A Bibliographic View on Constrained Clustering

Ludmila I. Kuncheva, Francis J. Williams, and Samuel L. Hennessey

School of Computer Science and Electronic Engineering
Bangor University, Bangor, Gwynedd
LL57 1UT, United Kingdom
l.kuncheva@bangor.ac.uk

Abstract. A keyword search on constrained clustering on Web-of-Science returned just under 3,000 documents. We ran automatic analyses of those, and compiled our own bibliography of 183 papers which we analysed in more detail based on their topic and experimental study, if any. This paper presents general trends of the area and its sub-topics by Pareto analysis, using citation count and year of publication. We list available software and analyse the experimental sections of our reference collection. We found a notable lack of large comparison experiments. Among the topics we reviewed, applications studies were most abundant recently, alongside deep learning, active learning and ensemble learning.

Keywords: Constrained clustering · Pairwise constraints · Semi-supervised learning · Clustering survey

1 Introduction

Constrained clustering is a topic in semi-supervised machine learning. The success of a clustering result is often judged by an external criterion, be it user satisfaction or a match with a predefined structure or class labels. Constrained clustering is aimed at improving the quality of the resultant partition. Pairwise constraints are most widely used. Must Link (ML) constraints between points \( A \) and \( B \) state that \( A \) and \( B \) must be in the same cluster, while Cannot Link (CL) constraints state that \( A \) and \( B \) must not be in the same cluster.

Constraints can be derived from a labelled data set, introduced manually by a user, or may come from a real life unsupervised problem. Consider the following example. We are interested in recognising individual animals from a video feed. At the start, there is no information about each individual animal. However, the total number of identities can be assumed within a small range. This will define the number of clusters. Suppose that bounding boxes with all animals in the video have been extracted, and features representations have been created thereof. Clustering the data may give us the animal identities. However, we have further information that can be included to improve the clustering. First, animals which are in the same frame of the video cannot be the same identity. This generates CL constraints. Additionally, tracking software may provide trajectories across frames for the same animal. This will generate ML constraints.

There are many more ways to introduce constraints in clustering, e.g., cluster size, diameter, cardinality, density, feature-specific logical expressions, and so on.

The available surveys on constrained clustering are a few years old now [53], [44], [60], [21], focus on a specific topic [54], or only marginally cover constrained clustering [141].

This paper presents a snapshot of the state-of-the-art in constrained clustering. We acknowledge that scientific impact may not be faithfully represented by citation count [26], [132]. However, in the absence of a better metric, we provide illustrative bibliometrics and Pareto analysis. While previous surveys report the details of the works they cover, here we look to identify tendencies, progress, prevalence, and presence of the different themes and topics.

The rest of the paper is organised as follows. Section 2 gives a general bibliographic metric of the area. In Section 3 we create a rough taxonomy of the constrained clustering topics throughout the literature. Section 4 shows a Pareto front analysis of the constrained clustering literature and its topics. Section 5 displays a list of some available software used for constrained clustering. Section 6 analyses and summarises the experimental studies carried out in the literature.
2 General bibliographic remarks

To create an overview of the importance and development of constrained clustering, we carried out publication search using the Web-of-Science platform [https://www.webofscience.com/](https://www.webofscience.com/) The reported results are valid as of 20th May 2022. However, the number of publications in the years before 2021 is unlikely to change if a search is carried out at a later date. We tried two combinations of keywords: (1) “constrained clustering” (CC), and (2) “constrained clustering” OR (“semi-supervised” AND clustering) (CC or SS). The quotation marks indicate that the word combination was kept intact. The search was on ‘Topic’, which includes title, abstract and keywords. Figure 1 plots the number of publications over the years from 1995 to 2021.

![Graph of publication counts](image)

(a) Original count  
(b) Scaled on “machine learning”  
(c) Scaled on “clustering”

Fig. 1: Number of publications retrieved from Web-of-Science a of 10th May 2022 using two keyword combinations: (1) “constrained clustering” (CC), and (2) “constrained clustering” OR (“semi-supervised” AND clustering) (CC or SS).

The total number of publications on both CC and CC or SS (subplot (a)) is rising along time but this could be due to the overall trend of rising number of publications in the world. To correct for this effect, we also recovered the number of publications on “machine learning” (ML) and on “clustering” (C). Subplots (b) and (c) show the scaled number of publications obtained by dividing the original counts in subplot (a) to the respective counts for ML and C. Interestingly, according to the ML scaling, the interest in constraint and semi-supervised clustering seems to decline, especially from 2016 onward. This is likely a result of the ongoing boom of publications on deep learning, which overpower other branches of machine learning. Indeed, in terms of clustering only (subplot (c)), constrained and supervised clustering both seem to have a steady upward trend. Arguably, the term ‘semi-supervised clustering’ is more popular and the publication proportion is increasing more notably compared to that for ‘constrained clustering’.

In view of the increasing publication counts, surveys and systematisation of the area of semi-supervised and constrained clustering would be beneficial to the research community.

Figure 2 displays a tree map diagram of the top eight areas represented within the 2,981 documents retrieved with the query “constrained clustering” OR (semi-supervised AND clustering) from Web of Science. The area distribution was obtained from Web-of-Science.

The diagram shows that the overwhelming majority of the publications are on the technology side although some application areas are also present (telecommunications, automation control systems and computational biology).

A keyword search was carried out of the retrieved papers (again, all fields were searched) using a set of relevant terms for constrained clustering. We chose not to include terms such as ‘graph’ or ‘optimisa-
Fig. 2: Top 8 areas of research according to Web-of-Science as of 20th May 2022. All fields were searched for query “constrained clustering” OR (semi-supervised AND clustering).

