Less is More: Non-Redundant Subspace Clustering

Ira Assent [◦] Emmanuel Müller [•] Stephan Günnemann [•] Ralph Krieger [•] Thomas Seidl [•]

[◦]Department of Computer Science
Aalborg University, Denmark
ira@cs.aau.dk

[•]Data Management and Data Exploration Group
RWTH Aachen University, Germany
{mueller, guennemann, krieger, seidl}@cs.rwth-aachen.de

ABSTRACT

Clustering is an important data mining task for grouping similar objects. In high dimensional data, however, effects attributed to the "curse of dimensionality", render clustering in high dimensional data meaningless. Due to this, recent years have seen research on subspace clustering which searches for clusters in relevant subspace projections of high dimensional data. As the number of possible subspace projections is exponential in the number of dimensions, the number of possible subspace clusters can be overwhelming.

In this position paper, we present our work on identifying non-redundant, relevant subspace clusters which reduce the result set to a manageable size. We discuss techniques for evaluating, visualizing and exploring subspace clusterings, and propose some directions for future work.

1. INTRODUCTION

Clustering groups data such that similar objects are within a group, while separating dissimilar ones. As the dimensionality of the data grows, distances are known to loose discriminative power [6]. Global dimensionality reduction removes dimensions which are considered irrelevant overall. However, relevance of attributes is not necessarily a global property. By reducing the dimensionality, only a single projection is retained and information in the remaining projections is lost.

In order to identify locally relevant subspace projections, research on subspace and projected clustering has focused on analyzing this relevance per cluster. The result R of these algorithms is therefore a set of clusters C_i with their respective subspace projections S_i: $R = \{(C_1, S_1), \ldots, (C_n, S_n)\}$, where each $S_i \subset D$, D being the dimensions of the full data space. As the number of possible subspace projections is exponential in the number of dimensions, the number of possible subspace clusters is typically overwhelming. Moreover, many clusters might be reported multiple times in different subspace projections, with little or no new information. This motivates our research on identifying only non-redundant, relevant subspace clusters.

2. NON-REDUNDANT SUBSPACE CLUSTERING

In our density-unbiased subspace clustering model (DUSC), we introduced the notion of redundant subspace clusters [3]. A redundant subspace cluster is defined as a cluster which is repeated (by a factor r in terms of the number of objects) in a higher dimensional subspace cluster. This is motivated by the observation that users are generally only interested in seeing a lower dimensional subspace cluster if the number of new objects in this lower dimensional projection is sufficiently large. We claim that more dimensions describe a more specific and thereby informative pattern than the trivial patterns in only few dimensions. DUSC provides a simple, but effective model for removing repeated results.

Another aspect to redundancy removal is that computing all subspace clusters in all subspace projections is computationally intensive. In our INSCY algorithm (indexing subspace clusters with in-process-removal of redundancy), we propose to avoid generation of redundant subspace clusters to reduce runtime [4]. The idea is to index approximate representations of potential subspace clusters. Furthermore, INSCY introduces a novel processing scheme by processing subspaces in depth-first order. The index is traversed such that high dimensional subspace clusters are detected first, and thus, depth-first processing enables pruning of low dimensional redundant representations. Using our conservative approximation introduced in our efficient density-based subspace clustering (EDSC) approach, this algorithm is known to be efficient and complete, i.e. no non-redundant subspace cluster is missed [2].

In our most recent approach for non-redundant subspace clustering, we have proposed a more general notion of redundancy. The RESCU (relevant subspace clustering) model [11], is based on the idea of a global redundancy elimination. In contrast to the redundancy notion [3] which compares only two subspace clusters at a time (locally), RESCU determines an overall optimal clustering result R. We formulate a global optimization problem that selects those subspace clusters which are both non-redundant (i.e. are not repeated in other clusters), and interesting (i.e. fulfill the desired clustering criteria to the best extent possible). Please note that redundancy in this case is defined with respect to all other subspace clusters in the result set. Thus, we tackle redundant subspace clusters which are merely split up into other subspace clusters in a similar subspace projection.

Finally, with OutRank [9], we have demonstrated the usefulness of non-redundant subspace clustering for the ranking of outliers hidden in subspace projections.

3. EVALUATION, EXPLORATION AND VISUALIZATION

Understanding the results of subspace clustering is a challenge in itself. Since clusters are detected with respect to different subspaces, they are difficult to represent and analyze. In order to facilitate the analysis and evaluation of different subspace clusters, and to compare them in a meaningful manner, we have developed OpenSubspace, an open source framework containing implementations of major subspace and projected clustering algorithms [10]. This framework can be easily extended with further algorithms, making it easy to adapt to the needs of its users. This framework has been used successfully to perform a thorough performance evaluation of state-of-the-art approaches in [12].

