

Discovering Multiple Clustering Solutions: Grouping Objects in Different Views of the Data

Emmanuel Müller[◦] Stephan Günnemann[•] Ines Färber[•] Thomas Seidl[•]

[◦]Karlsruhe Institute of Technology
emmanuel.mueller@kit.edu

[•]RWTH Aachen University
{guennemann, faerber, seidl}@cs.rwth-aachen.de

Abstract—Traditional clustering algorithms identify just a single clustering of the data. Today’s complex data, however, allow multiple interpretations leading to several valid groupings hidden in different views of the database. Each of these multiple clustering solutions is valuable and interesting as different perspectives on the same data and several meaningful groupings for each object are given. Especially for high dimensional data, where each object is described by multiple attributes, alternative clusters in different attribute subsets are of major interest.

In this tutorial, we describe several real world application scenarios for multiple clustering solutions. We abstract from these scenarios and provide the general challenges in this emerging research area. We describe state-of-the-art paradigms, we highlight specific techniques, and we give an overview of this topic by providing a taxonomy of the existing clustering methods. By focusing on open challenges, we try to attract young researchers for participating in this emerging research field.

Keywords:

data mining; disparate clustering; alternative clustering; subspace clustering; multi-view clustering

Tutorial Slides:

<http://dme.rwth-aachen.de/DMCS>

I. MOTIVATION

In today’s applications, data is collected for multiple analysis tasks. Thus, for each object one gathers many measurements in one large and high dimensional database to provide a large variety of information. In such scenarios one typically observes that each object can participate in various groupings, i.e. objects fit in different roles. For example, in customer segmentation, we observe for each customer multiple possible behaviors which should be detected as clusters. In other domains, such as sensor networks each sensor node can be assigned to multiple clusters according to different environmental events. In gene expression analysis, objects should be detected in multiple clusters due to the various functions of each gene. In general, multiple groupings are desired as they characterize different (unknown) views of the data. In this tutorial we focus on clustering paradigms to detect such *multiple clustering solutions* and provide a thorough discussion on specific approaches found in the literature.

In particular, we highlight the difference to traditional clustering techniques: In general, clustering techniques group similar objects in one group and separate dissimilar objects in different groups. However, traditional instantiations (e.g. the

well known *k*-means algorithm) provide only a *single clustering solution*. For a cluster analysis of the aforementioned complex data, they show two main drawbacks: (1) They aim at a single partitioning of the data and assign each object to exactly one cluster. (2) They output a single clustering (e.g. one set of *k* clusters) forming the resulting groups of objects.

In contrast, we discuss the principle of *multiple clustering solutions*. For one data set multiple sets of clusters (so-called *clusterings*) are specified, i.e. each object is clustered w.r.t. multiple views on the database. This provides more insights than only a single solution. Ideally, each clustering describes a different view on the data. As main objectives for multiple clusterings we observe:

- Each object is grouped in multiple clusters, representing different perspectives on the data.
- The result consists of many alternative solutions. Users may choose one or use multiple of these solutions.
- Clusterings differ to a high extend, and thus, each of these solutions provides additional knowledge.

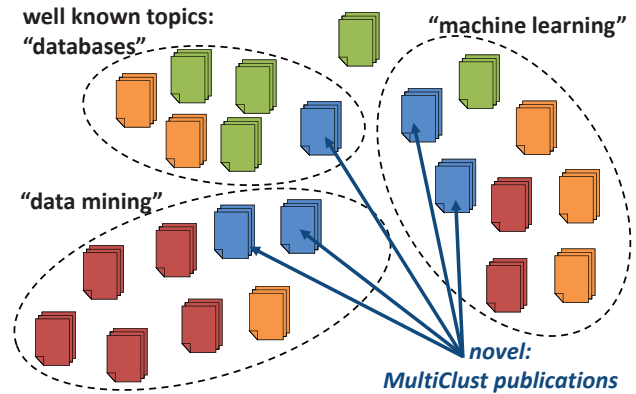


Fig. 1. Example: multiple clustering solutions

As mentioned before, these objectives are motivated by various application scenarios. Let us highlight this in one example: Considering topic analysis, publications concerning the topic of clustering can be grouped in multiple ways (cf. Figure 1). One specific partitioning is obvious to all of us: The origin of such publications in “databases”, “machine learning” and “data mining” are well known and form three major clusters. However, publications can be grouped w.r.t. various other perspectives (e.g. publications aiming at novel theoretical

