DICE: Density-based Interactive Clustering and Exploration

Daniyal Kazempour[1] Maksim Kazakov Peer Kröger[1] Thomas Seidl[1]

Abstract: Clustering algorithms are mostly following the pipeline to provide input data, and hyperparameter values. Then the algorithms are executed and the output files are generated or visualized. We provide in our work an early prototype of an interactive density-based clustering tool named DICE in which the users can change the hyperparameter settings and immediately observe the resulting clusters. Further the users can browse through each of the single detected clusters and get statistics regarding as well as a convex hull profile for each cluster. Further DICE keeps track of the chosen settings, enabling the user to review which hyperparameter values have been previously chosen. DICE can not only be used in scientific context of analyzing data, but also in didactic settings in which students can learn in an exploratory fashion how a density-based clustering algorithm like e.g. DBSCAN behaves.

Keywords: Density-based clustering. Interactive. Exploration. Hyperparameters.

1 Introduction

The density based clustering algorithm DBSCAN [Es96] enables the detection of arbitrarily shaped clusters which are density connected. The density is regulated through two hyperparameters, one which represents the neighborhood range of data points, denoted with ε-range and another one which states how many points have to be located within this ε-range, namely $minPts$. With both of the hyperparameters the users can control how dense the clusters shall be which are detected. However, finding a good hyperparameter setting is a tedious task. With a classic DBSCAN implementation one would re-run the algorithm with different parameters, visualize the output and repeat the whole procedure. It may be facilitated by utilizing a method like OPTICS [An99] which generates a plot making it easier to identify suitable hyperparameter settings. A data mining framework like e.g. ELKI [Sc15] provides besides a rich selection of data mining algorithms and index structures, and a high scale of configurability, visualizations and additional information to the discovered clusters. With DICE we aim to deliver a system which gives *immediate* feedback on which the users can intervene, based on provided information, inspecting like e.g. single clusters, or keeping track of changes made. Such interactive tools have been developed like e.g. VISA [As07] in context of subspace clustering. In one of our previous works [Ka18] we presented an interactive tool which makes it possible not only to set one hyperparameter setting for

[1] Ludwig-Maximilians-Universität München, Lehrstuhl für Datenbanksystem und Data Mining, Oettingenstraße 67, 80538, München, {kazempour,kroeger,seidl}@dbs.ifi.lmu.de

doi:10.18420/btw2019-42

the clustering algorithm in the beginning but also set different parameter values at different iteration steps, going back and forth, exploring different outcomes. We aim to target in this work with DICE the interactive clustering in context of density based methods. Our major contributions are: (1) An interactive version of a density-based clustering method, (2) Inspecting single clusters getting statistics and additional information such as convex hulls of a clusters and (3) Keeping track of hyperparameter changes made.

2 DICE - Interaction scheme

DICE abides to the following interaction scheme as seen in Figure 1. First data and initial ε and $minPts$ parameters are provided and the algorithm is executed. To provide a glimpse of what happens behind the interactive user interface, in a initial step all pairwise distances among all points in the data set are computed. This follows the YOCDO-principle: You Only Compute Distances Once. Based on this distance matrix, filters are applied returning a new matrix where all pairwise entries which have a distance above ε threshold are set to zero, while those being below are set to one, yielding an adjacency matrix. On this matrix the connected components are computed. The users see on the main screen the clusters and outliers with the initially provided parameters. As a reaction the users can now change either one or both of the parameters and are provided with an immediate visualization. This visualization-(re)action cycle can be pursued indefinitely.

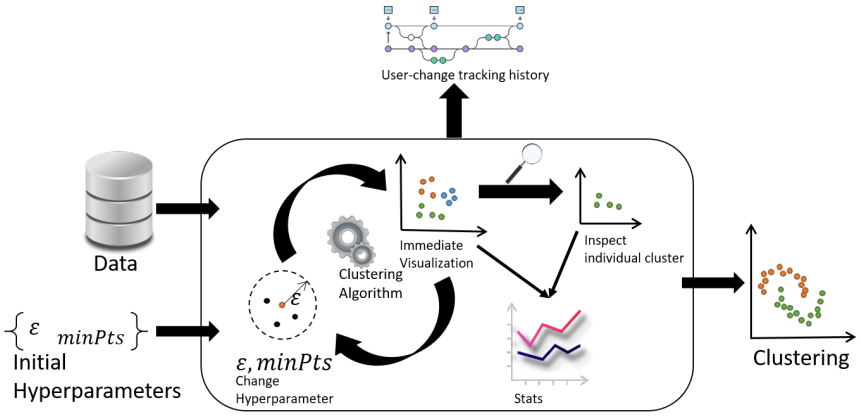

Fig. 1: Interaction scheme of DICE.

On this occasion it may happen that the users loose track of which hyperparameter settings have been explored so far. For this purpose track-records are provided enabling the users to see which e.g. ε and which $minPts$ parameters have been chosen in the previous steps, further it is shown how many clusters emerged at which parameter settings. If the users wish, they can slide through each of the detected clusters being displayed in the main window separately. Here histograms are provided, one which represents on the vertical axis the number of data points and on the horizontal axis, specific distances to other data points. The intention is to show users the distribution of pairwise distances within a cluster. In a much simpler fashion (compared to e.g. OPTICS) it enables to detect further clusters within a cluster, by looking for hills and valleys within the histogram. Besides zooming into

single areas within a cluster, which is provided natively by matplotlib, we thought of another concept to focus on a cluster. In the single-cluster view, users can apply “convex hull“ and are provided first with a plot showing all layers of convex hulls of the inspected cluster. Then, by each time clicking on that button, one convex hull and the points belonging to it is removed from the view of that cluster, enabling users to inspect a cluster layer-by-layer. The entire described process scheme can be repeated until a desired clustering of the data is found.