Fig. 3: Selected keywords arranged by frequency of occurrence within the as 2,981 documents retrieved from Web of Science on the 20th May 2022.

As also identified by other surveys, k-means is the undisputed leader in the area of constrained clustering. ‘Pairwise’ and ‘spectral’ were also expected at the top of the diagram along with ‘kernel’ and ‘density’. Interestingly, ‘fuzzy’ and ‘ensemble’-based approaches have been mentioned often enough to secure places ahead of ‘deep learning’ and ‘active learning’. We should emphasise that we are presenting a snapshot of the area at the current moment, and the keyword frequency distribution may change in the future. Most likely, ‘deep learning’ and keywords coming from persistent application areas will climb up the diagram.

3 Topics

We collated references from the previous surveys, from Google search and Web of Science. We browsed the references to identify research focused primarily on constrained clustering, so we cast aside papers where the
main thrust was an application of where constrained clustering was a secondary topic. As a result, we offer the reader a collection of references available at https://github.com/LucyKuncheva/Semi-supervised-and-Constrained-Clustering/blob/main/ConstrainedClusteringReferences.pdf. We decided to store the reference list separately so that we have space for the bibliometric analyses. We used this collection to carry out our Pareto analysis, which we see as one of the main contributions of this paper.

Creating a taxonomy of the literature on constrained clustering is not straightforward. Below we offer a rough, non-mutually exclusive grouping, acknowledging that the topics are not the same level of the hierarchy. We were guided by the interest in a particular topic, as well as by topics identified in previous surveys [60].

1. **K-means variants.** [146], [145], [139], [143], [129], [88], [82], [18], [79], [133], [120], [38], [47], [6], [62], [177], [15], [22], [38], [102], [117], [92], [167]

2. **Spectral clustering.** [77], [39], [87], [5], [150], [148], [149], [176], [7], [109], [110], [33], [52], [173], [51], [34], [105], [106], [124]

3. **Other clustering algorithms.** hierarchical clustering [43], [156], [64], [65], [162], [72], DBSCAN [56], [112], density-based clustering [159], [128], [99], mean-shift [8], [140], neural network-based [81], multi-view and multi-source clustering [152], [14], [63], [29], or model selection [122]

4. **Cluster ensembles.** [4], [161], [151], [34], [84], [97], [168], [136], [169], [125], [30], [69], [54], [116], [50], [59], [85], [150], [160], [104], [170]

5. **Deep learning methods.** [84], [103], [33], [130], [129], [142], [155], [125], [95], [107], [172]

6. **Soft computing approaches.** fuzzy k-means [70], [2], [96], [57], [108], evidential k-means [9], [10], evolutionary approaches [116], [118], [60], [68], [67], [111], [108] ant colony optimisation [105]

7. **Learning a distance metric.** [134], [16], [17], [90], [157], [37], [25], [78], [166], [147], [103], [73], [1]

8. **Incorporating the constraints into the criterion function.** [49], [43], [121], [19], [76], [113], [61], [25], [20], [91], [35], [40], [42], [21], [30], [40], [22], [114], [135], [115], [119], [42], [41], [98], [99]

9. **Active learning, user interaction, incremental clustering.** [148], [163], [101], [104], [58], [3], [32], [12], [81], [36], [123], [131], [144], [162]

10. **Applications.**

In addition to the theoretical and algorithmic advances in constrained clustering, we came across a beautiful variety of applications, among which: analysis of RNA [137], [33] gene expression data analysis [151], medical imaging [154], EEG data analysis [55], vegetation classification [138], regionalisation using spatial contiguity constraints [27], [72], [94], [120], text and document clustering [169], [127], [28], [142], information retrieval [171], object and face clustering in video [153], [86], [29], [39], [38], [158], [11], tracking of moving objects using radar sensors [112], time series clustering [73], [98], tourism [24], financial analysis [74], [174], defect prediction [175], [100], cyber security [155], [83], [71], and malware clustering [56].

## 4 Pareto analysis

We assume that papers published more recently and papers with a large number of citations are generally more influential. These two criteria are not likely to be satisfied by a single paper. A compromise between year of publication and number of citations should be sought. Admittedly, most recent papers would not have had enough exposure to attract a large number of citations yet, but this does not impact on their future relevance. Pareto front contains all non-dominated alternatives. A paper \( x \) is in the Pareto front if there is no other paper \( y \) in the collection which is better than \( x \) on both criteria.

Figure 4 shows a scatterplot of the papers covered in this survey in the space (year-of-publication, log(\( K \))), where \( K \) is the number of citations according to Google Scholar, as of 20th May 2022. The Pareto front is marked with a solid red line, and the papers in it are listed under the figure. Some dots are larger than others to indicate that there are more than one paper published in the same year, with the same number of citations.
A Bibliographic View on Constrained Clustering

- A: [157], 3669 citations, Xing et al. (2002) Distance metric learning with application to clustering with side-information
- B: [25], 1114 citations, Bilenko et al. (2004) Integrating constraints and metric learning in semi-supervised clustering
- C: [91], 917 citations, Kulis et al. (2013) Metric learning: A survey
- D: [54], 251 citations, Dong et al. (2020) A survey on ensemble learning
- E: [137], 12 citations, Tian et al. (2021) Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data
- F: [171], 9 citations, Zhan et al. (2022) Learning discrete representations via constrained clustering for effective and efficient dense retrieval

Fig. 4: Pareto front of all reviewed publications (as of 20th May 2022).

It is interesting to notice that the earlier papers are mostly on learning a distance metric, more recent ones are on clustering methodologies (ensembles and deep learning), and the most recent two papers in the Pareto front are application-orientated.

