Identifying Entangled Data Points on Iteration Trajectories of Clusterings

Daniyal Kazempour[1] and Thomas Seidl[1]

Ludwig-Maximilians-Universität München, Munich, Germany
{kazempour,seidl}@dbs.ifi.lmu.de

Abstract. Clustering methods like Mean Shift or k-Means are executed and if necessary, re-started with different parameters. This standard approach of providing input data, executing an algorithm, observing the outcome and re-running the method with different parameters completely neglects the insights which can be gained considering to look at the core of an clustering algorithm. In this work we provide an approach to analyze the trajectories of data points within a run of a clustering method. Thereby we also introduce the concept of data point entanglement, revealing if subsets of data points have common trajectories over the iterations of an clustering algorithm. Further we introduce the idea of data point resilience, whose intuition is to reflect if a subset of data points maintains its entanglement even at different hyperparameter settings.

Keywords: Clustering · Cluster Member Entanglement · Cluster Trajectories Alignment · Cluster Entanglement Resilience.

1 Introduction

Within the past decades a rich set of various clustering methods has been developed, aimed at different tasks such as detecting convex clusters such as k-means [1], detecting densely connected clusters as in e.g. DBSCAN [3] or detecting relevant subspaces (axis parallel or arbitrarily oriented) such as CLIQUE [4] or 4C [5]. Few of the algorithms have been put into interactive context revealing the inner changes and therefore revealing to the users a wealth of information on which basis decisions, e.g. the choice of hyperparameters can be performed as in PARADISO [6]. Especially in context of clustering algorithms where in an iterative fashion points are re-assigned to their clusters such as in k-means or where the points roam to their modes, we are convinced that within the trajectories of their point movements information can be gained regarding the connection between data points and their robustness towards different hyperparameters. In this work we use the Mean Shift[2] algorithm as utilized in the interactive framework PARADISO. The core idea of the Mean Shift algorithm is that each data point roams towards the center (mode) accounting other data points within a specific window, which is in literature referred to as a bandwidth. The whole process is executed in an iterative fashion until convergence

is achieved. PARADISO uses the Mean Shift algorithm as a basis and provides the users access to the algorithms internals, rendering it possible to set different bandwidths at different iterations steps, moving back and forth. Further the users can explore alternative paths in the sense of choosing different bandwidths over different iterations yielding different results. As the major contribution of this work, we introduce concepts and definitions for entanglement of data point trajectories and the resilience of the entanglement regarding different hyper-parameters (here the bandwidth). Although plenty of related work regarding trajectory alignment and similarity measure exist, best to our knowledge no other work is yet published which elaborates on cluster iteration trajectory and entanglement resilience computation.

2 Data Point Entanglement

We take as an example to explain the concept of data point entanglement the two-moons data set consisting of fifty data points. The two-moons data set consists of two opposing moon-shaped two-dimensional figures as it can be seen in figure 1 top left figure. We execute the Mean Shift algorithm with a bandwidth of 0.3. In figure 1 it can be seen how the points are dragged into two modes over the ten iterations.