models vs. tackling scalability issues, considering distributed computation vs. centralized computation, and many more). There are many orthogonal perspectives to the same set of publications. Thus, clustering algorithms should output multiple clusterings that represent these different views. It would be great to have highly differing clusterings that show novel and unknown research trends. In our example, the interdisciplinary topics of “multiple clustering solutions” is such an alternative cluster hidden in today’s publications. Given the well known clustering of publications in the three basic research areas one should aim for the detection of such alternative clusterings. They provide a significant contribution for knowledge discovery as they highlight yet unknown perspectives. Thus, the research area of multiple clustering solutions enforces the detection of alternative and highly differing clusterings.

II. GENERAL CHALLENGES

Abstracting from such scenarios we observe several general challenges that can be derived out of the application demands. In general, they occur due to data integration and merging of multiple data sources. Users try to provide a more complete picture on each object. Huge databases are gathered adding more and more information into existing databases. They extend the stored information and may lead to huge data dumps. Relations between individual tables get lost and different views are merged into one universal view on the data. In these high dimensional data spaces many different views on the data may exist. However, they are hidden in low dimensional projections of the overall data space. Specific algorithms have been designed that cope differently with these hidden views.

Challenge 1: Hidden views in the data space

As first challenge, we observe the identification of *relevant views* (also named *projections*, *subspaces*, or *space transformations*). The goal for this paradigm is the detection of different clusterings revealed by different views on the data. Therefore, besides the mere object groupings, one challenge is the uncovering of different views (e.g. lower dimensional projections) on the data.

Challenge 2: Selection of the relevant views

Different views enable the detection of multiple clustering solutions in high dimensional data. However, they also pose novel challenges to clustering models and require novel objective functions to select these relevant views. Since the obtained clusterings should differ to a high extend, one challenge addresses the search for *highly differing clusterings* (also named *disparate clustering*, *alternative clustering*, or *orthogonal clustering*). In addition to (dis-)similarity functions between objects, as for traditional clustering, this requires novel (dis-)similarity functions between clusters and clusterings.

Challenge 3: Given knowledge about known clustering

The difference between clusterings is even more important if one has some given knowledge, e.g. about a trivial and well known clustering solution. Novel clustering results should reflect a different perspective on the data and should not output any of the already known cluster structures. For example,

the grouping of object pairs for the given vs. the detected clustering should be highly dissimilar to each other.

Challenge 4: Processing schemes for clustering

In contrast to the computation of a single clustering solution, multiple clusterings have to output a set of solutions. Each of these solutions might base on the previous ones in an iterative processing or might be the result of an optimization process that outputs all results simultaneously. Overall, the general challenge is to provide a processing scheme, which computes multiple clustering solutions that contain high quality clusters and are highly differing to each other.

Challenge 5: Number of multiple clusterings

An important issue is the number of solutions, which can be a parameter value or obtained by the optimization inside the algorithm. Many alternative solutions seem valuable for the user, however, to many solutions might provide only redundant results. Thus, redundancy elimination and automatic selection of only the most interesting solutions pose a major challenge to this research area.

Challenge 6: Flexibility w.r.t. clustering model

Based on several decades of research in clustering, one tries to be as general as possible with novel solutions. Thus, multiple clustering should be flexible w.r.t. the underlying clustering model. General processing schemes, data structures, and dissimilarity models are desired. Exchanging the underlying clustering definition (e.g. using density-based, hierarchical or spectral clustering) might be essential for the applicability of multiple clustering solutions.

