PARADISO: An Interactive Approach of Parameter Selection for the Mean Shift Algorithm

Daniyal Kazempour, Anna Beer, Johannes-Y. Lohrer, Daniel Kaltenthaler, Thomas Seidl
Lehrstuhl für Datenbanksysteme und Data Mining, Institut für Informatik,
Ludwig-Maximilians-Universität München
(kazempour,beer,lohrer,kaltenthaler,seidl)@dbs.ifi.lmu.de

ABSTRACT

Many algorithms have been developed for detecting clusters of various kinds over the past decades. However, just few attempts have been made to provide an interactive setting for the clustering algorithms. In this paper, we present PARADISO, an interactive Mean Shift method. It enables the user to get back to any arbitrary iteration point of the run observing the evolution of the clusters after each iteration and to set different bandwidth parameters. The user gets a clustering result with this method which emerged through multiple bandwidths while the user can see the full chain of effects of the chosen bandwidths over all iterations. Further, our method provides so-called Points-Shifted-Distance plots (PSD plots) for the Mean Shift algorithm which aim to facilitate the choice of a different bandwidth for the user. Beyond the mentioned features, PARADISO provides a visualization method which lets the user see the different bandwidth choices made in form of pathways.

CCS CONCEPTS

• **Information systems → Clustering and classification**;

KEYWORDS

Adaptive, Clustering, Interactive, Mean Shift, PARADISO, Points-Shifted-Distance Plots, Visualization

ACM Reference Format:
Daniyal Kazempour, Anna Beer, Johannes-Y. Lohrer, Daniel Kaltenthaler, Thomas Seidl. 2018. PARADISO: An Interactive Approach of Parameter Selection for the Mean Shift Algorithm. In *SSDBM '18: 30th International Conference on Scientific and Statistical Database Management, July 9–11, 2018, Bozen-Bolzano, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3221269.3221299

1 INTRODUCTION

The clustering algorithms developed over the past decades are intended to be run with fixed hyperparameters. If the domain expert wants to examine the outcome of the clustering with different parameters, the algorithm has to be re-run with different parameters. A user interaction of changing the parameters at arbitrary iteration points is, to the best of our knowledge, not provided by any of the originally developed clustering methods.

In our work, we present PARADISO (*Parameter-Adjusted Real-time Animated Data Interaction Software*), which implements the first interactive variant of the Mean Shift algorithm. It enables the domain expert to start the algorithm on a data set and after a finished run to jump back to any previous iteration point, modify the bandwidth parameter of Mean Shift, and continue the computation with the new bandwidth from the stopped iteration point. PARADISO delivers the potential to travel back in time and explore the effects of bandwidth modifications at different iterations. In order to augment the domain expert in making a decision for a different bandwidth, PARADISO is the first method to provide for the Mean Shift algorithm statistics for that purpose. Further, we provide a visualization which provides the domain expert an overview of which different paths have been chosen so far leading to the different Mean Shift results. Finally, our work serves as a proof-of-concept for interactive clustering algorithms in general by specifying concepts which are recommended to be fulfilled by future interactive clustering methods. The Mean Shift algorithm [4] is a clustering method which detects clusters by locating the maxima of a density function (mode) and thus is referred to as a mode-seeking algorithm. The choice for using Mean Shift as our candidate for the interactive setting has been made based on the fact that Mean Shift requires only one hyperparameter namely the bandwidth. To serve an intuition to this hyperparameter, one can think of the bandwidth as the radius in which the mode is determined. A large bandwidth would lead to a faster convergence but also all points collapsing to a single point and thus to a single cluster. In contrast, a small bandwidth would lead to a much slower convergence of each of the data points and thus to many small clusters. In our work, we also elaborate on the purpose of choosing different bandwidths at different iterations of the Mean Shift algorithm. You can download the code at: https://gitlab.lrz.de/ru49qap/paradiso.

As a use case we aim to utilize PARADISO to detect suitable regions for deploying loan offices for the Kiva project. The adequacy of the locations for such loan offices depends on the number of loan queries from the surrounding area. Applying the standard Mean Shift algorithm would yield potential office locations but would disregard locally dense areas in the surroundings and just drag their mode to a global center. Being able to choose various bandwidths at different iterations can yield office locations which also consider locally dense subareas.