Figure 5 shows the Pareto front for the ten individual topics and Figure 6 shows all the 10 Pareto Fronts together.

The plots indicate that more recent interest in constrained clustering is focused on applications, deep learning, soft computing, and active learning. On the other hand, developments on k-means and spectral clustering were less represented in the more recent literature. Compared to the overall Pareto front, the ensemble methods, deep learning, and the applications seem to be dominating the rest of the topics in terms of recent citation counts. Category 3, ‘other methods’ seems to be enjoying a healthy interest all throughout the years.

5 Available software

Table 1 shows a list of software for constrained clustering which we found. Our list is by no means exhaustive.
Fig. 5: Pareto front for the ten topics.

Fig. 6: Pareto front for the ten topics on the same graph.
Table 1: Available software for constrained clustering

| Algorithm                                      | Year | Link                                                                 |
|------------------------------------------------|------|----------------------------------------------------------------------|
| Auto-tuning spectral clustering                | 2022 | https://github.com/tango4j/Auto-tuning-Spectral-Clustering          |
| Binary optimisation constrained k-means (BCKM) | 2019 | https://github.com/intellhave/BCKM                                |
| Cluster fractional allocation matrix (CFAM)    | 2020 | https://github.com/dung321046/ConstrainedClusteringViaPostProcessing |
| Constrained deep adaptive clustering (CDAP)    | 2020 | https://github.com/thuviar/CDAC-plus                               |
| Constrained graph clustering                   | 2017 | https://github.com/Behrouz-Babaki/Pigeon                          |
| Constrained K-means                            | 2017 | https://github.com/NestorRV/constrained_kmeans                     |
| Constrained online face clustering (COFC)      | 2018 | https://github.com/ankuPRK/COFC                                  |
| Constraint satisfaction clustering             | 2022 | https://github.com/autonlab/constrained-clustering                 |
| COP/PC K-means                                 | 2019 | https://github.com/ashkanmradi/constrained-k-means                 |
| COPK-means                                     | 2017 | https://github.com/Behrouz-Babaki/COPKmeans                       |
| DCDS                                           | 2019 | https://github.com/leule/DCDS                                     |
| Deep Constrained Clustering                    | 2020 | https://github.com/blueocean92/deep_constrained_clustering        |
| Lpbox-ADMM                                     | 2021 | https://github.com/wubaoyuan/Lpbox-ADMM                           |
| MinSizeK-means                                 | 2021 | https://github.com/Behrouz-Babaki/MinSizeKmeans                    |
| MIPK-means                                     | 2017 | https://github.com/Behrouz-Babaki/MPKmeans                        |
| PC-SOS-SDP                                     | 2021 | https://github.com/antoniosudoso/pc-sos-sdp                        |
| repCONC                                        | 2022 | https://github.com/jingtaozhan_repconc                            |
| Spectral clustering with fair constraints      | 2019 | https://github.com/matthklein/fair_spectral_clustering             |
| SpectralNet                                    | 2020 | https://github.com/KlugerLab/SpectralNet                         |
6 Experimental studies

Our analysis of the literature revealed a marked lack of large comparative studies. We gathered and summarised the experiments from the cited works which reported an experiment; a total of 76 publications. For each experiment, we collated the list of data sets used, the algorithms compared, and the evaluation metrics.

6.1 Datasets

A total of 245 datasets were identified. Out of these, 179 were used only once, and 30 were used twice, demonstrating the deficiency in comparative studies. Figure 7 displays a histogram of the datasets that have been used more than six times throughout the experimental studies. The four most commonly used datsets were: Iris, Wine, Glass, and Ionosphere from the UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/]

![Fig. 7: Histogram of datasets that appeared more than six times in the constrained clustering literature](image)

To complement this result, Figure 8 shows a histogram of the number of datasets used in the experiments reported in the literature. It can be seen that only a few works use more than 10 datasets, which, again, points at the lack of adequate large-scale comparisons.

6.2 Algorithms

Figure 9 displays the algorithms that are used for comparison within the experiments. The most commonly used algorithms were: K-Means, COP-KMeans, and MPC-KMeans.

Figure 10 shows a histogram of the number of algorithms compared against used in the experiments. Only a handful of papers report comparisons between above 10 algorithms.

6.3 Evaluation Metrics

Figure 11 shows the most popular metrics used to evaluate and compare constrained clustering algorithms. All there metrics rely an an external labelled dataset, which reinforces the message that, in absence of a better gauging criterion, the fundamentally flawed approach of comparing partition labels with pre-assigned labels is most often applied.

7 Conclusion

This work presents a bibliographic snapshot of the work on constrained clustering, as of 20 May 2022. The main sources for our analyses were Google Scholar and Web-of-Science. We did systematic keyword search in Web-of-Science, and sources citation counts from Google Scholar.
The main finding of our work is a notable lack of extensive experimental comparisons between methods for constrained clustering. Such comparisons would be a useful tool for present and future researchers.

Shared software is the way forward in order to unify the approaches and facilitate comparisons. The visibility of research works which share their software is much better.

Our Pareto analysis demonstrated that application-orientated papers will likely dominate the future development of the area, where generic methodology and algorithms will make space for more idiosyncratic ones. In terms of methodology, ensemble learning, deep learning and active/incremental learning appear to be still at the forefront.

Acknowledgment

This work is supported by the UKRI Centre for Doctoral Training in Artificial Intelligence, Machine Learning and Advanced Computing (AIMLAC), funded by grant EP/S023992/1.

References

1. Abin, A.A., Bashiri, M.A., Beigy, H.: Learning a metric when clustering data points in the presence of constraints. Advances in Data Analysis and Classification 14(1), 29–56 (2020). https://doi.org/10.1007/s11634-019-00359-6
Fig. 10: Histogram of the number of datasets used in the experiments.

