MultiClust 2013: Multiple Clusterings, Multi-view Data, and Multi-source Knowledge-driven Clustering

[Workshop Report]

Ira Assent[1], Carlotta Domeniconi[2], Francesco Gullo[3], Andrea Tagarelli[4], Arthur Zimek[5]

[1]Dept. of Computer Science, Aarhus University, Denmark
ira@cs.au.dk

[2]George Mason University, USA
carlotta@cs.gmu.edu

[3]Yahoo Labs, Spain
gullo@yahoo-inc.com

[4]DIMES, University of Calabria, Italy
tagarelli@dimes.unical.it

[5]Ludwig-Maximilians-Universität München, Germany
zimek@dbs.ifi.lmu.de

ABSTRACT

In this workshop report, we give a summary of the MultiClust workshop held in Chicago in conjunction with KDD 2013. We provide an overview on the history of this workshop series and the general topics covered. Furthermore, we provide summaries of the invited talks and of the contributed papers.

1. INTRODUCTION

Multiple views and data sources require clustering techniques capable of providing several distinct analyses of the data. The cross-disciplinary research topic on multiple clustering has thus received significant attention in recent years. However, since it is relatively young, important research challenges remain. Specifically, we observe an emerging interest in discovering multiple clustering solutions from very high dimensional and complex databases. Detecting alternatives while avoiding redundancy is a key challenge for multiple clustering solutions. Toward this goal, important research issues include: how to define redundancy among clusterings; whether existing algorithms can be modified to accommodate the finding of multiple solutions; how many solutions should be extracted; how to select among far too many possible solutions; how to evaluate and visualize results; and eventually how to most effectively help the data analysts in finding what they are looking for. Recent work tackles this problem by looking for non-redundant, alternative, disparate, or orthogonal clusterings. Research in this area benefits from well-established related areas, such as ensemble clustering, constraint-based clustering, frequent pattern mining, theory on result summarization, consensus mining, and general techniques coping with complex and high dimensional databases. At the same time, the topic of multiple clustering solutions has opened novel challenges in these research fields.

Overall, this cross-disciplinary research endeavor has recently received significant attention from multiple communities. The MultiClust workshop is a venue to bring together researchers from the above research areas to discuss issues in multiple clustering discovery.

MultiClust 2013 was the 4th in a series of workshops. The first MultiClust workshop was an initiative of Xiaoli Fern, Ian Davidson, and Jennifer Dy and was held in conjunction with KDD 2010 [7]. The successful workshop series continued with the 2nd MultiClust workshop at ECML PKDD 2011 [10] and the 3rd MultiClust workshop at SIAM Data Mining 2012 [11]. Additionally, an upcoming special issue of the Machine Learning Journal is dedicated to the MultiClust topics. The aim of this special issue is to establish an overview of recent research, to increase its visibility, and to link it to closely related research areas.

Furthermore, in 2012, the 3Clust workshop was held in conjunction with PAKDD [6]. It had a slightly different perspective but is very related to the MultiClust workshop topics. Therefore, the organizers of 3Clust and some organizers of previous MultiClust workshops teamed up for the 4th

MultiClust workshop at KDD 2013, giving more emphasis not only on emerging issues in the areas of clustering ensembles, semi-supervised clustering, subspace/projected clustering, co-clustering, and multi-view clustering, but in particular on discussing new and insightful connections between these areas. The vision is to make progress towards a unified framework that reconciles the different involved variants of the clustering problem.

2. SUMMARY OF THE WORKSHOP

2.1 Invited Talks

At MultiClust 2013 we had two very inspiring invited talks, by Michael R. Berthold (University of Konstanz, Germany) and by Shai Ben-David (University of Waterloo, Canada).

Michael discussed his approach to learning in "Parallel Universes", with a focus on the application area of bio-chemical and medical research (drug discovery). In this area, data objects are represented in very different, heterogeneous feature spaces and complex data types such as molecular structures or sequences, resulting also in different notions of similarity. Objects that could be similar (and should be clustered) in one of the representations might be very dissimilar in another representation. At the same time, different data representations are of different quality and partly faulty, outdated, unreliable or just noisy.

