Mach Learn (2015) 98:1–5
DOI 10.1007/s10994-014-5445-0

EDITORIAL

MultiClust special issue on discovering, summarizing and using multiple clusterings

Emmanuel Müller · Ira Assent · Stephan Günnemann · Thomas Seidl · Jennifer Dy

Received: 14 April 2014 / Accepted: 29 April 2014 / Published online: 31 May 2014

Traditionally, clustering has focused on discovering a single grouping of the data. In many applications, however, data is collected for multiple analysis tasks. Several features or measurements provide complex or high dimensional information. In such data, one typically observes several valid groupings, i.e. each data object fits in different roles. In contrast to traditional clustering these *multiple clusterings* describe alternative aspects that characterize the data in different ways. Traditional single clustering solutions can thus be regarded as special cases of multiple clustering solutions, where only a single set of clusters represents one notion of intra-cluster similarity and inter-cluster dissimilarity. The generality of multiple clustering solutions allows capturing multi-faceted information in more than a single similarity notion, while making it more challenging to detect cluster structures.

E. Müller (✉)
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: emmanuel.mueller@kit.edu

E. Müller
University of Antwerp, Antwerpen, Belgium
e-mail: emmanuel.mueller@ua.ac.be

I. Assent
Aarhus University, Aarhus, Denmark
e-mail: ira@cs.au.dk

S. Günnemann
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: sguennem@cs.cmu.edu

T. Seidl
RWTH Aachen University, Aachen, Germany
e-mail: seidl@cs.rwth-aachen.de

J. Dy
Northeastern University, Boston, MA, USA
e-mail: jdy@ece.neu.edu

The topic of multiple clusterings itself shows multiple research aspects: several alternative solutions vs. a single consensus that integrates different views; given views in multi-source clustering vs. detection of novel views by feature selection and space transformation techniques; a virtually unlimited number of alternative solutions vs. a non-redundant output restricted to a small number of disparate clusterings.

Studies in multiple clustering solutions have identified novel challenges in a number of research fields. Examples from the machine learning and knowledge discovery communities include work on frequent itemset mining, ensemble mining, constraint-based mining, theory on summarization of results, or consensus mining to name only a few. Discovering multiple clusterings is an emerging topic that has received significant attention in recent years at top international conferences. The seminal paper on the topic won the best paper award at ICDM 2004 (Gondek and Hofmann 2004) with follow-up work establishing various research directions in meta-clustering (Caruana et al. 2006), alternative clustering (Bae and Bailey 2006), alternative clusterings using constraints (Davidson and Qi 2008), multi-view clustering via orthogonalization (Cui et al. 2007), non-redundant subspace clustering (Assent et al. 2008), disparate clustering (Hossain et al. 2010), and non-redundant spectral clustering views (Niu et al. 2010). All of these research directions on multiple clusterings provide a new way of looking at the clustering problem. Furthermore, as a cross-disciplinary research endeavor multiple clusterings research has recently received significant attention from multiple related communities. The *MultiClust* workshops at KDD 2010, ECML PKDD 2011, SDM 2012 and the tutorial on "Discovering Multiple Clustering Solutions" (Müller et al. 2012) presented at ICDM 2010, SDM 2011, ICDE 2012, ICML 2013 attracted substantial numbers of attendees with different backgrounds.

This special issue is thus organized as a dedicated platform to reflect recent achievements in multiple clustering research, showcasing also the relationship with related research areas, and serving as an incubator for future research. A total of 29 submissions were received, 11 of which were finally accepted for this special issue. Each accepted paper has gone through two to three rounds of reviewing, each round with at least three referees. They illustrate the many facets of this emerging research field, provide an update on mature high-quality work in this area, and establish connections to related research areas. The contents of this special issue cover a variety of aspects on multiple clustering solutions. We structure the contributions into three main categories based on a recent taxonomy (Müller et al. 2012): (1) Alternative Clustering, (2) Clustering in Subspace Projections and (3) Clustering Ensembles.

In general, *Alternative Clustering* aims at detecting an alternative grouping deviating from a given clustering solution provided by the user. Thus, two or more complementary views of the data are detected as alternative clustering solutions. Alternative clustering approaches are especially useful for application scenarios where some given clusters are available to guide the knowledge discovery process. *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. *Clustering Ensembles*, utilizes multiple clusterings to find a common consensus. In contrast to the previous methods, one assumes to have the knowledge about different sources or multiple clusterings given. Approaches in this paradigm try to combine multiple clusterings and derive novel knowledge out of this combination. The publications in this special issue enhance these three paradigms with several novel models and algorithms as follows.