Fig. 11: Histogram of the evaluation metrics most often used in the experimental comparisons.

2. Abin, A.A., Beigy, H.: Active constrained fuzzy clustering: A multiple kernels learning approach. Pattern Recognition 48(3), 953–967 (2015). https://doi.org/10.1016/j.patcog.2014.09.008

3. Abin, A.A., Vu, V.V.: A density-based approach for querying informative constraints for clustering. Expert Systems with Applications 161, 113690 (2020). https://doi.org/10.1016/j.eswa.2020.113690

4. Al-Razgan, M., Domeniconi, C.: Clustering ensembles with active constraints. Springer (2009)

5. Alzate, C., Suykens, J.A.: A regularized formulation for spectral clustering with pairwise constraints. In: International Joint Conference on Neural Networks. pp. 141–148 (2009). https://doi.org/10.1109/IJCNN.2009.517872

6. de Amorim, R.C.: Constrained clustering with Minkowski weighted K-means. In: IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI). pp. 13–17 (2012). https://doi.org/10.1109/CINTI.2012.6496753

7. Anand, R., Reddy, C.K.: Graph-based clustering with constraints. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 51–62 (2011). https://doi.org/10.1007/978-3-642-20847-8_5

8. Anand, S., Mittal, S., Tuzel, O., Meer, P.: Semi-supervised kernel mean shift clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(6), 1201–1215 (2014). https://doi.org/10.1109/TPAMI.2013.190

9. Antoine, V., Guerrero, J.A., Xie, J.: Fast semi-supervised evidential clustering. International Journal of Approximate Reasoning 133, 116–132 (2021). https://doi.org/10.1016/j.ijar.2021.03.008

10. Antoine, V., Quost, B., Masson, M.H., Denoeux, T.: CECM: Constrained evidential C-means algorithm. Computational Statistics & Data Analysis 56(4), 894–914 (2012). https://doi.org/10.1016/j.csda.2010.09.021

11. Arachchilage, S.W., Izquierdo, E.: Adaptive aggregated tracklet linking for multi-face tracking. In: IEEE International Conference on Image Processing (ICIP). pp. 1366–1370 (2020). https://doi.org/10.1109/ICIP48778.2020.9190829

