Editorial

Theory and Applications of Data Clustering

Costas Panagiotakis,¹ Emmanuel Ramasso,² Paraskevi Fragopoulou,³ and Daniel Aloise⁴

¹Department of Business Administration, TEI of Crete, 72100 Agios Nikolaos, Greece

²Department of Automatic Control & Micro-Mechatronic Systems and Applied Mechanics Department, FEMTO-ST Institute, UMR CNRS 6174-UBFC/ENSM/UTBM, 25000 Besançon, France

³Department of Informatics Engineering, TEI of Crete, 71004 Heraklion, Greece

⁴Department of Computer Engineering and Automation, Federal University of Rio Grande do Norte, 59072-970 Natal, RN, Brazil

Correspondence should be addressed to Costas Panagiotakis; cpanag@staff.teicrete.gr

Received 17 January 2016; Accepted 17 January 2016

Copyright © 2016 Costas Panagiotakis et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This special issue is particularly focused on fundamental and practical issues in data clustering [1–6]. Data clustering aims at organizing a set of records into a set of groups so that the overall similarity between the records within a group is maximized while minimizing the similarity with the records in the other groups. The data clustering is a state of the art problem with increasing number of applications. For decades, data clustering problems have been identified in many applications and domains such as computer vision and pattern recognition (e.g., video and image analysis for information retrieval, object recognition, image segmentation, and point clustering), networks (e.g., identification of web communities), databases and computing (facing privacy in databases), and statistical physics and mechanics (e.g., understanding phase transitions, vibration control, and fracture identification using acoustic emission data). In addition, several definitions and validation measures [3, 7] of data clustering problem have been used on different applications in engineering. For instance, the goal of the classical clustering problem is to find the clusters that optimize a predefined criterion while the goal of the microaggregation problem [8] is to determine the clusters under the constraint of a given minimum cluster size for masking microdata.

In this special issue, the selected papers focus on the topics of theory and applications of data clustering. They propose new methods that have been successfully applied on several clustering problems including image segmentation [9, 10], time series clustering [4], graph clustering (community detection) [11, 12], and (stock) recommendation systems [13, 14]. Image segmentation is a key step in many image analysis

and interpretation tasks. Finding semantic regions is the ultimate goal of segmentation for image understanding. It has become a necessity for many applications, such as content based image retrieval and object recognition. The goal of time series clustering is to partition time series into clusters based on similarity or distance criteria, so that time series in the same group are similar and dissimilar to the time series in the other groups. Concerning the community detection problem, it holds that networks are usually composed of subgroup structures, whose interconnections are sparse and the intraconnections are dense, which is called community structure. Detecting the community structure of a network is a fundamental problem in complex networks which presents many variations. Community detection is often a NP-hard problem and traditional methods for detecting communities in networks can be concluded into two categories: graph partitioning and hierarchical clustering. The recommender system tries to predict the behavior of a complex system by producing a list of recommendations. In stock recommendation that has become a hot topic, most of the methods try to integrate multiple technologies, such as data mining, machine learning, herd psychology, and other nontraditional technologies.

During the last decades, there have been published thousands of clustering algorithms [1]. The clustering methods can be classified into five major categories [2]: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. A partitioning method constructs (crisp or fuzzy) partitions of the data, where each partition represents a cluster. The

partition is called crisp if each object belongs to exactly one cluster or fuzzy if one object is allowed to belong to more than one cluster at the same time. Hierarchical clustering algorithms recursively find nested clusters either in agglomerative (bottom-up) mode or in divisive (top-down) mode. Agglomerative algorithms start with each point as a separate cluster and successively merge the most similar pair of clusters. On the contrary, divisive algorithms start with all the data points in one cluster and recursively divide each cluster into smaller clusters. In both cases, a hierarchical structure (e.g., dendrogram) is provided which represents the merging or dividing steps of the method. The density-based methods continue growing a cluster as long as its density (number of data objects in the “neighborhood”) exceeds a threshold. Concerning the grid-based methods, they quantize the object space into a finite number of cells that form a grid structure. Then, they use statistical attributes for all the data objects located in each individual cell and clustering is performed on the grid, instead of data objects themselves. Model-based methods assume a model for each of the clusters and attempt to best fit the data to the assumed model.