To summarize, we give a brief summary of the contributions in this work: (1) An interactive and easy to use variant of the Mean Shift algorithm. (2) The capability to set different bandwidths at different iteration points and discovering the outcome and every

SSDBM '18, July 9–11, 2018, Bozen-Bolzano, Italy
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6505-5/18/07…$15.00
https://doi.org/10.1145/3221269.3221299

iteration point. (3) *Points-Shifted-Distance plots* (PSD plots) illustrating the shifted distance of the data points at each iteration compared to the previous iteration. (4) visualization concept showing which different bandwidths have been selected emerging from different iteration points of their originated run.

2 RELATED WORK

A comparison of static and interactive methods show that users tend to prefer the interactive solution. For example, this is shown in studies by Callahan and Koenemann [3] who compares interactive visual tools with traditional static tools for online catalog browsing, or Combs and Bederson [5] who compare a zoomable image browser to a static image browser. The users report a higher user-friendliness and efficiency in interactive tools than in static interfaces. Jeong et al. [7] developed a system called *iPCA* that visualizes the results of *Principle Component Analysis* (PCA), a widely used mathematical technique for high dimension data analysis, using multiple coordinated views and a rich set of user interactions. Guo [6] achieved a similar approach with "a suite of coordinated visualization and computational components centered around [...] methods to facilitate a human-led exploration" [6] which provides the visualization of data with plots and clusters. Bhadra et al. [2] provided the first possibility to detect clusters by using animation. "As visualizations are changed in real-time by the user, the data-items also change their positions on the screen in real-time. This creates an animation on the screen. In this animation, clusters typically behave like semi-rigid bodies moving from the old positions to the new positions. This behaviour helps the user in identifying them." [2]. While there are many different approaches to increase the detection of clusters with different user-guided approaches, to the best of our knowledge, none of these methods allow the dynamic adjustment of parameters within one run of the Mean Shift algorithm.

3 INTERACTION CONCEPTS

From an end-user perspective a person may know about clustering in general but not about a specific clustering method (here Mean Shift). PARADISO reveals the workings of a clustering algorithm by which the end-user can actually implicitly learn how the algorithm works by interacting with it. The PSD-plots and the ability to go through every iteration step reveal the immediate effects of a hyperparameter being changed by the end-user.

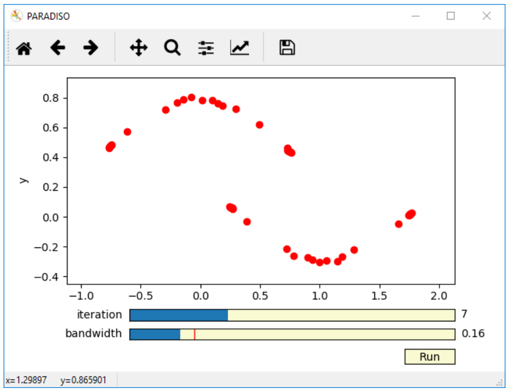

Figure 1: The main window of PARADISO with its two interaction elements.

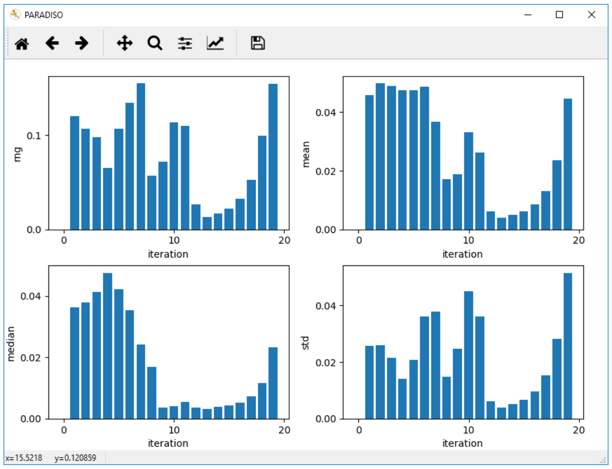

Figure 2: PSD plots displaying the development over the iterations of range (top-left), mean (top-right), median (bottom-left), and standard deviation values (bottom-right).

