Explain your clusters with words. The role of metadata in interactive clustering

Maciej Mozolewski[1], Samaneh Jamshidi[2], Szymon Bobek[1] and Grzegorz J. Nalepa[1]

[1]*Jagiellonian Human-Centered Artificial Intelligence Laboratory (JAHCAI) and Institute of Applied Computer Science, Jagiellonian University, Cracow, Poland*

[2]*Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Halmstad, Sweden*

Abstract

In this preliminary work, we present an approach for augmentation of clustering with natural language explanations. In clustering there are 2 main challenges: a) choice of a proper, reasonable number of clusters and b) cluster analysis and profiling. There is a plethora of technics for a) but not so much for b), which is in general a laborious task of explaining obtained clusters. In this work, we propose a method that aids experts in cluster analysis by providing iterative, human-in-the-loop methodology of generating cluster explanations. In a convincing example, we show how the process of clustering on a set of *objective variables* could be facilitated with textual *metadata*. In our case, images of products from online fashion store are used for clustering. Then, product descriptions are used for profiling clusters.

Keywords

XAI, clustering, metadata, Natural Language Processing, explanations, narratives

1. Introduction

Assigning labels to groups of similar objects is one of the ways how humans describe the world [**?**]. It begins with the notion that some phenomena or entities differ from each other and that they could be divided into distinct classes. Clarification of the differences between groups gets better and better along with the knowledge gained about the instances that form different groups. Finally, one is giving names to those different categories of entities. In essence, clustering in machine learning is no different process.

Clustering is an intrinsically subjective task and requires human assessment [**?**]. It is a purely statistical method which finds homogeneous groups of entities. It belongs to the family of unsupervised learning algorithms in contrast to classification or regression, which are supervised. At every step of this process, the user makes decisions based on her/his domain knowledge. First, the user needs to select features (variables) used by the algorithm. Secondly, the user selects the type of algorithm, similarity measures, number of clusters or size of the smallest one. Finally, she or he checks clusters by describing objects belonging to subsequent groups. It also follows that the process is iterative.

From our expertise in e-commerce and Industry 4.0, we often see distinctions between two types of data. There are *objective data* and the *subjective data* or *metadata*. For instance, in e-commerce popular approach for recommendations is based on finding users similar to each other in terms of interactions with products. Thus, *objective data* is composed of the behaviors of shoppers. The categories, titles and descriptions of the products form *metadata*, which is usually the result of the joint work of many employees of the e-store. For rolling steel factories, predictive maintenance models are derived mainly from *objective sensory data*, such as temperature, force, etc. Factory accounting data are *metadata*.

The more *objective* the data is, the more it is suited for modeling the phenomena, be it physical, business, sociological or psychological in nature. *Metadata* is more suitable for explaining the model to the user, convincing her or him, and prompting to make decisions and actions based on this knowledge. It is more prone to error because of its conventionality and subjectivity, but they speak to humans.

In this work we propose a method that allows for clustering dataset with *objective data*, and explain

IJCAI-ECAI 2022, the 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence, July 23–29, 2022 Messe Wien, Vienna, Austria

EMAIL: m.mozolewski@doctoral.uj.edu.pl (M. Mozolewski); samaneh.jamshidi@hh.se (S. Jamshidi); szymon.bobek@uj.edu.pl (S. Bobek); grzegorz.j.nalepa@uj.edu.pl (G. J. Nalepa)
URL: https://github.com/mozo64 (M. Mozolewski)
ORCID: 0000-0003-4227-3894 (M. Mozolewski); 0000-0001-7055-2706 (S. Jamshidi); 0000-0002-6350-8405 (S. Bobek); 0000-0002-8182-4225 (G. J. Nalepa)

CEUR Workshop Proceedings (CEUR-WS.org)

differences between clusters with *metadata*. We use XAI methods to explain differences between clusters using *metadata* which is perfectly understandable by humans, but may not be of sufficient quality to perform valid clustering. The selection of the most interpretable *metadata* is iterative and human-guided. In our example, we show how image-based clustering can be enhanced with textual description of clusters. We argue that such an approach can lead to better utilisation of *metadata* for cluster analysis purposes, which results in better understanding of clusters which is the final goal of every clustering task. Furthermore, it allows for checking the consistency between two or more possible instance representations (image and text) which might be crucial in domains that rely on both (e.g. e-commerce).