Learning (or clustering) in parallel universes is similar but also different in some crucial aspects from related approaches to clustering. If the data objects are represented in a high-dimensional but essentially homogeneous, numeric feature space, we have a global similarity measure that can be used for clustering. For multiple feature spaces of heterogeneous nature, many notions of similarity would be required. This is different from feature selection for clustering, where one would choose the most informative or useful subset of attributes. For a specific subset, there is usually no interpretation possible. Feature selection approaches select a subset of features from one, large universe and serve as preprocessing for subsequent learning algorithms. Similarly, projected clustering or subspace clustering selects subsets of features for each cluster however not as a preprocessing but as an integral step of the clustering procedure. Nevertheless, the features of the complete feature space are thought to belong to the same universe and the projected clustering or subspace clustering algorithm works on this complete, single feature space to select appropriate subsets. For clustering in parallel universes, different subsets of features are separated semantically from each other by their different nature in the first place. The most similar notion is multi-view or multi-represented learning; the idea there, however, is that the same concept can be learned in different representations and, especially in the setting of co-learning, the learning process in one representation can help or guide the learning process in another representation. In multi-instance learning, finally, the same object can have different representations in the same feature space, for example, a molecule can have different 3D confirmations.

For the specific approach of learning in parallel universes, Michael discussed some example approaches. Fuzzy cu-Means [13] is an adaptation from the fuzzy k-means family to the setting of parallel universes, where representations in some universes can be completely noisy. For the example of neighborgram clustering [5], Michael demonstrated the possibilities for the domain scientist to understand decisions of the algorithm and gain insights by an illustrative view of the results.

The other invited talk by Shai was focused on the gap between theory and practice in clustering. Although clustering is one of the most widely used tools in data analysis and exploration, it is not clear a priori what a good clustering is for a dataset. For many datasets, different clustering solutions can be equally meaningful. How to turn clustering in an actually well-defined task depends on the application, i.e., the domain expert can add some bias, expressing domain knowledge. How to formalize such a bias is the motivating question for the points Shai was taking in his talk [14; 1; 2]. In particular, he discussed (1) general properties of the input-output functionality of clustering paradigms, (2) quality measures for clusterings, and (3) measures for the clusterability of data.

1. In general terms, if we consider functions that take as input a dissimilarity function over some domain S (or, alternatively, a matrix of pairwise "distances" between points in the domain), and provide as output a partition of S, we would like to have properties that can distinguish "clustering" functions from other functions that output partitions. The ideal theory would define sets of properties to distinguish major clustering paradigms from each other. This could even work hierarchically. Shai showed examples of sets of properties defining single-linkage clustering, and sets of properties defining linkage clustering.

2. A different approach for defining clustering paradigms is given by measures of clustering quality. These can also be analyzed with an axiomatic approach. Shai names the properties "scale invariance", "consistency", "richness", and "isomorphism invariance" as a consistent set of properties and names many clustering quality measures that satisfy these axioms.

3. Finally, clusterability can be seen as applying clustering quality measures on optimal clustering solutions for a dataset.

As can be seen, the two invited talks covered very different perspectives, Michael taking a perspective from his practical application, Shai sharing thoughts from a theoretical point of view. This broadness of aspects is a good reflection of the scope of the MultiClust workshop series.

2.2 Research Papers

The contribution by Li et al. [9] discusses an approach to multi-view clustering based on a Markov Chain Monte Carlo sampling of relevant subspaces. A subspace is a subset of input features, and is considered to be a state of a Markov chain. The neighbors of a given state in the chain are the immediate subsets (one feature removed) and supersets (one feature added). The search in the chain is driven by the assessed quality of the clustering structure in the corresponding subspaces. Furthermore, in order to facilitate the discovery of diverse views of the data, the search is biased in favor of those subspaces that are dissimilar from the previously detected ones.