The very focus of our work lies in providing core concepts of interaction in context of interactive clustering. The main window of PARADISO provides three fundamental elements as it can be seen in Figure 1. Having the Mean Shift algorithm executed on a data set (here we use the well-known "two moons" data set which consists of two moon shaped clusters) the user has two elements which can be interacted with:

(1) The iteration slider which provides the user to move forward and backward through every single intermediate result.
(2) The bandwidth slider which gives the user the ability to change the bandwidth at any iteration step.

Considering a visualization of intermediate results and the points of interaction alone are insufficient for a basis in interactive clustering settings. As the user may recognize the changes and has the power to go through iterations and change parameters, the user is left without any information by not knowing when to interact in which way. This issue is addressed in our work by introducing a plot which provides aggregated information about data over every iteration. To be more precise we introduce the so-called *Points-Shifted-Distance plots* (PSD plots). The concept is inspired by the *OPTICS* plots [1].

The PSD plot is, to the best to our knowledge, the first of its kind, providing information for the choice of the bandwidth. The horizontal axis describes the iterations. The vertical axis describes first order statistics of how much the whole of the points in the data set have shifted in the current iteration, compared to the previous one. These can be seen in Figure 2. In our case, we provide the range, mean, median, and standard deviation values. The information that the PSD plots provide are the following:

- **Range values:** The range values indicate the differences between the lowest and the highest shifting distance. This difference tells the user about the range of which the points shift between the iterations. Having further information about the overall spread of the shifted distances, the standard deviation is the measure of choice.

- **Mean values:** A low mean value implies that the whole data set has shifted little, compared to the position of the data points in the previous iteration. In contrast, a large mean value implies the opposite effect of long distance shifts. This information is useful to increase the bandwidth if a faster colliding to a single mode is desired, or a lowering of the bandwidth to enforce a higher number of modes and thus a higher number of clusters.
- **Median values:** Compared to the mean, the median tells the user about the overall shifted distances without the bias by extremely small or large distance shifts. This can aid the user in e.g. detecting regions where the shifted distances remain mostly constant.
- **Standard deviation values:** From the standard deviation PSD plot we can obtain the information on how homogeneous the data points shift. A low variance in combination with e.g. a high mean value means that most of the points shift by an roughly same high distance, while a high standard deviation with a high mean implies that there are different subsets of data points which shift over different distances, implying that each subset has a different shifting speed towards its respective mode.

The visualization and the interaction elements provide the user with the tools to interact with clustering algorithms (here the mean shift). Further the PSD plots aid the user in taking decisions on the parameter at different iterations. While exploring the different bandwidths at different iterations it may very easily quite cumbersome to keep track which changes have been made at which iteration points. As a solution to this issue, in PARADISO, the user can go through the iterations, change the bandwidth, and create an alternative path of the run. Now the user has the option to either continue and perform different changes on the first run or to introduce another bandwidth at another (or the same) iteration point. The concept of creating different paths can be seen in Figure 3.

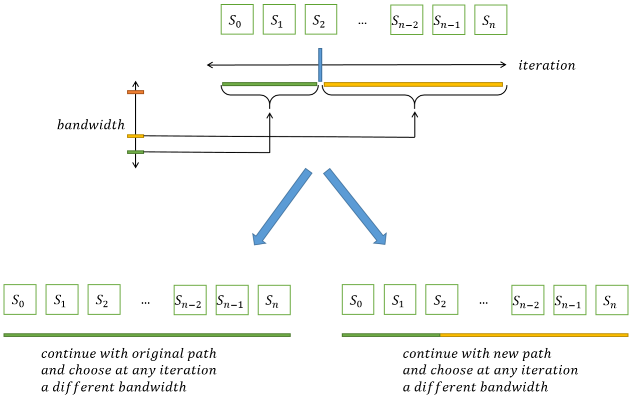

Figure 3: Concept of alternative paths.

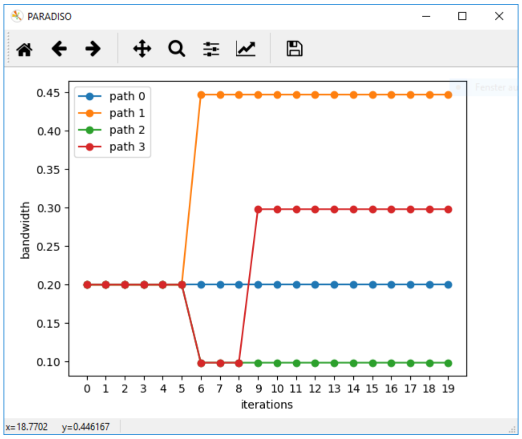

Figure 4: Alternative paths plot.

