predicted age in male children and full set of teeth

less carefully examined.

We will also present the equations for different models. The first equation which we will present is for Single-Variable Linear Regression.

$$1.14 * M2P$$

The second equation presents the Multiple-Variable Linear Regression model.

$$\begin{aligned} Age = &-0.30 * CP + 0.29 * M1P \\ &+ 0.24 * M2P + 0.40 * M3P + 0.47 * CL \\ &+ 0.16 * P1L + 0.25 * M2L - 2.78 \end{aligned} \tag{4}$$

Symbols are explained at the beginning of this section.

The last equation presented here was produced by SMO regression model. This model again works with the values normalised between $0 - 1$. This means that before application of the following equation one have to divide the development stage of all teeth by 13 (there are 14 development stages numbered $0, 1, \ldots 13$). The result is again number between 0 and 1 which represents age between 2.9 and 20.7 years.

$$\begin{aligned} Age = &-0.03 * I1P - 0.05 * I2P \\ &- 0.04 * CP + 0.06 * P1P + 0.011 * P2P \\ &+ 0.10 * M1P + 0.12 * M2P + 0.10 * M3P \\ &+ 0.09 * I1L - 0.05 * I2L + 0.10 * CL \\ &+ 0.11 * P1L - 0.04 * P2L + 0.03 * M1L \\ &+ 0.20 * M2L + 0.12 * M3L - 0.04 \end{aligned} \tag{5}$$

4.3 Age Modelling with Subset of Teeth

In this section we will create models from subset of teeth. We have chosen several subsets according to results of the feature selection and feature ranking methods. This is useful especially in archeology or forensics

Tab. 5

	Modelling Methods – RMS Error of model				
Subset of teeth	SMO Regression Method	Multi-variable Linear Regression	Single-variable Linear Regression	RBF Neural Network	Zero Regression
Left side only	1.06	1.37	1.07	1.12	3.29
Most significant teeth – left side only	0.98	0.96	1.15	1.14	3.28
Most significant teeth	1.15	1.13	1.15	1.04	3.28
Least significant teeth	1.31	1.28	1.29	1.34	3.28
Full set of teeth	0.93	0.95	1.15	0.92	3.28

Tab. 6

	Modelling Methods – RMS Error of model				
Subset of teeth	SMO Regression Method	Multi-variable Linear Regression	Single-variable Linear Regression	RBF Neural Network	Zero Regression
Left side only	1.19	1.19	1.42	1.18	3.18
Most significant teeth – left side only	1.13	1.11	1.38	1.2	3.47
Most significant teeth	1.39	1.35	1.38	1.40	3.47
Least significant teeth	1.44	1.45	1.46	1.56	3.47
Full set of teeth	1.13	1.11	1.38	1.08	3.47

when incomplete set of teeth is found. And it is an advantage to know how accurate the models are without some teeth. We will utilise information obtained in the first part of this section. The subsets we will test here are:

- Left side of the mandible – I1L, I2L, CL, P1L, P2L, M1L, M2L, M3L
- Most significant teeth identified above (left side only) – M1L, M2L, CL, M3L
- Most significant teeth identified above (both sides) – M1L, M2L, CL, M1P, M2P, CP
- The least significant teeth – I1L, I2L, P2L, I1P, I2P, P2P

The first set of teeth which we will test in this paper will be complete set of teeth from the left part of the mandible. For example [] states that both sides of mandible are developing with the same speed and for the age estimation one can easily replace the missing tooth with the corresponding tooth from the other part of the mandible and we want to test this theory and to find out how the models are affected. The second and third set represents the most significant teeth on one side only and then on both sides of mandible. The last set represent the least significant teeth. We were curious how big will be decrease of accuracy of models.

The table 5 shows RMS errors for female children for several different subsets. At the end of this table there is a row with results for full set of teeth for comparison.

First we will compare the full set of teeth to the teeth on the left side of mandible. The results for models with the left side only are in general slightly worse than models with full set of teeth. Our theory is that though differences between both sides of mandible are statistically insignificant there is difference between them which help models to decide.

The second set represents the most significant teeth on the right side of mandible. Results of this set are comparable to the full set of teeth. It shows that the most of of the age estimation is created from these teeth. The third row contains models with three most significant teeth on both sides of mandible. Models are a little bit worse than previous models. We explain this by third molars (M3P, M3L) missing in this dataset.

The fourth row – Least significant teeth – shows results for teeth which were marked as totally insignificant for the age estimation. As we expected, the results are much worse that for the full set of teeth. This show that removed teeth bear a lot of information which is missing now. On the other hand RMS Error of this model is half of the Zero Regression method. This means that these teeth even though insignificant they still contains information enough for decision. This is encouraging for forensics and anthropology since even the worst possible combination of teeth is able to produce some estimation.

The table 6 summarises results for male children. Results and comments are much similar to the results for the female children. With the exception that the best results are achieved by the model created from the full set of teeth. But the difference between this model and the most significant left-side subset is very small. In other word these models are comparable. And also comparable are to other models created from significant teeth (the first and third row in table 6). As we expected the model with the least significant teeth have the worst accuracy. Although the accuracy is significantly worse the models are still able to estimate age. And the estimation is twice better that the Zero Regression model. This shows that even the least significant teeth are able to model the age.

5 Conclusion

In this paper we have modelled dental age from teeth the teeth development stages of female and male children from Czech Republic between 2.5 – 20.5 . Development stages of teeth was identified according to Fanning, Moorees, Hunt method. First we have identified all three molars and the canine as the most significant teeth for the age estimation.

We have created models for teeth development using several well known modelling methods. The best model – RBF Neural Networks with 20 neurons in hidden layer – for female children achieved RMS error of 0.92 years. The model was created with the full set of teeth. In the best model the difference between modelled and actual chronological age of 80% female children is within range -1 – +1.2 years. The remaining 20% represents children whose teeth development is behind or ahead of their chronological age and errors of models.

For male children the models achieve slightly worse accuracy. The best model – again RBF Neural Network with 20 neurons in hidden layer – achieves RMS error of 1.08 years. The model was again created with full set of teeth. In the best model the difference between modelled and actual chronological age of 72% male children is within range -1 – +1.2 years. The reason for higher error in male children is not yet known. One possible reason is higher variance in teeth development in male children. The second may be less care in assigning development stages to the teeth.

The last experiment in this paper was to estimate the age with reduced set of teeth. We have selected four different sets of teeth which we have selected based on information from the first part of this paper. The results of three sets which includes the most significant teeth are comparable to each other and are also comparable to results of models with the full set of teeth. In both cases – female and male children – the models with the full set of teeth are the most accurate. This shows that even though that some teeth are not so significant for the age prediction they still cary some information which models can utilise.

Acknowledgments

This research is supported by Internal Grant "Improvement of data preprocessing module in FAKE GAME project" of Czech Technical University and the research program "Transdisciplinary Research in the Area of Biomedical Engineering II" (MSM6840770012) sponsored by the Ministry of Education, Youth and Sports of the Czech Republic.