Clusters in subspaces are detected using the Mean Shift algorithm, which is based on a non-parametric kernel density estimation approach. The quality of a subspace is measured in terms of the density of the clusters discovered therein. A weighting term, measuring the similarity with previously detected subspaces, is added to the density function, with the effect of favoring the sampling of subspaces dissimilar from one another. Two sampling processes are investigated: simulated annealing and greedy local search.
The preliminary results measuring clustering quality are encouraging. Scaling the proposed method to a large dimensionality, and the automatic identification of the number of views, are interesting open challenges for future directions the authors plan to pursue.

Babagholami-Mohamadabadi et al. [4] focus on the problem of distance-metric learning in a semi-supervised context. The problem consists in learning an appropriate metric distance for an input set of points based on a number of must-link and/or cannot-link constraints that are defined over the input points. The main novelty of the approach by Babagholami-Mohamadabadi et al. is that, unlike most existing methods, it can profitably take advantage of the data points that are not involved into any constraints. Based on this intuition, the authors develop a novel linear metric-learning method, which they also kernelize so to develop a non-linear version of the same method. The optimization strategy relies on the Deterministic Annealing EM (DAEM) algorithm, which allows for finding a local maximum of the proposed objective function.

Shiga and Mamitsuka [12] introduce a probabilistic generative approach to co-clustering that enables the embedding of auxiliary information. External information associated to both the rows and columns of the data matrix can be added and incorporated in the inference process. The parameters over the row and column clusters are learned via variational inference using an Expectation Maximization-style algorithm. The authors test the effectiveness of the proposed method using a gene expression dataset. They represent the auxiliary information as graphs that connect genes (or samples) known to be in the same cluster, according to the ground truth. Comparisons against unsupervised Bayesian co-clustering are in favor to the proposed technique, showing the positive effect of embedding the external information. Semi-supervised co-clustering is a relevant approach in a variety of applications, including text mining and recommender systems, where information regarding the users and products allows us to perform prediction for new users.

Kamishima and Akaho [8] distinguish "Absolute and Relative Clustering". This difference is intended to relate to the relationship between the data set and the clustering result. In absolute clustering, the decision to cluster two objects in the same cluster is independent of other objects. In relative clustering, the decision to cluster two objects in the same cluster is depending on other objects, i.e., the clustering task as a whole. The authors present several examples for their intuition. In the discussion, Shai Ben-David questioned the idea by assuming that the class of absolute clustering is probably empty. It would seem, however, that the authors' distinction can be an original approach to think about semi-supervised clustering, where pairwise instance-level constraints indeed specify desired decisions for pairs of objects independently of the remainder of that data set. How a (semi-supervised) clustering approach addresses such constraints would be a different question.

Spectral graph partitioning is the topic addressed in the short paper by Zheng and Wu [15]. The basic motivation for this study is to try overcome an accuracy issue in spectral modularity optimization — the repeated bisection process performed by a traditional spectral modularity optimization algorithm can fail in reaching global optimality due to its greedy nature. In order to take into account the global structure information in a graph, the spectral algorithm proposed by Zheng and Wu aims to find better, multisection divisions of the graph by extending the modularity matrix to a higher order and making use of orthogonal vectors of the Hadamard matrix for the representation of group assignments of the vertices in the graph divisions. The modularity matrix is randomly "inflated" to higher orders through the Kronecker product, as to coordinate with the orthogonal vectors. As a result, the graph can be cut into multiple sections directly. In sparse graphs, the time complexity is $O(K^4n^2)$ for a graph of n vertices, where K is the estimated number of communities.

3. CONCLUSIONS AND OUTLOOK

Clustering is a very traditional data mining task but at the same time provides many new challenges. The MultiClust workshop brings together researchers working at different aspects of the clustering problem with a particular focus on making use of multiple clustering solutions, envisioning a unified framework reconciling and integrating the different aspects of the clustering problem.
A continuation of the MultiClust workshop series is planned as a Mini-Symposium at SIAM Data Mining (SDM) 2014: `http://uweb.dimes.unical.it/multiclust2014/`.

