Project Proposal

Discovering Non-Redundant K-means Clusterings in Optimal Subspaces

Rahul Alapati*, Sathish Akula, Aditya Mahajan
Department of Computer Science and Software Engineering
Auburn University, Auburn, AL, 36849
* rza0037@auburn.edu

Objective of the Proposal

Data Clustering is a popular approach of grouping a set of data objects, in such a way that objects in the same group are more similar to each other than to those in other groups. Clustering on a huge object collection in high-dimensional space can often be done in more than one way, for instance, objects could be clustered by their shape or alternatively by their color. Non-redundant clustering addresses this class of problems.

The objective of this project is to investigate the challenge of discovering multiple interesting non redundant k-means clusterings in different subspaces. The basic idea is to find multiple mutually orthogonal subspaces and optimize the objective function of classical k-means in all of them. This would result in a rigorous mathematical treatment of the non-redundant clustering problem and thus an efficient algorithm, called NR-KMEANS[1]. In addition, a noise subspace, orthogonal to all the other subspaces would be introduced to capture all the unimodal variance in data. This property will help NR-KMEANS to outperform the state of the art non-redundant clustering methods, especially on high-dimensional data.

Algorithm

The existing state of the art approaches to non-redundant subspace clustering[2-5] suffer from one or more of the following drawbacks:

1. The algorithms may require many input parameters,
2. They may cause massive runtime and
3. They may produce massive amounts of results which are difficult to interpret.

Our algorithm NR-KMEANS (for Non-redundant K-means) addresses all the above mentioned drawbacks. The following are the key contributions of the NR-KMEANS algorithm:

- **Multiple interesting k-means Clusterings in Optimal Subspaces**: NR-KMEANS discovers multiple, non-redundant k-means clusterings in orthogonal subspaces.
- **Efficiency**: NR-KMEANS is easy to implement and compatible with many proposed extensions of k-means. It is fast, even without sophisticated performance optimizations.
- **Lightweight Parameterization**: The only input parameter of NR-KMEANS is the number of clusters for each subspace.
- **Noise Handling:** Unlike the existing non-redundant clustering approaches, NR-KMEANS includes the idea of noise subspace, which helps it outperform all the existing methods.

The NR-KMEANS algorithm is a simple extension of the well-known Lloyd's algorithm, with its alternating assignment and update steps. We extend the basic idea of classic k-means algorithm by making some assumptions and determine the cost function as follows:

$$F = \sum_{j=1}^{S} \sum_{i=1}^{k_j} \sum_{x \in C_{j,i}} ||P_j^T V^T x - P_j^T V^T \mu_{j,i}||^2$$

where S is the Number of Subspaces, k_j is the Number of Clusters in the j^{th} subspace, $C_{j,i}$ is the Objects of cluster i in subspace j, P_j is the Projection onto the j^{th} subspace, V is the orthogonal matrix of a rigid transformation x is the object of the dataset and $\mu_{j,i}$ is the original space mean of cluster i in subspace j.

We optimize this cost function with a modified version of Lloyd's algorithm.

Datasets to be evaluated

As a part of our experiments, we plan to evaluate our NR-KMEANS algorithm on the following datasets:

ALOI-2Sub[6], Stickfigures[7], Fruits[8], Syn3Sub, Spam and Shuttle.

ALOI-2Sub, Stickfigures, Fruits, Syn3Sub are special non-redundant datasets and contain more than one class labels. They can be obtained from http://dmm.dbs.ifi.lmu.de/downloads, as a part of the NR-KMEANS implementation package.

Spam and Shuttle can be obtained from the UCI repository[9] and they only contain a single set of class labels.

Evaluation Measures to be reported

As a result of our experiments, we plan to report the following evaluation measures:

Pair Counting F1-measure (pc-F1) to account for multiple label sets on both sides.

Average of the Variation Information metric ($\emptyset VI$) to measure the redundancy among the found subspaces.

We also plan to report the runtime of our algorithm in seconds, over the number of data points as well as the number of dimensions.

List of References

[1] Dominik Mautz, Wei Ye, Claudia Plant, and Christian Böhm. 2018. Discovering Non-Redundant K-means Clusterings in Optimal Subspaces. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). ACM, New York, NY, USA, 1973-1982.

[2] Dominik Mautz, Wei Ye, Claudia Plant, and Christian Böhm. 2017. Towards an Optimal Subspace for K-Means. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017. 365–373.

[3] Ying Cui, Xiaoli Z. Fern, and Jennifer G. Dy. 2007. Non-redundant Multi-view Clustering via Orthogonalization. In Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), October 28-31, 2007, Omaha,Nebraska,USA.133–142.

[4] Emmanuel Müller, Ira Assent, Stephan Günnemann, Ralph Krieger, and Thomas Seidl. 2009. Relevant Subspace Clustering: Mining the Most Interesting Nonredundant Concepts in High Dimensional Data. In ICDM 2009, The Ninth IEEE International Conference on Data Mining, Miami, Florida, USA, 6-9 December 2009. 377–386.

[5] Wei Ye, Samuel Maurus, Nina Hubig, and Claudia Plant. 2016. Generalized Independent Subspace Clustering. In Data Mining (ICDM), 2016 IEEE 16th International Conference on. IEEE, 569–578.

[6] http://aloi.science.uva.nl/

[7] Stephan Günnemann, Ines Färber, Matthias Rüdiger, and Thomas Seidl. 2014. SMVC: semi-supervised multi-view clustering in subspace projections. In The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014. 253–262.

[8] Juhua Hu, Qi Qian, Jian Pei, Rong Jin, and Shenghuo Zhu. 2015. Finding Multiple Stable Clusterings. In 2015 IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, NJ, USA, November 14-17, 2015. 171–180.

[9] https://archive.ics.uci.edu/ml/datasets.html

[10] Gabriela Moise and Jörg Sander. 2008. Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008. 533–541.

[11] Ira Assent, Ralph Krieger, Emmanuel Müller, and Thomas Seidl. 2008. INSCY: Indexing Subspace Clusters with In-Process-Removal of Redundancy. In Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008),

[12] Donglin Niu, Jennifer G. Dy, and Michael I. Jordan. 2010. Multiple Non-Redundant Spectral Clustering Views. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel. 831–838.

[13] Nina Hubig and Claudia Plant. 2017. Information-Theoretic Non-redundant Subspace Clustering. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 198–209.

[14] Eric Bae and James Bailey. 2006. COALA: A Novel Approach for the Extraction of an Alternate Clustering of High Quality and High Dissimilarity. In Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China.

[15] Ian Davidson and Zijie Qi. 2008. Finding alternative clusterings using constraints. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, 773–778.

[16] Xuan Hong Dang and James Bailey. 2010. A hierarchical information theoretic technique for the discovery of non linear alternative clusterings. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010. 573–582.

[17] Sen Yang and Lijun Zhang. 2016. Non-redundant multiple clustering by nonnegative matrix factorization. Machine Learning (2016), 1–18.

[18] Xuan-Hong Dang and James Bailey. 2010. Generation of Alternative Clusterings Using the CAMI Approach. In Proceedings of the SIAM International Conference on Data Mining, SDM 2010, April 29 - May 1, 2010, Columbus, Ohio, USA. 118–129.

[19] Nguyen Xuan Vinh and Julien Epps. 2010. minCEntropy: A Novel Information Theoretic Approach for the Generation of Alternative Clusterings. In ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010. 521–530.

[20] Stephan Günnemann, Ines Färber, and Thomas Seidl. 2012. Multi-view clustering using mixture models in subspace projections. In The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012. 132–140.