Machine Learning in sex determination

Some times, more complex doesn't mean better

•••

Juan F. Palomeque-Gonzalez
Director Atlas Archaeology Ltd.
Dep. Prehistoria UCM Madrid

First Idea

- I would like to apply the new advances in Machine Learning to a archaeological problem
- Should be a problem with enough data available, and with some traditional approaches
- And a problem relatively common, no something incredibly rare or unusual.

Problem to solve: Sex determination in skeletal remains

- Osteological Database: Goldman Osteological Dataset
- Frequent archaeological find
- Importance in research and historical interpretation

Database

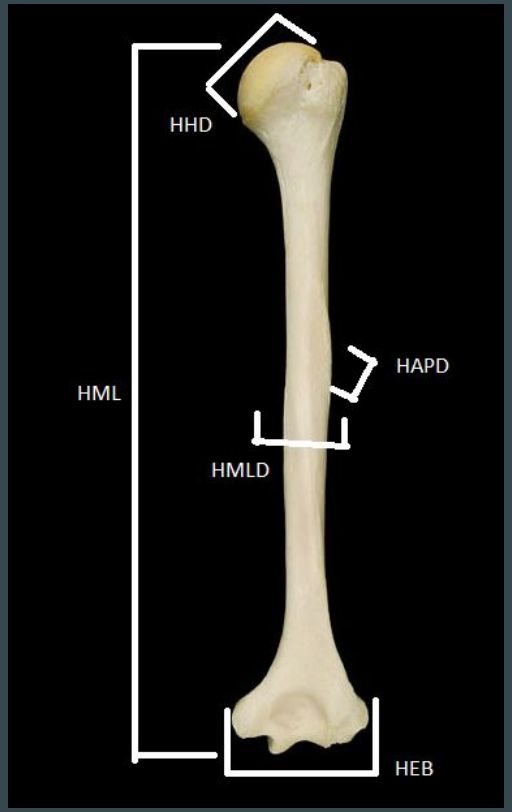
- Goldman Osteological Dataset (GO Dataset) (AUERBACH & RUFF, 2004) as of 13th December 2018.
- 1538 human skeletons, coming from diverse archaeological, anthropological or forensic collections.
- Measurement in all long bones, differentiated by side.
- Data about sex of the individuals.

Dataset processing

- We are going to use humerus measurements to try to determine the sex of the individuals using different machine learning techniques.
- In this case, there is not going to be side differentiation in the bones, so we are mixing humerus from both sides.
- Owing to some missing values in the original dataset, KNN imputation was used to complete the database.

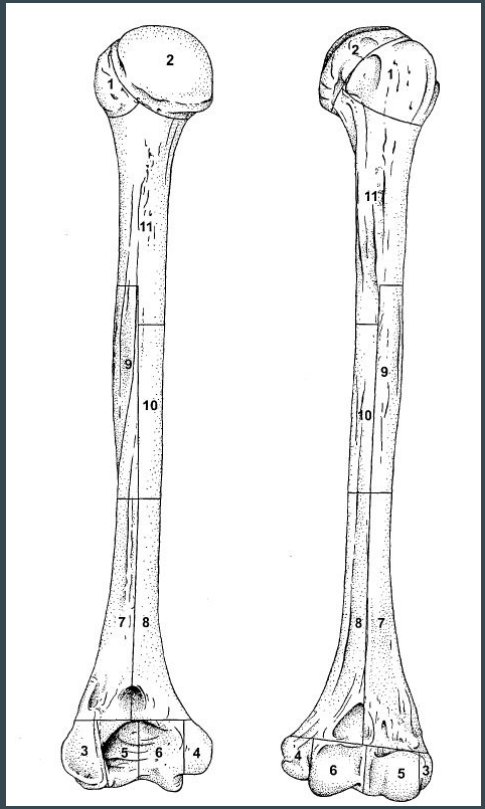
Measurements

- HML: Humerus maximum length
- HEB: Humerus Epicondylar Breadth
- HHD: Humerus Head Diameter
- HMLD: Humerus 50% Diaphyseal Mediolateral Diameter
- HAPD: Humerus 50% Diaphyseal Anteroposterior Diameter



Combinations

- Try to reproduce posible real archaeological finds
- All the measurements (HML, HEB, HHD, HMLD and HAPD)
- Only measurements in distal epiphysis and more than 50% of the diaphysis, HEB, HMLD and HAPD
- Only measurements in proximal epiphysis and more than 50% of the diaphysis, HHD, HMLD, HAPD)
- Only measurements in both epiphysis, HEB, HHD, HMLD, HAPD).



Techniques to test

Eleven different methods python package Scikit-learn

- Nearest Neighbors
- Linear Support Vector Machines (Linear SVM)
- Radial Basis Function Support Vector Machines (RBF SVM)
- Gaussian Process
- Decision Tree
- Random Forest (RF)
- Artificial Neural Network (ANN)
- AdaBoost
- Naive Bayes
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)

Test environment

- Cross-validation with 10 iterations was applied, and the average accuracy and standard error were calculated. (Kohavi, 1995).
- All the algorithms were running in the cloud computing service named Kaggle.com, to enable an easier comparison between them, and to help code sharing.