12. Awasthi, P., Balcan, M., Voevodski, K.: Local algorithms for interactive clustering. In: International Conference on Machine Learning. pp. 550–558 (2014). https://doi.org/10.48550/arXiv.1312.6724
37. Cohn, D., Caruana, R., McCallum, A.: Semi-supervised clustering with user feedback. Constrained Clustering: Advances in Algorithms, Theory, and Applications 4(1), 17–32 (2003). https://doi.org/10.1201/9781584889778.ch2
38. Covoes, T.F., Hruschka, E.R., Ghosh, J.: A study of K-means-based algorithms for constrained clustering. Intelligent Data Analysis 17(3), 485–505 (2013). https://doi.org/10.3233/IDA-130590
39. Cucuringu, M., Koutis, I., Chawla, S., Miller, G., Peng, R.: Simple and scalable constrained clustering: A generalized spectral method. In: 19th International Conference on Artificial Intelligence and Statistics. pp. 445–454 (2016)
40. Dao, T.B.H., Duong, K.C., Vrain, C.: A declarative framework for constrained clustering. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 419–434 (2013). https://doi.org/10.1007/978-3-642-40994-3_27
41. Dao, T.B.H., Khanh-Chuong Duong, Vrain, C.: Constrained clustering by constraint programming. Artificial Intelligence 244, 70–94 (2017). https://doi.org/10.1016/j.artint.2015.05.006
42. Dao, T., Vrain, C., Duong, K., Davidson, I.: A framework for actionable clustering using constraint programming. In: 22nd European Conference on Artificial Intelligence (ECAI). pp. 453–461 (2016). https://doi.org/10.3233/978-1-61499-672-9-453
43. Davidson, I., Ravi, S.S.: Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. In: European Conference on Principles of Data Mining and Knowledge Discovery. pp. 59–70 (2005). https://doi.org/10.1007/11564126_11
44. Davidson, I., Basu, S.: A survey of clustering with instance level constraints. ACM Transactions on Knowledge Discovery from Data w(1), 1–41 (2007)
45. Davidson, I., Ravi, S.: Clustering with constraints: Feasibility issues and the K-means algorithm. In: Proceedings of the 2005 SIAM international conference on data mining. pp. 138–149 (2005). https://doi.org/10.1137/1.9781611972757.13
46. Davidson, I., Ravi, S., Shamis, L.: A sat-based framework for efficient constrained clustering. In: Proceedings of the 2010 SIAM international conference on data mining. pp. 94–105 (2010)
47. Davidson, I., Wagstaff, K.L., Basu, S.: Measuring constraint-set utility for partitioning clustering algorithms. In: European Conference on Principles of Data Mining and Knowledge Discovery. pp. 115–126 (2006). https://doi.org/10.1007/11871637_15
48. Demiriz, A., Bennett, K.P., Bradley, P.S.: Using assignment constraints to avoid empty clusters in K-means clustering. Constrained Clustering: Advances in Algorithms, Theory, and Applications 201, 201–220 (2008).
https://doi.org/10.1007/9781584889977
49. Demiriz, A., Bennett, K.P., Embrechts, M.J.: Semi-supervised clustering using genetic algorithms. Artificial Neural Networks in Engineering (ANNIE) pp. 809–814 (1999)
50. Dimitriadou, E., Weingessel, A., Hornik, K.: A mixed ensemble approach for the semi-supervised problem. In: International Conference on Artificial Neural Networks. pp. 571–576 (2002). https://doi.org/10.1007/3-540-46084-5_93
51. Ding, S., Jia, H., Du, M., Xue, Y.: A semi-supervised approximate spectral clustering algorithm based on HMRF model. Information Sciences 429, 215–228 (2018). https://doi.org/10.1016/j.ins.2017.11.016
52. Ding, S., Qi, B., Jia, H., Zhu, H., Zhang, L.: Research of semi-supervised spectral clustering based on constraints expansion. Neural Computing and Applications 22(1), 405–410 (2013). https://doi.org/10.1007/s00521-012-0911-8
53. Dinler, D., Tural, M.K.: A survey of constrained clustering. Springer (2016). https://doi.org/10.1007/978-3-319-24211-8_3
54. Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Frontiers of Computer Science 14(2), 241–258 (2020). https://doi.org/10.1007/s11704-019-8208-z
55. Du, Y., Sun, B., Lu, R., Zhang, C., Wu, H.: A method for detecting high-frequency oscillations using semi-supervised K-means and mean shift clustering. Neurocomputing 350, 102–107 (2019). https://doi.org/10.1016/j.neucom.2019.03.055
56. Fang, Y., Zhang, W., Li, B., Jing, F., Zhang, L.: Semi-supervised malware clustering based on the weight of bytecode and API. IEEE Access 8, 2313–2326 (2019). https://doi.org/10.1109/ACCESS.2019.2962198
57. Fantoukh, N.I., Ismail, M.M.B., Bchir, O.: Automatic determination of the number of clusters for semi-supervised relational fuzzy clustering. International Journal of Fuzzy Logic and Intelligent Systems 20(2), 156–167 (2020). https://doi.org/10.5391/IJFIS.2020.20.2.156
58. Fantoukh, N.I., Ismail, M.M.B., Bchir, O.: Automatic determination of the number of clusters for semi-supervised relational fuzzy clustering. International Journal of Fuzzy Logic and Intelligent Systems 20(2), 156–167 (2020). https://doi.org/10.5391/IJFIS.2020.20.2.156
59. Fantoukh, N.I., Ismail, M.M.B., Bchir, O.: Automatic determination of the number of clusters for semi-supervised relational fuzzy clustering. International Journal of Fuzzy Logic and Intelligent Systems 20(2), 156–167 (2020). https://doi.org/10.5391/IJFIS.2020.20.2.156
60. Fantoukh, N.I., Ismail, M.M.B., Bchir, O.: Automatic determination of the number of clusters for semi-supervised relational fuzzy clustering. International Journal of Fuzzy Logic and Intelligent Systems 20(2), 156–167 (2020). https://doi.org/10.5391/IJFIS.2020.20.2.156
61. Fantoukh, N.I., Ismail, M.M.B., Bchir, O.: Automatic determination of the number of clusters for semi-supervised relational fuzzy clustering. International Journal of Fuzzy Logic and Intelligent Systems 20(2), 156–167 (2020). https://doi.org/10.5391/IJFIS.2020.20.2.156
61. Ganji, M., Bailey, J., Stuckey, P.J.: Lagrangian constrained clustering. In: Proceedings of the 2016 SIAM International Conference on Data Mining. pp. 288–296 (2016). https://doi.org/10.1137/1.9781611974348.33
62. Ge, R., Ester, M., Jin, W., Davidson, I.: Constraint-driven clustering. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 320–329 (2007). https://dl.acm.org/doi/10.1145/1281192.1281199