1. Alternative Clustering The paper "A Framework to Uncover Multiple Alternative Clusterings" by Xuan Hong Dang and James Bailey introduces a novel framework to discover multiple alternative clusterings of a dataset. The problem is formulated as an optimization of two objectives (combined in one function): the maximization of a mixture model to obtain accurate solutions, and the minimization of the information shared by the clustering solutions, in order to find diverse ones. The authors develop an Expectation Maximization (EM) algorithm to solve the optimization problem, which allows finding the clustering solutions in sequence. The EM based framework provides provable guarantees of convergence and is a principled way of discovering multiple alternate clusterings.

The paper "Subjectively Interesting Alternative Clusterings" by Kleanthis-Nikolaos Kontonasios and Tijl De Bie studies the problem of finding alternative clusterings as an iterative optimization scheme, where each new clustering tries to explain as much as possible of the data not yet explained by the previous clusterings. Specifically, the new clusterings are deemed interesting if they have high self-information (small probability) under the maximum entropy distribution taking into account the previously-found clusterings. The proposed method can handle arbitrary number of previous clusterings and have varying number of clusters in each step. The main contribution of this work is derived from an information theoretic viewpoint in data mining and is a very interesting instantiation of maximum-entropy to alternative clusterings. It can be re-interpreted as a form of "clustering on demand", where one or more new clusters can be produced so as to simultaneously maximize cluster quality and differentiation from previously computed clusters.

The paper "A Flexible Cluster-Oriented Alternative Clustering Algorithm for Choosing from the Pareto Front of Solutions" by Duy Tin Truong and Roberto Battiti proposes a technique for generating alternative clustering that lie on the Pareto front of optimality (not dominated by other clusterings in terms of quality and dissimilarity). This is a novel direction and a natural perspective in the area of alternative clusterings. The goal of alternative clustering is to find a high quality clustering solution that is different from one or more existing 'negative' clustering solutions. This goal is naturally captured by two objectives, the first aiming to maximize the quality of the clustering solution and the second trying to maximize the dissimilarity from the given 'negative' solution(s). This paper presents a multi-objective optimization approach for the alternative clustering problem based on evolutionary algorithms, which generates a collection of pareto-optimal clustering solutions.

2. Clustering in Subspace Projections The paper "Enriched Spatial Comparison of Clusterings Through Discovery of Deviating Subspaces" by Eric Bae and James Bailey proposes a method to compare two clusterings in the context of subspace cluster analysis. The authors introduce the novel problem of mining those subspaces in which two clusterings show similarity or strongly deviate. They introduce the cluster comparison measure ADCO for evaluating the subspace (or spatial) similarity of two clusters. This measure does not exhibit structure easily exploitable for efficient search, and the paper discusses in detail how the ADCO scores can be upper- and lower-bounded such that efficient mining is made feasible. For this purpose, the authors prove monotonicity properties of the bounds when traversing the subspace lattice of high-dimensional spaces. Hence, upper and lower bounds for ADCO allow for an incremental computation with substantial pruning possibilities when traversing the subspace lattice.

The position paper "On Meeting the Problem of Multiple Truths in Data from Clustering and Pattern Mining Perspectives" by Arthur Zimek and Jilles Vreeken reviews important topics and challenging issues in clustering and pattern mining. It uncovers some common traits in how the apparently different problems are currently approached in their corresponding

communities. The authors take on the grand task of drawing similarities between a large range of research tasks. The paper gives a brief overview of related areas of subspace, ensemble, alternative and multi-view clustering with a subsequence discussion on common aspects. The connections that are drawn contribute to a common understanding of a field, and allow the identification of open issues for future research efforts.

The paper "Unsupervised Feature Selection with Ensemble Learning" by Haytham Elghazel and Alex Aussem addresses the selection of features from high dimensional data, where redundant or irrelevant features are a typical issue. Several clusterings are evaluated to determine the relevance of subspaces during the feature selection process. The authors focus on integrating supervised and unsupervised methods of feature selection and reinforce the method through a consensus on multiple clusterings obtained on different set of features. In order to determine the importance of each feature sampling, random selection of features, and a recursive feature elimination method are proposed. A subset of relevant features for each cluster is selected using a technique based on the out-of-bag importance measure. A feature elimination scheme for this task is devised, which recursively removes the features with the lowest importance measure.

