DOI: 10.26485/978-83-60655-92-4/7 ISBN 978-83-60655-92-4

(HYPER)COMPLEX SEMINAR 2021
IN MEMORIAM OF
PROF. JULIAN ŁAWRYNOWICZ

Radosław A. Kycia, Agnieszka Niemczynowicz

CLUSTERIZATION IN ACTION - EDUCTIONAL REPORT

Abstract

This review paper summarizes the talk about the application of clustering algorithms in discovering underlying relations in the data.

Keywords and phrases: clustering; unsupervised learning; k-Means, Principal Component Analysis; Gaussian Mixture Model
Subject classification: [2000]

1 Introduction

Today's situation with abundant data gives an unprecedented opportunity to Machine Learning (ML) methods to discover some relations in these data that are usually invisible due to many factors. One of these factors is the large dimensionality of the data. The other factor can be an unsuitable frame of reference where the data are initially placed.

The classical Machine Learning methods splits into two main cases depending on the structure of the data:

- supervised learning,
- unsupervised learning.

The data can be imagined as a table with columns that describe some features of data objects and possibly some additional columns that describe classes. If the task is to construct a model that predicts classes for new records (rows) of data basis on the previous data with classes, then it falls into supervised learning. On the contrary, if the learning process aims to discover classes from the data without them, it is an example of unsupervised learning.

In this paper we present an ML pipeline that starts with raw data and helps to reason about the possible existence of classes for the data. These classes are visible as a group called clusters in some space to which data were transformed. Modern

Machine Learning methods allow the construction of many such transformations and extract cluster data. However, in this paper we describe the one way that was beneficially applied to our analyses of various phenomena [1, 2]

The paper is organized as follows: In the next section we describe the ML method used to analyze data, and the following section summarizes applications.

2 The method

There is a more or less precise way to proceed from raw data [3]. The first step is to clean data, which means resolving all issues with incomplete records, removing outliers if present, and converting the data to formats that can be easily processed. This step heavily depends on the phenomena which the data describe.

The next step for multidimensional data is to reduce their dimensionality without scarifying the information they carry by transforming it into a new space. One of the standard methods for such transformation is PCA (Principal Component Analysis) [3], which tries to perform a linear transformation on the initial dataset to the space in which axes are eigenvectors of the covariance matrix. Intuitively, this is based on the observation that the data/signal with small variance/change carries less information than the heavily varying one. The criterion for selecting the number of dimensions of such new space is the ratio of explained variance, which is carried to the new space by such truncation of dimensionality of the data. Typically, before PCA, some data standardization is performed to make all features distributed in the same range and with the same standard deviation.

Some analysis can be performed on transformed data to check if they tend to cluster around some centers. These clusters usually reflect some relations between data that are invisible before the initial scaling and transformation. There are many approaches to clustering. One of the standards is the k-Means algorithm [3]. It tries to associate points to the clusters by minimizing the variation within the clusters. The typical input for the algorithm is the desired number of clusters. In our approach, we used k-Means with a varying number of clusters to obtain the plot of SSE (within-clusters-sum-of-squares) vs. the number of clusters to identify the elbow point - the number of clusters at which increasing this number does not dramatically decrease SSE. This rule of thumb usually works well, and lower-dimensional data can be visually confirmed.

k-Means algorithm provides some clustering information; however, the good practice is to cross-check this result by performing different clustering algorithms on the transformed data. In our application we selected GMM (Gaussian Mixture Model) [3] to fit the data to the linear combination of Gaussians with parameters resulting from the fit. This model is better for interpretation since when clusters overlap, it provides information about the input of a border point to a given gaussian cluster. As an initial input of GMM we used the information about clusters from the k-Means algorithm, which is also typical practice in ML.

The pipeline looks as follows:

1. Clean data

2. Standarize data
3. Perform PCA
4. Determine optimal number of clusters using k-Means algorithm
5. Determine precise clustering info using GMM

Using this approach, we were able to distinguish usually invisible structures in various datasets that will be described in the following sections.

3 Generation Z

The first dataset is the questionnaire related to the motivation of Generation Z representatives at their workplace. Such data are essential nowadays since the GenZ representatives enter the market, and managers need to craft motivational systems to bind GenZ with the company and do their work as efficiently as possible.

In our research [2] using the aforementioned ML pipeline, some unexpected relation between the relation with boss, work-life balance, and engagement at work becomes visible. These relations were used to construct recommendations for managers.

4 Grape Oils

The second data set where the above ML pipeline proved to be helpful in the analysis of the parameters of oils obtained from grapes. The grapes were tagged with the names of the vines species produced from them. The analysis was restricted to the Czech Republic.

The clustering analysis indicated some similarities in specific grape seeds parameters [1]. These clusters can be used to control the variation of a given batch from the nominal ones. As a byproduct, these similarities can be used to check the similarity in the flavor of selected Czech vines - the information which is currently under verification.

5 Summary

Advanced ML methods allow distinguishing in the datasets relations which are invisible using simple transformations. Therefore they should be included in the standard toolset of any physicists, statisticians, or quality analysts.

Acknowledgements

RK work was supported by GACR grant number GA22-00091S, MUNI grant number MUNI/A/1092/2021, and Ministry of Education, Youth and Sports of the CR grant number 8J20DE004.

References

[1] Vladimír, M., Matwijczuk, A.P., Niemczynowicz, A. et al. *Chemometric approach to characterization of the selected grape seed oils based on their fatty acids composition and FTIR spectroscopy*, Sci Rep 11, 19256 (2021). https://doi.org/10.1038/s41598-021-98763-6

[2] Kycia, R.A., Niemczynowicz, A. Nieżurawska-Zajac, J. . *Towards the global vision of engagement of Generation Z at the workplace: Mathematical modeling*, 37th International Business Information Management Association Conference (IBIMA), pp. 6084-6095 (2021); ISBN: 978-0-9998551-6-4

[3] Raschka, S., Mirjalili, V., *Python Machine Learning - Second Edition: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*, Packt Publishing; 2nd edition 2017

Presented by Radosław A. Kycia, online,
during the Hypercomplex Seminar, Nov. 12th, 2021

Radosław A. Kycia, Department of Mathematics and Statistics, Masaryk Univeristy, The Czech Republic; Faculty of Computer Science and Telecommunications, Cracow University of Technology, Poland

Agnieszka Niemczynowicz, Faculty of Mathematics and Computer Science, University of Warmia and Mazury in Olsztyn, Poland

Klasteryzacja w akcji

S t r e s z c z e n i e

Ten artykuł przeglądowy podsumowuje dyskusję na temat zastosowania algorytmów klastrowania w odkrywaniu podstawowych relacji w danych.

Słowa kluczowe: klastrowanie; uczenie się nie-nadzorowane; k-średnie, analiza głównych składowych; Model mieszany Gaussa