Empirically exploring the effect of oxygen on the isotopic mapping of cremated and uncremated bones of a Central European Alpine passage.

Markus Mauder, Eirini Ntoutsi, Gisela Grupe, Peer Kroeger

Isotopic mapping has become an indispensable tool for the assessment of mobility and trade in the past. In this work we focus on the analysis of bioarchaeological finds derived from archaeological sites covering the transalpine passage. Our goal is to construct an isotopic map of the reference region and be able to locate the origin of new samples. To this end, data mining is employed that based on the available samples and the task at hand is able to derive task specific models like clustering, classification and regression models for data exploration, exploitation and prediction. There are several challenges for data mining in this context like the luck of sufficient samples, samples of different isotopic description, diagenetically altered samples etc.

In this work we focus on the integration of samples with different isotopic description in order to increase the size of the training set and therefore be able to extract more stable and general models. Our work is motivated by a problem at hand: part of our samples is uncremated whereas the rest are derived after cremation. Restring our analysis in only the cremated or uncremated samples is not the optimal solution given that the amount of samples is small and the derivation of the samples is a costly and time consuming procedure. Therefore, being able to combine the cremated with the uncremated samples for analysis purposes would be the optimal solution.

However, combining samples of different characteristics/ features is not straightforward. The straightforward approach on working in the common feature space translates into omitting some part of information that might be crucial for the analysis task. Consider for example the lung cancer prediction task; the information on whether a patient is a smoker or not is a crucial one since it has been found to be correlated with the class label. Therefore, omitting this information would result in a non-accurate model.

For the problem at hand, the missing information is the oxygen isotope which is not available in the cremated samples. Omitting the oxygen and working with the rest of the attributes, namely lead and strontium, is the straightforward approach in order to merge the two sets of samples, the cremated and the un-cremated ones. But, as discussed above, such an omission might lead to information loss. It is not clear what is the effect of oxygen, especially for the isotope fingerprint and the prediction model we want to build.

To this end, we present a comprehensive study on the effect of the inclusion/exclusion of features or feature-sets in the extracted data mining models. We evaluate the effect in terms of stability of the data mining models, which intuitively describes whether a model in the reduced feature space (i.e., without oxygen) “behaves” similarly to a model in the complete feature space (i.e., including oxygen). The notion of “behavior” here is quite abstract and depends on the task per se. In case of clustering, a similar behavior would be a similar partitioning of the population into clusters. In case of classification, a similar behavior would be same class predictions for an unseen instance. We formulate the notion of stability per task and show how the stability is affected as we change the feature space where the population is described. We also investigate whether and how the stability results correlate with feature selection and ranking.