The rest of the paper is organized as follows. In Section 2 we present current research in the area of interactive clustering and human-guided cluster analysis. The description of our method along with use-case studies is given in Section 3. Finally, we conclude our work and show perspectives of its further development is presented in Section 4.

2. Related works

In the survey on interactive clustering [**?**] authors have distinguished 2 groups of approaches in terms of how interaction occurs. In general, users can interact indirectly with the tool, by changing the parameters of the algorithm, or directly by giving feedback to the result of clustering. The parameters adjusted most frequently are the number of clusters and the similarity threshold [**?** **?**]. For topic modeling, users are given the option to select keywords and set their relative importance [**?**]. On the other hand, direct feedback might be realized by highlighting incorrect instances of splitting or merging the resulting clusters [**?**]. For textual data, users can provide the tool with blacklists of incorrect topic labels or set similar restrictions [**?**]. Based on this signal, the clustering tool learns user preferences and tries to incorporate the knowledge in the next iteration.

Explainable AI approaches have become particularly important, and although most work is generally focused on supervised learning, some works have been done to explain clusters. One of the most common methods for understanding clustering methods is visualization. By using low-dimensionality embedding and displaying them in two- or three-dimensional dimensions, one can get an overview of the clusters and their data. However, these visualizations are not always understandable and explainable.

The decision tree is one of the inherently interpretable algorithms. So one common way to explain models is to use decision trees. Nevertheless, the critical point for explaining the decision tree is its depth because decision trees with high depth no longer are interpreted, so we must pay attention to the depth of the tree produced. Using a small decision tree to divide a dataset into k clusters provides explainable clusters, but this approach has a trade-off between being explainable and accuracy. IMM algorithm [**?**] approximates k-means and k-median clustering by a threshold tree with k leaves. ExKMC [**?**] uses a threshold tree to provide an explainable k-mean clustering in which the number of tree leaves is greater than the number of clusters.

Besides, visualization or providing some conditions on features, using text data is reasonable to generate explanations to users. In [**?**] authors use captions of the images along with the images to create a more discriminative classification. In addition, they use this *metadata* to provide language explanation and generate a text description for each class. However, by blending textual and image modalities into one datset, authors limit the possibility of checking consistency between these two types of data and implicitly assume the correctness of possibly wrong image descriptions.

Similarly, in many other methods that aim at explaining differences between discovered clusters, the clustering task is transformed to classification one, and the classifier is then explained with available XAI methods such as LIME [**?**], Anchor [**?**], LUX [**?**], etc. One of the most recent implementations of such approach can be found in [**?**].

Another approach is given in [**?**], where authors present a toolkit for conformance checking between expert knowledge and automatic clustering. The differences between expert-based clustering and automated clustering are justified with XAI methods and the process is iterative. However, the explanations are not human-guided, and the expert has no impact on the way they are generated. In particular, it is not possible to provide additional *metadata* for explanations, nor modify the set of concepts that are used for explanations.

In all the cases the process is not iterative, nor human-guided. Finally, to the best of authors' knowledge, neither of the approaches known in the literature divides data into *objective* part with a good quality for cluster algorithms, but poor explanation capabilities and *meta-data* with possibly worse potential as clustering features, but better

explanation capabilities and possible inconsistencies with *objective* data that should be fixed. Addressing these issues was the primary motivation of our work that will be described in more detail in the following sections.

3. Cluster analysis with *metadata*

In this section, we will show how our method could be applied to real-case scenarios. We choose an example from the e-commerce field because the authors have experience working in this industry. Specifically, we work with online stores to provide them, among others, with recommendations of products to their end-users (clients).

In real-life scenarios, data about products are stored in product catalogs in shop databases, and most often exchanged with so-called product feeds (XML documents). We used a public dataset from Kaggle[1]. This dataset in terms of content resembles a product feed for an online store of a medium size product catalog. It consisted of 44000 products with category labels, titles, and images. For the code accompanying this example see GitHub repository[2].

As has been said before, we treat images as *objective data*. We used embeddings of images obtained via MobileNetV2 [**?**] as an input to clustering pipeline. The fully-connected layer at the top of the network was disregarded because we were not interested in the classification done by the model. The output of the final layer of the model was of length 20480. We used Singular Value Decomposition (SVD) with normalization to reduce the dimensionality of embeddings, leaving at least 90 percent of the variance.

