GMMbuilder – User-Driven Discovery of Clustering Structure for Bioarchaeology

Markus Mauder[1], Eirini Ntoutsi[2], and Yulia Bobkova[1]

[1] Ludwig-Maximilians-University Munich
{mauder@dbs,yulia.bobkova@campus}.lmu.de
[2] Leibniz Universität Hannover
ntoutsi@kbs.uni-hannover.de

Abstract. We present the *GMMbuilder* tool that allows domain scientists to build Gaussian Mixture Models (GMM) that adhere to domain specific knowledge. Domain experts use the tool to generate different models, extract stable object communities across these models and use these communities to interactively design a final clustering model that explains the data but also considers prior beliefs and expectations of the domain experts.

Keywords: bioarchaeology, isotopic mapping, Gaussian Mixture Models, interactive clustering, community detection, demo

1 Introduction

Nowadays data mining comprises an indispensable tool for social and humanity sciences. The *GMMbuilder* tool was developed in the context of the interdisciplinary research project FOR1670 [1]. The goal of the project is to build an isotopic fingerprint for bioarchaeological finds (human and animal remains) from excavation sites along the Inn-Eisack-Adige passage spanning Italy, Austria and Germany. The data consists of spatial information on the location of the finds and the ratios of oxygen, strontium and lead isotopes in the finds. Data mining methods were employed for the construction of a large scale isotopic map of the area to be used for differentiating between local and non-local finds and for the definition of the place of origin of the latter.

To be useful for origin prediction, the derived model must be based solely on isotopes, i.e., the spatial information of the finds should not be used for model building. Domain knowledge however suggests that the derived models should also depict spatial coherence. Intuitively, this means that finds coming from the same location should have similar isotope values. However, the task of building a model of plausible origins of the measured values is complicated by the noise introduced by the environment, range areas of animals, import of

---

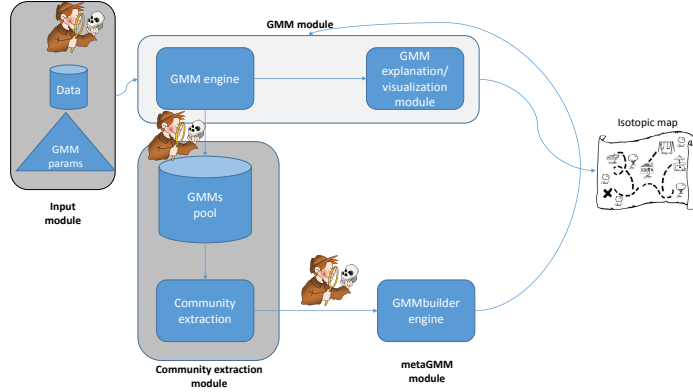[1] www.for1670-transalpine.uni-muenchen.de

Fig. 1. An overview of *GMMbuilder*

food, and further confounding factors. Additionally, displacement of live humans and animals and also animal trading in the past generates mixed measurements and spatial outliers. Over the course of many months various clustering models were developed, discussed with the domain experts and refined based on their feedback. This approach resulted in very slow turnaround. *GMMbuilder* was developed to allow model building and assessment in an integrated fashion with the final goal of deriving a model that fit the data well but it is also in accordance with domain knowledge (for example, spatial coherence of the models).

2 *GMMbuilder*

The final model is built by identifying strong object communities in the data and incorporating the models of these communities into the final clustering model. To derive the strongly connected components in the data we rely on unsupervised learning: In particular, we generate multiple clusterings from the data and we find object formations that are stable across many clusterings. The intuition is that similar objects should be clustered together across the different clusterings. The domain expert has a very active role in the whole process: from the selection of the clusterings from which the communities will be extracted to the selection of the communities that will form the basis for the final clustering model. Figure 1 depicts the *GMMbuilder* architecture, consisting of several modules that will be presented hereafter. As it is shown in this figure the role of the domain expert is vital.

GMM module The GMM module derives a clustering over a given dataset D. Domain knowledge suggests continuous values for the measurements, which can be best modeled as a mixture model of continuous distributions, like a Gaussian Mixture Model (GMM). Therefore, we applied the Expectation-Maximization (EM) algorithm [1] to extract a robust indication of the data's structure in an

unsupervised way. EM fits k multi-variate normal distributions over the given dataset, k is a user-defined input. The result is a soft-clustering; in our dataset though the assignment is typically fairly hard [2].