Results with complete bone

Classifier	Measurements	Accuracy	Standard_dev (+-)	Time (s)
RBF SVM	HML + HAPD	0.98	0.05	0.041187889
RBF SVM	HML + HMLD	0.96	0.03	0.032355795
AdaBoost	HML + HAPD	0.96	0.11	0.533332608
AdaBoost	HML + HMLD	0.94	0.08	0.511681744
Random Forest	HML + HAPD	0.9	0.14	0.182151562
Decision Tree	HML + HMLD	0.89	0.11	0.013340822
Decision Tree	HML + HAPD	0.89	0.14	0.014590186
Gaussian Process	HML + HMLD	0.88	0.11	12.06562505
Random Forest	HML + HMLD	0.88	0.07	0.168643684
Linear SVM	HEB + HHD	0.87	0.06	0.483543421

Results with proximal half of the bone

Classifier	Measurements	Accuracy	Standard_dev (+-)	Time
Linear SVM	HHD + HMLD	0.86	0.06	0.4509
Gaussian Process	HHD + HMLD	0.86	0.06	1191.799
LDA	HHD + HMLD	0.86	0.06	0.06301
QDA	HHD + HMLD	0.86	0.06	0.04153
Linear SVM	HHD + HAPD	0.86	0.06	0.4817
Random Forest	HHD + HAPD	0.86	0.07	0.2397
AdaBoost	HHD + HAPD	0.86	0.07	1.0966
Linear SVM	HHD + HMLD + HAPD	0.86	0.06	0.4828
Gaussian Process	HHD + HMLD + HAPD	0.86	0.06	1378.781
LDA	HHD + HMLD + HAPD	0.86	0.06	0.07273

Results with distal half of the bone

Classifier	Measurements	Accuracy	Standard_dev	Time
Linear SVM	HEB + HMLD	0.85	0.06	0.468612991
Linear SVM	HEB + HAPD	0.85	0.06	0.494
Gaussian Process	HEB + HAPD	0.85	0.07	1085.947
Random Forest	HEB + HAPD	0.85	0.06	0.1912
AdaBoost	HEB + HAPD	0.85	0.05	0.9893
RBF SVM	HEB + HMLD	0.84	0.06	3.346105789
Gaussian Process	HEB + HMLD	0.84	0.07	1025.181499
Decision Tree	HEB + HMLD	0.84	0.07	0.042554639
Random Forest	HEB + HMLD	0.84	0.07	0.183062724
AdaBoost	HEB + HMLD	0.84	0.06	0.910989408

Results with only both epiphysis

Classifier	Columns	Accuracy	Standard_dev (+-)	Time
Linear SVM	HEB + HHD	0.87	0.06	0.483543421
RBF SVM	HEB + HHD	0.87	0.06	3.72588986
Gaussian Process	HEB + HHD	0.87	0.06	1143.067841
Decision Tree	HEB + HHD	0.87	0.06	0.045386651
Random Forest	HEB + HHD	0.87	0.07	0.205463358
AdaBoost	HEB + HHD	0.87	0.06	1.156648346
Naive Bayes	HEB + HHD	0.87	0.07	0.020453713
LDA	HEB + HHD	0.86	0.07	0.079997056
Nearest Neighbors	HEB + HHD	0.85	0.06	0.034860382
Artificial Neural Net	HEB + HHD	0.64	0.08	550.4889243

Conclusions and future

- Using only “simple” humerus measurements and machine learning techniques can reach about 98% in accuracy in determination sex of skeletal remains. (85% in case of non-complete bone)
- That shows how this techniques can be used in a broad range of situations, no only in academic research, but in commercial or professional research too.
- The development of new ways to bring this methods to a wider public (from GUI to public access standard databases) can help in the advance of our knowledge about the human past. But this use can help in the creation of larger datasets.
- In general, more data could mean new techniques or uses of this methods in new scenarios. More standard databases and communication between commercial and academic archaeology.

Questions?

Thanks for your attention!

Bibliography

- Auerbach BM, & Ruff CB. 2004. Human body mass estimation: a comparison of “morphometric” and “mechanical” methods. American Journal of Physical Anthropology 125:331-342.
- Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).
- Avila, Julian, and Trent Hauck. 2017. Scikit-Learn Cookbook: Over 80 Recipes for Machine Learning in Python with Scikit-Learn. Packt Publishing Ltd.
- Garreta, Raul, and Guillermo Moncecchi. 2013. Learning Scikit-Learn: Machine Learning in Python. Packt Publishing Ltd.
- Jolly, Kevin. 2018. Machine Learning with Scikit-Learn Quick Start Guide: Classification, Regression, and Clustering Techniques in Python. Packt Publishing Ltd.

Bibliography II

- Kenyhercz, M. W., and N. V. Passalacqua. 2016. “Missing Data Imputation Methods and Their Performance With Biodistance Analyses.” In Biological Distance Analysis, 181–94.
- Knüsel, C. J., & Outram, A. K. (2004). Fragmentation: The Zonation Method Applied to Fragmented Human Remains from Archaeological and Forensic Contexts. Environmental Archaeology, 9(1), 85–98.
- Ogedengbe, O. O., Ajayi, S. A., Komolafe, O. A., Zaw, A. K., Naidu, E. C. S., & Okpara Azu, O. (2017). Sex determination using humeral dimensions in a sample from KwaZulu-Natal: an osteometric study. Anatomy & cell biology, 50(3), 180-186.