Understand your clusters: a link between the clustering data and explanatory meta-data

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

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

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

Abstract

In this preliminary work, we present an approach for clustering augmented with Natural Language Explanations. With clustering there are 2 main challenges: a) choice of a proper, reasonable number of clusters and b) cluster profiling. There is a plethora of technics for a) but not so much for b), which is in the general a laborious task of explaining obtained clusters. Clustering is in a sense art in that regard that it is an intuitive and iterative process. Therefore, XAI techniques are well suited in this area. 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, meta-data, SVD, T-SNE, Natural Language Understanding, Natural Language Processing, explanations, narratives

1. Introduction

Categorization is one of the ways how humans describe the world. To classify means to notice that some phenomena differ from each other. And more importantly how they differ. Finally, one is giving names to those different classes of entities. Clustering in machine learning in essence is a no different process.

Clustering methods are standard in many fields of human prosperity. In the advent of an ever-increasing amount of data, we use tools to automate the process, as manual clustering is too laborious. Indeed, there are many statistical, machine learning, and deep learning algorithms. The difficulty arises to convince end-users that derived clusters make sense. Will it be clustering sensor data in Industry 4.0 or behaviors of consumers, results need to be actionable. We believe that this is impossible if we do not explain what the algorithm has learned. It also follows that the process is iterative. We continually hypothesize about the significant differences between classes, then test results by looking at classified objects.

From our expertise in industry 4.0 and e-commerce, we often see distinctions between 2 types of data. There are "objective" data and the "subjective" data or "meta-data". For instance in e-commerce popular approach for recommendations is Collaborative Filtering. It is based on finding users similar to each other in terms of interactions with products. Thus, "objective" data are shoppers' behaviors. Categories, titles, and descriptions of products are "meta-data", which are usually the result of the joint work of many e-store employees. For rolling steel factories, predictive maintenance models are derived mainly from "objective" sensory data, like temperature, force, etc. Factory accounting data are "meta-data". 6-sigma quality standards lay somewhere in between of "objective-subjective" continuum.

This leads to the conclusion: The more "objective" data are, the more they are suited for modeling the phenomena, be it physical, business, sociological or psychological in nature. "Meta-data" are more suited for explaining the model to the user, convincing her or him, and prompt to make decisions and actions based on this knowledge. They are more prone to error, because of their conventionality and subjectivism, but they speak to humans. As a closing remark in this section, we want to pinpoint that humans seek agency. Clustering algorithms are unsupervised. XAI methods are well-fitted for the job of giving control, both in the clustering stage and profiling stage.

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

2. Related works

3. Explanations with textual meta-data

In this part, we will show how our framework 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 (https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small). 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.

Our workflow was based on this Kaggle notebook: https://www.kaggle.com/code/shubhijoshi/similar-image-finder-using-k-means/notebook, and we adapted it to our needs. As been said before, we treat images as "objective" data. We used embeddings of images obtained via MobileNetV2 (https://arxiv.org/pdf/1801.04381v4.pdf). 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 "meta-data" as defined in the previous section. In "step 0" framework 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 percent 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. As for "meta-data" which are textual, there are wrappers built on top of SpaCy and NLTK libraries. Users can contact text columns, lemmatize, remove stopwords and perform Tf-Idf vectorization. For numerical "meta-data", we found a way to incorporate them into textual explanations. For instance, the year could be recoded as the label "year2022", which will be easily interpreted along the pipeline. Other numerical variables could be recoded to low/medium/high bins, based on quartiles. Finally, the user constructs the "Pipeline" object and initializes it with 2 datasets: "objective" and "meta-data".

3.2. Assitance in clustering

The first step corresponds to running the unsupervised algorithm. Typically, the person who performs the analysis starts with the dilemma of choice of 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. See **??**. If data have an underlying structure, points representing

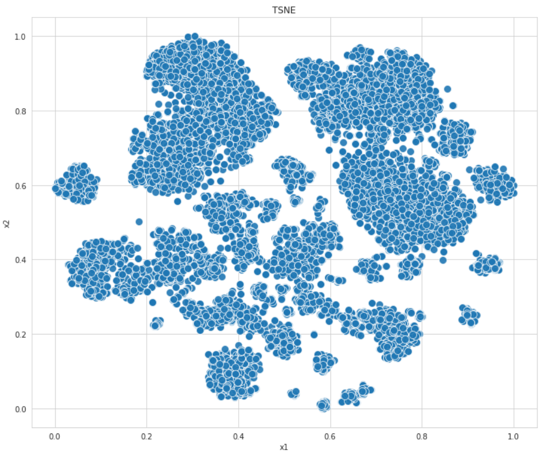


Figure 1: Preliminary visualisation of "objective" data in 2-D projection

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 "meta-data", like the label, observation id, and summary of description. The next clue is derived from the plot of the silhouette score **??**. 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 T-SNE is presented, this time with clusters colored different colors, which is depicted on **??**.