The definition of a metric that can be used to validate clusters of different densities and/or sizes is an open problem. In the literature, several clustering validity measures have been proposed to measure the quality of clustering [3, 7, 15]. In addition, using the clustering validity measures, it is possible to compare the performance of clustering algorithms and to improve their results by getting a local minima of them.

The papers, published in this special issue, have novelty and contain some interesting methods and applications on data clustering. We believe that the papers published in this special issue will motivate further research in the field of data clustering.

Acknowledgments

The guest editors wish to express their sincere gratitude to the authors and reviewers who contributed greatly to the success of this special issue. We would also like to thank the editorial board members of this journal for their support and help throughout the preparation of this special issue.

Costas Panagiotakis
Emmanuel Ramasso
Paraskevi Fragopoulou
Daniel Aloise

References

- [1] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.
- [2] T. W. Liao, “Clustering of time series data—a survey,” *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.
- [3] C. Panagiotakis, “Point clustering via voting maximization,” *Journal of Classification*, vol. 32, no. 2, pp. 212–240, 2015.
- [4] E. Ramasso, V. Placet, and M. L. Boubakar, “Unsupervised consensus clustering of acoustic emission time-series for robust damage sequence estimation in composites,” *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 12, pp. 3297–3307, 2015.
- [5] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, “NP-hardness of Euclidean sum-of-squares clustering,” *Machine Learning*, vol. 75, no. 2, pp. 245–248, 2009.
- [6] E. R. Hruschka, R. J. G. B. Campello, A. A. Freitas, and A. C. P. L. F. de Carvalho, “A survey of evolutionary algorithms for clustering,” *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 39, no. 2, pp. 133–155, 2009.
- [7] C.-H. Chou, M.-C. Su, and E. Lai, “A new cluster validity measure and its application to image compression,” *Pattern Analysis and Applications*, vol. 7, no. 2, pp. 205–220, 2004.
- [8] C. Panagiotakis and G. Tziritas, “Successive group selection for microaggregation,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 5, pp. 1191–1195, 2013.
- [9] C. Panagiotakis, H. Papadakis, E. Grinias, N. Komodakis, P. Fragopoulou, and G. Tziritas, “Interactive image segmentation based on synthetic graph coordinates,” *Pattern Recognition*, vol. 46, no. 11, pp. 2940–2952, 2013.
- [10] C. Panagiotakis, I. Grinias, and G. Tziritas, “Natural image segmentation based on tree equipartition, bayesian flooding and region merging,” *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2276–2287, 2011.
- [11] H. Papadakis, C. Panagiotakis, and P. Fragopoulou, “Distributed detection of communities in complex networks using synthetic coordinates,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2014, no. 3, Article ID P03013, 2014.
- [12] D. Aloise, G. Caporossi, P. Hansen, L. Liberti, S. Perron, and M. Ruiz, “Modularity maximization in networks by variable neighborhood search,” in *Graph Partitioning and Graph Clustering*, vol. 588, pp. 113–128, American Mathematical Society, 2013.
- [13] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.
- [14] E. J. de Fortuny, T. De Smedt, D. Martens, and W. Daelemans, “Evaluating and understanding text-based stock price prediction models,” *Information Processing & Management*, vol. 50, no. 2, pp. 426–441, 2014.
- [15] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory, “A taxonomy of visual cluster separation factors,” *Computer Graphics Forum*, vol. 31, no. 3, part 4, pp. 1335–1344, 2012.

Hindawi

Submit your manuscripts at
<http://www.hindawi.com>