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

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

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

[○] *Karlsruhe Institute of Technology (KIT), Germany*
emmanuel.mueller@kit.edu

I. Tutorial 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 methods. By focusing on open challenges, we try to attract young researchers for participating in this emerging research field.

Keywords: data mining; subspace clustering; orthogonal clustering; alternative clustering; multiple perspectives

Tutorial Slides: http://dme.rwth-aachen.de/DMCS

II. Motivation

In today's applications, data is collected for multiple analysis tasks. Thus, for each object one gathers many measurements in one 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 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.

III. Covered Paradigms in this Tutorial

Traditional Clustering approaches will be mentioned to highlight their drawbacks in the detection of multiple clustering solutions.

Subspace Clustering has its focus on detecting multiple clusters in arbitrary subspace projections of high dimensional data. Each subspace cluster is associated with an individual set of relevant dimensions in which this object grouping has been discovered. Subspace clustering allows objects to be part of multiple clusters but does not focus on different views of the data.

Orthogonal Clustering actively searches for multiple different views. For each object multiple clustering solutions are detected in these highly differing views. In general, these different views are characterized by orthogonal spaces. Focusing on these views, orthogonalization techniques have been developed to detect novel knowledge for each object.

Alternative Clustering aims at detecting an alternative grouping deviating from a given clustering solution provided by the user. Thus, two complementary views of the data are detected. These approaches are especially useful for recent application scenarios where some given clusters are available to guide the knowledge discovery process.

Consensus Clustering and other paradigms that utilize multiple clusterings to find a common consensus will be covered in an extended version of the tutorial.

IV. 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.