63. Ghasemi, Z., Khorshidi, H.A., Aickelin, U.: Multi-objective semi-supervised clustering for finding predictive clusters. Expert Systems with Applications 195, 116551 (2022). https://doi.org/10.1016/j.eswa.2022.11.055
64. Gilpin, S., Davidson, I.: Incorporating SAT solvers into hierarchical clustering algorithms: An efficient and flexible approach. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1136–1144 (2011). https://doi.org/10.1145/2020408.2020585
65. Gilpin, S., Davidson, I.: A flexible ILP formulation for hierarchical clustering. Artificial Intelligence 244, 95–109 (2017). https://doi.org/10.1016/j.artint.2016.05.009
66. González-Almagro, G., Luengo, J., Cano, J.R., García, S.: DILS: Constrained clustering through dual iterative local search. Computers & Operations Research 121, 104979 (2020). https://doi.org/10.1016/j.cor.2020.104979
67. González-Almagro, G., Luengo, J., Cano, J.R., García, S.: Enhancing instance-level constrained clustering through differential evolution. Applied Soft Computing 108, 107435 (2021). https://doi.org/10.1016/j.asoc.2020.107435
68. González-Almagro, G., Rosales-Pérez, A., Luengo, J., Cano, J.R., García, S.: ME-MEOA/DCC: Multiobjective constrained clustering through decomposition-based memetic elitism. Swarm and Evolutionary Computation 66, 100939 (2021). https://doi.org/10.1016/j.swevo.2021.100939
69. Greene, D., Cunningham, P.: Constraint selection by committee: An ensemble approach to identifying informative constraints for semi-supervised clustering. In: European Conference on Machine Learning. pp. 140–151 (2007). https://doi.org/10.1007/978-3-540-74958-5_16
70. Grira, N., Crucianu, M., Boujemaa, N.: Fuzzy clustering with pairwise constraints for knowledge-driven image categorisation. IEEE Proceedings-Vision, Image and Signal Processing 153(3), 299–304 (2006). https://doi.org/10.1049/ip-vis:20050060
71. Gu, Y., Li, K., Guo, Z., Wang, Y.: Semi-supervised K-means DDoS detection method using hybrid feature selection algorithm. IEEE Access 7, 64351–64365 (2019). https://doi.org/10.1109/ACCESS.2019.2917532
72. Guo, D.: Regionalization with dynamically constrained agglomerative clustering and partitioning (redcap). International Journal of Geographical Information Science 22(7), 801–823 (2008). https://doi.org/10.1080/13658810701674970
73. Guo, L., Gan, H., Xia, S., Xu, X., Zhou, T.: Joint exploring of risky labeled and unlabeled samples for safe semi-supervised clustering. Expert Systems with Applications 176, 114796 (2021). https://doi.org/10.1016/j.eswa.2021.114796
74. Han, Y., Wang, T.: Semi-supervised clustering for financial risk analysis. Neural Processing Letters 53(5), 3561–3572 (2021). https://doi.org/10.1007/s11063-021-10564-0
75. He, G., Pan, Y., Xia, X., He, J., Peng, R., Xiong, N.N.: A fast semi-supervised clustering framework for large-scale time series data. IEEE Transactions on Systems, Man, and Cybernetics: Systems 51(7), 4201–4216 (2019). https://doi.org/10.1109/TSMC.2019.2931731
76. Hiep, T.K., Duc, N.M., Trung, B.Q.: Local search approach for the pairwise constrained clustering problem. In: Proceedings of the Seventh Symposium on Information and Communication Technology. pp. 115–122 (2016). https://doi.org/10.1109/ISICT.2016.7872983
77. Hoi, S.C., Jin, R., Lyu, M.R.: Learning nonparametric kernel matrices from pairwise constraints. In: Proceedings of the 24th international conference on Machine learning. pp. 361–368 (2007). https://doi.org/10.1145/1273496.1273542
78. Hoi, S.C., Liu, W., Chang, S.F.: Semi-supervised distance metric learning for collaborative image retrieval and clustering. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 6(3), 1–26 (2010). https://doi.org/10.1145/1823746.1823752
79. Hong, Y., Kwong, S.: Learning assignment order of instances for the constrained K-means clustering algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39(2), 568–574 (2009). https://doi.org/10.1109/TSMCB.2008.2006641
80. Hosseini, M.J., Gholidpour, A., Beigy, H.: An ensemble of cluster-based classifiers for semi-supervised classification of non-stationary data streams. Knowledge and information systems 46(3), 567–597 (2016). https://doi.org/10.1007/s10115-015-0837-4
81. Hsu, Y.C., Kira, Z.: Neural network-based clustering using pairwise constraints. arXiv (2016). https://arxiv.org/abs/1608.00321
82. Huang, H., Cheng, Y., Zhao, R.: A semi-supervised clustering algorithm based on must-link set. In: International Conference on Advanced Data Mining and Applications. pp. 492–499 (2008). https://doi.org/10.1007/978-3-540-88192-6_48
83. Huda, S., Miah, S., Hassan, M.M., Islam, R., Yearwood, J., Alrubaian, M., Almogren, A.: Defending unknown attacks on cyber-physical systems by semi-supervised approach and available unlabeled data. Information Sciences 379, 211–228 (2017). https://doi.org/10.1016/j.ins.2016.09.041
84. Ienco, D., Pensa, R.G.: Semi-supervised clustering with multiresolution autoencoders. In: 2018 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2018). https://doi.org/10.1109/IJCNN.2018.8489353
85. Iqbal, A.M., Moh’d, A., Khan, Z.: Semi-supervised clustering ensemble by voting. In: The International Conference on Information and Communication Systems (ICICS) (2012)
86. Kalogeiton, V., Zisserman, A.: Constrained video face clustering using INN relations. In: 31st British Machine Vision Conference (BMVC). Virtual Event, UK (2020)
87. Kamvar, S.D., Klein, D., Manning, C.D.: Spectral learning. In: International Joint Conference of Artificial Intelligence. p. 10.1145/1756006.1756037 (2003)
88. Khashabi, D., Wieting, J., Liu, J.Y., Liang, F.: Clustering with side information: From a probabilistic model to a deterministic algorithm. arXiv (2015). https://doi.org/10.48550/arXiv.1508.06235
89. Khiari, M., Boizumault, P., Crémielleux, B.: Constraint programming for mining n-ary patterns. In: International Conference on Principles and Practice of Constraint Programming. pp. 552–567 (2010). https://doi.org/10.1007/978-3-642-15396-9_44
