A Generator for Subspace Clusters

Anna Beer, Nadine Sarah Schüler, and Thomas Seidl

LMU Munich
{beer,seidl}@dbs.ifi.lmu.de
n.schueler@campus.lmu.de

Abstract. We introduce a generator for data containing subspace clusters which is accurately tunable and adjustable to the needs of developers. It is online available and allows to give a plethora of characteristics the data should contain, while it is simultaneously able to generate meaningful data containing subspace clusters with a minimum of input data.

Keywords: Data Generator · Subspace Clustering · Reproducibility

1 Introduction

Developing algorithms in the field of data mining is usually an iterative process in which a main idea is implemented and then tested on several use-cases or experiments containing a ground truth. Depending on the results of those, the algorithm is modified and a loop of alternately testing and improving the algorithm starts. If the same data or only a few data sets are used in several iterations of this cycle, we create overfitting algorithms. The fields in which such subspace clusters can occur are manifold and especially for gene expression data or other data with medical background, clusters are most often found only in meaningful subspaces. Nevertheless, the number of labeled datasets is limited, and datasets containing labeled subspace clusters are rare. So, instead of using the few real world labeled datasets to develop and improve a subspace clustering algorithm, artificial datasets, of which the ground-truth is known by construction, are often used. Additionally, we can generate datasets in such a way, that they emphasize the advantages of the algorithm and help to detect diverse properties which possibly emerged in the development process. Data generators simplify the cumbersome process of constructing new datasets by hand, and allow building reproducible data sets, which are versatile enough to produce a non-overfitting algorithm in the above described development cycle. Nevertheless, there are only few publicly available data generators and none for generating data containing subspace clusters, even though some are used in diverse subspace clustering papers, as described in Section 2. Thus, we developed a generator for data containing subspace clusters, which allows to determine a multitude of parameters and is described in Section 3. Section 4 concludes this short paper and gives ideas for future work.

2 Related Work

The quality of most subspace clustering algorithms presented in the last years is shown using synthetic data, the construction of which is usually not well described or not reproducible at all. Most authors created very elementary data generators, leading to a multitude of generators with too little setting options to construct datasets with reasonably predictable characteristics. Looking at a multitude of subspace clustering related papers, we found the following to describe their data generation process best: SubClu [KKK04], SURFING [BPR+04], CLIQUE [AGGR98], which uses the generator described in [ZM97], and a review of diverse subspace clustering algorithms [PHL04]. Further, ResCu [MAG+09] and INSCY [AKMS08] use the same generator as [KKK04]. While all of those generators allow the user to set the number of points and dimensionality of the dataset as well as the number and dimensionality of clusters explicitly or implicitly, some crucial aspects are missing in each. E.g., the density or variance of clusters can be set in SubClu and SURFING, but not in [PHL04] or CLIQUE. CLIQUE constructs clusters differently to the other generators, as the user defines hypercubes in which the uniformly distributed points are more dense than in the surrounding areas. Surprisingly, generating data with noise is only provided by the generator from CLIQUE. The other generators construct, similarly to ours, some Gaussian distributed clusters and have different properties: in SubClu and SURFING, no cluster can be clustered in the full dimensional space, but the authors do not describe how this is reached. In [PHL04], the values of the relevant dimensions for each instance in a cluster can be restricted, leading to hypercube-shaped clusters.

MDCGen [IZFZ], which is probably the most recent and a very elaborated generator especially designed for multidimensional data and also subspace clustering, does not provide the possibility that a point can belong to multiple clusters at once. Additionally, there are data generators introduced independently from the field of subspace clustering, but to the best of our knowledge none of them is able to construct data containing subspace clusters of arbitrary dimensionality. [MLG+13] gives an overview over some data generators for big data benchmarking, like Hibench, LinkBench, CloudSuite, TPC-DS, YCSB, BigBench and BigDataBench, and BDGS. MUDD [SP04] is a generator similar to those. They are designed to create big data sets with similar properties as some given real world data, but users cannot specify enough details to be able to expose the advantages and disadvantages of their algorithms in development. RAIL [KBS19] is an interactive generator concentrating on producing linear correlated data, but allows only constructing 3-dimensional datasets containing 2-dimensional planes.

3 The Generator

In contrast to the data generators described in Section 2, the work here presented offers to define a plethora of characteristics of the dataset to be constructed while

simultaneously allowing to generate meaningful datasets containing subspace clusters without having to think about parameters too much: It requires only three parameters for the general set-up: The number of points n, the number of dimensions dim and m, a flag determining if it is possible for a point to belong to more than one subspace cluster. If given only those three parameters, we proceed as follows: To restrict the number of subspaces generated we use a fixed number of cluster centers, determined by a random number k between 1 and $\sqrt{n}$. The clusters are then randomly allocated to a number of subspaces $< k$ and the number of points as well as the number of dimensions of all subspaces clusters is drawn randomly from a uniform distribution within the given limits.

