Discovering Non-Redundant K-means Clusterings in Optimal Subspaces (Extended Abstract)

Presentation of work originally published in the Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining

Dominik Mautz[1] Wei Ye[2] Claudia Plant[3] Christian Böhm[4]

Keywords: clustering; k-means; subspace; non-redundant

A huge object collection in high-dimensional space can often be meaningfully clustered in more than one way. For instance, objects could be clustered by their shape or alternatively by their color. Each grouping represents a different view of the data set. The new research field of *non-redundant clustering* addresses this type of problems. In our paper [Ma18], we follow the approach that different, non-redundant *k*-means-like clusterings may exist in different, arbitrarily oriented subspaces of the high-dimensional space. Minor assumptions about the orthogonality of the subspaces enable a particularly rigorous mathematical treatment of the non-redundant clustering problem and thus a particularly efficient algorithm, which we call Nr-Kmeans (for non-redundant *k*-means).

Figure 1 shows an example of a non-redundant clustering task. Given pictures of four objects—from the *Amsterdam Library of Object Images* (ALOI)— taken from different viewing angles and illumination temperatures could either be clustered by their shape into round and cylindrical objects or alternatively by their color into red and green objects. Each grouping represents a different view of the data set and is equally valid. From a mathematical perspective, there are two different low-dimensional subspaces, each exhibiting an interesting clustering structure. The clusterings in the subspaces are mutually non-redundant, i. e. each object belongs to different clusters in different subspaces. Classical clustering algorithms are not suited to capture these distinct views and may find only one of the possible partitions or a hard to interpret mixture of different clusterings.

The proposed non-redundant clustering algorithm Nr-Kmeans tackles this problem with a simple idea: find multiple mutually orthogonal subspaces—that may be arbitrarily oriented within the full space—such that the objective function of classical *k*-means is optimized in all of them. Both, the subspaces and the clusterings within are optimized simultaneously and influence each other during optimization. The only parameters needed are the expected number of clusters within each subspace. The orthogonality between subspaces ensures that the discovered clusterings represent different views on the data providing mutually

[1] Ludwig-Maximilians-Universität München, Munich, Germany, mautz@dbs.ifi.lmu.de
[2] University of California, Santa Barbara, CA, USA, weiye@cs.ucsb.edu
[3] Faculty of Computer Science, University of Vienna, Vienna, Austria
ds:UniVie, University of Vienna, Vienna, Austria, claudia.plant@univie.ac.at
[4] MCML, Ludwig-Maximilians-Universität München, Munich, Germany, boehm@dbs.ifi.lmu.de

doi:10.18420/inf2019_32

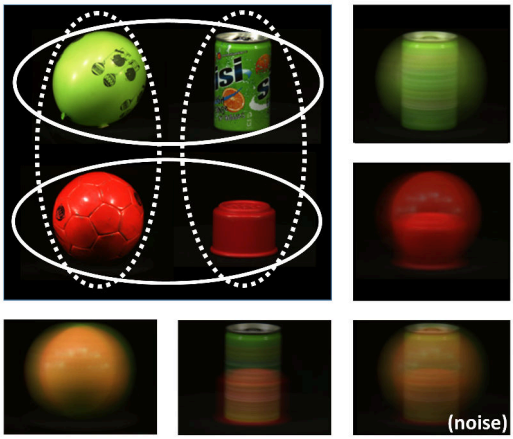

Fig. 1: Multiple clustering possibilities of objects according to color and shape. The smaller images show the corresponding average images.

non-redundant information and further allows for an efficient optimization procedure. The dimensionality of each subspace is determined automatically and the subspaces are well suited for visualization and further analysis, as they reveal the relationships between the individual clusters of a clustering. Inheriting from k-means, the result of Nr-Kmeans includes interpretable cluster centers, as displayed in Figure 1. In addition, our technique introduces a noise subspace, orthogonal to the other subspaces. The noise subspace captures all the unimodal variance in the data that is not interesting for clustering. This property allows Nr-Kmeans to prune away subspace dimensions without any clustering information and helps to outperform existing methods, especially on high-dimensional data.

Furthermore, it is possible to extend Nr-Kmeans with many other proposed k-means extensions in a straightforward manner, for instance, extensions exploiting the triangle inequality to speed up the assignment step, or we can simply initialize cluster centers within the subspaces using k-means++ or account for outliers with k-means--.

In our experiments, we show that Nr-Kmeans is a fast algorithm that, at the same time, yields results of a very high clustering quality with a high non-redundancy. Further, we show that the simultaneous optimization of both clustering and subspace is superior to an incremental clustering extraction procedure harnessed by some of the comparison methods. In short, we can say that Nr-Kmeans outperforms the comparison methods and discovers highly relevant combinations of subspaces and clusterings.

Bibliography

[Ma18] Mautz, Dominik; Ye, Wei; Plant, Claudia; Böhm, Christian: Discovering Non-Redundant K-means Clusterings in Optimal Subspaces. In (Guo, Yike; Farooq, Faisal, eds): Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018. ACM, pp. 1973–1982, 2018.