90. Klein, D., Kamvar, S.D., Manning, C.D.: From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. Tech. rep., Stanford (2002)
91. Kulis, B.: Metric learning: A survey. Foundations and Trends® in Machine Learning 5(4), 287–364 (2013). https://doi.org/10.1561/2200000019
92. Kulis, B., Basu, S., Dhillon, I., Mooney, R.: Semi-supervised graph clustering: a kernel approach. Machine learning 74(1), 1–22 (2009). https://doi.org/10.1007/s10994-008-5084-4
93. Kulshreshtha, P., Guha, T.: An online algorithm for constrained face clustering in videos. In: IEEE International Conference on Image Processing (ICIP). pp. 2670–2674 (2018). https://doi.org/10.1109/ICIP.2018.8451343
94. Kupfer, J.A., Gao, P., Guo, D.: Regionalization of forest pattern metrics for the continental united states using contiguity constrained clustering and partitioning. Ecological Informatics 9, 11–18 (2012). https://doi.org/10.1016/j.ecoinf.2012.02.001
95. Lafabregue, B., Weber, J., Gantarowski, P., Forestier, G.: Deep constrained clustering applied to satellite image time series. In: Proceedings of MACLEAN: MAChine Learning for EArth Observation Workshop co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD) (2019)
96. Lai, D.T., Miyakawa, M., Sato, Y.: Semi-supervised data clustering using particle swarm optimisation. Soft Computing 24(5), 3499–3510 (2020). https://doi.org/10.1007/s00500-019-04114-z
97. Lai, Y., He, S., Lin, Z., Yang, F., Zhou, Q., Zhou, X.: An adaptive robust semi-supervised clustering framework using weighted consensus of random K-means ensemble. IEEE Transactions on Knowledge and Data Engineering 33(5), 1877–1890 (2019). https://doi.org/10.1109/TKDE.2019.295259E
98. Lampert, T., Diao, T.B.H., Lafabregue, B., Serrette, N., Forestier, G., Crémilleux, B., Vrain, C., Gantarowski, P.: Constrained distance based clustering for time-series: A comparative and experimental study. Data Mining and Knowledge Discovery 32(6), 1663–1707 (2018). https://doi.org/10.1007/s10618-018-0573-y
99. Lelis, L., Sander, J.: Semi-supervised density-based clustering. In: Ninth IEEE International Conference on Data Mining, pp. 842–847 (2009). https://doi.org/10.1109/ICDM.2009.143
100. Li, C., Zhou, J.: Semi-supervised weighted kernel clustering based on gravitational search for fault diagnosis. ISA transactions 53(5), 1534–1543 (2014). https://doi.org/10.1016/j.isatra.2014.05.019
101. Li, H., Wang, Y., Li, Y., Xiao, G., Hu, P., Zhao, R., Li, B.: Learning adaptive criteria weights for active semi-supervised learning. Information Sciences 561, 286–303 (2021). https://doi.org/10.1016/j.ins.2021.01.045
102. Li, T., Ding, C., Jordan, M.I.: Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In: Seventh IEEE International Conference on Data Mining (ICDM). pp. 577–582 (2007). https://doi.org/10.1109/ICDM.2007.98
103. Li, X., Yin, H., Zhou, K., Zhou, X.: Semi-supervised clustering with deep metric learning and graph embedding. World Wide Web 23(2), 781–798 (2020). https://doi.org/10.1016/j.isatra.2014.05.019
104. Li, Y., Wang, Y., Yu, D.J., Ye, N., Hu, P., Zhao, R.: ASCENT: Active supervision for semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering 32(5), 868–882 (2019). https://doi.org/10.1109/TKDE.2019.2897307
105. Li, Z., Liu, J., Tang, X.: Pairwise constraint propagation by semidefinite programming for semi-supervised classification. In: Proceedings of the 25th international conference on Machine learning. pp. 576–583 (2008). https://doi.org/10.1145/1390156.1390229
106. Li, Z., Liu, J., Tang, X.: Constrained clustering via spectral regularization. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 421–428 (2009). https://doi.org/10.1109/CVPR.2009.5206852
107. Lin, T.E., Xu, H., Zhang, H.: Discovering new intents via constrained deep adaptive clustering with cluster refinement. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 8360–8367 (2020). https://doi.org/10.1609/aaai.v34i05.6353
108. Liu, H., Huang, S.t.: Evolutionary semi-supervised fuzzy clustering. Pattern Recognition Letters 24(16), 3105–3113 (2003). https://doi.org/10.1016/S0167-8655(03)00177-6
109. Lu, Z., Ip, H.H.: Constrained spectral clustering via exhaustive and efficient constraint propagation. In: European Conference on Computer Vision. pp. 1–14 (2010). https://doi.org/10.1007/978-3-642-15567-3_1
110. Luo, J., Ma, H., Zhou, D.: A Pareto ensemble based spectral clustering framework. Future Generation Computer Systems 37, 789–809 (2020). https://doi.org/10.1016/j.future.2021.07.019
111. Mueller, M., Kramer, S.: Integer linear programming models for constrained clustering. In: International Conference on Discovery Science. pp. 159–173 (2010). https://doi.org/10.1007/978-3-642-16184-1_12
112. Pelleg, D., Baras, D.: K-means with large and noisy constraint sets. In: European Conference on Machine Learning. pp. 674–682 (2007). https://doi.org/10.1007/978-3-540-74958-5_67
113. Pourrajabi, M., Moulavi, D., Campello, R.J., Zimek, A., Sander, J., Goebel, R.: Model selection for semi-supervised clustering. In: Proceedings of the 17th International Conference on Extending Database Technology (EDBT). pp. 331–342 (2014). https://doi.org/10.5441/002/edbt.2014.31
114. Prabakara Raj, S.R., Ravindran, B.: Incremental constrained clustering: A decision theoretic approach. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 475–486 (2013). https://doi.org/10.1007/978-3-642-40319-4_41
115. Rutayisire, T., Yang, Y., Lin, C., Zhang, J.: A modified cop-kmeans algorithm based on sequenced cannot-link set. In: International Conference on Rough Sets and Knowledge Technology. pp. 217–225 (2011). https://doi.org/10.1007/978-3-642-44225-4_30
116. Shental, N., Bar-hillel, A., Hertz, T., Weinshall, D.: Computing gaussian mixture models with EM using equivalence constraints. In: Advances in Neural Information Processing Systems (NIPS). pp. 465–472 (2003)
153. Wu, B., Zhang, Y., Hu, B.G., Ji, Q.: Constrained clustering and its application to face clustering in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3507–3514 (2013). https://doi.org/10.1109/CVPR.2013.450