In this section, we will present tools dedicated to data scientists who would like to perform clustering. We propose a 2-step clustering loop, which consists of k-means clustering and textual explanations of clusters. Data preparation also could be performed more than once, if needed. For the sake of simplicity, we call it "step 0" in this work.

3.1. Data preparation

The method requires 2 types of data: *objective* and *metadata* as defined in the previous section. In "step 0" method provides users with helper functions to prepare both types of data. For *objective data* there is a function that performs a reduction of dimensionality via SVD followed by normalization. It works on any numerical data, which could be as well as one-hot variables and continuous real values (floats). User sets the percentage of explained variance left after SVD reduction. The optimal count of new dimensions could be determined automatically by our algorithm. This is done by probing different dimension counts with *scipy.optimize* package, so the user does not need to do this manually. Regarding *metadata* which is textual, there are wrappers built on top of the SpaCy[3] and NLTK[4] libraries. Users can contact text columns, lemmatize, remove stopwords and perform Tf-Idf vectorization. For numerical *metadata*, we found a way to incorporate them into textual explanations. For instance, the year could be re-coded as the label "year2022", which will be easily interpreted along the pipeline. Other numerical variables could be re-coded to low/medium/high bins, based on quartiles. Finally, the user constructs the "Pipeline" object and initializes it with 2 datasets: *objective* and *metadata*.

3.2. Assistance in clustering

The first step corresponds to running the unsupervised clustering algorithm. Typically, the person performing the analysis starts with the dilemma of choosing the number of clusters. It can be resolved with her/his background knowledge, intuition, practicality prerequisites, or just a trial and error approach. To give our users a hint in this regard, we use the T-SNE [**?**] 2-dimensional projection of the data. At the moment, this is a solely visual clue. It is depicted in Figure 1. If data have an underlying structure, points representing observations will cluster, which would be observed on the chart. As T-SNE on massive data could be resource intensive, the default is to run this process on random subsample and cache results. Additionally, users can apply textual labels to the T-SNE chart, plotted on a subsample of data, to avoid cluttering the chart. Labels could represent the most important pieces of *metadata*, such as the label, the observation id, and summary of description. The next clue is derived from the silhouette score on a plot in the Figure 2. The range of the number of clusters to be tested is provided in accordance with the previous clue. To speed-up computations, this plot could be obtained on a random subset, and results are cached for further reference. For now, the user interprets the plot on her/his own. Finally, clustering with k-means is performed on all observations. Visualization with

[1] See: https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small

[2] See: https://github.com/mozo64/xai-survey/blob/sklearn-text-clustering-example/src/example1-clustering-products-fashion.ipynb

[3] See: https://spacy.io/

[4] See: https://www.nltk.org/

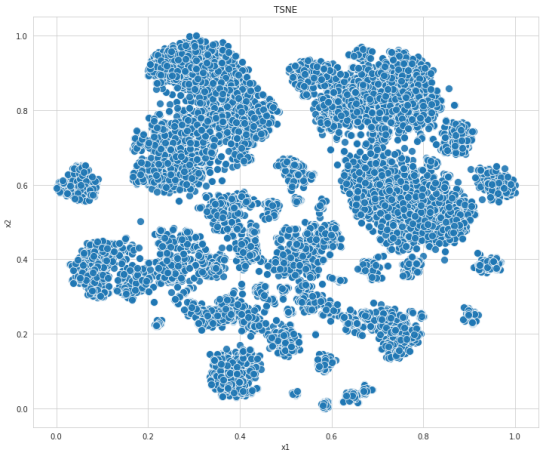

Figure 1: Preliminary visualisation of *objective data* in 2-D projection with T-SNE dimensionality reduciton.

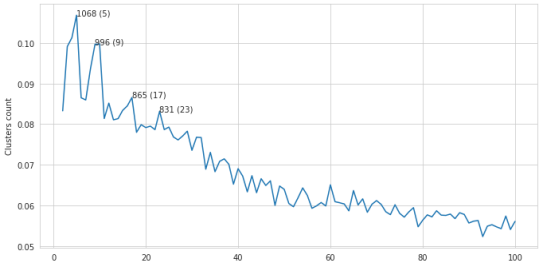

Figure 2: Silhouette score with cluster counts and values

T-SNE is presented, this time with clusters colored in different colors, which is depicted in Figure 3.