Acknowledgements

We would like to thank all the authors. Their creativity made the workshop a success. Also all participants shared their thoughts in our discussions. In particular we thank the invited speakers for sharing their insights and experience in our topic. Last but not least we acknowledge the good work and effort of all members of the program committee. They provided high quality reviews for our submissions in a tight schedule. The members of the program committee in alphabetical order are:

- James Bailey, University of Melbourne, Australia
- Ricardo J. G. B. Campello, University of São Paulo, Brazil
- Xuan-Hong Dang, Aarhus University, Denmark
- Ines Färber, RWTH Aachen University, Germany
- Wei Fan, IBM T. J. Watson Research Center and IBM CRL, USA
- Ana Fred, Technical University of Lisbon, Portugal
- Stephan Günnemann, CMU, USA
- Dimitrios Gunopulos, University of Athens, Greece

- Michael E. Houle, NII, Japan
- Emmanuel Müller, KIT, Germany
- Erich Schubert, LMU Munich, Germany
- Thomas Seidl, RWTH Aachen University, Germany
- Grigorios Tsoumakas, Aristotle University of Thessaloniki (AUTh), Greece
- Giorgio Valentini, University of Milan, Italy
- Jilles Vreeken, University of Antwerp, Belgium

4. REFERENCES

[1] M. Ackerman and S. Ben-David. Clusterability: A theoretical study. *Journal of Machine Learning Research - Proceedings Track*, 5:1–8, 2009.

[2] M. Ackerman, S. Ben-David, and D. Loker. Towards property-based classification of clustering paradigms. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *NIPS*, pages 10–18. Curran Associates, Inc., 2010.

[3] I. Assent, C. Domeniconi, F. Gullo, A. Tagarelli, and A. Zimek, editors. *4th MultiClust Workshop on Multiple Clusterings, Multi-view Data, and Multi-source Knowledge-driven Clustering, in conjunction with the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA*, 2013.

[4] B. Babagholami-Mohamadabadi, A. Zarghami, H. A. Pourhaghighi, and M. T. Manzuri-Shalmani. Probabilistic non-linear distance metric learning for constrained clustering. In Assent et al. [3], pages 4:1–4:8.

[5] M. Berthold, B. Wiswedel, and D. Patterson. Interactive exploration of fuzzy clusters using neighborgrams. *Fuzzy Sets and Systems*, 149(1):21–37, 2005.

[6] C. Domeniconi, F. Gullo, and A. Tagarelli, editors. *The First International Workshop on Multi-view data, High Dimensionality, and External Knowledge: Striving for a Unified Approach to Clustering, in conjunction with PAKDD 2012, Kuala Lumpur, Malaysia*, 2012.

[7] X. Z. Fern, I. Davidson, and J. G. Dy. MultiClust 2010: discovering, summarizing and using multiple clusterings. *SIGKDD Explorations*, 12(2):47–49, 2010.

[8] T. Kamishima and S. Akaho. Absolute and relative clustering. In Assent et al. [3], pages 6:1–6:6.

[9] G. Li, S. Günnemann, and M. J. Zaki. Stochastic subspace search for top-k multi-view clustering. In Assent et al. [3], pages 3:1–3:6.

[10] E. Müller, S. Günnemann, I. Assent, and T. Seidl, editors. *2nd MultiClustWorkshop: Discovering, Summarizing and Using Multiple Clusterings, in conjunction with ECML PKDD 2011, Athens, Greece*, 2011.

[11] E. Müller, T. Seidl, S. Venkatasubramanian, and A. Zimek, editors. *3rd MultiClust Workshop: Discovering, Summarizing and Using Multiple Clusterings, in conjunction with SIAM Data Mining 2012, Anaheim, CA*, 2012.

[12] M. Shiga and H. Mamitsuka. Variational Bayes coclustering with auxiliary information. In Assent et al. [3], pages 5:1–5:4.

[13] B. Wiswedel and M. R. Berthold. Fuzzy clustering in parallel universes. *International Journal of Approximate Reasoning*, 45(3):439–454, 2007.

[14] R. Zadeh and S. Ben-David. A uniqueness theorem for clustering. In J. Bilmes and A. Y. Ng, editors, *UAI*, pages 639–646. AUAI Press, 2009.

[15] H. Zheng and J. Wu. Spectral graph multisection through orthogonality. In Assent et al. [3], pages 7:1–7:6.