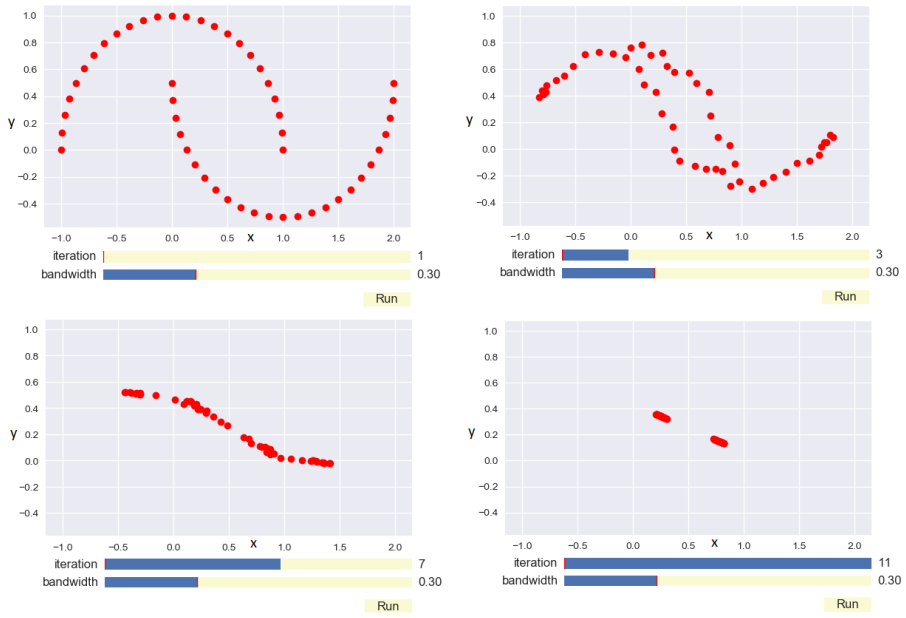

Fig. 1: two-moons Mean Shift intermediate results at different iterations

Definition 1. *Storing the position of each of the data points $p \in \mathcal{D}$ at all iterations $i \in \mathcal{I}$ yields a trajectory $\theta \in \mathcal{T}$ for each data point. Two data points p_i, p_j are entangled if $d_{DTW}(\theta_{p_i}, \theta_{p_j}) \leq \tau$ where the smaller the threshold τ is chosen, the more entangled are the two data points.*

The core intuition behind definition 1 is that the more similar the trajectories of two data points are, the more entangled the two observed data points. As a measure of similarity of trajectories we have utilized dynamic time warp (DTW). In context of time series analysis DTW is widely used to determine how well given time series align. Due to the brevity of this paper, we remain using here DTW for the rest of this work, yet want to emphasize that further research regarding other distance measures is a vital topic. Having the DTW distance between two data point iteration trajectories, we define a distance matrix as follows:

Definition 2. *An entanglement matrix $\boldsymbol{M}$ is a distance matrix where each of its elements $m_{ij} \in \boldsymbol{M}$ is a pairwise distance $d_{DTW}(\theta_{p_i}, \theta_{p_j})$ between two data point trajectories $\theta_{p_i}, \theta_{p_j}$ as stated in definition 1.*

Such an entanglement matrix can be seen in figure 2 which is based on the two-moons data set mentioned before. Here we can see all pairwise data point trajectory similarities, where a dark color implies a high similarity between two data point trajectories and a bright color means the opposite. As the entanglement matrix is symmetric, we can omit either the lower or upper triangle. Further we can also exclude the diagonal from our observations as a data point trajectory is always maximum similar to itself.

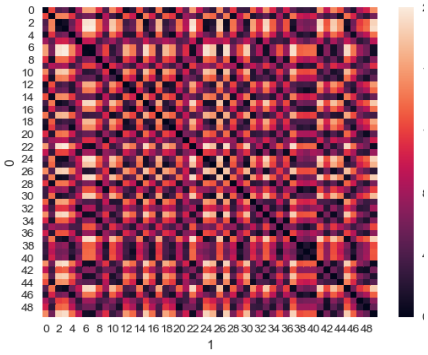

Fig. 2: An entanglement matrix

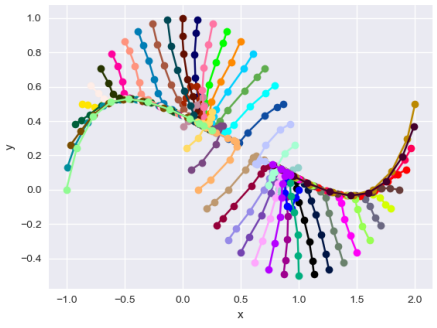

Fig. 3: Trajectories of each of the two-moons data points

Visualizing the trajectories of each data point results in a pattern as seen in figure 3 where the movement of the data points is directed towards two centers (modes).

3 Entanglement Resilience

Based on the data from the entanglement matrix from figure 2 we have selected three tuples of data points which have an very low DTW distance on their trajectories, namely $p_{38} = (0.5, -0.366025404), p_{39} = (0.965925826, 0.258819045), p_{40} = (0.617316568, -0.423879533)$. We computed their scores at a bandwidth of 0.5, 0.4, 0.3, 0.2, 0.1 and 0.05 on the two-moons data set. In figure 4 it becomes immediately visible that the low DTW distance and thus the high entanglement between (p_{38}, p_{40}) remains mostly the same while on the other two data point pairs the DTW distance increase massively at lower bandwidths weakening their entanglement. This phenomenon of a stable level DTW distance we define as follows:

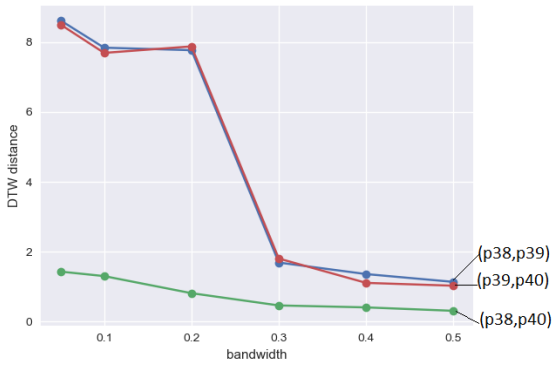

Fig. 4: DTW distance of data point tuples at different bandwidth runs of the Mean Shift algorithm on the two-moons data set.

Definition 3. *The entanglement between two data points p_i, p_j is coined with the term resilience ρ if the variance of the DTW distance of their respective trajectories at different hyperparameter settings $\lambda \in \Lambda$ is small, meaning that the resulting value ρ is close to zero.*

$$\rho_{(\theta_{p_i}, \theta_{p_j})_\Lambda} = var(d_{DTW}(\theta_{p_i}, \theta_{p_j})_\Lambda) = \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} (d_{DTW_\lambda}(\theta_{p_i}, \theta_{p_j}) - \mu_\Lambda)^2 \quad (1)$$

Based on the equation above the resilience of the entanglement between two data point trajectories can be computed. Intuitively spoken, it reflects the variance of the distance of two data point trajectories at different hyperparameters λ regarding the mean which is defined as the average of the distance of two data point trajectories over all chosen hyperparameters λ: $\mu_\Lambda :=$

$\frac{1}{|\Lambda|}\sum_{\lambda\in\Lambda} d_{DTW_\lambda}(\theta_{p_i}, \theta_{p_j})$. Having such a measure, the immediate question is what it can be used for. The entanglement resilience can serve as an indication for a subset of data points which have a strong relation to each other regardless the choice of different hyperparameters. Applying definition 3 to our example of the data points trajectory tuples (p_{38}, p_{39}), (p_{39}, p_{40}) and (p_{38}, p_{40})leads to the following values: $\rho_{(p_{38},p_{39})} = 13.51$, $\rho_{(p_{39},p_{40})} = 13.66$ and $\rho_{(p_{38},p_{40})} = 0.23$. This small value supports our observation from figure 4 that the trajectories of the data points 38 and 40 remain similar to each other at different hyperparameters.

4 Conclusion

In our work we have introduced two core concepts, namely the data points entanglement and the entanglement resilience. Both are primarily aimed to be used in context of interactive clustering tools such as PARADISO. They can be either used to automate steps within an algorithm or to provide additional information to the users, serving as a basis for their decisions in the interactive setting. Future work encompasses the extension and application on high-dimensional settings and a broader range of clustering algorithms at which this concept is applicable. Further the choice of the threshold τ as mentioned in definition 1 requires further investigations approaching questions such as: How τ should be chosen, and if this parameter can be learned or estimated. One of the major motivations of this work is to catalyst the development of methods which reveal patterns that take place 'under the hood' of the clustering algorithms, leveraging the understanding of such methods.

References

1. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Band 1. University of California Press, 1967, S. 281–297
2. Comaniciu, Dorin; Peter Meer. "Mean Shift: A Robust Approach Toward Feature Space Analysis". IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE. 24 (5): 603–619.(May 2002)
3. Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M., eds. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.
4. Agrawal, R., Gehrke, J., Gunopulos, D. et al. Automatic Subspace Clustering of High Dimensional Data. Data Min Knowl Disc (2005) 11: 5.
5. Böhm, Christian; Kailing, Karin; Kröger, Peer; Zimek Arthur. Computing Clusters of Correlation Connected Objects. Proc. ACM Int. Conf. on Management of Data (SIGMOD), pp. 455-466, Paris, France. 2004.
6. Kazempour, Daniyal; Beer, Anna; Lohrer, Johannes-Y.; Kaltenthaler, Daniel; Seidl, Thomas. PARADISO: An Interactive Approach of Parameter Selection of the Mean Shift Algorithm. SSDBM 2018.