DipTransformation: Enhancing the Structure of a Dataset and thereby improving Clustering (Extended Abstract)

Presentation of work originally published in the Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM 2018).

Benjamin Schelling[1] Claudia Plant[1 2]

Keywords: Clustering, Dip-test, Dataset-Transformation

The clustering of a data set depends strongly on the structure it contains. A data set might have a well-defined structure, but this does not necessitate good clustering results. If the structure is hidden in an unfavourable scaling, clustering usually fails. Confronted with a data set one cannot quite cluster, usually this would lead to a new clustering method which is capable of dealing with the new and problematic type of data set, but this is not always necessary. The aim of the DipTransformation is to enhance the data set by re-scaling and transforming its features and thus emphasizing and accentuating its structure. If the structure is sufficiently clear, clustering algorithms - even well-established ones - will perform far better. To the best of our knowledge, there are currently no methods besides DipTransformation that have the goal of enhancing structure.

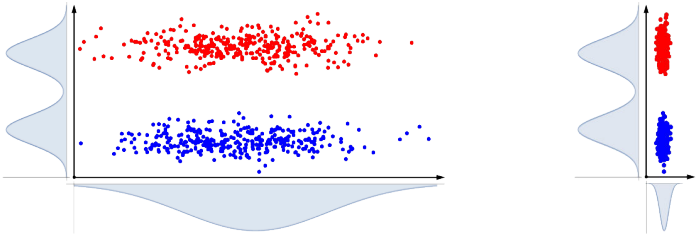
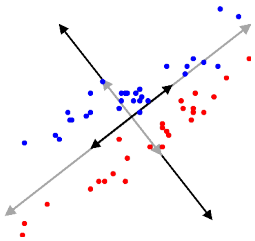


Fig. 1: A simple, synthetic data set before (a) and after (b) scaling it with its dip values. The dip value are used to tell us how much structure a feature contains and how relevant it is for clustering.

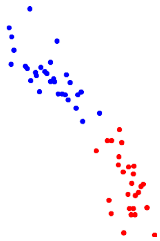
DipTransformation makes it possible to compensate for the unfortunate scaling of the features with the help of the Dip test [1]. The Dip test measures the amount of structure in a feature. Take a look at Fig. 1. It is a very simple data set, consisting of two Gaussian

[1] Faculty of Computer Science, University of Vienna, Vienna, Austria
[2] ds:UniVie, University of Vienna, Vienna, Austria

doi:10.18420/inf2019_31



(a) The Whiteside-data set and the main direction of interest.



(b) After the DipTransformation.

Fig. 2: We find the direction with the most structure (black: how much structure is found, grey: how large it is originally scaled) and scale the features according to it. It is now very easy to cluster.

distributed clusters. It should be very easy to cluster, but many algorithms (e.g. k-means) have massive difficulties with it, due to the scaling. Measuring the amount of structure of the features with the dip test and re-scaling the features leads to Fig. 1.b, a very easy to cluster data set. The heuristic here is that a feature is scaled relative to how much structure is found. The horizontal axis has barely any structure, it is uni-modal, and thus is scaled such that it has only a very small extension, which means that it has no great influence on clustering, as the values are all very similar. The feature with high structure – it is clearly multi-modal, i.e. has clusters one can distinguish from each other – is scaled such that this feature has a high influence on clustering.

The DipTransformation is not limited to axes-parallel re-scaling. Using a cleverly devised search strategy, it can automatically find non-axis-parallel features with high dip values, which it rescales as explained. One such example is shown in Fig. 2. The Whiteside-data set is a real-world data set that is difficult to handle for many clustering approaches, due to its clusters which are tricky to differentiate. Most approaches fail completely, but after the transformation, it is almost trivial.

In conclusion, we developed a technique that can improve the structure of a data set and thus its clustering. We show in [2] that this is true by testing it extensively on various data sets, all of which become far easier to cluster for various standard and state-of-the-art clustering approaches. DipTransformation assumes no data distribution, is deterministic, basically parameter-free and quite fast compared to various clustering approaches. It can thus be used as a pre-clustering step, that enhances the data set, and the clustering algorithm can be selected according to user preferences.

Bibliography

[1] Hartigan, J. A., Hartigan, P. M., *The Dip Test of Unimodality*, The Annals of Statistics, 1985.

[2] Schelling, B., Plant, C., *DipTransformation: Enhancing the Structure of a Dataset and Thereby Improving Clustering*, ICDM, 2018.