Being capable of creating such alternative paths, it is vital that the user keeps track of the alternative paths that have been investigated. In order to provide the user a tool for the full control over paths, we introduce – as the third and last component – the concept of an alternative paths plot. In the paths plot in Figure 4 we can see e.g. the origin path (path 0) which has been generated throughout all iterations by the bandwidth 0.2. A different bandwidth of 0.45 has been chosen at iteration 5, describing a branching path (path 1) which has been created by the two bandwidths 0.2 and 0.45 at different iteration points. Further, path 2 has been generated also branching off at iteration 5 to a lower bandwidth of 0.1 for three iterations and branching again off to a bandwidth of 0.3 beginning at iteration 9. With this paths plot the user is provided with a path revealing which choices have already been made at different iteration points. Path 3 exposes an interesting difference compared to the other paths. It first branches off from path 0, but again branches after iteration 8.

Conclusively, we have described three elements with which we define the very foundation for any interactive clustering algorithm, as seen in Figure 5. Via controlling iterations and controlling the algorithm specific iterations, the user performs an action based on the visualization of statistics through plots (here our PSD plot). This action leads to an reaction which is the visualization of intermediate results (the results at each iteration), on which in turn the user makes an observation, leading to eventually further actions. In order to keep track of the modifications PARADISO provides the paths plot by which users can follow at which iterations they have made changes to the hyperparameters.

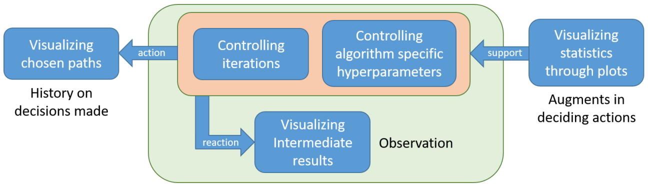
Visualizing chosen paths
History on decisions made
action
Controlling iterations
Controlling algorithm specific hyperparameters
support
Visualizing statistics through plots
Augments in deciding actions
reaction
Visualizing Intermediate results
Observation

Figure 5: Interaction model for clustering algorithms

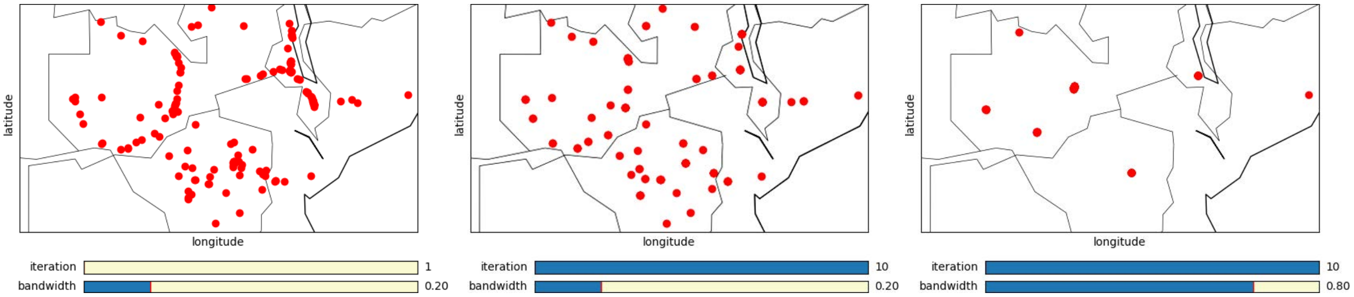

Figure 6: Distribution of loan requests on a magnified region in Africa via Kiva: The unmodified raw data (left), applied Mean Shift after 10 iterations with bandwidth 0.2 (center), and applied Mean Shift after 10 iterations with dynamic bandwidth (0.2 until iteration 2, 0.8 afterwards) (right).

