University of Vienna
Faculty of Computer Science
Prof. Claudia Plant
Martin Perdacher

Data Mining

WS 2019/20

Programming Assignment 1: Clustering High Dimensional Data

Implement one of the following algorithms and compare your results with the algorithms implemented in Environment for DeveLoping KDD-Applications Supported by Index-Structures (ELKI):

Correlation Clustering

- CASH[1]
- COPAC[5]
- ERiC: Exploring Relationships among Correlation Clusters[4]
- HiCO: Mining Hierachies of Correlation Clusters[6]
- LMCLU[11]
- ORCLUS: Arbitrarily ORiented projected CLUSter generation[8]

Subspace (axis-parallel) clustering algorithms

- ~~CLIQUE~~[9]
- DiSH: Detecting Subspace cluster Hierachies[3]
- DOC: Density-based Optimal projective Clustering[14]
- HiSC: Finding Hierarchies of Subspace Clusters[2]
- P3C: A Robust Projected Clustering Algorithm[13]
- PreDeCon: Subspace Preference weighted Density Connected Clustering [10]
- ~~PROCLUS: PROjected CLUStering~~[7]
- SUBCLU: Density connected Subspace Clustering[12]

Include the following in your documentation:

- Provide the pseudo code of your algorithm.
- Describe the algorithm in general.
- You can find example datasets at the ELKI website:
  `https://elki-project.github.io/datasets/`

- Evaluate your clustering result with the metrics implemented in the python `scikit-learn` pacakge, in the clustering performance evaluation section:

  `http://scikit-learn.org/stable/modules/clustering.html`

Literatur

[1] ACHTERT, E., BÖHM, C., DAVID, J., KRÖGER, P., AND ZIMEK, A. Robust clustering in arbitrarily oriented subspaces. In Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA (2008), pp. 763–774.

[2] ACHTERT, E., BÖHM, C., KRIEGEL, H., KRÖGER, P., MÜLLER-GORMAN, I., AND ZIMEK, A. Finding hierarchies of subspace clusters. In Knowledge Discovery in Databases: PKDD 2006, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, September 18-22, 2006, Proceedings (2006), pp. 446–453.

[3] ACHTERT, E., BÖHM, C., KRIEGEL, H., KRÖGER, P., MÜLLER-GORMAN, I., AND ZIMEK, A. Detection and visualization of subspace cluster hierarchies. In Advances in Databases: Concepts, Systems and Applications, 12th International Conference on Database Systems for Advanced Applications, DASFAA 2007, Bangkok, Thailand, April 9-12, 2007, Proceedings (2007), pp. 152–163.

[4] ACHTERT, E., BÖHM, C., KRIEGEL, H., KRÖGER, P., AND ZIMEK, A. On exploring complex relationships of correlation clusters. In 19th International Conference on Scientific and Statistical Database Management, SSDBM 2007, 9-11 July 2007, Banff, Canada, Proceedings (2007), p. 7.

[5] ACHTERT, E., BÖHM, C., KRIEGEL, H., KRÖGER, P., AND ZIMEK, A. Robust, complete, and efficient correlation clustering. In Proceedings of the Seventh SIAM International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA (2007), pp. 413–418.

[6] ACHTERT, E., BÖHM, C., KRÖGER, P., AND ZIMEK, A. Mining hierarchies of correlation clusters. In 18th International Conference on Scientific and Statistical Database Management, SSDBM 2006, 3-5 July 2006, Vienna, Austria, Proceedings (2006), pp. 119–128.

[7] AGGARWAL, C. C., PROCOPIUC, C. M., WOLF, J. L., YU, P. S., AND PARK, J. S. Fast algorithms for projected clustering. In SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, USA. (1999), pp. 61–72.

[8] AGGARWAL, C. C., AND YU, P. S. Finding generalized projected clusters in high dimensional spaces. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA. (2000), pp. 70–81.

[9] AGRAWAL, R., GEHRKE, J., GUNOPULOS, D., AND RAGHAVAN, P. Automatic subspace clustering of high dimensional data for data mining applications. In SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA. (1998), pp. 94–105.

[10] BÖHM, C., KAILING, K., KRIEGEL, H., AND KRÖGER, P. Density connected clustering with local subspace preferences. In Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 2004), 1-4 November 2004, Brighton, UK (2004), pp. 27–34.

[11] HARALICK, R. M., AND HARPAZ, R. Linear manifold clustering in high dimensional spaces by stochastic search. Pattern Recognition 40, 10 (2007), 2672–2684.

[12] KAILING, K., KRIEGEL, H., AND KRÖGER, P. Density-connected subspace clustering for high-dimensional data. In Proceedings of the Fourth SIAM International Conference on Data Mining, Lake Buena Vista, Florida, USA, April 22-24, 2004 (2004), pp. 246–256.

[13] MOISE, G., SANDER, J., AND ESTER, M. P3C: A robust projected clustering algorithm. In Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China (2006), pp. 414–425.

[14] PROCOPIUC, C. M., JONES, M., AGARWAL, P. K., AND MURALI, T. M. A monte carlo algorithm for fast projective clustering. In Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, Madison, Wisconsin, USA, June 3-6, 2002 (2002), pp. 418–427.