Another issue in evaluating subspace and projected clustering is also found in traditional clustering: being unsupervised, there is generally no ground truth. This makes it difficult to assess the qualitative performance. In addition to traditional clustering, however, there are many different aspects that could be considered in comparing a result set with e.g. synthetic data as ground truth. This has led to a wide variety of quality measures being used in the literature, making these reported results incomparable. In our Pleiades system [5], now integrated in OpenSubspace, we included several measures for the evaluation and comparison of subspace and projected clustering to assess the performance of various algorithms under all of these measures.

For visual exploration of subspace clustering results, we have devised techniques that allow users to gain an overview over the result, and to inspect detailed properties as needed in our VISA (Visual Subspace Clustering Analysis) approach [1]. By using a color and size coding scheme, along with a bracketing method, users can visually understand the effects of parameterizations of different algorithms. This makes it possible to e.g. select the right level of redundancy in subspace clustering in order to include all desired information, but without being overwhelmed. Our Morpheus system [8], also integrated in OpenSubspace, builds on the VISA methods to provide users with the possibility of navigating through the subspace clustering results, both in a 2-dimensional or 3-dimensional representation. This makes it possible to explore the subspace clusters interactively, and to select detailed information on interesting patterns.

4. FUTURE WORK

Open research questions in non-redundant subspace clustering arise with respect to the notion underlying redundancy as such. As discussed above, we have studied approaches that compare subspace clusters in a pair-wise approach or more generally by defining an overall optimization criterion. An issue that still needs to be resolved is the instantiation of any generalized optimization criterion. Depending on the data, domain-specific notions of interestingness or novelty (which can be considered the opposite of redundancy) arise. Balancing the degree of acceptable redundancy and interestingness is typically a non-trivial task. For some applications a strict removal of clusters is not desired. Since multiple views on the objects are possible, i.e. each object can belong to several clusters, redundancy elimination could misleadingly identify just a single view. Thus, recent approaches such as OSCLU [7] determine multiple and orthogonal groupings within the data. Whether an alternative grouping represents novel knowledge or describes redundant and trivial information, is difficult to assess and requires further new methods in the area of non-redundant subspace clustering. We claim that interactive techniques, which base on visualization of redundancy, and which are devised specifically for comparative analysis could be beneficial in advancing this research issue. This implies that alternative solutions to a given subspace clustering can be shown to enable users to make an informed choice.

Other issues that are important for research in this area are the runtime behavior of more complex models, such as the RESCU technique. Clearly an overall optimization problem is difficult to solve in practice, and more efficient algorithmic solutions are necessary. This is crucial for being able to compute alternative subspace clusterings for interactive exploration within reasonable response times.

Acknowledgment

This work has been supported in part by the UMIC Research Centre, RWTH Aachen University, Germany.

5. REFERENCES

[1] I. Assent, R. Krieger, E. Müller, and T. Seidl. VISA: visual subspace clustering analysis. *SIGKDD Expl., Visual Analytics*, 9(2):5–12, 2007.

[2] I. Assent, R. Krieger, E. Müller, and T. Seidl. EDSC: efficient density-based subspace clustering. In *CIKM*, pages 1093–1102, 2008.

[3] I. Assent, R. Krieger, E. Müller, and T. Seidl. DUSC: Dimensionality unbiased subspace clustering. In *ICDM*, pages 409–414, 2007.

[4] I. Assent, R. Krieger, E. Müller, and T. Seidl. INSCY: Indexing subspace clusters with in-process-removal of redundancy. In *ICDM*, pages 719–724, 2008.

[5] I. Assent, E. Müller, R. Krieger, T. Jansen, and T. Seidl. Pleiades: Subspace clustering and evaluation. In *ECML PKDD*, pages 666–671, 2008.

[6] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbors meaningful. In *ICDT*, pages 217–235, 1999.

[7] S. Günnemann, E. Müller, I. Färber, and T. Seidl. Detection of orthogonal concepts in subspaces of high dimensional data. In *CIKM*, pages 1317–1326, 2009.

[8] E. Müller, I. Assent, R. Krieger, T. Jansen, and T. Seidl. Morpheus: Interactive exploration of subspace clustering. In *KDD*, pages 1089–1092, 2008.

[9] E. Müller, I. Assent, U. Steinhausen, and T. Seidl. Outrank: ranking outliers in high dimensional data. In *DBRank at ICDE*, pages 600–603, 2008.

[10] E. Müller, I. Assent, S. Günnemann, T. Jansen, and T. Seidl. OpenSubspace: An open source framework for evaluation and exploration of subspace clustering algorithms in WEKA http://dme.rwth-aachen.de/opensubspace. 2009.

[11] E. Müller, I. Assent, S. Günnemann, R. Krieger, and T. Seidl. Relevant Subspace Clustering: mining the most interesting non-redundant concepts in high dimensional data. In *ICDM*, pages 377–386, 2009.

[12] E. Müller, S. Günnemann, I. Assent, and T. Seidl. Evaluating clustering in subspace projections of high dimensional data. *PVLDB*, 2(1):1270–1281, 2009.