III. DIFFERENT PARADIGMS IN OUR TAXONOMY

As first step for characterization and overview of existing approaches, we provide a taxonomy of paradigms and methods. With the tutorial we cover several clustering paradigms and highlight their differences in the detection of multiple clustering solutions. We propose a novel taxonomy, which characterizes each approach according to multiple criteria that have been derived from the mentioned challenges in the previous section. Based on this taxonomy we characterize the main clustering models found in the literature: Despite others, we discuss techniques for modeling views in the data, defining similarity measures between clusterings, and detecting alternatives to given knowledge. As primary characteristic in our taxonomy we distinguish each clustering algorithm according to the underlying data space by using three taxonomic classes.

Primary characteristic of our taxonomy

- Perspective w.r.t. the underlying data spaces:
 - original data space
 - orthogonal space transformations
 - different subspace projections

As first paradigm, multiple clusterings have been proposed working in the original data space [3], [2], [11], [17], [8], [9], [26]. Most of these techniques focus on the distinction of different clustering solutions. Both iterative and simultaneous approaches have been proposed for computation of disparate

clusterings. In this area, enhanced techniques have been developed for the detection of alternative solutions vs. some given knowledge.

In parallel to this research direction several researchers have focused on space transformations, i.e. orthogonalization of space [10], [24], [6], [7]. As key idea they perform clustering and orthogonalization steps in an interleaved fashion. Based on a given clustering solution one computes orthogonal spaces. While original data space is more appropriate for the given clusters, the orthogonal space reveals novel cluster structures. More and more clusterings can be computed in an iterative processing.

A third paradigm focuses on the selection of subspace projections [1], [25], [18], [21], [19], [15], [14], [20], [4], [23]. Clusters are detected w.r.t. a set of relevant dimensions. Thus, each cluster is described by a set of clustered objects and an individual subspace projection. In most cases, a simultaneous processing computes a large number of redundant subspace clusters. Further optimization techniques have to be applied for the selection of the final clustering result. Only few techniques in this area have considered given knowledge or disparate clustering models.

Overall, we structure the tutorial according to this primary taxonomy but there are further perspectives revealing different characteristics to be discussed. Each of these perspectives copes with one of the challenges 1 – 6 in the detection of multiple clustering solutions and has been addressed by different approaches. A selection of secondary characteristics is listed below. Together with our primary characterization, they highlight the main research directions found in the literature. In Figure 2 we give a brief summary of presented techniques w.r.t. their characterization in all of these taxonomic classes.

Secondary characteristics of our taxonomy

- Perspective w.r.t. given knowledge:
 - no clustering given
 - one or multiple clusterings given
- Perspective w.r.t. cluster computation:
 - iterative computation
 - simultaneous computation
- Perspective w.r.t. view detection:
 - views are given
 - detect individual view per cluster
 - detect common view per clustering

As main focus of the tutorial, we emphasize general challenges w.r.t. all taxonomic classes, distinguish the different solutions and summarize the open challenges not yet addressed in the literature. Thus we do not only cover a discussion of basic solutions but also identify open issues to be addressed in the future. Especially, these open issues might attract young researchers for participating in this emerging research field.

IV. OPEN CHALLENGES

By discussing the different perspectives on multiple clustering solutions, we derive several open research questions.

Especially, we highlight that most approaches provide enhanced solutions only in one of the mentioned perspectives. They propose very specific solutions to a single challenge (cf. Section II). For example, some techniques focus only on the iterative detection of a single alternative to a given clustering. Thus, they miss to optimize the overall result set. Clearly, such specialized solutions do not address all challenges in this research area. Unfortunately, more general techniques tackling several challenges in more abstract and flexible solutions are very rare and are yet to be developed. In our tutorial we discuss essential combinations of challenges to be tackled in the future:

- Challenge 1 & 2:
Enhanced view selection w.r.t. (dis-)similarity measures
- Challenge 3 & 4:
Simultaneous computation with given knowledge
- Challenge 3 & 5:
Optimal coverage of alternative clusterings
- Challenge 2 & 6:
Exchangeable cluster models by decoupling view selection and cluster detection

Let us discuss only the last combination in more details. In most techniques, view detection and multiple clusterings are tightly bound to the underlying cluster definition. However, such specialized algorithms are hard to adapt (e.g. to application demands). A general aim is to provide a decoupling of such tight bounds. In particular, the selection of views as proposed by many subspace clustering algorithms could be decoupled from the underlying clustering models. Some *subspace search* techniques have proposed first ideas into this direction [4], [23]. However, still some dependencies are incorporated in these techniques. In contrast, an ideal subspace search should utilize common objectives of view selection, independent of the underlying cluster definition.