Users can specify the properties of the data further by giving information for every subspace S, namely the number of points, dimensionality, and number of clusters in S. Additionally, the variance of each cluster can be given. Figure 1 shows how subspaces can be distributed. In this example, there are four different subspaces, of which the first contains two clusters, the second and third contain one cluster each, and the fourth contains three clusters. The last two points belong to no cluster at all, while all other points are in two clusters in different subspaces. If $m = false$, a point may only belong to exactly one cluster, we insert the given subspaces into the $n \times dim$ matrix as long as there are sufficient points not assigned yet. Points and dimensions not assigned to belong to a certain subspace cluster, are filled with uniformly distributed noise data and a 0 in the label-matrix giving the cluster-assignments. If $m = true$, subspaces are first assigned in the same way as described above, before points are assigned to a second subspace and obtain a second cluster membership (see Figure 1). This is again assigned by going through the points and if there are enough unassigned dimensions to meet the requested subspace dimensionality this point will become a member of the second subspace in addition to the first one. When the points belong to a subspace cluster, the values are drawn from a appropriate multidimensional Gaussian distribution function, the center and standard deviation of which can be given by users. The remaining values are again drawn from a uniform distribution function. Uniformly distributed noise points can be added. Our generator is online available under https://github.com/NanniSchueler/SubCluGen.git and outputs the data matrix as well as the label matrix.

4 Conclusion

In summary, we introduced a data generator especially designed for subspace clusters. It expects only three parameters: the size and dimensionality of the dataset as well as as boolean value determining if a point can belong to clusters in different subspaces. With that a fast construction of data is possible. Simultaneously, reproducible datasets with very specific properties can be designed by users to test algorithms they are developing for diverse characteristics. The generator is easy to use and we plan to extend it with even more possibilities, like, e.g., non-axis parallel subspace clusters or other distributions instead of Gaussian, in future work. Also a combination with RAIL or some of the men-

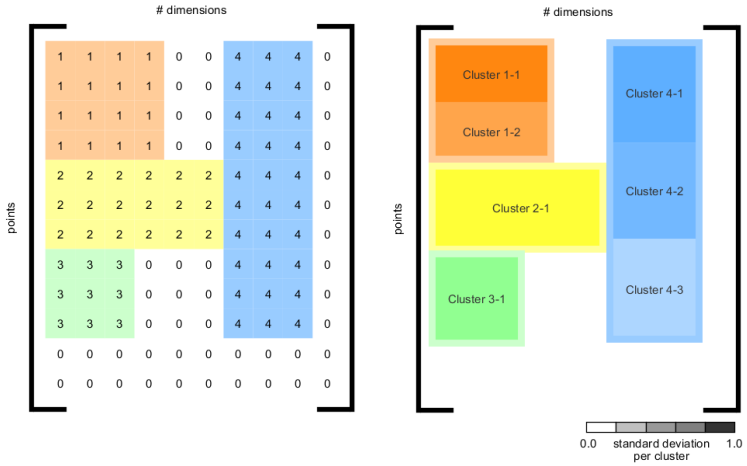

Fig. 1: Left: Label-Matrix as given by the generator as output, 0 implies uniformly distributed data, where numbers 1 to 4 imply the subspace affinity. On the right, the exact cluster affinity can be seen as well as the variance of the clusters implied by colour saturation.

tioned generators taking real world data into account could deliver a variety of reproducible datasets containing the desired properties for testing and developing.

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

[AGGR98] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. *Automatic subspace clustering of high dimensional data for data mining applications*, volume 27. ACM, 1998.

[AKMS08] Ira Assent, Ralph Krieger, Emmanuel Müller, and Thomas Seidl. Inscy: Indexing subspace clusters with in-process-removal of redundancy. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 719–724. IEEE, 2008.

[BPR+04] Christian Baumgartner, Claudia Plant, K Railing, H-P Kriegel, and Peer Kroger. Subspace selection for clustering high-dimensional data. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, pages 11–18. IEEE, 2004.

[IZFZ] Félix Iglesias, Tanja Zseby, Daniel Ferreira, and Arthur Zimek. Mdcgen: Multidimensional dataset generator for clustering. *Journal of Classification*, pages 1–20.

[KBS19] Daniyal Kazempour, Anna Beer, and Thomas Seidl. Data on rails: On interactive generation of artificial linear correlated data. In *International Conference on Human-Computer Interaction*, pages 184–189. Springer, 2019.

[KKK04] Karin Kailing, Hans-Peter Kriegel, and Peer Kröger. Density-connected subspace clustering for high-dimensional data. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 246–256. SIAM, 2004.

[MAG+09] Emmanuel Müller, Ira Assent, Stephan Günnemann, Ralph Krieger, and Thomas Seidl. Relevant subspace clustering: Mining the most interesting non-redundant concepts in high dimensional data. In *2009 Ninth IEEE International Conference on Data Mining*, pages 377–386. IEEE, 2009.

[MLG+13] Zijian Ming, Chunjie Luo, Wanling Gao, Rui Han, Qiang Yang, Lei Wang, and Jianfeng Zhan. Bdgs: A scalable big data generator suite in big data benchmarking. In *Advancing Big Data Benchmarks*, pages 138–154. Springer, 2013.

[PHL04] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explorations Newsletter*, 6(1):90–105, 2004.

[SP04] John M Stephens and Meikel Poess. Mudd: a multi-dimensional data generator. In *ACM SIGSOFT Software Engineering Notes*, volume 29, pages 104–109. ACM, 2004.

[ZM97] Mohamed Zaït and Hammou Messatfa. A comparative study of clustering methods. *Future Generation Computer Systems*, 13(2-3):149–159, 1997.