3. Clustering Ensembles The paper "Metacluster-based Projective Clustering Ensembles" by Francesco Gullo, Carlotta Domeniconi, and Andrea Tagarelli proposes a new projective clustering ensembles formulation for consensus clustering. Their formulation ensures that a consensus cluster is put in relation with at least one cluster from each ensemble member, and that each cluster within the ensemble is assigned to at least one cluster in the projective consensus clustering. They then propose an enhancement of a prior two-stage, metaclustering-based approach in which greater interaction between cluster and feature optimization is promoted by forcing cluster centers to be associated with at least one cluster from each solution of the ensemble.

The paper "Unsupervised Ensemble Minority Clustering" by Edgar Gonzàlez and Jordi Turmoet addresses a very relevant data analysis problem of identifying the clusters in the presence of noise, which is particularly challenging when the number of points to be clustered is relatively small compared to the amount of noise. In order to approach this issue, the paper develops an approach that uses a combination of multiple weak clusterings in a novel way. Two novel methods to generate weak clusterings are proposed. The first uses Bregman divergence distance, and the second extends a method of random splits of the data set. A new procedure is devised that allows determining the values of the adjustable parameters of the algorithm in an unsupervised manner.

The paper "Comparative Study of Matrix Refinement Approaches for Ensemble Clustering" by Natthakan Iam-On and Tossapon Boongoen discusses a comparative study on summarizing multiple clusterings. As a survey it provides a nice overview on ensemble clustering methods, with an emphasis on approaches based on matrix refinements. Additionally, the authors evaluate both the ensemble member generation phase as well as the aggregation phase where a final result is represented based on the ensemble members. In each step they evaluate different methods providing a comparison of 7 matrix refinement approaches coupled with 8 consensus functions. The evaluation is based on 13 different datasets and methods are evaluated using several evaluation measures.

The paper "Greedy Learning of Latent Tree Models for Multidimensional Clustering" by Teng-Fei Liu, Nevin Zhang, Peixian Chen, April Hua Liu, Leonard Poon, and Yi Wang presents a novel technique for fast learning approximative Latent Tree Models (LTM). Previous learning methods had the disadvantages of long runtime in areas of days / weeks. Therefore, the authors improved the runtime by inventing a novel algorithm following a

bottom-up search for growing LTMs. The authors show that their idea of checking when to stop expanding the working subset by applying LTM to the working subset is very effective for multidimensional clustering.

The paper "Probabilistic Consensus Clustering using Evidence Accumulation" by André Lourenco, Samuel Rota Bulò, Nicola Rebagliati, Ana Fred, Mário Figueiredo, and Marcello Pelillo proposes a probabilistic model for consensus clustering based on evidence accumulation. After obtaining the co-association matrix, the authors build a generative model for consensus clustering. In particular, the number of times two data points are co-clustered is generated from a binomial distribution. The consensus clustering results are not limited to the crisp clustering, but could be mixed membership vectors. Further, the paper also gives another interpretation of the proposed algorithm from Bregman divergence perspective.

We are grateful to all the authors who submitted to this special issue for their high quality work and their interest in discussing their approaches as part of this special issue. We would also like to thank the reviewers for their thoughtful and useful comments that further improved the quality of the articles. We are indebted to the MLJ editor in chief, Peter Flach, for his support and encouragement all the way, and to the staff at the editorial office for their assistance.

Research in multiple clusterings is clearly receiving growing interest. With this special issue, we hope to contribute to its visibility, and to encourage more researchers from related fields to join the effort in exploring multiple views on the increasingly complex data of today.

References

Assent, I., Krieger, R., Müller, E., Seidl, T. (2008). INSCY: Indexing subspace clusters with in-process-removal of redundancy. In: ICDM, pp. 719–724.

Bae, E., Bailey, J. (2006). Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In: ICDM, pp. 53–62.

Caruana, R., Elhawary, M. F., Nguyen, N., Smith, C. (2006). Meta clustering. In: ICDM, pp. 107–118.

Cui, Y., Fern, X. Z., Dy, J. G. (2007). Non-redundant multi-view clustering via orthogonalization. In: ICDM, pp. 133–142.

Davidson, I., Qi, Z. (2008). Finding alternative clusterings using constraints. In: ICDM, pp. 773–778.

Gondek, D., Hofmann, T. (2004). Non-redundant data clustering. In: ICDM.

Hossain, M. S., Tadepalli, S., Watson, L. T., Davidson, I., Helm, R. F., Ramakrishnan, N. (2010). Unifying dependent clustering and disparate clustering for non-homogeneous data. In: SIGKDD.

Müller, E., Günnemann, S., Färber, I., Seidl, T. (2012). Discovering multiple clustering solutions: Grouping objects in different views of the data. In: ICDE, pp. 1207–1210.

Niu, D., Dy, J., Jordan, M. (2010). Multiple Non-Redundant Spectral Clustering Views. In: ICML.