3 Demonstration outline

In this section we elaborate on the “screenplay“ of the demo. We first start DICE with a given data set and initial hyperparameter settings. This will open the main window first as seen in Figure 2 (center). Since DICE is an interactive tool, we will make the demo session interactive and strongly encourage the attendants to interact with the tool or to suggest e.g. which hyperparameters to choose. Since we may try out many different hyperparameter settings we will highlight the issue of loosing-track which hyperparameters have been chosen. In a second step we reveal the hyperparameter tracking capabilities of DICE as seen in Figure 2 (left and right).

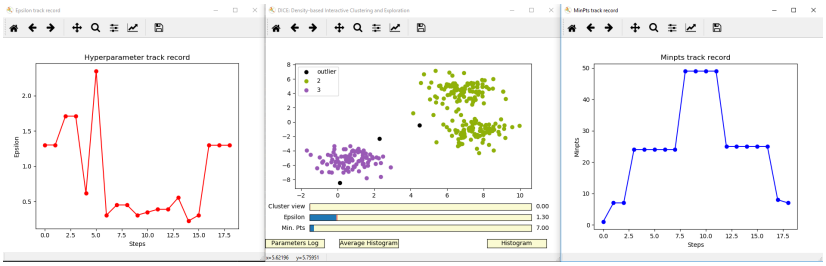

Fig. 2: Main window of DICE (center) with the track record for different ε hyperparameter settings (left) and different $minPts$ hyperparameter settings (right).

In a third step we go deeper into the analysis process by inspecting single clusters more in detail as it can be seen in Figure 3 (left). Choosing a cluster, we inspect it by first looking at the clusters histogram highlighting which information can be obtained (Figure 3 (right). Further we show that even on single cluster level, we can change the hyperparameters to observe in how far they affect the currently observed cluster. In the fourth and last step we demonstrate how DICE determines the convex hull layers being represented in a separate plot (Figure 3 (center)). By clicking repeatedly on the “Convex Hull“ button the participants can observe how like the layers of an onion the convex hulls are removed.

4 Concluding remarks and Outlook

Being aware that DICE is in an early prototype stage, it provides various points of interaction and visualization. DICE enables to inspect single clusters being supported with histograms

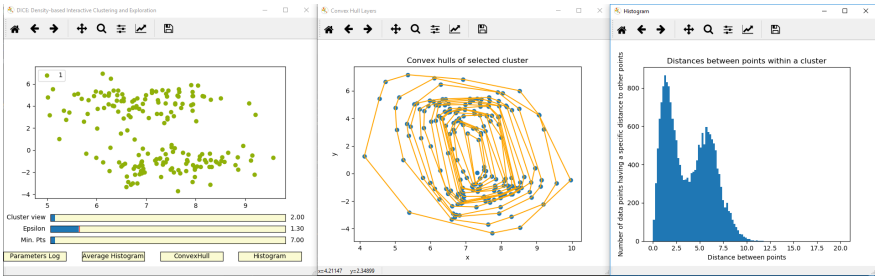

Fig. 3: Visualization of a single cluster (left) with its convex hull layer plot (center) and with a distance plot (right).

on frequencies of pairwise distances. In its current stage it is not suitable for high-volume or high-dimensional data sets. For such further research on the computation schemes are necessary regarding the high-volume aspect. For the high-dimensionality aspect a careful choice of visualization techniques and histograms is required. Beyond the two mentioned aspects different approaches such as minimum spanning trees (MST) for each cluster may be added, since they are (compared to convex hulls) applicable in high-dimensional data. In context of didactic means, DICE is a tool which enables students to analyze data as well as to learn and understand how density based methods work.

Acknowledgement

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content.

References

[An99] Ankerst, M.; Breunig, M. M.; Kriegel, H.-P.; Sander, J.: OPTICS: Ordering Points to Identify the Clustering Structure. SIGMOD Rec. 28/2, pp. 49–60, June 1999, ISSN: 0163-5808.

[As07] Assent, I.; Krieger, R.; Müller, E.; Seidl, T.: VISA: Visual Subspace Clustering Analysis. SIGKDD Explor. Newsl. 9/2, pp. 5–12, Dec. 2007, ISSN: 1931-0145.

[Es96] Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.: A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In. KDD'96, Portland, Oregon, pp. 226–231, 1996.

[Ka18] Kazempour, D.; Beer, A.; Lohrer, J.; Kaltenthaler, D.; Seidl, T.: PARADISO: an interactive approach of parameter selection for the mean shift algorithm. In: SSDBM 2018, Bozen-Bolzano, Italy, July 09-11, 2018. 26:1–26:4, 2018.

[Sc15] Schubert, E.; Koos, A.; Emrich, T.; Züfle, A.; Schmid, K. A.; Zimek, A.: A Framework for Clustering Uncertain Data. PVLDB 8/12, pp. 1976–1979, 2015.