Challenge 7:

Scalability w.r.t. database size and dimensionality

In addition to these open issues, we observe two orthogonal challenges that found only minor attention in this research area. First, scalability w.r.t. database size and dimensionality is one of these issues. Since the first subspace clustering technique [1] several enhanced models have been developed. However, scalability to large and high dimensional databases is still an open research issue to be addressed for this research area. A recent evaluation study [20] has shown major scalability drawbacks for state-of-the-art subspace clustering techniques. Thus, recently scalability gets more attention and some scalable techniques for very large and high dimensional datasets have been proposed [5], [13], [22].

Challenge 8:

Comparability and quality assessment

Second, we observe open challenges in comparability and quality assessment. In recent years, the importance of comparison studies and repeatability of experimental results is increasingly recognized in the databases and knowledge discovery communities. VLDB initiated a special track on *Experiments*

| publication | space | processing | given know. | # clusterings | subspace detection | flexibility |
|-------------|-------------|--------------|------------------|---------------|--------------------|-------------------|
| [3] | original | | | $m \geq 2$ | | exchangeable def. |
| [2] | original | iterative | given clustering | $m = 2$ | | specialized |
| [12] | original | iterative | given clustering | $m = 2$ | | specialized |
| [17] | original | simultaneous | no | $m \geq 2$ | | specialized |
| [16] | original | simultaneous | no | $m = 2$ | | specialized |
| [8] | original | simultaneous | no | $m \geq 2$ | | specialized |
| [10] | transformed | iterative | given clustering | $m = 2$ | dissimilarity | exchangeable def. |
| [24] | transformed | iterative | given clustering | $m = 2$ | dissimilarity | exchangeable def. |
| [6] | transformed | iterative | given clustering | $m \geq 2$ | dissimilarity | exchangeable def. |
| [1] | subspace | | no | $m \geq 2$ | no dissimilarity | specialized |
| [25] | subspace | | no | $m \geq 2$ | no dissimilarity | specialized |
| [19] | subspace | simultaneous | no | $m \geq 2$ | no dissimilarity | specialized |
| [21] | subspace | simultaneous | no | $m \geq 2$ | no dissimilarity | specialized |
| [15] | subspace | simultaneous | no | $m \geq 2$ | dissimilarity | specialized |
| [14] | subspace | simultaneous | given clustering | $m \geq 2$ | dissimilarity | specialized |
| [4] | subspace | | no | $m \geq 2$ | no dissimilarity | specialized |
| [23] | subspace | | no | $m \geq 2$ | dissimilarity | exchangeable def. |

Fig. 2. Overview of techniques w.r.t. our taxonomy

and Analyses and conferences such as SIGMOD and SIGKDD have established guidelines for repeatability of scientific experiments in their proceedings. Authors are encouraged to provide implementations and data sets. While these are important contributions towards a reliable empirical research foundation, there is still a lack of open source implementations, benchmark data, and evaluation criteria for multiple clusterings. In particular, the community should strive for a common quality assessment to establish a fair and comparable evaluation of multiple clustering solutions.

V. OVERVIEW OF TUTORS' RESEARCH INTERESTS

Our main research interests cover efficient data mining techniques, non-redundant and orthogonal clustering in subspace projections as well as clustering of complex data. In the past years, we initiated the open-source initiative *OpenSubspace*, a unified repository of subspace clustering paradigms. Especially, in combination with our recent comparative evaluation study, it provides a general benefit for the research community. With this tutorial we reveal the relations between several recent mining paradigms and initiate common research directions on this emerging topic.