154. Xia, K., Gu, X., Zhang, Y.: Oriented grouping-constrained spectral clustering for medical imaging segmentation. Multimedia Systems 26(1), 27–36 (2020). https://doi.org/10.1007/s00530-019-00626-8

155. Xian, G.: Cyber intrusion prevention for large-scale semi-supervised deep learning based on local and non-local regularization. IEEE Access 8, 55526–55539 (2020). https://doi.org/10.1109/ACCESS.2020.2981162

156. Xiao, W., Yang, Y., Wang, H., Li, T., Xing, H.: Semi-supervised hierarchical clustering ensemble and its application. Neurocomputing 173, 1362–1376 (2016). https://doi.org/10.1016/j.neucom.2015.09.009

157. Yan, R., Zhang, J., Yang, J., Hauptmann, A.G.: A discriminative learning framework with pairwise constraints for video object classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(4), 578–593 (2006). https://doi.org/10.1109/TPAMI.2006.65

158. Yan, S., Wang, H., Li, T., Chu, J., Guo, J.: Semi-supervised density peaks clustering based on constraint projection. International Journal of Computational Intelligence Systems 14(1), 140–147 (2021). https://doi.org/10.2991/ijcis.d.201102.002

159. Yang, F., Li, T., Zhou, Q., Xiao, H.: Cluster ensemble selection with constraints. Neurocomputing 235, 59–70 (2017). https://doi.org/10.1016/j.neucom.2017.01.001

160. Yang, T., Pasquier, N., Precioso, F.: Semi-supervised consensus clustering based on closed patterns. Knowledge-Based Systems 235, 107599 (2022). https://doi.org/10.1016/j.knosys.2021.107599

161. Yang, W., Wang, X., Lu, J., Dou, W., Liu, S.: Interactive steering of hierarchical clustering. IEEE Transactions on Visualization and Computer Graphics 27(10), 3953–3967 (2020). https://doi.org/10.1109/TVCG.2020.2995100

162. Yang, W., Wang, X., Lu, J., Dou, W., Liu, S.: Interactive steering of hierarchical clustering. IEEE Transactions on Visualization and Computer Graphics 27(10), 3953–3967 (2020). https://doi.org/10.1109/TVCG.2020.2995100

163. Yang, Y., Tan, W., Li, T., Ruan, D.: Consensus clustering based on constrained self-organizing map and improved Cop-Kmeans ensemble in intelligent decision support systems. Knowledge-Based Systems 32, 101–115 (2012). https://doi.org/10.1016/j.knosys.2011.08.011

164. Yang, Y., Teng, F., Li, T., Wang, H., Wang, H., Zhang, Q.: Parallel semi-supervised multi-ant colonies clustering ensemble based on mapreduce methodology. IEEE Transactions on Cloud Computing 6(3), 857–867 (2015). https://doi.org/10.1109/TCC.2015.2511724

165. Yin, X., Chen, S., Hu, E., Zhang, D.: Semi-supervised clustering with metric learning: An adaptive kernel method. Pattern Recognition 43(4), 1320–1333 (2010)

166. Yin, X., Chen, S., Hu, E., Zhang, D.: Semi-supervised clustering with metric learning: An adaptive kernel method. Pattern Recognition 43(4), 1320–1333 (2010)

167. Yu, Z., Kuang, Z., Liu, J., Chen, H., Zhang, J., You, J., Wong, H.S., Han, G.: Adaptive ensemble of semi-supervised clustering solutions. IEEE Transactions on Knowledge and Data Engineering 29(8), 1577–1590 (2017). https://doi.org/10.1109/TKDE.2017.2695615

168. Yu, Z., Luo, P., You, J., Wong, H.S., Leung, H., Wu, S., Zhang, J., Han, G.: Incremental semi-supervised clustering ensemble for high dimensional data clustering. IEEE Transactions on Knowledge and Data Engineering 28(3), 701–714 (2016). https://doi.org/10.1109/TKDE.2015.2499200

169. Yu, Z., Wong, H.S., You, J., Yang, Q., Liao, H.: Knowledge based cluster ensemble for cancer discovery from biomolecular data. IEEE transactions on nanobioscience 10(2), 76–85 (2011). https://doi.org/10.1109/TNB.2011.2144997

170. Zhan, J., Mao, J., Liu, Y., Guo, J., Zhang, M., Ma, S.: Learning discrete representations via constrained clustering for effective and efficient dense retrieval. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM). pp. 1328–1336 (2022). https://doi.org/10.1145/3488856.3498443

171. Zhang, H., Basu, S., Davidson, I.: A framework for deep constrained clustering - Algorithms and advances. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 57–72 (2019). https://doi.org/10.1007/978-3-030-46150-8_4

172. Zhang, T., Ando, R.K.: Analysis of spectral kernel design based semi-supervised learning. Advances in neural information processing systems 18, 1601–1608 (2005). https://doi.org/10.1109/ICPR.2005.2137

173. Zhang, X., Zheng, Y., Ye, X., Peng, Q., Wang, W., Li, S.: Clustering with implicit constraints: A novel approach to housing market segmentation. Transactions in GIS 26(2), 585–608 (2022). https://doi.org/10.1111/tgis.12878
175. Zhang, Z.W., Jing, X.Y., Wang, T.J.: Label propagation based semi-supervised learning for software defect prediction. Automated Software Engineering 24(1), 47–69 (2017). https://doi.org/10.1007/s10515-016-0194-x

176. Zhi, W., Wang, X., Qian, B., Butler, P., Ramakrishnan, N., Davidson, I.: Clustering with complex constraints - Algorithms and applications. In: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence. pp. 1056–1062 (2013)

177. Zhigang, C., Xuan, L., Fan, Y.: Constrained K-means with external information. In: 2013 8th International Conference on Computer Science and Education. pp. 490–493 (2013). https://doi.org/10.1109/ICCSE.2013.6553960