GMMbuilder – User-Driven Discovery of Clustering Structure for Bioarchaeology

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

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

Abstract. We present *GMMbuilder*, a tool that allows domain scientists to build Gaussian Mixture Models (GMM) that adhere to domain specific constraints like spatial coherence. Domain experts use this 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

Data mining has become an indispensable tool for social and the humanities. The *GMMbuilder* tool was developed in the context of the interdisciplinary research project FOR1670[1] that aims at building 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 to differentiate local from non-local finds and to define the place of origin of the latter.

To be useful for origin prediction, the derived model must be based solely on isotopes, i.e. supplementary information like the spatial origin of the finds should not be used for model building. Domain knowledge however suggests that the derived models should also be spatially coherent. 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 food, and further confounding factors. Additionally, displacement of live humans and animals and also animal trading in the past generates mixed

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

B. Berendt et al. (Eds.): ECML PKDD 2016, Part III, LNAI 9853, pp. 8–11, 2016.
DOI: 10.1007/978-3-319-46131-1_2

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 was characterized by very slow turnaround. *GMMbuilder* was developed to allow model building and assessment in an integrated fashion and allow for immediate feedback by domain experts. The result is a model that fits the data well but it is also in accordance with domain knowledge (for example, spatial coherence of the models).

2 *GMMbuilder*

The 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.

Clusterer module. The Clusterer 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, the Expectation-Maximization (EM)

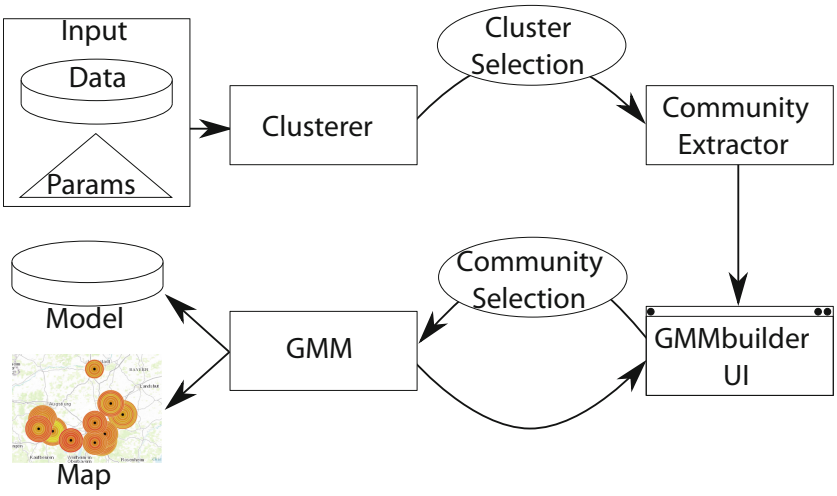

Fig. 1. An overview of *GMMbuilder*. Oval shapes depict user interaction.

algorithm [1] was applied 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 and Cluster Selection modules. 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. 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 cluster members and the distribution of the isotope values in each cluster. The result of this step is a set of user-accepted clusterings $\mathbb{C}$.

Community Extractor 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 $\mathbb{C}$:

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

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 the final 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.

GMM 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. 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 inform 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. This new GMM is used to re-evaluate the membership probability of each data point in our dataset D and a new 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 the user's decision, the community is presented to the user by depicting the spatial distribution of the community members and their feature distribution. The former is shown in a map, the latter as parallel coordinates (c.f. Fig. 2). The user can then select another community c' to evaluate. Again a Gaussian model of its objects will be extracted and added to the GMM. The old component c and the new component c' will be used to re-evaluate

the membership probability of all points in D. This is an iterative process, the user can add or remove communities and directly inspect the effect on the final clustering. 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 example 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 presented to the user. The user can interactively choose which of these communities should be part of the final clustering model. User decisions are reflected in the final model so the user can directly inspect the effect of her decisions and proceed accordingly by removing or adding certain components. *GMMbuilder* is a web-based tool. A screenshot of the interactive model building step is shown in Fig. 2.

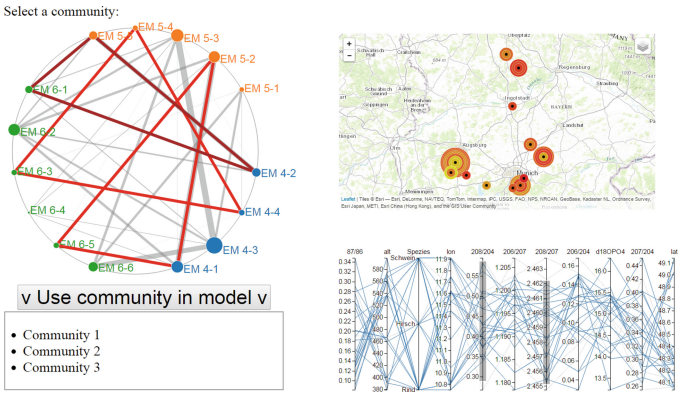

Fig. 2. *GMMbuilder*: Interactive GMM building - inspecting one of the communities found in all three clusterings (orange, green and blue). (Color figure online)

References

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Ser. B (Methodological) **39**(1), 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: 27th International Conference on Scientific and Statistical Database Management (2015)

http://www.springer.com/978-3-319-46130-4

Machine Learning and Knowledge Discovery in Databases
European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part III
Berendt, B.; Bringmann, B.; Fromont, E.; Garriga, G.; Miettinen, P.; Tatti, N.; Tresp, V. (Eds.)
2016, XXII, 307 p. 119 illus., Softcover
ISBN: 978-3-319-46130-4