Forensic Science International 279 (2017) 72–82

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Technical Note

Novel age estimation model based on development of permanent teeth compared with classical approach and other modern data mining methods

Michal Štepanovský[a], Alexandra Ibrová[b,*], Zdeněk Buk[c], Jana Velemínská[b]

[a] Department of Computer Systems, Faculty of Information Technology, Czech Technical University in Prague, Thákurova 9, 160 00 Prague, Czech Republic
[b] Department of Anthropology and Human Genetics, Faculty of Science, Charles University, Viničná 7, 128 43 Prague, Czech Republic
[c] Department of Theoretical Computer Science, Faculty of Information Technology, Czech Technical University in Prague, Thákurova 9, 160 00 Prague, Czech Republic

ARTICLE INFO

Article history:
Received 14 March 2017
Received in revised form 25 July 2017
Accepted 4 August 2017
Available online 14 August 2017

Keywords:
Age estimation
Data mining
Model
Population-specific standards

ABSTRACT

In order to analyze and improve the dental age estimation in children and adolescents for forensic purposes, 22 age estimation methods were compared to a sample of 976 orthopantomographs (662 males, 314 females) of healthy Czech children and adolescents aged between 2.7 and 20.5 years. All methods are compared in terms of the accuracy and complexity and are based on various data mining methods or on simple mathematical operations. The winning method is presented in detail.

The comparison showed that only three methods provide the best accuracy while remaining user-friendly. These methods were used to build a tabular multiple linear regression model, an M5P tree model and support vector machine model with first-order polynomial kernel. All of them have mean absolute error (MAE) under 0.7 years for both males and females. The other well-performing data mining methods (RBF neural network, K-nearest neighbors, Kstar, etc.) have similar or slightly better accuracy, but they are not user-friendly as they require computing equipment and the implementation as computer program. The lowest estimation accuracy provides the traditional model based on age averages (MAE under 0.96 years). Different relevancy of various teeth for the age estimation was found. This finding also explains the lowest accuracy of the traditional averages-based model.

In this paper, a technique for missing data replacement for the cases with missing teeth is presented in detail as well as the constrained tabular multiple regression model. Also, we provide free age prediction software based on this wining model.

1. Introduction

Age estimation methods for the purposes of forensic practice require high accuracy of results. In children and adolescents one of the most stable markers for age estimation is the development of dentition. There are various limitations for age estimation from dental remains, for review see Refs. [1,2], but the developmental stages of teeth highly correlate with the actual chronological age of children and adolescents, as genetics plays a more important role than environment in mineralization of teeth [1]. There are various methods for calculating the age of an individual from mineralization stages of dentition (e.g. [3–5]) which are traditional, easy to use and provide a decent level of accuracy. A number of authors developed modifications of these methods in order to increase the accuracy, adjust the tables for specific populations or to develop a more complex approach (e.g. [6–8]). Also, with new computing methods and advanced computing equipment, more sophisticated approaches using data mining methods were adopted for improvement of age predictions [9–11]. However, the more complex and sophisticated the method becomes, the less user-friendly it is. The data mining methods (DMM) may provide better results than traditional methods, but they also usually require special computing equipment, computer programmers or a specialist in the field.

The questions are—do the complex and sophisticated methods provide an improvement of results at such level that they are worth engaging in forensic practice? And are there DMM methods that provide good results while remaining user-friendly? Therefore, the main aim of this paper is to compare 22 methods based on either DMM or on simple mathematical operations in terms of their

* Corresponding author.
E-mail address: ibrovaa@natur.cuni.cz (A. Ibrová).

http://dx.doi.org/10.1016/j.forsciint.2017.08.005
0379-0738/

accuracy, complexity and applicability and also to present the winning method in detail.

2. Material and methods

This study was based on a sample of 976 orthopantomographs taken of 662 boys and 314 girls of Czech nationality aged between 2.7 and 20.5 years. Panoramic X-ray films were performed on individuals treated at three clinics in Prague, Czech Republic (Stomatology Clinic of the Faculty Hospital Královské Vinohrady, Pediatric Stomatology Clinic of the Faculty Hospital Motol and Stomatology Clinic of the First Faculty Hospital in Prague). Imaging was carried out mainly in the 1990s (range from 1981 to 2007). The selected children were born between 1973 and 2004. They were healthy and without any growth disorders. The number of individuals related to the class intervals of width 5 years are provided in Table 1.