4 USE CASE

Having elaborated on our concept of interactive clustering, we now proceed with a use case showing one possible field of application on real world data. For this purpose, we use the Kiva[1] loan data set from Kaggle[2]. This data set contains the locations of people asking for loans on the Kiva platform. As we focus on the continent of Africa, the data set encompasses around 3.000 queries related to loans and contains the longitude and latitude information including an assignment of the queries to their respective country (cf. Figure 7).

On the first image in Figure 6 we can see a magnified region of the African continent. In Figure 6 (left) we can see the original distribution of the loan queries. Figure 6 (center) shows the modes after iteration 10 using a bandwidth of 0.2. It can be seen that there are certain regions (modes) where most of the requests come from. In Figure 6 (right) we can see a result at iteration 10 where the result emerged from a change of the bandwidth after iteration 2 to a bandwidth of 0.8. This increase of the bandwidth leads to fewer modes which encompass larger areas of loan queries. This information can be useful for establishing the so-called Kiva offices as there are currently offices only located in Nairobi, Kenya, according to their website.

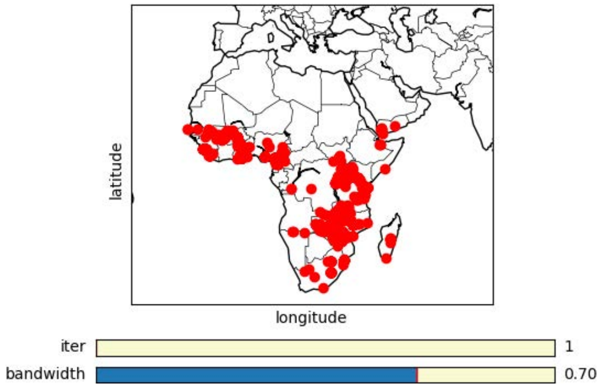

Figure 7: Distribution of loan requests on the African continent via Kiva.

5 CONCLUSION AND FUTURE PROSPECTS

In our work we introduced PARADISO, the first interactive version of the MeanShift clustering algorithm. We further introduced elements such as the PSD plot and the alternative path plot. In our use case, we provided one scenario where the user can interactively see the effects on modifying the parameter of the algorithm and different iteration points. With the mentioned elements we further defined a core concept which is, as we think, applicable to most of the clustering algorithms being modified to an interactive setting.

However our work here is not done yet, as there are various aspects by which the core interaction concept can be enhanced. Further, there is room for improvement regarding the efficiency and the aspect of dealing with high dimensional data which brings its own challenges in the context of interactiveness.

It is of special interest to develop mechanisms for memory management of the intermediate states regarding the aspects of how many or which states should be stored and presented to the user. In context of data provenance our method can be used to replay specific clustering and capture the lineage of clustering results.

We hope that this work motivates and fosters the research and design of further interactive clustering methods, as especially the demand on explainability of the results produced by data mining methods increases. A demand, that we are convinced, can be addressed by interactive approaches.

REFERENCES

[1] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. OPTICS: ordering points to identify the clustering structure. In *ACM Sigmod record*, Vol. 28. ACM, 49–60.

[2] Debangshu Bhadra and Ashim Garg. 2001. Interactive Visual Framework for Detecting Clusters of a Multidimensional Dataset. (2001).

[3] Ewa Callahan and Jürgen Koenemann. 2000. A comparative usability evaluation of user interfaces for online product catalog. In *Proceedings of the 2nd ACM conference on Electronic commerce*. ACM, 197–206.

[4] Yizong Cheng. 1995. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17, 8 (Aug 1995), 790–799. DOI: http://dx.doi.org/10.1109/34.400568

[5] Tammara TA Combs and Benjamin B Bederson. 1999. Does zooming improve image browsing?. In *Proceedings of the fourth ACM conference on Digital libraries*. ACM, 130–137.

[6] Diansheng Guo. 2003. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization* 2, 4 (2003), 232–246.

[7] Dong Hyun Jeong, Caroline Ziemkiewicz, Brian Fisher, William Ribarsky, and Remco Chang. 2009. iPCA: An Interactive System for PCA-based Visual Analytics. In *Computer Graphics Forum*, Vol. 28. Wiley Online Library, 767–774.

[1] Kiva is a non-profit platform which enables people to lend money via the Internet to people with low income. http://www.kiva.org/

[2] http://www.kaggle.com/mithrillion/kiva-challenge-coordinates/data