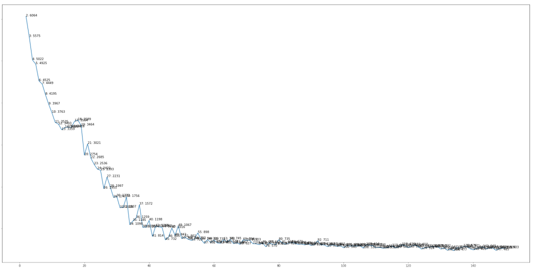


Figure 2: Silhouette score with cluster counts and values

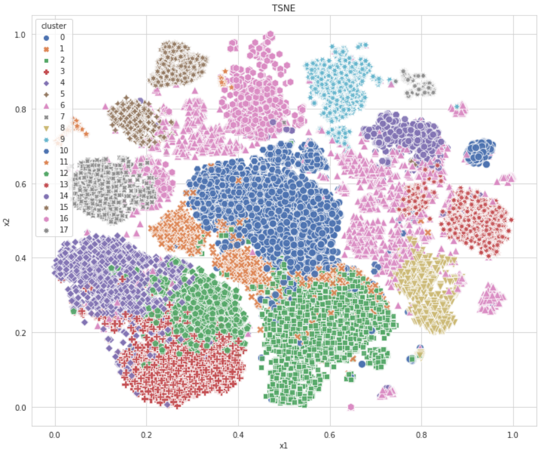


Figure 3: Preliminary visualisation of "objective" data in 2-D projection

3.3. Interactive explanations

The second step is to explain clusters, so the person who performs data analysis can assess the result. We would like to give users agency in refining 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 for training 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 plot **??**. 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 if data scientists should define visualization function and pass into Pipeline class API. Moreover, there is random observation with the LIME explainer for a given class.

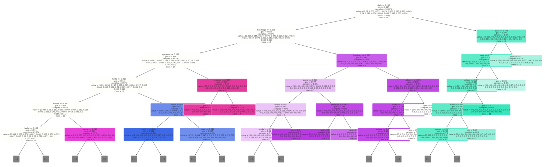


Figure 4: Decision tree classifier which explains how clusters differ in terms of metadata

Figure 5: Example products

4. Summary

We believe that the XAI methods are most useful when they support human work. The main purpose of this paper is to show that XAI can be put into practice when it is based on 2 foundations. The first is the need for a dialogue between the machine learning system and its user, taking into account the needs and predispositions of the latter. The second is the observation that different classes of data are helpful to vary degrees in grouping objects themselves and in providing explanations. Our approach

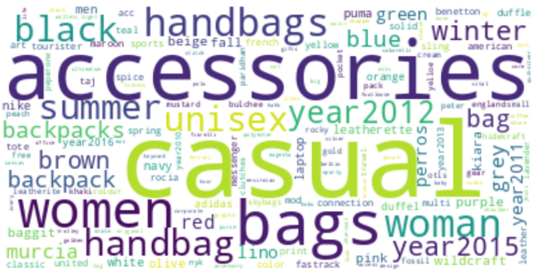
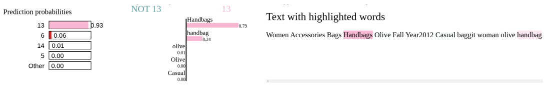


Figure 6: Word cloud for a given category



Women Accessories Bags Handbags Olive Fall Year2012 Casual baggit woman olive handbag

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

is obviously not new, but it is worth emphasizing if one wants to create useful solutions for machine learning practitioners. We have developed a pipeline that relates to the cluster analysis method, as this technique is widely used with a broad spectrum of applications in science, industry, business, and marketing. Furthermore, it is easy to show that it is actually a technical extension of the natural process of describing reality. It is also clearly an iterative process, so it is a form of a dialogue between humans and algorithms, and understanding is even more important than, for example, in regression techniques.

In future work we would like to improve our framework with several extensions. We will focus on automatically proposing number of clusters based both on embedding features with technics similar to T-SNE and metrics like silhouette score. We want to test other techniques clustering than k-means. For instance hierarchical clustering could more suited 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.