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

Maciej T. Mozolewski[1], Samaneh Jamshidi[2] and Szymon Bobek[1]

[1] *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 clustering augmented with Natural Language Explanations. With clustering there are 2 main challenges: a) choice of 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 general laborious task of explaining obtained clusters. Clustering is in a sense art in that regard that it is intuitive and iterative process. Therefore, XAI techniques are well suited in this area. On a convincing example we show how 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, NLU,

1. Introduction

Categorization is a one of ways how humans describe the world. To classify means to notice that some phenomenas 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 an essence is no different process.

Clustering methods are common 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-user that derived clusters make sense. Will it be clustering sensors data in Industry 4.0 or behaviours of consumers, results needs to be actionable. We believe that this is not possible if we do not explain what 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 distinction 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 with Collaborative Filtering. It is based on finding users similar to each other's in terms of interactions with products. Thus, "objective" data are shoppers behaviours. 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 factory, predictive maintenance models are derived mainly from "objective" sensory data, like temperature, force etc. Factory accounting data are "meta-data". 6-sigma quality standards lays somewhere in between of "objective-subjective" continuum.

This leads to the conclusion: The more "objective" data are, the more they are suited for modelling the phenomena, being it physical, business, sociological or psychological in nature. "Meta-data" are more suited for explaining the model to 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 human. 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 on 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 scenario. We choose example from e-commerce filed because authors have experience of 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 shops databases, and most often exchanged with so-called product feeds (xml documents). We used 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 online store of 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). Fully-connected layer at the top of the network was disregarded, because we were not interested in 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 dimensionality of embeddings, leaving at least 90 percent of variance.

In this section we will present tool dedicated for data scientists who would like to perform clustering. We propose 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 previous section. In "step 0" framework provides user with helper functions to prepare both types of data. For "objective" data there is function which performs 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. 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 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. User can concat text columns, lemmatize, remove stopwords and perform Tf-Idf vectorisation. For numerical "meta-data", we found a way to incorporate them in textual explanations. For instance year could be recoded as 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, user constructs "Pipeline" object and initialise it with 2 datasets: "objective" and "meta-data".

3.2. Assitance in clustering

First step corresponds to running unsupervised algorithm. Typically, person who performs the analysis starts with dilemma of choice of the number of clusters. It can be resolved with her/his background knowledge, intuition, practicality prerequisites or just trial and error approach. To give our users hint in this regard, we use T-SNE 2-dimensional projection of the data. At the moment, this is solely visual clue. See **??**. If data have underlying structure, points representing observations will cluster, which would be observed on the chart. As T-SNE on massive data could be resources intensive, the

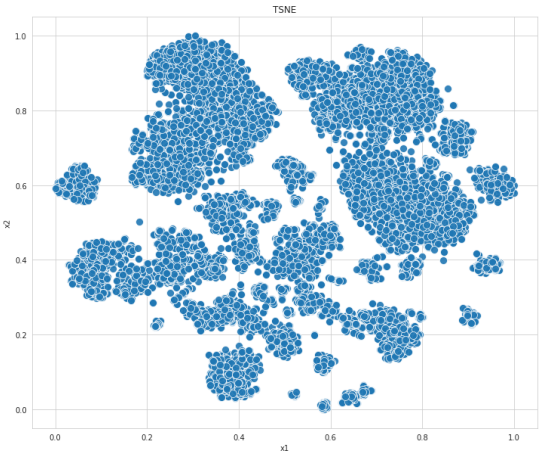

Figure 1: Preliminary visualisation of "objective" data in 2-D projection

default is to run this process on random subsample and cache results. Additionally, user can apply textual labels to T-SNE chart, plotted on subsample of data, to avoid cluttering the chart. Labels could represent most important pieces of "meta-data", like label, observation id, summary of description. Next clue is derived from plot of silhouette score **??**. Range of number of clusters to be tested is provided

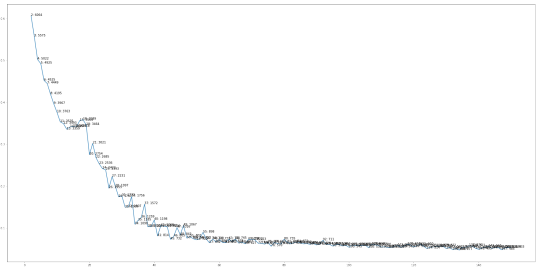

Figure 2: Silhouette score with cluster counts and values

in accordance to previous clue. To speed-up computations, this plot could be obtained on random subset and results are cached for further reference. For now, user interprets the plot on her/his own. Finally, clustering with k-means is performed on all observations. Visualisation with T-SNE is presented, this time with clusters colored with different colors, which is depicted on **??**.

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

3.3. Interactive explanations

Second step is to explain clusters, so person who performs data analysis can assess the result. We would like to give user agency in refining explanations. Thus, we provide her or him with possibility to influence explanations with extending stopwords with his own terms. On the other hand, we initialise whitelist with keywords like "year2022", defined "step 0". Then we use Tf-Idf vectoriser, taking into account aforementioned lists. Vectors are used for training decision tree classifier. The size of list of additional terms is under control of a user. She or he can change it and interactively observe the changes on a plot **??**. The last stage is plot of

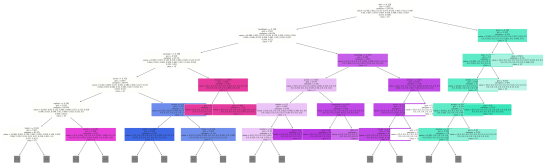

Figure 4: Decision tree classifier which explains how clusters differ in terms of metadata

world cloud of each cluster, using the same Tf-Idf vectoriser. Plots are accompanied by examples of observations, if data scientist should define visualisation function and pass into Pipeline class API. Moreover, there is random observation with LIME explainer fo a given class.

Figure 5: Example products

Figure 6: Word cloud for a given category

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

4. Summary