ACKNOWLEDGMENT

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

REFERENCES

- [1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in *SIGMOD*, 1998, pp. 94–105.
- [2] E. Bae and J. Bailey, "Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity," in *ICDM*, 2006, pp. 53–62.
- [3] R. Caruana, M. F. Elhawary, N. Nguyen, and C. Smith, "Meta clustering," in *ICDM*, 2006, pp. 107–118.
- [4] C.-H. Cheng, A. W. Fu, and Y. Zhang, "Entropy-based subspace clustering for mining numerical data," in *KDD*, 1999, pp. 84–93.
- [5] R. L. F. Cordeiro, A. J. M. Traina, C. Faloutsos, and C. T. Jr., "Finding clusters in subspaces of very large, multi-dimensional datasets," in *ICDE*, 2010, pp. 625–636.
- [6] Y. Cui, X. Z. Fern, and J. G. Dy, "Non-redundant multi-view clustering via orthogonalization," in *ICDM*, 2007, pp. 133–142.
- [7] Y. Cui, X. Z. Fern, and J. G. Dy, "Learning multiple nonredundant clusterings," *TKDD*, vol. 4, no. 3, 2010.
- [8] X. H. Dang and J. Bailey, "Generation of alternative clusterings using the cami approach," in *SDM*, 2010, pp. 118–129.
- [9] X. H. Dang and J. Bailey, "A hierarchical information theoretic technique for the discovery of non linear alternative clusterings," in *SIGKDD*, 2010.
- [10] I. Davidson and Z. Qi, "Finding alternative clusterings using constraints," in *ICDM*, 2008, pp. 773–778.
- [11] D. Gondek and T. Hofmann, "Conditional information bottleneck clustering," in *Clustering Large Data Sets at ICDM*, 2003, pp. 36–42.
- [12] D. Gondek and T. Hofmann, "Non-redundant data clustering," in *ICDM*, 2004.
- [13] F. Gullo, C. Domeniconi, and A. Tagarelli, "Advancing data clustering via projective clustering ensembles," in *SIGMOD Conference*, 2011, pp. 733–744.
- [14] S. Günnemann, I. Färber, E. Müller, and T. Seidl, "ASCLU: Alternative subspace clustering," in *MultiClust at KDD*, 2010.
- [15] S. Günnemann, E. Müller, I. Färber, and T. Seidl, "Detection of orthogonal concepts in subspaces of high dimensional data," in *CIKM*, 2009, pp. 1317–1326.
- [16] M. S. Hossain, S. Tadepalli, L. T. Watson, I. Davidson, R. F. Helm, and N. Ramakrishnan, "Unifying dependent clustering and disparate clustering for non-homogeneous data," in *SIGKDD*, 2010.
- [17] P. Jain, R. Meka, and I. S. Dhillon, "Simultaneous unsupervised learning of disparate clusterings," in *SDM*, 2008, pp. 858–869.
- [18] K. Kailing, H.-P. Kriegel, and P. Kroeger, "Density-connected subspace clustering for high-dimensional data," in *SDM*, 2004, pp. 246–257.
- [19] G. Moise and J. Sander, "Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering," in *KDD*, 2008, pp. 533–541.
- [20] E. Müller, S. Günnemann, I. Assent, and T. Seidl, "Evaluating clustering in subspace projections of high dimensional data," *PVLDB*, vol. 2, no. 1, pp. 1270–1281, 2009.
- [21] 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*, 2009, pp. 377–386.
- [22] E. Müller, I. Assent, S. Günnemann, and T. Seidl, "Scalable density-based subspace clustering," in *CIKM*, 2011.
- [23] D. Niu, J. Dy, and M. Jordan, "Multiple Non-Redundant Spectral Clustering Views," in *ICML*, 2010.
- [24] Z. Qi and I. Davidson, "A principled and flexible framework for finding alternative clusterings," in *KDD*, 2009, pp. 717–726.
- [25] K. Sequeira and M. J. Zaki, "Schism: A new approach for interesting subspace mining," in *ICDM*, 2004, pp. 186–193.
- [26] N. X. Vinh and J. Epps, "mincentropy: a novel information theoretic approach for the generation of alternative clusterings," in *ICDM*, 2010.