3.3. Interactive explanations

The second step is to explain the clusters so that the person performing the data analysis can assess the result. We would like to give users the freedom to refine explanations. Thus, we provide her or him with the possibility to influence explanations by extending stopwords with his own terms. On the other hand, we initialize the whitelist with keywords like "year2022", defined in "step 0". Then we use the Tf-Idf vectorizer, taking into account the aforementioned lists. Vectors are used to train decision tree classifiers. The size of the list of additional terms is under the control of a user. She or he can change it and interactively observe the changes in a Figure 6. Furthermore, example observations are presented in Figure 4, word clouds that describe

Figure 3: Preliminary visualisation of *objective data* in 2-D projection with T-SNE dimensionality reduciton. Colors denote clusters discovered with a usage of *objective* data.

Figure 4: Example images of products form the *objective* data that were assigned to the same cluster.

Figure 5: Word cloud for a category of products presented in Fig. 4 generated with *meta-data*.

clusters in Figure 5 and LIME [**?**] explanation for one instance of *metadata* in Figure 7.

The last stage is a plot of the word cloud of each cluster, using the same Tf-Idf vectorizer. Plots are accompanied by examples of observations.In addition, a user is presented an explanation generated with LIME for a random instance from a given class. The whole process is iterative, and the expert decides on its convergence.

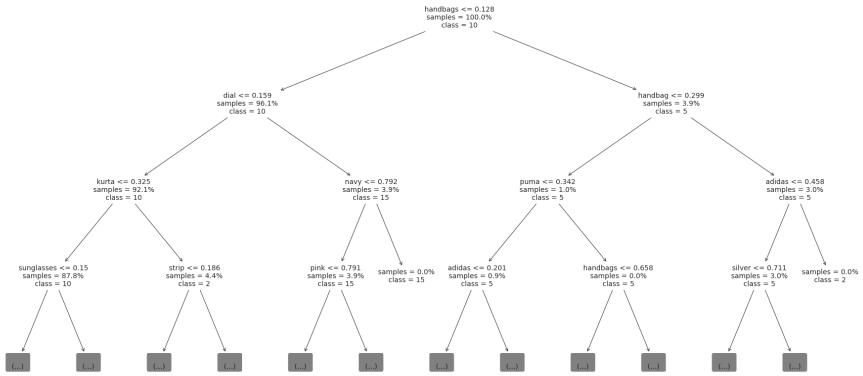

Figure 6: Decision tree classifier which explains how clusters differ in terms of *metadata*.

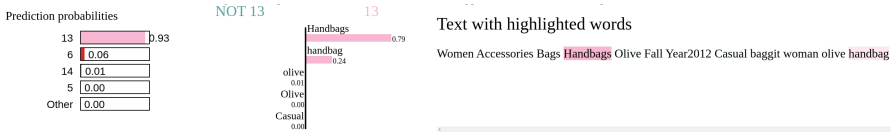

Figure 7: LIME explanation for observation in category "handbags".

4. Summary

In this work, we presented the method that allows for explaining clusters with concepts that could be more human-readable than the data which was used as an input to clustering algorithm. We based our method on the observation that different types of data are suitable in different degrees to clustering and explaining tasks. We demonstrated the feasibility of our approach on the e-commerce example, where images were treated as input for clustering and textual descriptions of images as basis for cluster descriptions.

In future work we would like to improve our method with several extensions. We will focus on automatically proposing number of clusters based on both embedding features with methods similar to T-SNE and metrics like silhouette score. We want to test clustering techniques other than k-means. For instance, hierarchical clustering could be more suitable in e-commerce, where taxonomies of products are multilayer. Word clouds could be replaced with topic analysis with Latent Dirichlet Allocation or techniques derived from Natural Language Generation. Another interesting direction is to construct explanations with other modalities, like visual, by something more sophisticated than presenting example images. It could be done for instance with image captioning.

Acknowledgments

The work of Szymon Bobek has been additionally supported by a HuLCKA grant from the Priority Research Area (Digiworld) under the Strategic Programme Excellence Initiative at the Jagiellonian University (U1U/P06/NO/02.16).

The work of Samaneh Jamshidi was supported by CHIST-ERA grant CHIST-ERA-19-XAI-012 funded by Swedish Research Council.