Input module & the GMM pool The selection of the input data D for the GMM model is crucial as it affects the derived clustering model. Therefore, we rely on the domain experts to decide which of the generated models are acceptable or not. The decision is based on their expertise, however in order to facilitate their task, we provide a detailed clustering description, in terms of the spatial projection of the clustering and the isotope distribution of the different clusters (c.f. Figure **??**). The result of this step is a set of user-accepted clusterings, the so-called *clustering pool* $\mathbb{C}$.

Stable community extraction module By examining the different clusterings, we can identify objects that are frequently assigned to the same cluster. We call such object formations "stable" communities. More formally, a stable community c consists of a set of points $p \in D$ that are clustered into the same cluster across multiple clusterings $\mathbb{C}$:

$$c(C_1, C_2, \ldots, C_n) = \{p \mid p \in C_{1,i} \wedge p \in C_{2,j} \wedge \ldots \wedge p \in C_{n,m}\}$$

where $C_{i,j}$ is the set of points in the jth cluster in clustering C_i.

The idea is to use these strong components as building blocks for clustering, because their members have shown a strong adhesion to each other over a range of clusterings and therefore, they are more likely to represent a cluster in any final model-based clustering.

metaGMM module The stable communities extracted from the previous step which indicate strong connections in their data objects might not agree with domain experts' prior beliefs and expectations. For example, a community might consist of objects which are close in the isotopic space, but their spatial coordinates are far apart. Such a case is very probable in our application due to mobility and trading in the past. Since the domain experts are interested in an isotopic clustering model that is also spatially coherent, the aforementioned community is not a good "seed" for the *GMMbuilder*.

Therefore, we rely again on the domain expert to decide which of the detected stable communities should take part in the final model generation. When the expert selects a community c to evaluate, a Gaussian model of its objects is extracted and added to the GMM pool. This new GMM is used to re-evaluate the membership probability of each data point in our dataset D. In that sense, a new temporal clustering is created based on c's model. The user can directly inspect the results and decide whether it is a good or bad model for final clustering. To support user's decision, the community "profile" is presented to the user; the profile consists of the spatial distribution of the community members and their isotope characteristics, namely, variance and covariance. For the former, we use a map, for the later parallel coordinates visualization (c.f., Figure **??**. The user

can then select another community c' to evaluate. This is an iterative process, the user can add or remove communities and directly inspect the effect in the final clustering. The process stops when the user is satisfied with the results. The output of this step is a set of user-accepted communities from which Gaussian models are extracted. All points in D will be assigned to these models, deriving the final clustering.

3 Demo scenario

In our running scenario we generate different clusterings by varying the number of clusters for the EM algorithm. The users can inspect the individual clusterings, with the help of the clustering statistics and visualization window, and select those that should contribute to the final model. The stable communities, derived from the selected clusterings, will be shown to the user. The user can interactively choose which of these communities should be part of the final clustering model. User decisions are reflected to the final model so the user can directly inspect the effect of her decisions and proceed accordingly by removing or adding certain components.
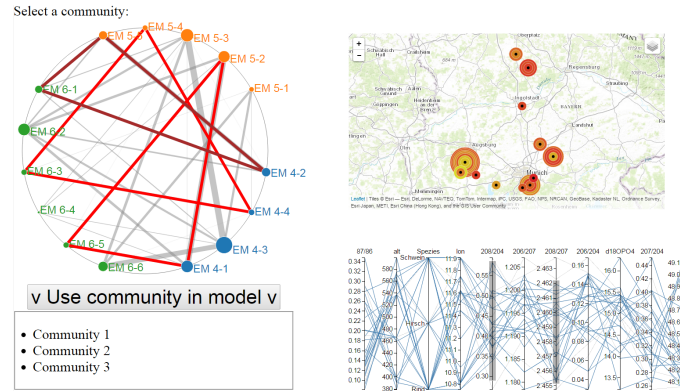

Fig. 2. *GMMbuilder*: Interactive GMM building - inspecting one of communities found in three clusterings (orange, green and blue)

References

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological) pp. 1–38 (1977)
2. Mauder, M., Ntoutsi, E., Kröger, P., Grupe, G.: Data mining for isotopic mapping of bioarchaeological finds in a central european alpine passage. In: SSDBM (2015)