Development of each tooth was divided into 14 sub-stages, and each stage was assigned a numerical value ranging from 1 to 14 [3]. "Initial cusp formation" was denoted as stage 1, the "Coalescence of cusps" as stage 2 and so forth until the last stage "Apical closure complete" as stage 14. Stage 0 was used when no data was available. Table 2 summarizes tooth development stages. Data processing and analysis were performed using software tools Matlab [http://www.mathworks.com/products/matlab] and Weka [http://www.cs.waikato.ac.nz/ml/weka]. The mean absolute error and root mean squared error for all presented models (#1–#22) was estimated by using 5-fold cross-validation, where the models are completely build upon the training set and no information from the testing set is involved during the training phase. Hyper-parameters of used models were tuned on the training set only.

Data mining methods are able to directly use tooth development stages as an input. Nevertheless, the tooth development stage represents an ordinal categorical variable with a nonlinear monotonic relationship to the dental age of an individual. Therefore, we also investigated the possibility of replacing tooth development stages by the representative median (or average) age before creating the model. The median (or average) age was computed from all individuals of representative population who have the same mineralization stage for the same tooth type. This potentially eliminates the nonlinear relationship and transforms the tooth development stage into a ratio-scaled continuous variable. Models using this transformation are referred as "tabular". Other models, namely models #3, #5, #7, #9, #11, #13, #15, #17 and #19, use tooth development stages directly as an input and produce estimated age of an individual.

This section continues with short descriptions of principles of each analyzed age estimation model:

Model #1: Tabular model based on age averages, e.g. [3,12], is a widely-used classical method of age estimation in forensic practice due to its simplicity. The model uses tables containing the average age of all individuals from a representative population who have an equally developed specific tooth type. These tables can be found e.g. in Smith [12]. Age estimation of unknown individual is realized by estimating the developmental stage of each available tooth from an X-ray image, looking up the age for each estimated stage in the tables and computing the average value of age. This means that each tooth has the same contribution/weight for the final age estimation. To make the results in this paper comparable, our own tables of age averages from our dataset were computed. Moreover, the lower and the upper 5% of samples in all development stages were removed before computing the averages to avoid existing anomalies. Removal of anomalies was used only for generating tables; afterwards, the model was tested using the full dataset.

Table 1
Grouped frequency distribution of study sample.

Age	0–5 years	5–10 years	10–15 years	15–20 years	over 20 years
Males	25	320	258	58	1
Females	20	152	99	43	0

Table 2
Modification of developmental stages from Moorrees et al. (1963).

Meaning	Coding	
	Single-rooted teeth	multi-rooted teeth
–	0	0
Initial cusp formation C_i	1	1
Coalescence of cusps C_{co}	2	2
Cusp outline complete C_{oc}	3	3
Crown $^1/_2$ complete $C_{r1/2}$	4	4
Crown $^3/_4$ complete $C_{r3/4}$	5	5
Crown complete C_{rc}	6	6
Initial root formation R_i	7	7
Initial cleft formation C_{li}	–	8
Root length $^1/_4$ $R_{1/4}$	8	9
Root length $^1/_2$ $R_{1/2}$	9	10
Root length $^3/_4$ $R_{3/4}$	10	11
Root length complete R_c	11	12
Apex $^1/_2$ closed $A_{1/2}$	12	13
Apical closure complete A_c	13	14

Model #2: Tabular model based on age medians could be considered as an alternative to model #1, where the only difference is that medians are used instead of age averages.

Model #3: Multiple linear regression model (MLR) [13] is based on a method that approximates dental age by a linear equation. The model directly uses tooth development stages, thus avoiding age tables as is the case of model #1. In this model the collinear attributes were removed, and attribute selection using the Akaike information metric was used to remove attributes with the smallest standardized coefficient if this improves the final model.

Model #4: Constrained tabular multiple linear regression model (TMLR) is similar to MLR (model #3) but uses the transformation of input data as described above and only non-negative coefficients. This model is easy to use and overcomes two main drawbacks of the basic model based on age averages (see model #1). First, it uses medians which are much more robust comparing to the average, and thus it is better suited for skewed data and for avoiding outliers [14,15]. Second, instead of computing the average from estimations related to individual teeth it uses a linear model (linear combination) reflecting tooth age estimation capabilities. In fact, computing the average is a special case of linear combination where each term is multiplied by exactly the same constant. Thus, this model can be considered as generalization of model #1, except that it uses medians instead of averages. For this model, we compare two versions—version A and version B. In the version A, the collinear attributes were removed and a greedy method was used for the attribute selection using the Akaike information metric. Moreover, the teeth producing negative coefficients in the created model were simply not included. This guarantees the ordering of the model outputs with respect of increasing tooth development stages—i.e. higher development stage results in higher estimated age. In the version B, we use the algorithm implemented in Matlab by lsqnonneg, which is a function designed to solve non-negative least-squares problem and it is based on algorithm described in Ref. [16]. This implementation belongs to active-set methods [17]. Usually, only a subset of constraints is active at the solution, i.e. the corresponding regression coefficient would be negative if unconstrained. If the true active set is known, the solution will simply be